TESE DE DOUTORADO

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM

CIÊNCIA DA COMPUTAÇÃO

"*SEnsembles* – uma abordagem para melhorar a qualidade das correspondências de instâncias disjuntas em estudos observacionais explorando características idênticas e *ensembles* de regressores"

ALUNO: Sergio Ricardo Borges Junior
ORIENTADOR: Prof. Dr. Ricardo Rodrigues Ciferri
COORIENTADORA: Profa. Dra. Marilde T. P. Santos

São Carlos
Dezembro/2016

CAIXA POSTAL 676
FONE/FAX: (16) 3351-8233

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

SENSEMBLES – UMA ABORDAGEM PARA MELHORAR A QUALIDADE DAS CORRESPONDÊNCIAS DE INSTÂNCIAS DISJUNTAS EM ESTUDOS OBSERVACIONAIS EXPLORANDO CARACTERÍSTICAS IDÊNTICAS E *ENSEMBLES* DE REGRESSORES

SERGIO RICARDO BORGES JUNIOR

ORIENTADOR: PROF. DR. RICARDO RODRIGUES CIFERRI

COORIENTADORA: PROFA. DRA. MARILDE TEREZINHA PRADO SANTOS

São Carlos - SP
Dezembro/2016

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

SENSEMBLES – UMA ABORDAGEM PARA MELHORAR A QUALIDADE DAS CORRESPONDÊNCIAS DE INSTÂNCIAS DISJUNTAS EM ESTUDOS OBSERVACIONAIS EXPLORANDO CARACRETÍSTICAS IDÊNTICAS E ENSEMBLES DE REGRESSORES

SERGIO RICARDO BORGES JUNIOR

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Doutor em Ciência da Computação, área de concentração: Engenharia de Software / Banco de Dados / Interação Humano-Computador.
Orientador: Prof. Dr. Ricardo Rodrigues Ciferri.
Coorientadora: Profa. Dr. Marilde Terezinha Prado Santos.

São Carlos - SP
Dezembro/2016

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia

Programa de Pós-Graduação em Ciência da Computação

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a defesa de Tese de Doutorado do candidato Sérgio Ricardo Borges Jr., realizada em 16/12/2016.

Prof. Dr. Ricardo Rodrigues Ciferri
(UFSCar)

Prof. Dr. Estevam Rafael Rafael Hruschka Jr.
(UFSCar)

Prof. Dr. Renato Bueno
(UFSCar)

Prof. Dr. André Carlos Ponce de Leon Ferreira de Carvalho
(USP)

****

Prof. Dr. Luiz Guilherme Dácar da Silva Scorzafave
(USP)

Certifico que a sessão de defesa foi realizada com a participação à distância do membro Prof. Dr. Luiz Guilherme Dácar da Silva Scorzafave e, depois das arguições e deliberações realizadas, o participante à distância está de acordo com o conteúdo do parecer da comissão examinadora redigido no relatório de defesa da aluna Sérgio Ricardo Borges Jr.

Prof. Dr. Ricardo Rodrigues Ciferri
Presidente da Comissão Examinadora
(UFSCar)

Dedicatória

À Minha Querida e Amada Esposa Patrícia Borges

"A você, minha princesa, que sempre esteve ao meu lado nesta longa caminhada".

Aos meus pais,

"que me ensinaram a viver com consciência, dignidade e responsabilidade".

Aos professores Prof. Dr. Ricardo Rodrigues Ciferri e

Profa. Dra. Marilde Terezinha Prado Santos,

"que sempre me inspiraram".

AGRADECIMENTO

Ao nosso bom Deus.

À minha esposa Patrícia Borges,
"*pela paciência, companheirismo e incentivo nas horas difíceis*".

Aos meus orientadores Prof. Dr. Ricardo Rodrigues Ciferri e Profa. Dra. Marilde T. Prado Santos,
"obrigado por todos os ensinamentos".

À Profa. Dra. Cristina Dutra de Aguiar Ciferri,
"que me ajudou a entender os desafios da correspondência de instâncias".

Aos professores Dr. Luiz Guilherme Dácar da Silva Scorzafave e Dr. Walter Belluzzo Júnior,
"que me apresentaram um novo horizonte: os estudos observacionais".

Ao Prof. Dr. Estevam Rafael Hruschka Junior,
"que me indicou o uso de ensembles".

Aos professores Dra. Marcela Xavier Ribeiro, Dr. Renato Bueno e
Dr. André Carlos Ponce de Leon de Carvalho,
"que apontaram alguns equívocos na Qualificação".

Ao Sr. Waldomiro Barioni,
"que cedeu o seu precioso tempo para me ajudar com a Regressão Logística".

Ao Prof. Dr. Ricardo Cerri,
"que me fez refletir sobre o aprendizado de máquina".

À Dra. Ana Paula Bortoletto Martins,
"por ter cedido o conjunto de dados do PBF que utilizou em seu doutorado".

Aos professores Dr. Waldir Barros Fernandes Jr. e Dr. Carlos Magnus Carlson Filho,
"pelo incentivo e apoio ao projeto".

Aos professores Dr. Adriano Luís Simonato e Me. Mariângela *Cazetta,*
"pela paciência em me explicar as fórmulas matemáticas".

GBD
Grupo de Banco de Dados
UFSCar

Aos colegas do Grupo de Banco de Dados,
"nos quais sempre me espelhei".

ppgcc
Programa de Pós-Graduação em Ciência da Computação

Ao Programa de Pós-Graduação em Ciência da Computação (PPG-CC)
do Departamento de Computação (DC) da UFSCar,
"pelo título de mestre e doutor em Ciência da Computação".

"Tem mais presença em mim o que me falta".

Manoel de Barros

RESUMO

Introdução. Os conjuntos de dados manipulados em estudos observacionais possuem instâncias pertencentes a dois grupos distintos (i.e. grupo de tratamento e grupo de controle), as quais são comparadas para estimar o efeito do tratamento sobre os resultados. Para isso, em uma das abordagens, chamada de *Propensity Score Matching* (PSM), estima-se o escore de propensão para as instâncias de ambos os grupos e, em seguida, efetua-se a correspondência dessas instâncias com base nos valores dos escores de propensão. O escore de propensão é a probabilidade de atribuição de um tratamento com base nas características observadas (por exemplo, renda, sexo e idade). Neste contexto, a regressão logística é amplamente utilizada para estimar o escore de propensão e há uma ampla variedade de métodos de correspondência de instâncias. **Objetivo.** Esta pesquisa de doutorado tem como objetivo principal investigar alternativas computacionais para melhorar a qualidade das correspondências de instâncias em conjuntos de dados que são manipulados em estudos observacionais. **Metodologia.** Investigou-se técnicas que estimam o escore de propensão e métodos para se efetuar a correspondência das instâncias em estudos observacionais. Assim, foi possível investigar como as características idênticas das instâncias poderiam ser exploradas em um novo processo de correspondência e, como *ensembles*, mais precisamente, *bagging*, *random forest e boosting*, poderiam substituir a regressão logística ao estimar os escores de propensão das instâncias, no contexto do processo de PSM. **Proposta.** Esta pesquisa propõe uma nova abordagem no contexto do processo PSM, denominada "*SEnsembles*", que visa melhorar a qualidade da correspondência das instâncias com base em 2 processos principais, os quais utilizam técnicas que considerem em separado as características idênticas das instâncias e os *ensembles* de regressores, mais precisamente, *bagging*, *random forest e boosting*. **Resultados.** A abordagem proposta "*SEnsembles*" melhorou a qualidade da correspondência de instâncias para a maioria dos *calipers* utilizado (zero, 0,05, 0,10, 0,15, 0,20, 0,25 e 0,30) quando comparada ao *baseline Nearest Neighbor Matching* (*NNM*). Com base nos experimentos, quando houve ganho, a técnica que separa as características idênticas das instâncias proporcionou ganhos de até 53,8% na qualidade da correspondência, com média de 12,1% de melhoria e 2,7% de redução média do número de pares de instâncias correspondidas. Já a técnica que substituiu a regressão logística pelos *ensembles* proporcionou as melhores correspondências com o *caliper* zero e com os valores 0,20, 0,25 e 0,30, com ganhos de até 36,3% e, com média de 12,7% de melhoria e 7,6% de redução do número de pares de instâncias correspondidas.

Palavras-chave: estudos observacionais, correspondência de instâncias, escore de propensão, *Propensity Score Matching*, *ensembles, regressores.*

ABSTRACT

Introduction. The datasets used in observational studies have instances belonging to two distinct groups (i.e. treatment group and control group), which are compared in order to estimate the effect of the treatment over the results. For such, in one of the approaches, called Propensity Score Matching (PSM), the propensity score for the instances of both groups is estimated and, subsequently, the correspondence of these instances is performed based on the values for the propensity score. The propensity score is the probability of attribution of a treatment based on the observed characteristics (e.g. income, sex and age). In this context, the logistic regression is widely used to estimate the propensity score and there is an great variety of instance correspondence methods. **Objective.** This doctor´s thesis has as its main objective to investigate computational alternatives in order to improve the quality of the instance correspondence in datasets that are manipulated in observational studies. **Methodology.** Techniques that estimate the propensity score and methods to perform the instance correspondence in observational studies were investigated. Thus, it was possible to investigate how the identical characteristics of the instances could be exploited in a new process to perform correspondence and, how ensembles could substitute the logistic regression by estimating the propensity scores of the instances, in the context of the PSM process. **Proposal.** This thesis proposes a new approach in the context of the PSM process, called “SEnsembles”, which aims to improve the quality of instance correspondence based on two main processes, which use techniques that separately consider the identical characteristics of the instances and the ensembles of regressors, more precisely, bagging, random forest and boosting. **Results.** The proposed approach “*SEnsembles*” improves the quality of the instance correspondence for the majority of calipers used (i.e. zero, 0.05, 0.10, 0.15, 0.20, 0.25 and 0.30) when compared to the baseline *Nearest Neighbor Matching* (NNM). Based on the experiments, when there was an improvement over the baseline, the technique that separates the identical characteristics of the instances presented improvements of up to 53.8% in the quality of correspondence, with an average of gains of 12.1%; and only 2.7% of average in the reduction of the number of pairs of instances matched. The technique which substituted the logistic regression for ensembles of regressors, in turn, presented the best correspondence with the caliper zero and with the values 0.20, 0.25 and 0.30, with improvements of up to 36.3% and an average of gains of 12.7%; and a slightly reduction of 7.6% in the number of pairs of instances matched.

Keywords: observational studies, instance correspondence, propensity score, Propensity Score Matching, ensembles, regressors.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AIBF – Avaliação de Impacto do Programa Bolsa Família

BAG – *Bagging*

CART – *Classification and Regression Trees*

ECS – Engenharia de Correspondência com *Slicer*

ECE – Engenharia de Correspondência com *Ensembles*

FBS – *Feature-Based Similarity*

FEA – Faculdade de Economia, Administração e Contabilidade

GBD/UFSCar – Grupo de Banco de Dados da Universidade Federal de São Carlos

GNU – *General Public License*

INEP – Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

IPTW – *Inverse Probability of Treatment Weighting*

MDS – Ministério do Desenvolvimento Social e Combate à Fome

NNM – *Nearest Neighbor Matching*

PBF – Programa Bolsa Família

POF – Pesquisa de Orçamentos Familiares

PS – *Propensity Score*

PSA – *Propensity Score Analysis*

PSM – *Propensity Score Matching*

RF – *Random Forest*

RPART – *Recursive Partitioning and Regression Trees*

SAEB – Sistema de Avaliação da Educação Básica

SE – *Standard Error*

SEnsembles – *Slicer and Ensembles*

SMCG – Seleção da Melhor Correspondência Global

SMCL – Seleção da Melhor Correspondência Local

OMG – *Object Management Group*

USP – Universidade do Estado de São Paulo

SUMÁRIO

Capítulo 1

INTRODUÇÃO

1.1 Contextualização

Os ensaios clínicos randomizados são considerados ideais para estimar os efeitos de tratamentos, intervenções ou exposições, sobre os resultados (Austin, 2011). Os pesquisadores ao conduzirem esses ensaios possuem o controle total do experimento e os indivíduos em estudo são alocados de maneira aleatória em dois grupos distintos. Assim, indivíduos que receberão o tratamento são alocados no grupo de tratamento e, os indivíduos que serão utilizados para comparação e não receberão o tratamento são alocados no grupo de controle.

O efeito ou impacto do tratamento pode ser estimado comparando-se diretamente os indivíduos do grupo de tratamento e os indivíduos não tratados do grupo de controle, pois, nestes casos, a característica aleatória do experimento e o controle do pesquisador garantem que esses grupos sejam similares.

A similaridade aqui abordada está relacionada à distribuição dos valores das covariáveis (ou atributos) nos grupos de tratamento e de controle, ou seja, se os grupos possuem a mesma distribuição das covariáveis, então os seus indivíduos são similares e os grupos estão balanceados entre si. Por exemplo, em um ensaio clínico randomizado sobre o uso de um medicamento para a doença hemoglobinopatias CC, as covariáveis podem se referir ao sexo, idade, e raça. Neste caso, para grupos com a mesma distribuição de idade, sexo e raça, os seus indivíduos são similares.

Entretanto, a realização de experimentos randomizados nem sempre é viável ou ética (SHADISH, STEINER, 2010). Por exemplo, não é possível comparar os

efeitos do Programa Bolsa Família (PBF) sobre um determinado aspecto, por exemplo, qual o impacto da participação no PBF para as famílias em relação ao consumo de alimentos uma vez que os indivíduos beneficiários (grupo de tratamento) e não beneficiários (grupo de controle) não são selecionados ao acaso, ou seja, de maneira aleatória. Adicionalmente, supondo que seja possível a criação de um grupo de beneficiários do PBF por meio de dados disponibilizados pelo governo, ainda é preciso a criação de um grupo de não beneficiários com as mesmas características (i.e. com a mesma distribuição dos valores das covariáveis) dos indivíduos do grupo de beneficiários.

O exemplo apresentado do PBF ilustra que a seleção dos indivíduos não pode ocorrer ao acaso e o pesquisador não possui o controle total do experimento. Nestes casos, o efeito do tratamento sobre os resultados pode ser estimado por meio de um estudo observacional, o qual é definido como sendo uma investigação empírica com o objetivo de elucidar as relações de causa e efeito, quando não é viável ou possível o uso de experimentação controlada (COCHRAN, 1965).

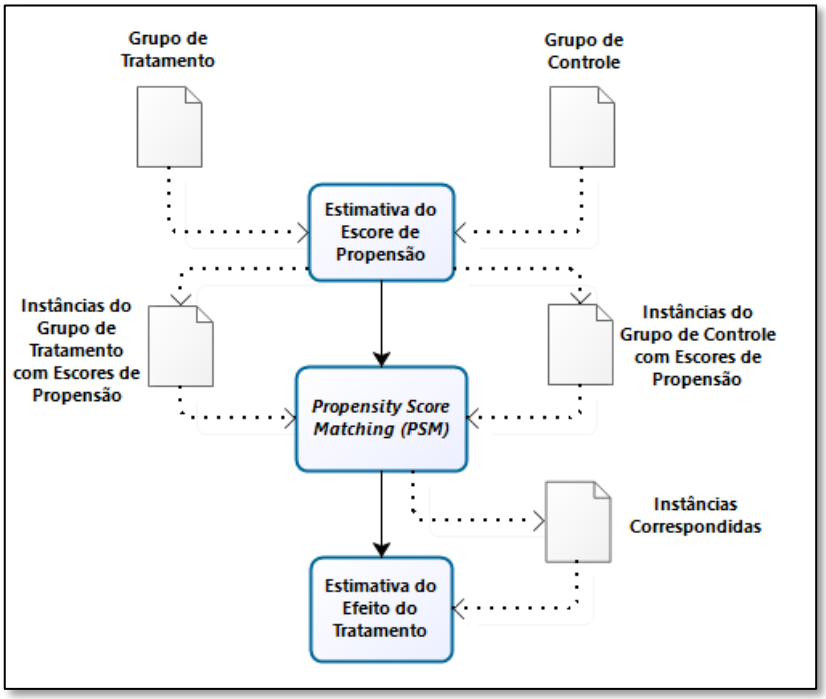
Nos estudos observacionais é pouco provável que as covariáveis (características, atributos) dos indivíduos dos grupos de tratamento e de controle estejam balanceadas inicialmente, ou seja, é pouco provável que esses indivíduos sejam similares considerando os valores de suas covariáveis. Diante desse problema, Rosenbaum e Rubin (1983) propuseram um método para gerar grupos de tratamento e de controle balanceados, com base em uma medida de probabilidade, denominada Escore de Propensão, que é estimado para cada indivíduo do estudo em questão.

O escore de propensão é definido como sendo a probabilidade condicional de atribuição de um determinado tratamento, dado um conjunto de covariáveis observadas. Adicionalmente, Rosenbaum e Rubin (1983) também demonstraram que indivíduos com os mesmos valores de escore de propensão possuem a mesma distribuição das covariáveis, ou seja, esses indivíduos são similares considerando-se as covariáveis observadas.

A estimativa do escore de propensão é a primeira etapa para se avaliar o efeito ou impacto de um determinado tratamento sobre os resultados. Os pesquisadores, após estimarem os escores de propensão dos indivíduos (instâncias) dos grupos de tratamento e de controle, podem adotar um dos métodos de correspondência de indivíduos (instâncias), na etapa conhecida como *Propensity*

Score Matching, na qual as instâncias são correspondidas (pareadas) com base no valor do escore de propensão, conforme ilustrado na Figura 1.1.

Figura 1.1 – Etapas para avaliar o efeito ou impacto de um tratamento.



Fonte: Elaborado pelo autor.

Após o PSM, os pesquisadores estimam o efeito do tratamento por meio de testes estatísticos, tal como teste de hipóteses, comparando as médias da variável resultado das instâncias de ambos os grupos. Por exemplo, supondo que a variável resultado seja a renda dos indivíduos, pode-se aplicar um teste de hipóteses para verificar se há diferenças significativas entre a média das instâncias do grupo de tratamento com a média das instâncias do grupo de controle e, dessa maneira, avaliar se houve impacto de um determinado programa sobre as rendas dos indivíduos participantes.

É importante ressaltar que há outras três abordagens, além do PSM, as quais utilizam o escore de propensão para estimar o efeito do tratamento: *stratification* (ou *subclassification*), *inverse probability of treatment weighting* (IPTW) e *covariate adjustment* (AUSTIN, 2011; ROSENBAUM, RUBIN,1983). Porém, essas abordagens não são investigadas no presente trabalho por não serem objetos de estudo da pesquisa.

A próxima seção ilustra o uso da correspondência de instância em aplicações do mundo real no âmbito dos estudos observacionais, os quais estão presentes em várias áreas do conhecimento humano. No entanto, recomenda-se, caso já tenha conhecimento, a leitura da seção seguinte, que trata sobre a motivação do presente trabalho.

1.2 Aplicações do Mundo Real

Estudos observacionais são amplamente realizados em várias áreas do conhecimento. Em uma rápida pesquisa no *Google Scholar* sobre o clássico artigo de Rosenbaum e Rubin (1983) foram encontradas, no início de novembro de 2016, 16.346 citações. Apenas para comparação, essa mesma pesquisa foi realizada no início de 2012 por Li (2012) e foram obtidas 7.300 citações. Já na época do exame de qualificação de doutorado, em dezembro de 2014, o número já era de 12.239 citações.

Dentre as áreas que possuem pesquisas realizadas por meio de estudos observacionais, destacam-se: economia (SMITH, TODD, 2005; IMBENS, 2004; DEHEJIA, WAHBA, 1999), medicina (AUSTIN, et al., 2012; YE, KASKUTAS, 2009; WOLFE, MICHAUD, 2004; D'AGOSTINO, 1998), finanças (CAMPELLO, GRAHAM, HARVEY, 2010), sociologia (GRODSKY, 2007; GANGL, 2006), política (HO et al., 2007, ARCENEAUX et al., 2006), educação (PFEFFERMANN, LANDSMAN, 2011; HONG, YU, 2008; STAFF, et al. 2008) e psicologia (MCCAFFREY, 2013, MCCAFFREY, 2004).

No Brasil, vários trabalhos são desenvolvimentos utilizando o método proposto por Rosenbaum e Rubin (1983), destacando-se, dentre eles, a avaliação do impacto do Programa Bolsa Família (AIBF I) em 2005 (MDS, CEDEPLAR, 2014) e em 2009 (AIBF II) (MDS, 2014).

Na AIBF I foram entrevistadas 15.426 famílias, em 269 municípios alocados em três macrorregiões: Nordeste, Norte Urbano e Centro Oeste, Sul e Sudeste. Na segunda rodada (AIBF II) foram entrevistas 11.433 famílias, das 15.426 da primeira rodada (MDS, 2014).

Os resultados da AIBF I foram apresentados em quatro dimensões relacionadas aos domicílios pesquisados: gasto domiciliar (alimentação, habitação, vestuário, transporte, saúde, educação e despesas diversas), educação (frequência escolar, progressão e evasão), trabalho (ocupação, procura por emprego e transições ocupacionais) e, por fim, a capacidade da mulher em participar das decisões domésticas (MDS, CEDEPLAR, 2014).

Os domicílios foram separados em três grupos para avaliação do impacto dos indicadores acima, conforme a seguir:

a) Grupo de tratamento: constituído por domicílios que declararam receber benefício do PBF;
b) Grupo de Controle 1 (C1): composto pelos domicílios que declararam receber outros benefícios na época da pesquisa (vale gás, cartão alimentação, etc.);
c) Grupo de Controle 2 (C2): constituído pelos domicílios que declararam nunca ter recebido nenhum benefício, independente de serem inscritos em algum programa público.

Os dois grupos de controle foram criados com a justificativa de comparar dois tipos distintos de resultados do PBF. Assim, comparou-se o grupo de beneficiários do PBF com beneficiários de outros programas sociais (Grupo C1) e, na segunda comparação, o grupo de beneficiários do PBF com o Grupo C2, que não receberam nenhum tipo de benefício.

Para comparar os grupos de tratamento e de controle foi necessário estimar o escore de propensão para os domicílios de ambos os grupos. O escore de propensão, neste caso, representa a probabilidade do domicílio de ter recebido o PBF com base nas características observadas. Em seguida, os domicílios foram combinados (pareados) com base no escore de propensão, ou seja, cada domicílio do grupo de tratamento foi pareado com o domicílio mais similar no grupo de controle. Para isso, foi utilizado o método de correspondência pelo “vizinho mais próximo” (Seção 3.2.3). Por fim, o impacto do programa, ou melhor, o efeito médio do tratamento (recebimento do PBF) foi estimado por meio da diferença dos resultados entre os domicílios pareados (MDS, CEDEPLAR, 2014).

Na AIBF II voltou-se a entrevistar os mesmos domicílios, classificando-os entre beneficiários e não beneficiários do PBF. Porém, não foi possível replicar o

questionário da AIBF I em 26% dos domicílios, aproximadamente, pois as famílias não foram encontradas nos endereços indicados. Segundo o MDS (2014), com a AIBF II foi possível entender o impacto do PBF nas condições de vidas das famílias beneficiadas, comparando as características observadas de 2005 e 2009, com intuito de responder duas questões principais:

> (i) as famílias beneficiárias do PBF estão em melhores condições de vida em 2009 do que estavam em 2005? e (ii) as famílias beneficiárias estão em melhores condições de vida do que estavam em 2005 por causa do PBF? MDS (2014).

Para responder as questões levantadas, foi realizada uma avaliação de impacto do PBF que comparou os grupos de beneficiários e não beneficiários das duas pesquisas AIBF I e AIBF II, conforme o recebimento do PBF. A Tabela 1.1 apresenta os números dos domicílios das duas pesquisas, classificando-os conforme a situação do recebimento do PBF. Nota-se que a quantidade de beneficiários de 2009 é bem superior à quantidade de 2005 e que 929 famílias deixaram de receber o benefício em 2009 em comparação ao ano de 2005.

Tabela 1.1 – Domicílios AIBF I x AIBF II.

		Grupos AIBF I (2005)		
		Intervenção	**Controle 1 (C1)**	**Controle 2 (C2)**
		Beneficiários	Não beneficiários e inscritos no Cadastro Único.	Não Beneficiários
Grupos AIBF II (2009)	Beneficiários	1.844	1.121	1.707
	Não Beneficiários	929	1.352	1.302

Fonte: Adaptado de MDS (2014).

Para avaliar o impacto do PBF foram considerados os dados das duas pesquisas (AIBF I e AIBF II). Nesta avaliação, objetivou-se utilizar todos os dados de beneficiários de 2009 (1.844 + 1.121 + 1.707), comparando-os com não beneficiários inscritos no Cadastro Único de 2009 (1.352) e os não beneficiários do Grupo de Controle 2 da AIBF I (1.302). Deste modo, as 929 famílias que não recebiam benefícios em 2009 foram eliminadas da avaliação (MDS, 2014).

Para comparação dos grupos foi utilizado uma ponderação pelo escore de propensão. Assim, os escores de propensão foram utilizados para atribuir pesos aos domicílios, em uma abordagem IPTW (*Inverse Probability of Treatment Weighting)*, na qual os domicílios não beneficiários que possuem características mais similares aos beneficiários receberam pesos mais altos, enquanto que os domicílios com características menos similares aos beneficiários recebem pesos menores.

A abordagem de ponderação pelo escore de propensão buscou considerar diferenças nas amostras das duas pesquisas (AIBF I e AIBF II), uma vez que nem todos os domicílios da AIBF I foram pesquisados novamente na AIBF II. Para avaliar o impacto foram utilizados métodos comparativos utilizando as médias das características dos domicílios beneficiários e não beneficiários das pesquisas em 2005 e 2009.

Por fim, um conjunto de dados do PBF foi utilizado no presente trabalho. Esse conjunto de dados é descrito com maiores detalhes na metodologia no Capitulo 2.

1.3 Motivação

O programa Observatório da Educação (INEP, 2016a) foi criado em 2006 pelo governo federal para fomentar o desenvolvimento de estudos e pesquisas na área da educação. Um grupo de pesquisadores da FEA/USP Ribeirão Preto, coordenado pelo prof. Dr. Walter Belluzzo Júnior, desenvolveu o projeto denominado “Uma Análise da Evolução e dos Determinantes do Desempenho Escolar no Brasil”, (INEP, 2016b), o qual buscou analisar a evolução e os determinantes do desempenho escolar no Brasil. Para isso, esse projeto foi dividido em duas etapas.

Na primeira etapa buscou-se avaliar os diferenciais de desempenho escolar entre os alunos da rede pública e privada, com base nos dados sobre cor, raça, renda, entre outros e, na segunda, buscou-se avaliar as desigualdades de desempenho escolar dos alunos do ensino fundamental do Estado de São Paulo. Em ambas as etapas foram utilizadas bases de dados disponibilizadas pelo INEP, tais como: SAEB (Sistema de Avaliação da Educação Básica), Censo Escolar, Prova Brasil, entre outras.

Como os pesquisadores da FEA/USP necessitavam reunir e comparar as informações das bases de dados disponibilizadas pelo INEP, uma parceira entre a FEA/USP e Departamento da Computação da UFSCar foi firmada para que pesquisadores da área banco de dados pudessem apoiar a integração dessas bases de dados, uma vez não possuíam identificadores em comum e não tinham sido planejadas para uma integração.

Neste contexto, iniciou-se a presente pesquisa, cujo objetivo inicial era integrar bases de dados governamentais. Porém, com a investigação do tema, chegou-se aos estudos observacionais, nos quais os pesquisadores necessitam comparar grupos de tratamento e controle mais similares quanto possível. Neste sentido, observou-se que vários trabalhos na literatura investigavam aspectos computacionais que apoiam o processo de correspondência de instâncias. Em virtude disso, foi realizada uma investigação das etapas de estimativa do escore de propensão e de PSM.

Vários trabalhos na literatura apresentam alternativas à regressão logística para estimar o escore de propensão, destacando-se, dentre elas, o trabalho de Setoguchi et al. (2008), que comparou a aplicação de redes neurais, árvores de regressão com e sem “poda”, com a regressão logística. No entanto, Setoguchi et al. (2008) não avaliaram o uso dos métodos *ensembles* de regressores em substituição à regressão logística. Além disso, utilizou apenas conjuntos de dados construídos de maneira sintética e, portanto, não utilizado dados reais.

Já uma comparação de *ensembles* de regressores foi realizada por Lee et al. (2010). Os *ensembles* que foram utilizados substituíram a regressão logística na estimativa do escore de propensão por meio de árvores de regressão. Os autores concluíram que *ensembles* forneceram excelente desempenho em termos de balanceamento de covariáveis e estimativas de efeito de tratamento. Porém, Lee et al. (2010) não avaliaram o uso de *ensembles* de regressores no processo de PSM, mas sim, em uma abordagem com ponderação pelo escore de propensão. Além disso, os conjuntos de dados utilizados foram baseados nos conjuntos de dados utilizados por Setoguchi et al. (2008), os quais foram constituídos apenas por dados sintéticos.

Recentemente, os trabalhos descritos por McCaffrey et al. (2013) e Watkins et al. (2013) também propuseram o uso de *ensembles* de regressores para gerar uma ponderação das instâncias baseada no escore de propensão.

No trabalho de McCaffrey et al. (2013) foi utilizado *ensembles* de regressores, mais precisamente, *boosting*, para gerar uma ponderação baseada no escore de propensão em múltiplos tratamentos. Os resultados demonstraram que regressão logística foi superada por *boosting* ao tentar balancear mais do que dois grupos de tratamento. Ainda, segundo McCaffrey et al. (2013), uma das vantagens do método *boosting* está relacionado ao seu processo iterativo, que pode levar ao melhor balanceamento dos grupos de tratamento e de controle, ou seja, pode produzir grupos mais similares.

Já Watkins et al. (2013) aplicaram *ensembles* de regressores, mais precisamente, *random forest* e *bagging*, também baseados em árvore de regressão para estimar o escore de propensão, bem como para comparar o desempenho de tais métodos com a regressão logística. Os autores concluíram que *random forest* e *bagging* produziram melhor balanceamento das covariáveis com aumento da precisão em relação à regressão logística.

Em relação do uso de *ensembles* de regressores no processo de PSM, observou-se que não há muitos trabalhos na literatura. Em um recente trabalho, Austin e Small (2014) propuseram dois métodos que utilizam amostras *bootstrap* (*bagging*) para estimar a variabilidade da estimativa do efeito do tratamento.

No primeiro método as amostras foram compostas por pares de instâncias já correspondias no processo de PSM, enquanto que no segundo, as amostras foram obtidas a partir da amostra original e os escores de propensão foram estimados separadamente em cada amostra.

Diante do exposto, nota-se que *ensembles* de regressores são citados como alternativas úteis à regressão logística por produzirem melhores distribuições das covariáveis em abordagens que ponderam as instâncias pelos escores de propensão. Porém, não foi encontrado nenhum trabalho na literatura que os empregam em uma estratégia para apoiar o processo de PSM e, isso, levantou alguns questionamentos principais:

a) A substituição da regressão logística por *ensembles* de regressores, mais precisamente, *bagging*, *boosting* e *random forest*, no contexto do processo de PSM, pode melhorar o balanceamento dos grupos de tratamento e de controle com a geração de correspondências mais similares?

b) Em quais situações o uso dos *ensembles* de regressores é vantajoso?

c) Será que uma abordagem que explore, além da substituição da regressão logística por *ensembles* de regressores, as características idênticas das instâncias, pode produzir correspondências mais similares?

Diante do exposto, destaca-se, como a principal motivação da pesquisa, a busca por correspondências de instâncias mais similares, com a especificação de uma nova abordagem no âmbito do processo de PSM, que considere a substituição da regressão logística por *ensembles* de regressores e que explore as características idênticas das instâncias.

1.4 Hipóteses da Tese

As hipóteses de pesquisa a serem comprovadas são:

Hipótese 1: “Uma abordagem para correspondência de instâncias, no contexto do processo de PSM, que considere *ensembles* de regressores, mais especificamente, *bagging, boosting e random forest*, pode permitir correspondências de instâncias mais similares, a partir de conjuntos de dados com características de disjunção de instâncias”.

Hipótese 2: “Uma estratégia para correspondência de instâncias que trate separadamente as características idênticas das instâncias pode permitir correspondências mais similares de instâncias e melhorar a qualidade final do processo de correspondência”.

Ressalta-se que as hipóteses acima estão relacionadas aos processos responsáveis por gerar as correspondências de instâncias da nova abordagem, a qual foi denominada de “*SEnsembles*” em virtude de seus processos.

1.5 Objetivos

A presente pesquisa de doutorado tem como objetivo investigar alternativas computacionais para melhorar a qualidade das correspondências de instâncias em conjuntos de dados que são manipulados em estudos observacionais, os quais possuem como característica principal a disjunção de instâncias. Assim, busca-se obter grupos de tratamento e de controle mais balanceados e com instâncias correspondidas mais similares, utilizando-se de técnicas que considerem as características idênticas das instâncias e *ensembles* de regressores em substituição à regressão logística.

Os objetivos específicos delineados para esta pesquisa são:

- Investigar como *ensembles* de regressores, mais precisamente, *bagging*, *boosting* e *random forest*, podem atuar em uma estratégia para estimar os escores de propensão das instâncias no contexto do processo de PSM e, como utilizá-los em conjunto com os métodos de correspondências já existentes;
- Investigar em quais situações os *ensembles* de regressores (*bagging, boosting e random forest*) podem melhorar o balanceamento dos grupos de instâncias (tratamento e controle);
- Investigar como as características idênticas das instâncias pode propiciar a elaboração de uma estratégia que melhore a qualidade da correspondência das instâncias e em quais situações;
- Propor e implementar uma nova abordagem, denominada nesta tese de "*SEnsembles*", para correspondência de instâncias;
- Validar a abordagem proposta "*SEnsembles*" por meio de experimentos que resultem na verificação de suas estratégias.

1.6 Escopo da Pesquisa

A presente pesquisa apresenta algumas delimitações para determinar o seu escopo. Essas delimitações são relacionadas aos métodos para correspondências de instâncias e aos *ensembles* de regressores. Porém, antes de descrevê-las, é importante ressaltar que os termos correspondência e pareamento possui o mesmo significado no presente trabalho e, em muitas ocasiões ao longo do texto, o uso do termo “correspondência” foi usado com o sentido de representar um conjunto de pareamentos de instâncias.

Já em relação aos métodos de correspondência (ou pareamento), os esforços foram direcionados para melhorar a correspondência 1:1 sem substituição de instâncias, isto é, uma instância do grupo de tratamento é pareada (ou correspondida), com somente uma instância do grupo de controle e, uma instância do grupo de controle pode ser pareada com apenas uma instância do grupo de tratamento (sem substituição). Esse escopo se justifica pela busca em se utilizar estratégias que empreguem métodos para correspondências considerados simples, que tratam do pareamento 1:1 sem substituição.

Ressalta-se que o uso do *caliper*, que é um limitador de distância entre instâncias para se efetuar a correspondência, foi considerado. Com isso, foi possível verificar o comportamento da nova abordagem para correspondência, “*SEnsembles”*, com pareamentos efetuados considerando-se os seguintes *calipers*: 0, 0,05, 0,10, 0,15, 0,20, 0,25 e 0,30. Ressalta-se que o *caliper* com valor igual zero permite a correspondência de duas instâncias sem considerar uma determinada distância mínima entre elas.

Já em relação aos *ensembles* de regressores, o enfoque da pesquisa foi direcionado para os *ensembles* mais utilizados na literatura recente para estimar os escores de propensão das instâncias, destacando-se: *bagging*, *random forest* e *boosting*. Assim, buscou-se elaborar uma estratégia com esses *ensembles* para substituir a regressão logística ao estimar os escores de propensão no contexto do processo de PSM.

Por fim, é importante destacar que as delimitações foram impostas para permitir direcionar mais adequadamente os esforços, atingir os objetivos propostos e comprovar, de forma integral ou parcial, as hipóteses da tese.

1.7 Organização do Trabalho

O presente trabalho está organizado da seguinte forma:

- O Capítulo 2 apresenta a fundamentação teórica necessária a presente pesquisa, a qual aborda os assuntos sobre integração de dados, escore de propensão, *Propensity Score Matching* (PSM) e, por fim, *ensembles* de regressores;

- O Capítulo 3 apresenta a revisão bibliográfica contendo os trabalhos que descrevem os principais métodos de correspondência de instâncias e, também, os principais trabalhos que utilizam métodos de aprendizado de máquina em estudos observacionais, mas com enfoque voltado aos e*nsembles* de regressores para estimar o escore de propensão e, suas aplicações no contexto do processo de PSM;

- O Capítulo 4 apresenta a metodologia adotada no desenvolvimento da pesquisa, a qual descreve o tipo de pesquisa, métodos utilizados, conjuntos de dados usados (aplicações-alvo), métricas adotadas, recursos aplicados, plano de trabalho e, por fim, as formas de acompanhamento da pesquisa;

- O Capítulo 5 descreve a nova abordagem proposta, denominada "*SEnsembles"*, para a correspondência de instâncias;

- O Capítulo 6 apresenta a validação da abordagem "*SEnsembles"* com a análise comparativa dos resultados de uma ampla gama de experimentos que buscaram validar a abordagem proposta;

- Já o Capítulo 7 apresenta a conclusão do presente trabalho, descrevendo a comprovação das hipóteses da pesquisa, análise dos objetivos frente aos resultados alcançados, contribuições obtidas e sugestões de trabalhos futuros;

- Por fim, são apresentadas as referências utilizadas na elaboração desta tese, Glossário, Apêndice A, o qual apresenta os processos ECS e ECE da abordagem proposta "*SEnsembles*" representados em uma notação para modelagem de processos e, por último, o Anexo A, o qual apresenta a codificação para geração de conjuntos de dados sintéticos descrita por Lee et al. (2010) e utilizada nesta tese.

Capítulo 2

FUNDAMENTAÇÃO TEÓRICA

2.1 Considerações Iniciais

Este capítulo tem como objetivo descrever a fundamentação teórica necessária para o entendimento da presente pesquisa, destacando os temas sobre Integração de Dados, Escore de Propensão, *Propensity Score Matching* (PSM) e, por fim, *Ensembles*, com destaque para os métodos *bagging*, *boosting* e *random forest*.

2.2 Integração de Dados

Na literatura, os trabalhos sobre integração de dados são categorizados em integração em nível de esquema (SAGI, GAL, 2013; BERNSTEIN et al., 2011; RAHM, 2011; RAHM, BERNSTEIN, 2001; SPACCAPIETRA, PARENT, 1994; SPACCAPIETRA, PARENT, DUPONT, 1992) e integração em nível de instância (DOAN, HAVELY, IVES, 2012; CHRISTEN, 2012; DORNELES, GONÇALVES, MELLO, 2010).

A seguir, serão apresentadas as seções sobre integração em nível de esquema e em nível de instância. Adicionalmente, incluiu-se também uma Seção sobre integração de instâncias em estudos observacionais, que é o foco desta pesquisa.

2.2.1 Integração em Nível de Esquema

As abordagens de integração em nível de esquema são caracterizadas pela criação de mapeamentos semânticos entre esquemas com base nos elementos que os compõem, tais como: atributos, classes ou tipos-entidade, tipos de dados, restrições, entre outros (BERNSTEIN et al., 2011; RAHM, BERNSTEIN, 2001). Por exemplo, em um esquema E1 o nome de uma pessoa pode ser mantido em dois atributos: primeiro nome e sobrenome, enquanto que, em outro esquema E2, o nome pode ser mantido apenas em um único atributo denominado nome.

Dessa forma, as abordagens de integração em nível de esquema devem estabelecer as correspondências entre os elementos dos esquemas a serem integrados, identificar as similaridades existentes e resolver os conflitos. Segundo Ciferri (2002), os conflitos em nível de esquema podem ser classificados em três grupos: conflitos de nome, conflitos semânticos e conflitos estruturais.

Os conflitos de nomes são pertinentes aos nomes utilizados para representar os diferentes elementos que compõe os esquemas a serem integrados. Os conflitos de nomes podem resultar em problemas de sinônimos, quando nomes diferentes são aplicados para representar o mesmo elemento em esquemas diferentes e, em problemas de homônimos, quando o mesmo nome é aplicado a diferentes elementos dos esquemas a serem integrados, porém com significados distintos.

Um exemplo do problema de sinônimos ocorre quando, em um esquema A, o nome Produto é utilizado para representar todos os produtos à venda, enquanto que em um esquema B, o mesmo elemento é chamado de Material. Já um problema de homônimos, ocorre em um esquema A, quando o termo Funcionário é utilizado para representar todos os funcionários contratados da empresa e que possuem vínculos empregatícios, enquanto no esquema B, o termo funcionário é utilizado para representar os funcionários terceirados, que não possuem vínculos empregatícios.

Já o conflito semântico ocorre quando um mesmo elemento é modelado de maneira diferente representando conjuntos que se sobrepõe. Por exemplo, em um esquema A, o elemento Equipamento representa todos os equipamentos de uma empresa, incluindo, computadores, impressoras, copiadores, entre outros, enquanto

que em um esquema B, o elemento Equipamento representa somente os computadores que a empresa possui.

Por último, o conflito estrutural ocorre quando o mesmo conceito é representado em estruturas diferentes nos esquemas a serem integrados. Por exemplo, em um esquema A, o nome de uma pessoa é armazenado em apenas um atributo, enquanto que no esquema B, o nome é armazenado em dois atributos, nome e sobrenome. Um segundo exemplo do conflito estrutural pode ocorrer quando, em um esquema A, é utilizado uma hierarquia de superclasse/subclasse para representar os dados dos clientes e funcionários, enquanto que em um esquema B, esses dados são mantidos separadamente em dois tipos entidade diferentes.

Como visto, os conflitos devem ser resolvidos por meio da correspondência entre os esquemas. Segundo Doan e Havely (2005), a correspondência de esquemas pode ser realizada baseada em dois métodos: regras e aprendizado de máquina. Em geral, as regras exploram as informações sobre os elementos dos esquemas e proporcionam um rápido e conciso método de capturar o conhecimento do usuário sobre o domínio. Já os métodos baseados em aprendizado de máquina exploram não somente as informações sobre os elementos dos esquemas, mas também sobre as instâncias. Outra questão relevante em relação às abordagens de integração de esquemas é a aplicação em diversos modelos de dados, destacando-se o modelo entidade relacionamento e entidade relacionamento estendido (PARENT, SPACCAPIETRA, 1992), o modelo relacional (IBRAHIM et al., 2014; KARASNEH et al., 2009) e o modelo XML (GONG et al., 2012; AL-GHANIM, NOAH, SEMBOK, 2011).

2.2.2 Integração em Nível de Instância

As abordagens de integração em nível de instância são caracterizadas pela identificação de instâncias com base nos valores de seus atributos ou no contexto de aplicação (DOAN, HALEVY, IVES, 2012; CHRISTEN, 2012). Dessa forma, neste nível de integração busca-se identificar quais as instâncias que representam a mesma entidade do mundo real.

Várias abordagens têm sido desenvolvidas para solucionar os problemas de integração de instâncias. Segundo Kou (2008), as abordagens podem ser baseadas na similaridade das características (dos atributos) das instâncias (*Feature-Based Similarity - FBS*) e em contexto.

Nas abordagens FBS são utilizadas técnicas que empregam funções de similaridade para comparar valores dos atributos das instâncias. Assim, as instâncias que possuem a mesma similaridade são consideradas a mesma entidade no mundo real.

Já nas abordagens baseadas em contexto, além de empregar funções de similaridade, também analisam informações derivadas do contexto ou do domínio de aplicação para apoiar a identificação das instâncias correspondentes. Por exemplo, pode-se analisar a instituição do autor de um artigo mantido em duas bibliotecas digitais com nomes de autoria diferentes para determinar se é a mesma pessoa, ou seja, a mesma instância.

Como visto, a integração de dados em nível de instância busca identificar quais instâncias mantidas em bases de dados diferentes, ou seja, em esquemas diferentes, representam a mesma entidade no mundo real. Como passos genéricos para a solução do problema de integração de instâncias, têm-se: (i) agrupamento de instâncias similares em um mesmo grupo; (ii) identificação da instância representativa dentre as instâncias de um grupo; e (iii) integração dos dados entre as instâncias de um mesmo grupo visando gerar uma instância com dados completos e corretos para a instância representativa do grupo. O passo 3 consiste em copiar valores das instâncias para a instância representativa de forma que todos os seus atributos possuam valores (não sejam vazios) e possuam valores corretos. Como nota-se, o resultado do processo de integração a geração de N instâncias a partir de M instâncias, onde M > N, sendo que para cada uma das N instâncias, um subconjunto m das M instâncias foi usado para gerá-la. Assim, o processo de integração consiste em combinar instâncias (2 ou mais) em uma única instância.

2.2.3 Integração de instâncias em Estudos Observacionais

A integração de instâncias em estudos observacionais pode ser considerada um caso particular do problema de integração de dados em nível de instância que se diferencia pelo fato de que as instâncias a serem integradas (relacionadas entre si) não representam a mesma entidade no mundo real, mas são similares tanto quanto possível. Portanto, a integração de instâncias em estudos observacionais visa o pareamento entre pares de instâncias de dois conjuntos de dados distintos (com dados disjuntos entre si), mas com características similares entre as instâncias. Assim, ao contrário do processo de integração genérico, não se visa uma combinação de duas ou mais instâncias em uma única instância e, sim, a correspondência entre pares de instâncias de conjuntos distintos.

Em geral, as instâncias nos estudos observacionais estão alocadas em dois grupos disjuntos, instâncias que receberam um determinado tratamento (ou intervenção, benefício, dentre outros fatores) e instâncias que não receberam o tratamento. Porém, essas instâncias precisam ser combinadas (pareadas) para se estimar o efeito do tratamento sobre os resultados.

Entretanto, diferentemente dos estudos randomizados, nos quais os pesquisadores possuem total controle do experimento e garantem que os indivíduos tratados e não tratados são comparáveis, nos estudos observacionais os indivíduos dos grupos de tratamento e de controle não são diretamente comparáveis. Por exemplo, como comparar os efeitos do Programa Bolsa Família (PBF) sobre algum aspecto, se os indivíduos beneficiários são selecionados por um órgão governamental com base em critérios previamente definidos? Ou ainda, supondo que seja possível formar um grupo de não beneficiários do Programa Bolsa Família, como compará-los aos beneficiários do programa, uma vez que a seleção não foi aleatória e o pesquisador não teve o controle do experimento? Então, como garantir que esses grupos sejam similares?

Para exemplificar essa dissimilaridade entre os grupos de tratamento e de controle, considere a Figura 2.1 que ilustra dois grupos de indivíduos do clássico

conjunto de dados usado por Lalonde (1986), no qual se investigou o efeito de um programa de treinamento de trabalhadores.

Figura 2.1 – Exemplo de conjunto de dados utilizado em estudos observacionais.

Beneficiários

	row.names	treat	age	educ	black	hispan	married	nodegree	re74	re75	re78
111	NSW111	1	20	9	1	0	0	1	6083.99400	0.00000	8881.66500
112	NSW112	1	17	9	0	1	0	1	445.17040	74.34345	6210.67000
113	NSW113	1	20	12	1	0	0	0	989.26780	165.20770	0.00000
114	NSW114	1	18	11	1	0	0	1	858.25430	214.56360	929.88390
115	NSW115	1	27	12	1	0	1	0	3670.87200	334.04930	0.00000

Não Beneficiários

	row.names	treat	age	educ	black	hispan	married	nodegree	re74	re75	re78
300	PSID115	0	20	11	0	0	1	1	5822.94100	3532.30600	11075.56000
301	PSID116	0	29	12	0	0	1	0	14288.93000	3503.66100	8133.40700
302	PSID117	0	23	12	0	0	1	0	14347.71000	3482.17700	3818.44500
303	PSID118	0	20	11	1	0	0	1	0.00000	3480.38700	5495.66500
304	PSID119	0	42	7	0	0	1	1	4324.10200	3457.11300	9856.43600

Fonte: Elaborado pelo autor.

Os indivíduos beneficiários eram conhecidos pelo programa, mas os indivíduos não beneficiários foram obtidos a partir de outra fonte de dados. Porém, como estimar o efeito do tratamento com base no rendimento do ano de 1978 (resultado), se o pesquisador não teve o controle do experimento e os grupos não são similares?

Nesta perspectiva, Rosenbaum e Rubin (1983) introduziram o método conhecido como *Propensity Score Matching* (Seção 2.4), com o objetivo de combinar instâncias do grupo de tratamento com instâncias mais similares do grupo de controle, com base no escore de propensão, que foi definido como sendo a probabilidade condicional de atribuição de um determinado tratamento, dado um conjunto de covariáveis observadas (Seção 2.3).

É importante ressaltar que a integração em estudos observacionais não está voltada para a verificação se duas instâncias representam a mesma entidade no mundo real. Porém, nesses estudos, objetiva-se formar grupos de tratamento e de controle com instâncias similares. Neste sentido, entende-se que a integração nesses casos não seja o termo mais adequado a ser utilizado. Assim, no presente

trabalho, o termo integração foi substituído por correspondência, que melhor representa a necessidade dos pesquisadores que demandam suas pesquisas por meio de estudos observacionais.

Contudo, o escopo da presente pesquisa não abordou as estimativas do efeito do tratamento, que ocorre após a etapa de PSM. No entanto, recomenda-se a leitura do trabalho de Steiner e Cook (2013) como leitura inicial.

2.3 Escore de Propensão

O escore de propensão (*propensity score*) foi definido por Rosenbaum e Rubin (1983) como sendo a probabilidade condicional de atribuição de um determinado tratamento, dado um conjunto de covariáveis observadas. Considere N unidades i (i = 1, 2, ..., N) e z_i = 1, se a unidade i está no grupo de tratamento e, z_i = 0, se a unidade i está no grupo de controle. Assume-se também x_i ($x_i = x_1, x_2, ..., x_3$) como sendo o conjunto de covariáveis observadas para cada unidade i, na qual todos os valores de x_i são anteriores ao tratamento, assim, o escore de propensão para cada unidade i é denotado por:

$$e(x) = \text{pr}(z = 1 | x) \tag{2.1}$$

onde se supõe que dado x, z é independente:

$$\text{pr}(z_1, \dots, z_n | x_1, \dots, x_n) = \prod_{i=1}^{N} e(x_i)^{z_i} \{1 - e(x_i)\}^{1-z_i} \tag{2.2}$$

É importante ressaltar que Rosenbaum e Rubin (1983) demonstraram que o escore de propensão é um escore balanceado, ou seja, para um conjunto de unidades (instâncias) i com o mesmo valor de escore de propensão, a distribuição das covariáveis observadas é mesma entre as unidades tratadas e não tratadas. Em outras palavras, as instâncias com o mesmo valor do escore de propensão,

mantidas em dois grupos disjuntos (tratamento e controle), são similares, considerando-se as covariáveis observadas.

A estimativa do escore de propensão poder ser realizada por vários métodos (MCCAFFREY et al., 2004). Porém, a regressão logística é a técnica mais usada (WESTREICH, et al., 2010; D'AGOSTINO, 1998), a qual permite estimar o escore de propensão com bases nas covariáveis observadas (variáveis independentes). O resultado da regressão é um valor entre 0 e 1, que representa a probabilidade (propensão) de um individuo pertencer ao grupo de tratamento (variável dependente) dado suas características (covariáveis observadas).

O modelo logístico é descrito pela Equação (2.3) (STRAUSS, 1992):

$$f(z) = \frac{1}{1 + e^{-z}} \tag{2.3}$$

ou, pela equação (2.4):

$$f(z) = \frac{e^{z}}{1 + e^{z}} \tag{2.4}$$

onde $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$. Assim, substituindo z em (2.3), tem-se:

$$f(z) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}} \tag{2.5}$$

ou ainda, substituindo em (2.4), tem-se:

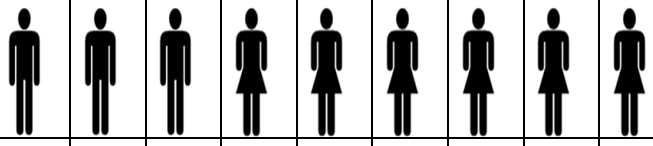
$$f(z) = \frac{e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}} \tag{2.6}$$

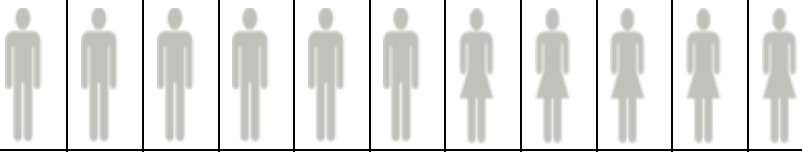
onde x_1, x_2, ..., x_n são covariáveis independentes, ou seja, as covariáveis observadas, e β_0, β_1, ..., β_n, são coeficientes de regressão estimados com base nos valores das covariáveis observadas das unidades i (ANDERSON, SWEENEY, WILLIAMS, 2003).

Para exemplificar a estimativa do escore de propensão por meio da regressão logística, considere o exemplo apresentado por Littnerova et al. (2013), nos quais os

grupos de tratamento e de controle são constituídos por indivíduos do sexo masculino e feminino. A Figura 2.2 ilustra tais indivíduos com suas respectivas idades.

Figura 2.2 – Grupo de Tratamento e de Controle.

	Grupo de Tratamento								
Sexo									
idade	49	52	61	55	58	62	62	63	66

	Grupo de Controle										
Sexo											
idade	52	52	55	61	63	93	53	62	62	63	87

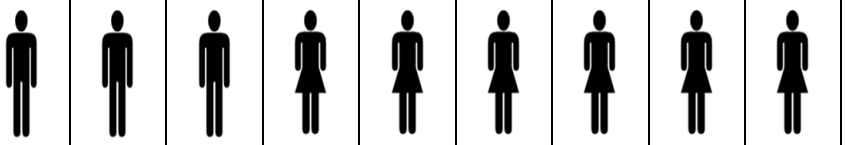
Fonte: Adaptado de LITTNEROVA et al. (2013).

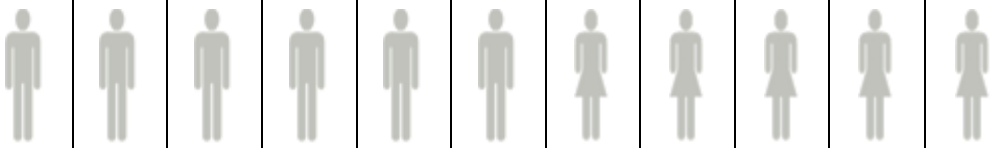
Estimando-se os coeficientes de regressão β_0, β_1 e β_2 sendo x_1 = sexo e x_2 = idade e, substituindo-os em (2.6), tem-se:

$$f(\mathrm{z} = \mathrm{tratado}) = \frac{\exp(5.60 - 1.34 * sexo - 0.08 * idade)}{1 + \exp(5.60 - 1.34 * sexo - 0.08 * idade)} \qquad (2.7)$$

Os valores do escore de propensão estimados a partir dos coeficientes de regressão e com base nos valores da covariáveis dos indivíduos são apresentados na Figura 2.3.

Figura 2.3 – Estimativa do Escore de Propensão.

	Grupo de Tratamento								
Sexo									
idade	49	52	61	55	58	62	62	63	66
Escore de Propensão	0,5062	0,4417	0,2666	0,7001	0,6430	0,5604	0,5604	0,5390	0,4743

	Grupo de Controle										
Sexo											
idade	52	52	55	61	63	93	53	62	62	63	87
Escore de Propensão	0,4417	0,4417	0,3790	0,2666	0,2342	0,0224	0,7351	0,5604	0,5604	0,5390	0,1281

Fonte: Elaborado pelo autor.

Após a estimativa do escore de propensão, deve-se estimar o efeito do tratamento. Para isso, os indivíduos do grupo de tratamento devem ser comparados com indivíduos do grupo de controle que possuem as mesmas características, ou seja, estes indivíduos devem ser similares considerando as covariáveis observadas.

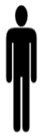
Um dos métodos que permite definir quais são os indivíduos correspondentes nos grupos de tratamento e de controle é o método PSM, o qual é descrito a seguir.

2.4 *Propensity Score Matching* (PSM)

O processo de PSM permite combinar instâncias do grupo de tratamento com instâncias do grupo de controle que compartilham um valor similar de escore de propensão (ROSENBAUM, RUBIN, 1983, 1985).

A abordagem mais comum usada no PSM é formar pares de instâncias que combine uma instância do grupo de tratamento com uma instância do grupo de controle. Para exemplificar esse processo de pareamento, considere as instâncias dos grupos de tratamento e de controle apresentados na Figura 2.3, para os quais foram estimados seus respectivos escores de propensão por meio de regressão logística. Aplicando-se o pareamento pelo algoritmo “vizinho mais próximo”, ou seja, cada instância do grupo de tratamento será combinada com uma instância do grupo de controle com o valor do escore de propensão mais próximo, tem-se o conjunto de pares conforme a Figura 2.4.

Figura 2.4 – Pareamento das instâncias dos grupos de tratamento e de controle.

Sexo														
idade	49	52		52	61		61	63		55	53		58	62
Escore Propensão	0,5062	0,4417		0,4417	0,2666		0,2666	0,2342		0,7001	0,7351		0,6430	0,5604

Sexo											
idade	62	62		62	63		63	52		66	55
Escore Propensão	0,5604	0,5604		0,5604	0,5390		0,5390	0,4417		0,4743	0,3790

Fonte: Elaborado pelo autor.

No pareamento acima duas instâncias do grupo de controle não foram correspondidas com nenhuma instância do grupo de tratamento, uma vez que o grupo de tratamento era menor que o grupo de controle. As instâncias não pareadas são de dois indivíduos, sendo um de cada sexo, com idade de 93 (masculino) e 87 (feminino) e escores de propensão de 0,0224 e 0.0128, respectivamente.

Como visto, o objetivo do processo de correspondência é encontrar o melhor balanceamento dos grupos de tratamento e de controle, ou seja, busca-se encontrar

grupos de tratamento e de controle os mais similares possíveis, considerando-se as covariáveis observadas.

Para exemplificar o balanceamento dos grupos de tratamento e de controle, antes e depois da correspondência das instâncias, considere parte (10 instâncias) do conjunto de dados Lalonde (1986) conforme a Figura 2.5. Note que as instâncias estão classificadas entre tratadas (1) e não tratadas (0) com base na coluna *treat.*

Figura 2.5 – Conjunto de Dados Lalonde (1986): parte das instâncias.

	row.names	treat	age	educ	black	hispan	married	nodegree	re74	re75	re78
1	PSID310	0	23	11	0	0	1	1	8910.7450	0.0000	4183.444
2	NSW156	1	25	13	1	0	0	0	12362.9300	3090.7320	0.000
3	PSID214	0	20	11	0	0	1	1	2547.0470	1099.2580	0.000
4	PSID97	0	24	12	1	0	1	0	8573.7520	4293.1940	0.000
5	PSID192	0	23	13	0	0	0	0	601.4949	1394.6610	4975.505
6	PSID343	0	36	11	0	0	1	1	1404.7940	0.0000	0.000
7	NSW122	1	18	10	1	0	0	1	0.0000	798.9079	9737.154
8	PSID147	0	37	14	0	0	1	0	18501.3600	2638.9350	13429.580
9	PSID268	0	16	9	1	0	0	1	0.0000	277.5000	3983.951
10	NSW54	1	25	12	1	0	0	0	0.0000	0.0000	2348.973

Fonte: Elaborado pelo autor.

A Figura 2.6 apresenta uma codificação da correspondência de instâncias, na qual foi utilizada o pacote *MatchIt* (HO et. al., 2011) da linguagem *R* (R FOUNDATION, 2016) e o método de correspondência pelo “vizinho mais próximo” (*nearest neighbor matching - NNM*) , bem como a regressão logística (*logit*) para estimar os escores de propensão.

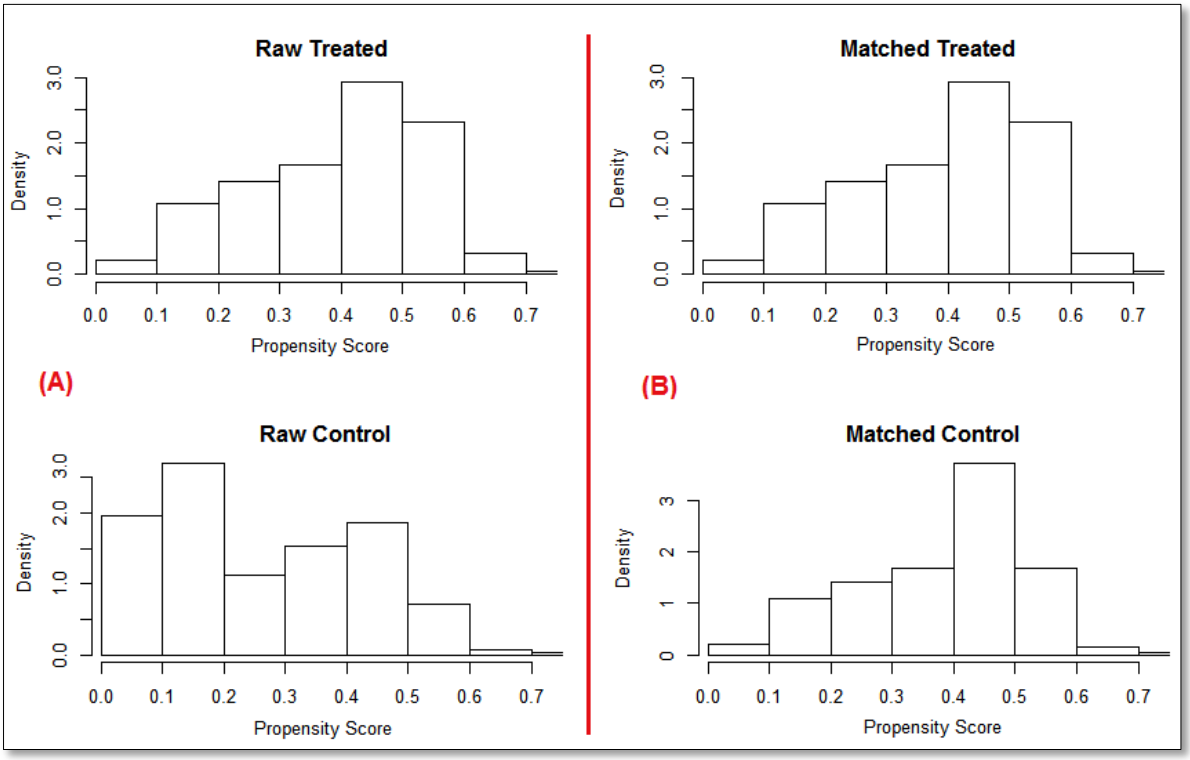
Figura 2.6 – Método de Correspondência de Instâncias.

```
# Pacote MatchIt
require(MatchIt)

# Definição do dataset
data (lalonde, package = "MatchIt")


# Correspondência pelo vizinho mais próximo (nearest neighbor matching)
# Distance: regressão logística (logit)
pairs_matched = matchit(treat ~ age + educ + married + nodegree + re74 + re75,
                        data=lalonde, method = "nearest", distance="logit")

# Plotagem dos histogramas
plot(pairs_matched, type="hist")
```

Fonte: Elaborado pelo autor.

A Figura 2.7 apresenta um resumo de balanceamento em duas partes. A parte (A) e (B) ilustram as distribuições dos escores de propensão antes e depois da correspondência, respectivamente.

Figura 2.7 – Histogramas de Balanceamento.



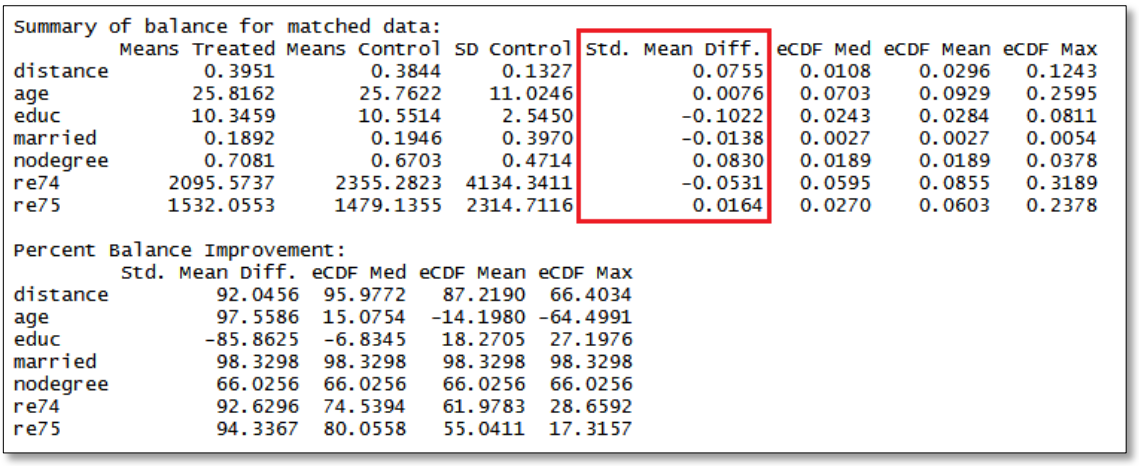
Fonte: Obtida pelo autor a partir da linguagem R (R FOUNDATION, 2016).

Os histogramas da parte (A) dos grupos de tratamento e de controle são diferentes, uma vez que esses grupos ainda não estão balanceados. Já na parte (B), os histogramas são mais similares, o que demonstra melhor balanceamento entre os grupos.

Um das medidas mais comum no diagnostico de balanceamento é a diferenças entre as médias das covariáveis do grupo de tratamento e de controle, dividido pelo desvio padrão do grupo de tratamento, a qual é conhecida como diferença normalizada da média (*standardized difference in means*) (STUART, 2010).

A Figura 2.8 apresenta a diferença da média (em destaque) para todas as covariáveis (atributos) do conjunto de dados Lalonde (1986), incluindo a média normalizada do escore de propensão (*distance*).

Figura 2.8 – Sumário do balanceamento após a correspondência.



```
Summary of balance for matched data:
         Means Treated Means Control SD Control Std. Mean Diff. eCDF Med eCDF Mean eCDF Max
distance        0.3951        0.3844     0.1327          0.0755   0.0108    0.0296   0.1243
age            25.8162       25.7622    11.0246          0.0076   0.0703    0.0929   0.2595
educ           10.3459       10.5514     2.5450         -0.1022   0.0243    0.0284   0.0811
married         0.1892        0.1946     0.3970         -0.0138   0.0027    0.0027   0.0054
nodegree        0.7081        0.6703     0.4714          0.0830   0.0189    0.0189   0.0378
re74         2095.5737     2355.2823  4134.3411         -0.0531   0.0595    0.0855   0.3189
re75         1532.0553     1479.1355  2314.7116          0.0164   0.0270    0.0603   0.2378

Percent Balance Improvement:
         Std. Mean Diff. eCDF Med eCDF Mean eCDF Max
distance         92.0456  95.9772   87.2190  66.4034
age              97.5586  15.0754  -14.1980 -64.4991
educ            -85.8625  -6.8345   18.2705  27.1976
married          98.3298  98.3298   98.3298  98.3298
nodegree         66.0256  66.0256   66.0256  66.0256
re74             92.6296  74.5394   61.9783  28.6592
re75             94.3367  80.0558   55.0411  17.3157
```

Fonte: Elaborado pelo autor.

Segundo Stuart (2010), a definição do melhor balanceamento é complexa. No entanto, em relação à diferença normalizada da média pode-se selecionar um método de correspondência que produz a menor diferença considerando a maior quantidade de covariáveis, bem como um método que minimiza grandes diferenças (maior do que 0,25).

Como visto, ao combinar instâncias dos grupos de tratamento e de controle, objetiva-se formar grupos de instâncias similares com base no escore de propensão, uma vez que instâncias que possuem valores aproximados (ou exatos) do escore de propensão são mais semelhantes entre si do que instâncias com valores de escore de propensão não pareados (AUSTIN, 2014).

2.5 *Ensembles*

Um *Ensemble* é um conjunto de modelos, classificadores ou regressores, que são combinados de alguma maneira para predizer um novo caso, ou seja, um novo exemplo (DIETTERICH, 2000). Se um *ensemble* emite uma predição de um valor discreto (classe), por exemplo, ‘jogar’ ou ‘não jogar’, o problema é considerado de classificação. Porém, se a predição foi um valor contínuo, o problema é considerado de regressão.

Ao construir um *Ensemble* objetiva-se que a combinação de vários modelos possa melhorar a predição, com base nas respostas dos membros que o compõe, formando assim, um comitê.

Nesta tese, investigou-se como os *ensembles* que manipulam o conjunto de treinamento, destacando-se: *bagging* (BREIMAN, 1996) e *boosting* (FREUND, SCHAPIRE, 1997) e, os que manipulam as características de entrada, mas precisamente, *random forest* (BREIMAN, 2001), podem substituir a regressão logística para estimar o escore de propensão das instâncias em uma abordagem no contexto do processo de PSM. Dessa forma, os *ensembles* nesta tese foram utilizados para emitirem um valor continuo no intervalo [0 1] para cada instância, os quais foram utilizados como sendo os escores de propensão das instâncias dos grupos de tratamento e de controle. Em outras palavras, o problema pelo qual os *ensembles* estão sendo utilizados pode ser considerado de regressão.

2.5.1 *Bootstrap aggregating (bagging)*

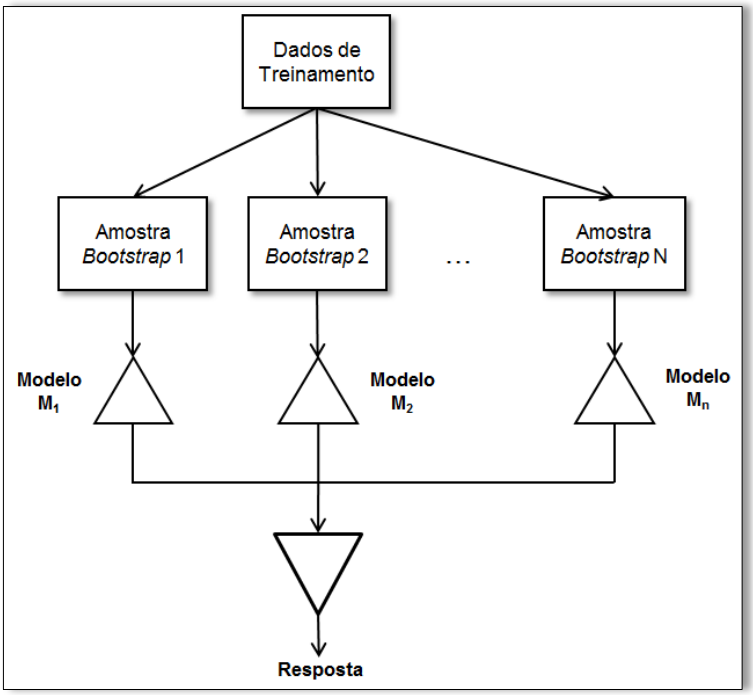
***B**ootstrap **agg**regati**ng** (bagging)* (BREIMAN, 1996) é um *ensemble* utilizado para agregar várias predições de modelos independentes que são treinados a partir de amostras *bootstrap* (EFRON, TIBSHIRANI, 1993), as quais são obtidas a partir do conjunto de treinamento.

Cada amostra *bootstrap* é utilizada para treinar um modelo e a resposta do conjunto é emitida com base na combinação das respostas dos modelos individuais. Em geral, a decisão final é obtida por votação majoritária de classes no caso de classificação, ou pela média dos valores no caso de regressão.

É importante ressaltar que as instâncias (casos) do conjunto de dados possuem a mesma probabilidade de serem incluídas em alguma amostra (i.e. possuem o mesmo peso). No entanto, como as amostras são obtidas com substituição, algumas instâncias podem ser incluídas em mais de uma amostra, enquanto que outras podem ser omitidas do conjunto de treinamento (DIETTERICH, 2000).

A Figura 2.9 ilustra de forma geral o funcionamento do *ensemble bagging.*

Figura 2.9 – Visão geral do *ensemble bagging*.



Fonte: Autoria do Autor.

O algoritmo do *ensemble bagging* que foi adaptado a partir Kuncheva (2004) é apresentado a seguir.

Algoritmo 2.1 – *Bagging.*

Fase de Treinamento:

1. Inicialize os parâmetros:
   D = 0, *ensemble.*
   L, número de modelo (classificador ou regressor).

2. Para k = 1..., L
   Obtenha uma amostra *bootstrap* S_k a partir de Z (conjunto de dados treinamento).
   Construa um modelo (classificador ou regressor) D_k usando S_k como o conjunto de treinamento.
   Adicione o modelo ao *ensemble*, $D = D\ U\ D_k$

3. Retorne D (*ensemble*).

Fase de Teste:

4. Execute D_1, ..., D_L para uma entrada x

5. No caso de classificação, escolha a classe com maior número de votos como sendo a classe de x ou, no caso de regressão, faça a média com base nos valores D_1, ..., D_L para x.

Fim.

Na fase de treinamento, o algoritmo consiste em gerar S_k amostras *bootstrap* que são construídas randomicamente a partir dos dados de treinamento *Z*. Para cada amostra S_k *é* construído um modelo que é adicionado ao *ensemble.* Na fase de teste, para uma dada instância x, é emitida uma resposta do *ensemble* com base nas respostas individuais dos modelos. Assim, no caso de classificação, é escolhida a classe com maior número de votos e, no caso de regressão, a média dos valores emitidos pelos modelos.

2.5.2 *Boosting*

Boosting é um método similar ao bagging no sentido de criar um *ensemble* com base nas predições de modelos independentes (classificadores ou regressores), porém é um processo iterativo, no qual em cada iteração é criado um modelo com base em uma amostra de treinamento e as instâncias preditas de forma errada possuem os pesos aumentados para a próxima iteração do processo.

Dois dos primeiros algoritmos efetivos de *boosting* foram propostos por Schapire (1990) e Freund (1995). Esses autores, no mesmo ano de 1995, também propuseram o algoritmo "Adaboost" (***Ada****ptive* ***Boost****ing)* em uma conferência europeia (FREUND, SCHAPIRE, 1995). Porém, a citação mais referenciada desse trabalho é de 1997 (FREUND, SCHAPIRE, 1997). Além disso, outras versões do AdaBoost também foram propostas pelos autores, destacando-se os algoritmos AdaBoost.M1, AdaBoost.M2 (FREUND, SCHAPIRE, 1996), entre outros.

Nesta tese foi utilizado algoritmo *Gradiente Boost* (FRIEDMAN, 2001), que se baseia na minimização de uma função de custo (perda) e utiliza o método de otimização por descida do gradiente, conforme descrito seguir.

Algoritmo 2.2 – *Gradient Boosting* (FRIEDMAN, 2001).

Entrada:

Conjunto de dados de entrada: $(x, y)_{i=1}^{N}$
Número de iterações: T
Função de custo (perda): $\Psi(y_i, f)$

Início

Inicialize $\hat{f}(x)$ com uma constante,

$$\hat{f}(x) \leftarrow \arg min_{f(x)} \sum_{i=1}^{N} \Psi(y_i, f(x)) \tag{2.8}$$

Para t = 1 ..., T

1. Calcule o gradiente negativo

$$z_i = -\frac{\partial}{\partial f(\mathrm{x}_i)} \Psi(y_i, f(\mathrm{x}_i))|_{f(\mathrm{x}_i)=\hat{f}(\mathrm{x}_i)} \tag{2.9}$$

2. Ajuste um modelo de regressão, $g(\mathrm{x})$, predizendo z_i a partir das covariáveis x_i

3. Escolha um gradiente de descida como:

$$\rho = arg \min_{\rho} \sum_{i=1}^{N} \Psi(y_i, \hat{f}(\mathrm{x}_i) + \rho g(\mathrm{x}_i)) \tag{2.10}$$

4. Atualize a função $\hat{f}(\mathrm{x})$:

$$\hat{f}(\mathrm{x}) \leftarrow \hat{f}(\mathrm{x}) + \rho g(\mathrm{x}) \tag{2.11}$$

Fim.

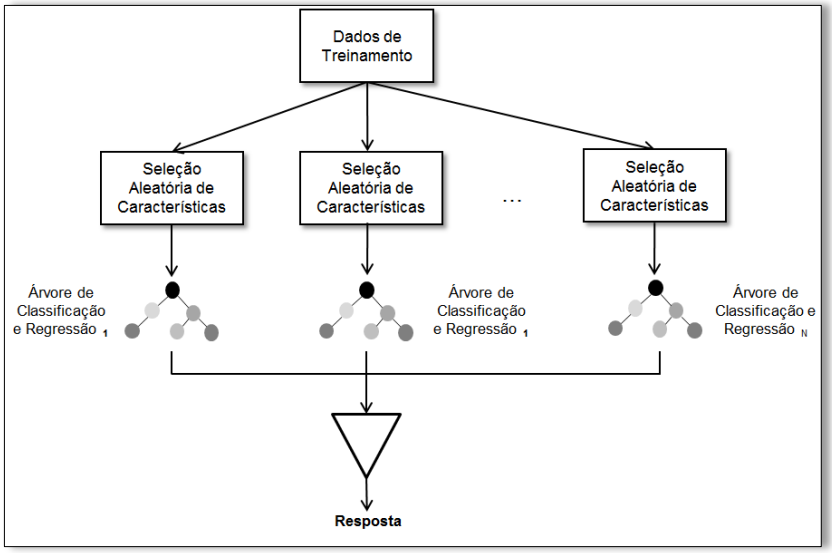
É importante ressaltar que o algoritmo *Gradient Boosting* foi escolhido por ser o mesmo algoritmo utilizado no trabalho de Lee et al. (2010), no qual utilizou-se uma abordagem que ponderou as instâncias utilizando como pesos os escores de estimados por *ensembles* de regressores.

2.5.3 *Random Forest*

O método *random forest* (floresta aleatória) foi desenvolvido por Breiman (2001) para combinar vários modelos, especificamente árvore de classificação e/ou regressão, com base em vetores de características (covariáveis), os quais são gerados de maneira aleatória e independente a partir do conjunto de dados.

Segundo Breiman (2001), as florestas aleatórias podem diferenciar-se na maneira que os vetores de características são utilizados na geração das árvores. Na abordagem conhecida como *Forest-RI* (*random input*) um vetor de características é selecionado de forma aleatória a partir dos dados de treinamento e, a árvore é construída considerando-se somente as características já selecionadas, ou seja, aquelas presentes no vetor de características. A Figura 2.10 ilustra de maneira geral esse processo.

Figura 2.10 – *Random Forest.*

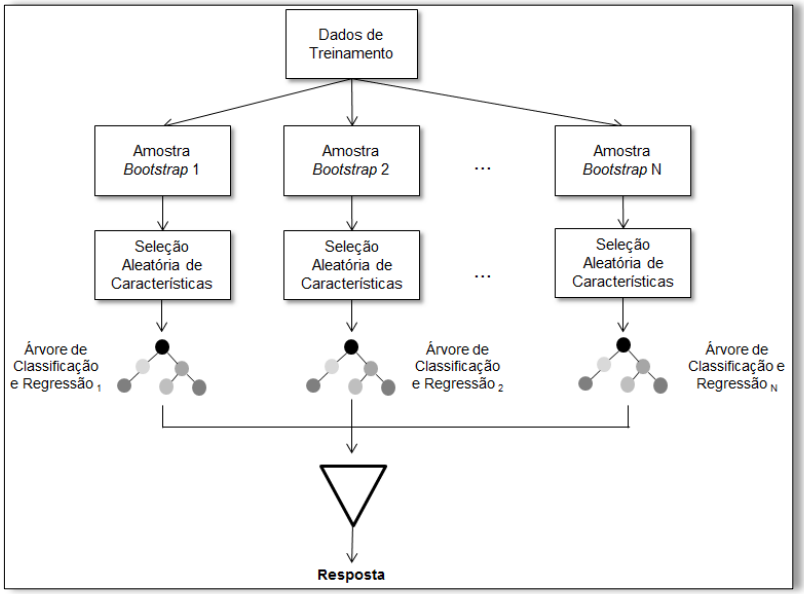


Fonte: Autoria do Autor.

A abordagem *Forest-RI* pode ser utilizada quando o número de características é suficiente grande. No entanto, quando esse número é pequeno, a seleção aleatória se torna difícil uma vez que não existem muitas características para serem selecionadas. Nesses casos, em cada nó da árvore gera-se uma combinação linear de características a partir dos dados de treinamento. Assim, nesta abordagem, chamada de *Forest-RC* (*random combination*), ao invés de considerar apenas um

único vetor de características em cada árvore, como na abordagem *Forest-RI*, consideram-se várias combinações de características geradas em cada nó. É importante ressaltar que Breiman (2001) sugere o uso de amostras *bootstrap* com *random forest* conforme ilustrado na Figura 2.11.

Figura 2.11 – *Random Forest* com amostras *bootstrap*.



Fonte: Elaborado pelo autor.

O algoritmo *Random Forest* para classificação e regressão com amostras *bootstrap* é o seguinte (LIAW, WIENER, 2002).

Algoritmo 2.3 – *Random Forest* com amostras *bootstrap*.

Entrada:

n_{Tree} (número de árvores) e conjunto de dados.

Início

1. Obtenha n_{Tree} amostras *bootstrap* a partir do conjunto de dados original.
2. Para cada amostra, desenvolva um modelo (uma árvore de classificação ou regressão), com a seguinte modificação: ao invés de escolher a melhor divisão entre todas as características (covariáveis), faça uma amostragem aleatória m_{try} e escolha a melhor divisão entre elas.
3. Faça a predição de novos dados agregando as predições das árvores n_{Tree} (i.e. a maioria dos votos no caso de classificação e a média no caso de regressão).

Fim.

As amostras *bootstrap* são obtidas a partir do conjunto de dados e, em seguida, as árvores de classificação ou regressão são geradas com seleção aleatória das características em cada amostra. Por último, as árvores são combinadas para emitir a predição do *ensemble*, seja por meio de uma classe, quando utilizado como classificador ou, um valor, quando utilizado como regressor.

Como visto, as florestas aleatórias permitem a criação de árvore de classificação e regressão com entradas aleatórias. Segundo Breiman (2001), a aleatoriedade introduzida garante que as árvores sejam minimamente correlacionadas, mantendo a força do conjunto. Por fim, ainda segundo Breiman (2001), *random forest* possuem boa precisão e são relativamente robustas em relação à *outlines* e ruídos, mais rápidas que *bagging* e *boosting*, além de serem simples e facilmente paralelizadas.

2.6 Considerações Finais

O presente capítulo descreveu a fundamentação teórica necessária para o desenvolvimento da pesquisa, destacando-se, dentre os assuntos abordados, os tópicos sobre integração de dados, escore de propensão, PSM e *ensembles* de regressores, uma vez que esses conceitos foram fundamentais para a elaboração da abordagem para correspondência de instâncias "*SEnsembles*".

Nesta tese, foram utilizadas as implementações dos *ensembles* disponíveis em pacotes da linguagem *R* (R FOUNDATION, 2016). Assim, foi utilizado o pacote *ipred* (*bagging*), *gbm* (*boosting*) e *randomforest* (*Random Forest*).

A seguir apresenta-se o Capítulo 3 o qual aborda a revisão da literatura separada em duas seções principais: métodos de correspondência de instâncias e métodos de aprendizagem de máquina aplicados aos estudos observacionais, mas com enfoque principal voltado para *ensembles* de regressores.

Capítulo 3

TRABALHOS CORRELATOS

3.1 Considerações Iniciais

O presente capítulo aborda a revisão da literatura separando os trabalhos correlatos em dois assuntos principais. No primeiro grupo, (Seção 3.2) são descritos os principais métodos para se efetuar correspondência de instâncias que estão relacionados ao presente trabalho e ao contexto do processo de PSM. Já no segundo grupo (3.3) são descritos os trabalhos que aplicam aprendizado de máquina em estudos observacionais, com enfoque nos trabalhos que utilizam *ensembles* de regressores para estimar o escore de propensão (3.2.1) ou que utilizam *ensembles* no processo de PSM (3.3.2).

3.2 Métodos para Correspondência de Instâncias

Esta seção descreve os principais métodos utilizados para se efetuar correspondências (*matching*) de instâncias no contexto do processo de PSM, destacando-se: *exact matching*, *subclassification* e *NNM*, os quais serão exemplificados por meio do pacote *MathIt* (HO et al., 2011) da linguagem *R* (R FOUNDATION, 2016) e o conjunto de dados Lalonde (1986).

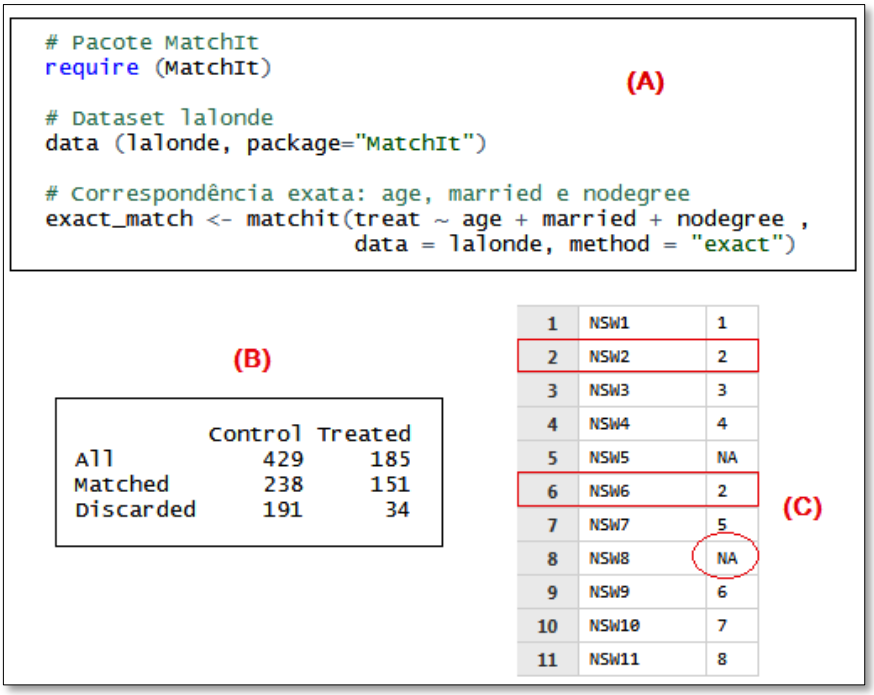
3.2.1 *Exact Matching*

O método *exact matching* permite a criação de subclasses contendo instâncias de ambos os grupos de tratamento e de controle com valores iguais para as covariáveis selecionadas, ou seja, uma subclasse conterá somente instâncias que possuem valores idênticos para todas as covariáveis.

A quantidade de subclasses a serem criadas depende da quantidade de covariáveis selecionadas e também dos seus respectivos valores. Por exemplo, considere que em um processo de correspondência exata seja selecionada somente a covariável sexo, a qual possui dois valores distintos: masculino e feminino. Neste caso, após a correspondência das instâncias, serão formadas duas subclasses, uma contendo todas as instâncias contendo pessoas do sexo masculino e outra com pessoas do sexo feminino. É importante ressaltar que um número elevado de covariáveis selecionadas pode aumentar a quantidade de instâncias que não possuem correspondentes, uma vez que se torna mais difícil encontrar duas instâncias iguais.

A Figura 3.1 ilustra um processo de correspondência exata com três covariáveis selecionadas do conjunto de dados Lalonde (1986): *age*, *married* e *nodegree* (parte A).

Figura 3.1 – Correspondência com Exact Matching.



Fonte: Elaborado pelo autor.

Na parte (B) é possível visualizar o total de instâncias que foram combinadas, ou seja, que possuem correspondentes e, também, o total de instâncias não combinadas. Já a parte (C) apresenta uma relação de instâncias e suas respectivas subclasses. As instâncias da linha 2 e 6 foram correspondidas e elas estão na mesma subclasse (subclasse 2), enquanto que a instância da linha 8 foi descartada, pois não há nenhuma outra instância idêntica a ela, ou seja, não possui nenhum correspondente.

Como visto, a correspondência exata implica em formar subclasses de instâncias idênticas com base nas covariáveis selecionadas. Porém, quando há muitas covariáveis pode ocorrer no aumento da quantidade de instâncias não correspondidas, o que é uma desvantagem desse método.

Na presente pesquisa o método *exact matching* foi introduzido em um dos processos da abordagem "*SEnsembles*" para separar instâncias exatamente idênticas, ou seja, instâncias com características iguais. Esse processo será detalhado na descrição da abordagem "*SEnsembles*" no Capítulo 5.

3.2.2 *Subclassification*

O método subclassificação (COCHRAN, 1968) permite criar subclasses (estratos) de instâncias que possuem a distribuição dos valores das covariáveis o mais próximo possível, ou seja, as instâncias são alocadas em subclasses com base em uma distância, por exemplo, o escore de propensão, de maneira que as instâncias que possuem valores aproximados sejam alocadas na mesma subclasse. A Figura 3.2 apresenta um exemplo desse método com 6 subclasses. Além disso, foi utilizado o escore de propensão para estimar a distância entre as instâncias (parte A). A parte (B) da figura apresenta as subclasses com as respectivas quantidades de instâncias.

Figura 3.2 – Correspondência com Subclassificação.

(A)

```
# Subclassficação
# Covariáveis selecionadas: age, hispan, educ e black

subclass_match <- matchit(treat ~ age + hispan + educ + black, data = lalonde,
                          distance = "logit", method = "subclass", subclass=6)
# Resumo:
summary (subclass_match)
```

(B)

```
 Sample sizes by subclasses:
         Subclass 1 Subclass 2 Subclass 3 Subclass 4 Subclass 5 Subclass 6
 Treated         31         31         30         27         33         33
 Control        347         19         11          9         12         31
 Total          378         50         41         36         45         64
```

Fonte: Elaborado pelo autor.

Entretanto, qual é o número ideal de subclasses? Segundo Rosenbaum e Rubin (1985), a criação de 5 subclasses remove ao menos 90% do viés no efeito do tratamento, pois todas as covariáveis são utilizadas na estimativa do escore de propensão. Segundo Stuart (2010), o número usual de subclasses é entre 5 e 10, mas Lunceford e Davidian (2004) afirma que esse número deve ser maior, entre 10 e 20, em amostras maiores.

O método mais sofisticado de subclassificação é denominado *full matching*, o qual seleciona a quantidade de classes automaticamente de forma a permitir a menor distância entre as instâncias que compõe cada classe. Para isso, em cada classe haverá apenas um instância do grupo de controle e um ou mais instâncias do grupo de tratamento ou, uma instâncias do grupo de controle e um ou mais do grupo de tratamento (ROSENBAUM, 1991).

A Figura 3.3 ilustra uma subclassificação utilizando o método *full matching* (parte A). Na parte (B) visualiza-se 3 subclasses (54, 43, 65). As subclasses 43 e 54 foram formadas com uma instância do grupo de tratamento, enquanto que a subclasse 65 foi constituída com duas instâncias desse grupo. Porém, observa-se que os escores de propensão entre as instâncias em cada subclasse são bem próximos.

Figura 3.3 – *Full Matching.*

```
# Subclassificação - Full Matching:

# Covariáveis selecionadas: age, hispan, educ e black

subclass_match <- matchit(treat ~ age + hispan + black + educ , data = lalonde,
                          distance = "logit", method = "full")
```

(A)

	row.names	treat	age	educ	black	hispan	distance	subclass
1	NSW9	1	22	16	1	0	0.71310111	54
2	PSID330	0	22	16	1	0	0.71310111	54
3	PSID69	0	30	17	1	0	0.70578731	54
4	PSID15	0	22	14	1	0	0.69273510	43
5	PSID217	0	32	16	1	0	0.69107887	43
6	NSW75	1	28	15	1	0	0.68965784	43
7	PSID234	0	19	13	1	0	0.68906615	43
8	NSW182	1	25	14	1	0	0.68596880	65
9	NSW12	1	21	13	1	0	0.68453500	65
10	PSID381	0	21	13	1	0	0.68453500	65

(B)

Fonte: Elaborado pelo autor.

Como visto, o método de subclassificação permite criar subclasses com instâncias de ambos os grupos de tratamento e de controle, mas com destaque para o método *full matching*, que minimiza as distâncias entre as instâncias em cada subclasse. Por fim, Steiner e Cook (2013) afirmam que uma das maiores vantagens da subclassificação é a facilidade de implementação em *softwares* estatísticos.

3.2.3 *Nearest Neighbor Matching (NNM)*

O método de correspondência de instâncias conhecido como “vizinho mais próximo” (*Nearest Neighbor Matching - NNM*) (RUBIN, 1973) permite que cada instância do grupo de tratamento seja combinada ao menos com uma instância do grupo de controle com base na menor distância entre elas. A distância entre as instâncias pode ser estimada por vários métodos, mas o escore de propensão estimado pela regressão logística é um das medidas mais utilizadas.

O método *NNM* inicia-se selecionando uma instância do grupo de tratamento e, para esta instância, seleciona-se uma instância do grupo de controle mais similar, ou seja, com a menor distância. A seleção das instâncias do grupo de tratamento pode ocorrer de forma aleatória ou ordenada (ascendente ou descendente) e deve ser informada como entrada para o método de correspondência.

A abordagem mais simples do método *NNM* é a formação de pares de instâncias 1:1, ou seja, cada par é composto por uma instância do grupo de tratamento e uma instância do grupo de controle. A Figura 3.4 ilustra essa abordagem, na qual utilizou o escore de propensão e seleção aleatória das instâncias do grupo de tratamento (parte A). Na parte (B) observa-se que todas as instâncias do grupo de tratamento foram combinadas (pareadas) com instâncias do grupo de controle. No entanto, 244 instâncias do grupo de controle foram descartadas, pois não foram combinadas com nenhuma instância do grupo de tratamento.

Figura 3.4 – Correspondência com *NNM.*

```
# Correspondência pelo vizinho mais próximo:

nearest_match <- matchit(treat ~ age + hispan + black+ educ + re74 + re75,
                         distance="logit", data = lalonde, method = "nearest")

# Resumo:
summary(nearest_match)
```

(A)

```
 Sample sizes:
           Control Treated
 All           429     185
 Matched       185     185
 Unmatched     244       0
 Discarded       0       0
```

(B)

Fonte: Elaborado pelo autor.

A Figura 3.5 ilustra quatro pares de instâncias resultantes do método de correspondência aplicado. Os pares são compostos por duas instâncias, sendo uma de cada grupo (tratamento e controle), as quais possuem os valores mais aproximados (menor distância) do escore de propensão (*distance*).

Figura 3.5 – Resultado da correspondência pelo método *NNM*.

	row.names	treat	age	educ	black	hispan	married	nodegree	re74	re75	re78	distance
1	PSID15	0	22	14	1	0	1	0	748.43990	11105.37000	18208.55000	0.78942058
2	NSW9	1	22	16	1	0	0	0	0.00000	0.00000	2164.02200	0.76907159
3	NSW75	1	28	15	1	0	0	0	0.00000	0.00000	9598.54100	0.75458684
4	PSID48	0	25	12	1	0	1	0	295.84930	6942.87100	461.05070	0.74597322
5	NSW121	1	29	14	1	0	0	0	0.00000	679.67340	17814.98000	0.74382019
6	PSID399	0	27	14	1	0	0	0	0.00000	0.00000	10122.43000	0.74004346
7	NSW178	1	33	11	1	0	0	1	0.00000	7867.91600	6281.43300	0.74060784
8	PSID330	0	22	16	1	0	0	0	2564.68000	0.00000	116.74040	0.72906574

Fonte: Elaborado pelo autor.

Como visto, nas correspondências acima houve um grande descarte de instâncias do grupo de controle (244), mas, segundo (STUART, 2010), esse descarte não diminui a precisão, uma vez que as instâncias mais similares de ambos os grupos foram combinadas.

Em uma segunda abordagem do método de correspondência pelo vizinho mais próximo considera-se correspondência do tipo k:1, na qual uma instância do grupo de tratamento é combinada com K instâncias do grupo de controle. A Figura 3.6 ilustra essa abordagem com K igual a 2 instâncias.

Figura 3.6 – Correspondência *NNM* com K=2.

```
# Correspondência pelo vizinho mais próximo (K=2:1):

nearest_match <- matchit(treat ~ age + hispan + black+ educ + re74 + re75,
                 distance="logit", data = lalonde, method = "nearest", ratio = 2)

# Resumo:
summary(nearest_match)
```

(A)

```
Sample sizes:
           Control Treated
All            429     185
Matched        370     185
Unmatched       59       0
Discarded        0       0
```

(B)

	row.names	1	2
1	NSW1	PSID134	PSID214
2	NSW2	PSID151	PSID57
3	NSW3	PSID187	PSID21
4	NSW4	PSID253	PSID109
5	NSW5	PSID334	PSID315
6	NSW6	PSID188	PSID115
7	NSW7	PSID375	PSID105
8	NSW8	PSID157	PSID180
9	NSW9	PSID15	PSID395
10	NSW10	PSID361	PSID168

(C)

Fonte: Elaborado pelo autor.

Na parte (A) observa-se que foram consideradas 2 instâncias do grupo de controle (*ratio* = 2) para cada instância do grupo de tratamento. Na parte (B) visualiza-se que a quantidade de descarte (59) foi menor em relação à correspondência *NNM* do tipo 1:1 (244). E, na parte (C), visualizam-se as instâncias correspondidas.

No exemplo acima, a quantidade de instâncias do grupo de controle foi suficiente para atender a quantidade mínima estipulada para a correspondência (k = 2). No entanto, quando a quantidade é insuficiente pode-se adotar uma estratégia com substituição, na qual uma instância do grupo de controle pode ser correspondida com mais do que uma instância do grupo de tratamento.

A correspondência com substituição permite que as instâncias do grupo de controle que são similares a muitas instâncias do grupo de tratamento possam ser correspondidas mais do que uma vez. Porém, é necessário monitorar a quantidade, pois pode ocorrer que somente uma pequena quantidade dessas instâncias seja selecionada (DEHEJIA, WAHBA, 1999). A Figura 3.7 ilustra uma correspondência considerando 4:1 com substituição (*replace* = *TRUE*) (parte A).

Figura 3.7 – Correspondência com Substituição.

```
# Correspondência pelo vizinho mais próximo (K=4:1) com substituição:

nearest_match <- matchit(treat ~ age + hispan + black+ educ + re74 + re75,
                 distance="logit", data = lalonde, method = "nearest",
                 ratio = 4, replace=TRUE)

# Resumo:
summary(nearest_match)
```

(A)

```
Sample sizes:
          Control Treated
All           429     185
Matched       166     185
Unmatched     263       0
Discarded       0       0
```

(B)

	row.names	1	2	3	4
1	NSW1	PSID368	PSID408	PSID375	PSID191
2	NSW2	PSID366	PSID391	PSID136	PSID31
3	NSW3	PSID293	PSID380	PSID416	PSID386
4	NSW4	PSID388	PSID389	PSID374	PSID368
5	NSW5	PSID400	PSID376	PSID398	PSID110
6	NSW6	PSID372	PSID140	PSID370	PSID277
7	NSW7	PSID386	PSID373	PSID293	PSID380

(C)

Fonte: Elaborado pelo autor.

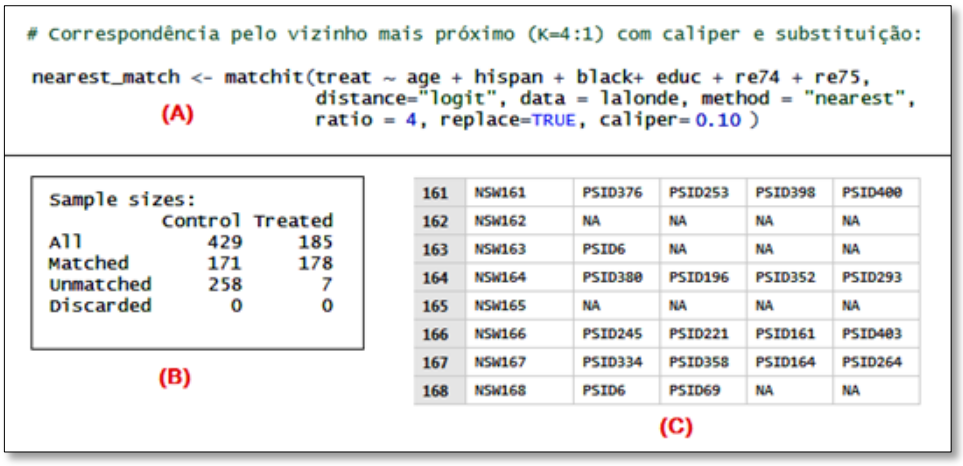
Na parte (B) da figura acima observa-se que, mesmo utilizando uma estratégia com substituição, apenas 166 instâncias do grupo de controle foram correspondidas, ou seja, a maioria das instâncias (233) foi descartada, o que demostra que há instâncias do grupo de controle que são mais similares às instâncias do grupo de tratamento. Essa similaridade pode ser observada na parte (C), na qual as instâncias das linhas 3 e 7 compartilham 3 instâncias do grupo de controle.

É importante ressaltar que a correspondência K:1 sem restrições podem levar a combinações ruins, ou seja, “pobres”, em situações que não há instâncias do grupo de controle similares (próximos) a uma instância do grupo de tratamento, ou seja, as instâncias do grupo de controle possuem valores muito distantes do escore de propensão de uma instância do grupo de tratamento (STUART, 2010). Para minimizar esse problema pode-se adotar uma distância máxima (*caliper*) como limite para a combinação (COCHRAN, RUBIN, 1973). Assim, para cada instância do grupo de tratamento será selecionada somente instâncias do grupo de controle que contemplem a distância máxima estipulada.

Segundo Rosenbaum (2010), o valor mais usual para o *caliper* é 20% sobre o desvio padrão dos valores do escore de propensão, mas pode ser muito em alguns casos. Por isso, o ideal é iniciar o *caliper* com 20% e reduzi-lo à medida que se avalia o balanceamento do escore de propensão.

A parte (A) da Figura 3.8 ilustra uma correspondência com *caliper* igual a 10% sobre o desvio padrão dos valores do escore de propensão. Já na parte (B) nota-se que 7 instâncias do grupo de tratamento não foram correspondidas com nenhuma instância do grupo de controle, pois não houve nenhuma instância do grupo de controle que estivesse comtemplando a distância mínima imposta pelo *caliper*, como aconteceu com as instâncias das linhas 162 e 165 da parte (C). Além disso, pelo mesmo limite imposto, as instâncias das linhas 163 e 168 foram correspondidas com menos instâncias do que as demais.

Figura 3.8 – Correspondência com Caliper e Substituição.



```
# Correspondência pelo vizinho mais próximo (K=4:1) com caliper e substituição:

nearest_match <- matchit(treat ~ age + hispan + black+ educ + re74 + re75,
                         distance="logit", data = lalonde, method = "nearest",
                         ratio = 4, replace=TRUE, caliper= 0.10 )
```

(A)

```
Sample sizes:
          Control Treated
All           429     185
Matched       171     178
Unmatched     258       7
Discarded       0       0
```

(B)

161	NSW161	PSID376	PSID253	PSID398	PSID400
162	NSW162	NA	NA	NA	NA
163	NSW163	PSID6	NA	NA	NA
164	NSW164	PSID380	PSID196	PSID352	PSID293
165	NSW165	NA	NA	NA	NA
166	NSW166	PSID245	PSID221	PSID161	PSID403
167	NSW167	PSID334	PSID358	PSID164	PSID264
168	NSW168	PSID6	PSID69	NA	NA

(C)

Fonte: Elaborado pelo autor.

A ordem em que as instâncias do grupo de tratamento são correspondidas pode alterar a qualidade da correspondência (STUART, 2010). Para minimizar esse problema, Rosenbaum (1989) propôs o método *Optimal Matching* com objetivo de minimizar a distância global entre as instâncias correspondidas. Porém, segundo Gu e Rosenbaum (1993), *Optimal Matching* reduz a distância entre os pares, mas não é superior para obter o melhor balanceamento entre os grupos de tratamento e de controle. Assim, se o objetivo é obter o melhor balanceamento entre os grupos, *greedy matching* pode ser suficiente (STUART, 2010), como o *NNM*.

Como visto, a correspondência pelo vizinho mais próximo exige algumas definições importantes, que modificam o resultado da correspondência, tais como:

- Qual a medida de distância a ser utilizada?
- Qual é o método de seleção das instâncias do grupo de tratamento?
- Cada instância do grupo de tratamento será correspondida com quantas instâncias do grupo de controle?
- Cada instância do grupo de controle será correspondida com quantas do grupo de tratamento?
- Será adotado algum *caliper*? Qual?

Por fim, ressalta-se que uma única execução do método *NNM* não é praticada, uma vez que é possível aplicar variações com intuito de encontrar a melhor estratégia para um determinado estudo (para um conjunto de dados). Já na presente pesquisa, o método *NNM* é utilizado nas duas principais estratégias da "*SEnsembles*", sempre com correspondência 1:1 sem substituição, pois é uma das maneiras mais usadas e também mais simples de efetuar a correspondência das instâncias.

3.2.4 Comparações dos métodos para correspondência de instância

Recentemente, Austin (2014) comparou dois métodos de correspondência de instância, *Optimal Matching* Rosenbaum (1989) e *NNM* (RUBIN, 1973), para formar pares de instâncias com base nos valores dos escores de propensão. Em relação ao método *NNM* considerou-se estratégias com e sem substituição, *caliper* e quatro diferentes maneiras para seleção das instâncias do grupo de tratamento.

Segundo Austin (2014), a comparação realizada foi a primeira que buscou comparar métodos com substituição com outros métodos comumente utilizados, como o *NNM* sem substituição e *Optimal Matching.* Como resultado, o autor sugere que para muitas situações, o método *NNM* com *caliper* e sem substituição pode ser usado para formar pares de instância dos grupos de tratamentos e de controle com valores similares do escore de propensão. Além disso, o autor desencorajou o uso de *NNM* com substituição, pois não resultou em estimativa com menor viés se comparado com os melhores desempenhos do método *NNM* sem substituição e com *caliper.*

Adicionalmente, um dos resultados obtidos por Austin (2014) ao comparar *Optimal Matching* e *NNM*, também já havia sido encontrado por Gu e Rosenbaum (1993), no qual esses métodos obtiveram desempenho similar ao tentar balancear as covariáveis. Esse resultado, também foi encontrado em um trabalho recente de Austin e Small (2014), o que sugere nenhuma outra investigação futura neste sentido.

Contudo, o trabalho de Austin (2014), ao comparar os métodos com base no escore de propensão e, ao sugerir o método de *NNM* com *caliper* e sem substituição, provê um direcionamento da melhor forma para aplicar o PSM para a formação de pares de instâncias de ambos os grupos de tratamento e de controle.

Porém, ressalta-se que o uso de *caliper* limita o número de pares de instâncias correspondidas, o que em muitos casos, pode não ser desejado.

Na presente pesquisa foram adotadas as sugestões descritas no trabalho de Austin (2014), com a elaboração de uma abordagem no contexto do processo de PSM que considere a formação de pares de instância sem substituição e, provendo o uso de *calipers* se desejado pelo usuário especialista.

A seguir serão descritos os principais trabalhos que aplicam os métodos de aprendizado de máquina no contexto de estudos observacionais, mas com enfoque em trabalhos que utilizam *ensembles* de regressores para estimar o escore de propensão ou que utilizam *ensembles* no processo de PSM.

3.3 Aprendizado de Máquina em Estudos Observacionais

A presente seção aborda os principais trabalhos que aplicam os métodos de aprendizagem de máquina no contexto dos estudos observacionais, mais precisamente, o uso de *ensembles* de regressores para estimar o escore de propensão.

Inicialmente, a pesquisa começou a ser delineada a partir do trabalho de Westreich et al. (2010), o qual comparou quatro alternativas à regressão logística para estimar o escore de propensão: redes neurais, árvores de regressão, *support vector machine* e *boosting*. Nesse trabalho, os autores concluíram que *boosting* mostrava-se promissor no contexto de PSA (*Propensity Score Analysis*).

Em seguida, pesquisou-se na literatura os trabalhos correlatos ao trabalho de Westreich et al. (2010), sendo encontrado o trabalho de Setoguchi et al. (2008), que comparou os métodos de aprendizado de máquina (redes neurais e árvore de classificação e regressão) para estimar o escore de propensão, além de compará-los com a regressão logística. Como resultado, Setoguchi et al. (2008) concluíram que redes neurais tenderam a produzir menor viés em cenários com não aditividade e não linearidade. No entanto, os autores concluíram que novos estudos seriam necessários em cenários realísticos, uma vez que utilizaram um conjunto de dados que foi construído somente com dados sintéticos.

Com o entendimento dos trabalhos de Westreich et al. (2010) e Setoguchi et al. (2008), fez o levantamento dos trabalhos mais recentes que utilizam métodos de aprendizado de máquina em estudos observacionais, obtendo-se assim, um conjunto de artigos para aprofundamento de estudos. Nesses artigos observou-se um avanço do uso de *ensembles* de regressores para estimar o escore de propensão, e um número muito pequeno de trabalhos que aplicam *ensembles* aliados ao processo de PSM.

Diante do exposto, faz-se necessário aqui a descrição dos trabalhos que utilizam *ensembles* de regressores para estimar o escore de propensão das instâncias e, também, o uso de desses *ensembles* aplicados ao processo de PSM, conforme apresentado pelas duas próximas seções, respectivamente.

3.3.1 *Ensembles* para estimar o escore de propensão

O trabalho que merece maior destaque dentre os pesquisados foi proposto por Lee et al. (2010), por apresentar uma comparação de vários métodos baseados em árvores de classificação e regressão (CART) em uma abordagem de ponderação pelo escore de propensão.

A ponderação pelo escore de propensão é uma abordagem que utiliza o escore de propensão para atribuir pesos as instâncias dos grupos de tratamento e de controle. Lee et al.(2010) assumiram uma abordagem para atribuição de pesos as instâncias, na qual as instâncias do grupo de tratamento receberam peso 1, enquanto que cada instância do grupo de controle recebeu como peso o inverso do complemento do escore de propensão, ou seja, $p_i/(1-p_i)$, onde p_i é o escore de propensão estimado para cada instância i.

Já em relação à estimativa do escore de propensão, além da regressão logística, Lee et al. (2010) utilizaram classificadores *based-tree*, ou seja, regressores baseados em árvores de regressão, destacando-se: CART, CART com “poda”, *bagging com* CART (*bagged* CART), e dois *ensembles* de regressores: *random forest* e *boosting* com CART (*boosted* CART). Para avaliar o desempenho de tais métodos e compará-los com a regressão logística, os autores utilizaram 3 conjuntos de dados contendo 500, 1000 e 2000 instâncias, respectivamente, as quais foram

obtidas a partir de um conjunto de dados já utilizado por Setoguchi et al. (2008), mas com uma pequena modificação.

Na avaliação, Lee et al. (2010) utilizaram 5 métricas:

- ASAM (***A**verage **S**tandardized **A**bsolute **M**ean*): é uma media de distância utilizada para avaliar o balanceamento das covariáveis com base na diferença média normalizada da média das covariáveis;
- *Bias*: diferença do efeito do tratamento obtido em relação ao verdadeiro efeito do tratamento introduzido no conjunto de dados (-0.4);
- *Standard Error* (SE): desvio padrão do efeito do tratamento estimado;
- 95 *percent confidence interval* (CI) *coverage*: porcentagem na qual o intervalo de confiança de 95% incluiu o verdadeiro efeito do tratamento;
- *Weights*: medida utilizada para verificar a distribuição dos pesos para instâncias do grupo de controle. A ponderação das instâncias pode ser afetada quando há muitas instâncias com pesos extremos (0 e 1).

Em relação aos resultados obtidos por Lee et al. (2010), destacam-se:

- CART e CART com ”poda” produziram maiores valores de ASAM do que outros métodos, ou seja, não proveram consistente balanceamento das covariáveis;
- *Ensembles* de regressores (*bagged* CART, *random forest* e *boosted* CART) produziram menores valores de ASAM em todos os cenários verificados, o que indicou melhor desempenho do balanceamento das covariáveis;
- CART e CART com “poda” obtiveram viés (*Bias*) mais alto em relação aos *ensembles* de regressores e porcentagem mais baixa de cobertura do intervalo de confiança de 95%. Já *boosted* CART apresentou melhor porcentagem de cobertura do intervalo de confiança: maior que 98,6%;
- Regressão logística produziu maiores erros (SE) quando comparada a outros métodos;
- Regressão logística e *random forest* produziram um número relativamente grande de pesos extremos, enquanto que CART e *bagged* CART produziram um pequeno número.

Com base nos resultados obtidos, Lee et al. (2010) concluíram que os ensembles de regressores avaliados produziram um bom desempenho em relação ao balanceamento de covariáveis e estimativa do efeito do tratamento e, ainda, que boosted CART e random forest obtiveram desempenho superior, sugerindo assim, que estes dois ensembles de regressores sejam considerados em futuras abordagens de estimativas do escore de propensão.

Recentemente, outros trabalhos também avaliaram o uso de *ensembles* de regressores relacionados à regressão logística, destacando-se os trabalhos propostos por Watkins et al. (2013), McCaffrey et al. (2013) e Ellis et al. (2013).

Watkins et al (2013) investigaram a aplicação de *ensembles* de regressores, mais precisamente, *random forest* e *bagging*, em comparação à regressão logística, para estimar os escores de propensão e aplicá-los em uma abordagem de ponderação pelo escore de propensão. Os escores de propensão foram estimados por meio de quatros métodos: *random forest, bagging*, um único regressor baseado em árvore de regressão e, a regressão logística.

Em seguida, as instâncias foram ponderadas recebendo os seguintes pesos: 1/escore de propensão, para instâncias do grupo de tratamento e, 1/(1- escore de propensão), para instâncias do grupo de controle. Para avaliar o balanceamento das covariáveis foi calculada a diferença normalizada da média. Por fim, para um dos quatro métodos avaliados foi estimado o efeito do tratamento, que buscou avaliar o efeito da terapia física e ocupacional sobre a habilidade motora de crianças com baixo peso em idade pré-escolar. Como resultado, os autores concluíram que *random forest* e *bagging* produziram melhor balanceamento das covariáveis em relação à regressão logística. E, ainda, que *ensembles* de regressores são alternativas úteis à regressão logística para controle o confundimento em estudos observacionais. É importante ressaltar que o confundimento ocorre quando há desajuste na comparabilidade dos grupos de tratamento e de controle. Por exemplo, um grupo é mais idoso que outro ou fuma mais, ou seja, as covariáveis que produz o confundimento estão distribuídas de forma desigual nos grupos de tratamento e de controle.

Já McCaffrey et al. (2013) propuseram uma abordagem para aplicar *ensembles de* regressores, mas precisamente, *boosting*, em uma abordagem baseada na ponderação pelo escore de propensão em múltiplos tratamentos. Os autores apresentaram estimativas do escore de propensão para múltiplos

tratamentos usando regressão logística e *boosting*. Para isso, os autores introduziram covariáveis *dummy* para indicar a participação das instâncias em cada programa de tratamento. Em seguida, *boosting* com árvores de regressão foram aplicados em cada programa de tratamento para estimar o escore de propensão das instâncias. Os resultados demonstraram que regressão logística multinomial foi superada por *boosting* ao tentar balancear mais do que dois grupos de tratamento. E ainda, segundo McCaffrey et al. (2013), uma das vantagens do método *boosting* está relacionado ao seu processo iterativo, que pode levar ao melhor balanceamento dos grupos de tratamento e de controle.

Por fim, Ellis et al. (2013) ao examinarem os efeitos dos métodos de estimativa sobre balanceamento das covariáveis, verificaram que a regressão logística proporcionou melhor balanceamento que BCART (*boosting* com árvores de classificação e regressão) independente do método de aplicação usado (*matching* ou ponderação). No entanto, os autores sugeriram que novos estudos de simulação são necessários para comparar os modelos de estimativa de escore de propensão em condições variadas.

Para uma melhor visualização dos trabalhos descritos acima, a Tabela 3.1 apresenta um resumo com a descrição dos nomes dos autores, objetivo do trabalho, métodos utilizados, conclusão e a abordagem com base no contexto utilizado.

Tabela 3.1 – Uso de Ensembles para estimar o escore de propensão.

Trabalho	Objetivo	Métodos	Conclusão	Abordagem
Setoguchi et al. (2008)	Comparar o desempenho de redes neurais, CART e regressão logística.	Redes neurais e CART	Redes neurais proveram estimativas com viés (Bias). A estimativa do efeito do tratamento foi robusta quando utilizado a Regressão logística.	PSM*
Westreich et al. (2010)	Comparar os métodos de aprendizado de máquina e regressão logística.	Redes neurais, CART, *support vector machine* e *boosting*.	*Boosting* com CART mostra-se promissor no contexto de PSA	Nenhuma**
Lee et al. (2010)	Comparar os métodos baseados em CART e regressão logística.	CART, pruned CART *bagging* CART, *random forest* e *boosting* CART.	*Ensembles* fornecem excelente desempenho em termos de balanceamento de covariáveis e estimativas de efeito de tratamento.	IPTW***

Trabalho	Objetivo	Métodos	Conclusão	Abordagem
McCaffrey et al. (2013)	Apresentar um tutorial para estimar o escore de propensão em múltiplos tratamentos.	*Boosting (generalized boosted models)*	*Boosting* superou a regressão logística ao tentar balancear mais do que dois grupos de tratamento.	IPTW***
Watkins et al. (2013)	Comparar os métodos *random forest, bagging* e regressão logística.	*Random forest* e *bagging*	*Random forest* e *bagging* produziram melhor balanceamento das covariáveis e são úteis para o controle do confundimento em estudos observacionais.	IPTW***
Ellis et al. (2013)	Comparar *boosting* com regressão logística.	*Boosting* CART	A regressão logística proporcionou melhor balanceamento das covariáveis em cenários com "*main effects*".	IPTW***

* PSM: *Propensity Score Matching.*

** Nenhuma: os autores apenas fizeram levantamento da literatura sem aplicação prática.

*** IPTW: *Inverse Probability of Treatment Weighting.*

Fonte: Elaborado pelo autor.

Como observado, os trabalhos indicam que *ensembles* de regressores constituem-se uma alternativa útil à regressão logística para estimar o escore de propensão, com bom desempenho ao balancear as covariáveis observadas (Lee et al.,2010; MCCAFFREY et al., 2013; WATKINS et al. 2013), e para o controle do confundimento (WATKINS et al. 2013).

Diante do exposto, a presente pesquisa buscou aliar os *ensembles* regressores para estimar os escores de propensão ao processo de PSM, para produzir melhor balanceamento dos grupos de tratamento e de controle.

A seguir são descritos os trabalhos que aplicam *ensembles* em conjunto com o processo de PSM.

3.3.2 *Ensembles* aplicados ao processo de PSM.

Na literatura há muitos trabalhos que aplicam *ensembles* de regressores ao processo de PSM. Neste contexto, ressalta-se o trabalho de Austin e Small (2014), que propuseram duas abordagens que utilizam amostras *bootstrap* (*bagging*) com PSM para estimar a variabilidade da estimativa do efeito do tratamento.

Austin e Small (2014) consideraram três métodos de correspondência sem substituição nas duas abordagens propostas. O primeiro método utilizado foi o "vizinho mais próximo" para a formação de pares do tipo 1:1, com seleção aleatória das instâncias do grupo de tratamento. O segundo método foi o "vizinho mais próximo" com *caliper* de 0,2 sobre o desvio padrão do escore de propensão e, o terceiro, foi o método *Optimal Matching* (ROSENBAUM, 1989).

Em relação às duas abordagens propostas por Austin e Small (2014), na primeira, as amostras foram compostas por pares de instâncias já correspondidas no processo de PSM por um dois três métodos de correspondência mencionados. Já na segunda abordagem, as amostras foram obtidas a partir da amostra original. Dessa forma, os escores de propensão foram estimados para cada instância em cada amostra e, em seguida, as instâncias contidas em cada amostra foram correspondidas.

Nas duas abordagens, os efeitos do tratamento foram estimados em cada amostra. Como resultado, os autores apontam que a primeira abordagem, com amostras *bootstrap* formadas com pares de instâncias já correspondidas, resultou em estimativas de *standard error* (SE) mais próximas ao desvio padrão empírico da distribuição amostral, o que demonstra melhor aplicabilidade ao estimar a variabilidade do efeito do tratamento.

Por fim, observa-se que os *ensembles* de regressores são utilizados, principalmente, em abordagens que ponderam as instâncias pelo escore de propensão, diferentemente, da abordagem "*SEnsembles*" proposta nesta tese.

3.4 Considerações Finais

Este capítulo descreveu a revisão da literatura separando os trabalhos em duas seções principais. Na primeira seção foram descritos os principais métodos para correspondência de instâncias no contexto do PSM e, na segunda, foram descritos os principais trabalhos que empregam *ensembles* de regressores para estimar o escore de propensão e ao processo de PSM. No próximo capítulo será descrita a metodologia usada na pesquisa.

Capítulo 4

METODOLOGIA

4.1 Considerações Iniciais

Este capítulo descreve a metodologia usada no desenvolvimento da pesquisa, destacando-se os métodos utilizados, as métricas adotadas, os conjuntos de dados manipulados, os recursos aplicados, o plano de trabalho das atividades realizadas e, por fim, as formas de acompanhamento dessas atividades.

Ressaltar-se que está tese segue o método de pesquisa Hipotético-Dedutivo (LAKATOS, MARCONI, 2008), pois apresenta uma hipótese para um problema e, por meio de um processo dedutivo, busca-se uma solução e, em seguida, a comprovação ou refutação da hipótese com base na experimentação.

4.2 Métodos

Para alcançar os objetivos pretendidos pelo presente trabalho foram adotadas algumas estratégias, destacando-se, dentre elas: a compreensão inicial dos principais conceitos necessários para o entendimento do tema, revisão bibliográfica, investigação dos métodos de correspondência de instâncias aplicados em estudos observacionais, investigação dos métodos *ensembles* de regressores e suas aplicações e, ainda mais especificamente, suas aplicações no âmbito dos estudos

observacionais, investigação de como as características idênticas das instâncias frente aos métodos para correspondência de indivíduos poderiam resultar em correspondências mais similares e, por fim, uma investigação sobre a linearidade dos conjuntos de dados.

A revisão bibliográfica foi uma estratégia essencial à pesquisa, pois buscou encontrar trabalhos correlatos (ou relacionados) na literatura, bem como propiciar a atualização constante do estado da arte em relação aos assuntos abordados. Os temas pesquisados são relacionados à correspondência de instâncias e aos métodos *ensembles* de regressores, ambos aplicados em estudos observacionais.

Desde o início da pesquisa algumas "*strings* de busca" foram utilizadas, tais como: *entity matching*, *entity resolution, instance matching*, estudos observacionais*, propensity score*, *propensity score matching*, *causal effects*, *matching strategies*, *propensity score estimation*, *ensembles* (*bag*, *boosting* e *random forest*), *learning machine,* regressão logística, as quais foram executas nas seguintes bibliotecas digitais: IEEE, ACM, Scopus, DBLP, JSTOR, Pubmed, Econlit e Repec. O material pesquisado, até agosto de 2016, possui 4.53 GB distribuídos em 3.702 arquivos em 316 pastas. Entretanto, como o fechamento do escopo do projeto, a principal "*string* de busca" foi definida como sendo: "*propensity* score" e *ensemble*.

A investigação dos métodos de correspondência de instâncias aplicados em estudos observacionais foi realizada com base na literatura e, também, por meio da execução dos principais métodos da linguagem *R* (R FOUNDATION, 2016), o qual possui alguns pacotes com métodos para correspondência de instâncias. Neste contexto, também foi necessário compreender o funcionamento da regressão logística, cujos escores de propensão são estimados na maioria dos trabalhos e, por fim, o processo de PSM. Em outras palavras, foi necessária a compreensão de como as instâncias são correspondidos em estudos observacionais e, quais os principais métodos utilizados para isso. Destaca-se aqui o pacote *MatchIt* (HO et al., 2011) da linguagem *R* (R FOUNDATION, 2016), que possui vários métodos de correspondência de instância e, em especial, o método *NNM* (RUBIN, 1973)*,* o qual fornece várias configurações para se efetuar as correspondências, tais como: o uso de limitadores de distância (*calipers*), substituição ou não de instâncias, entre outros.

Em seguida, iniciou-se a investigação dos métodos *ensembles* (*bagging*, *random forest* e *boosting*) e as suas aplicações em estudos observacionais. Com essa investigação observou-se que os métodos *ensembles* são raramente utilizados

em abordagens relacionadas ao processo de PSM, mas são aplicados em abordagens IPSW, as quais ponderam as instâncias utilizando-se seus escores de propensão. Assim, foi necessária uma investigação para aliá-los ao processo de PSM com o propósito de apoiar a correspondências de instâncias e, neste caso, verificou-se que os *ensembles* de regressores poderiam compor uma nova abordagem para correspondência de instâncias, mais precisamente, ao utilizá-los em substituição à regressão logística para estimar os escores de propensão.

Paralelamente a investigação dos *ensembles*, iniciou-se uma investigação para se descobrir como as características idênticas das instâncias poderiam ser exploradas para melhorar a similaridade das correspondências e, essa investigação resultou em uma das duas estratégias da nova abordagem para correspondências de instâncias.

Com o entendimento dos métodos para correspondências, dos *ensembles* e as suas aplicações em estudos observacionais, bem como da elaboração de uma estratégia que considerasse as características idênticas das instâncias, iniciou-se a definição de uma nova abordagem no contexto do processo PSM com dois processos principais e concomitantes, os quais são responsáveis por gerar as correspondências das instâncias com estratégias diferenciadas entre si, sendo uma delas baseada nas características idênticas das instâncias e processos de separação e junção do conjunto de dados e, a outra, baseada na substituição da regressão logística por *ensembles* de regressores (*bagging, random forest* e *boosting*) ao estimar os escores de propensão.

Por último, uma investigação também foi realizada para avaliar as correspondências de instâncias em conjuntos de dados com covariáveis com maiores e menores linearidade/aditividade, conforme os cenários descritos por Lee et al. (2010). Porém, os resultados dessa investigação não foram conclusivos e por isso não estão presentes nesta tese. Contudo, idealiza-se que análises prévias da linearidade/aditividade das covariáveis do conjunto de dados possam contribuir na definição de quais métodos poderiam produzir melhores correspondências de instâncias.

Por fim, a nova abordagem para correspondência de instâncias, nomeada de "*SEnsembles*", foi desenvolvida utilizando-se a linguagem *R* (R FOUNDATION, 2016), o qual também possibilitou a realização de experimentos que permitiram a validação dessa abordagem e de seus processos.

4.3 Métricas Adotadas

Para a validação da abordagem *SEnsembles* buscou-se adotar métricas que se justificam pela necessidade de se medir o balanceamento das covariáveis dos grupos de tratamento e de controle. Para tanto, foi utilizada a métrica ASAM (***A**verage **S**tandardized **A**bsolute **M**ean **D**istance*) conforme descrita por Lee et al. (2010). Dessa forma, para cada covariável observada foi calculada a diferença absoluta dos valores do grupo de tratamento e de controle, dividido pelo desvio padrão dos valores do grupo de tratamento. Em seguida, foi realizada a média dos valores obtidos, conforme ilustrado na Equação (4.1).

$$\text{ASAM} = \text{Mean}|(\frac{\bar{x}_{Tratamento} - \bar{x}_{Controle}}{\sigma_{Tratamento}})| \tag{4.1}$$

onde $\bar{x}_{Tratamento}$ é a média dos valores de cada covariável das instâncias do grupo de tratamento; $\bar{x}_{Controle}$ é a média dos valores de cada covariável das instâncias do grupo de tratamento e, $\sigma_{Tratamento}$, é o desvio padrão dos valores de cada covariável das instâncias do grupo de tratamento. O valor da métrica ASAM indica a diferença dos valores das covariáveis das instâncias do grupo de tratamento e de controle e, quanto menor, mais similares eles são (que é o objetivo desejável).

Ressalta-se também a adoção de duas outras métricas, o número de pares de instâncias resultantes de um processo de correspondência e, o número de instâncias descartadas do grupo de tratamento, ou seja, instâncias do grupo de tratamento que poderiam ser correspondidas, mas não houve nenhum correspondente no grupo de controle. Em geral, como pode ser observado na Seção 3.2.3, isso acontece quando se utiliza limitadores de distância (*calipers)* ao se estabelecer os pares de instâncias.

Além das métricas mencionadas (i.e. ASAM, número de pares de instâncias correspondidas e número de instâncias descartadas do grupo de tratamento), que são consideradas métricas absolutas, pois são obtidas diretamente após um processo de correspondência de instâncias, também, adotou-se, em alguns

experimentos que buscaram comparar a abordagem "*SEnsembles*" com o método *NNM*, métricas relativas, as quais utilizaram como valores referenciais os valores das métricas absolutas obtidas pelo *NNM* ao se efetuar as correspondências. Assim, as métricas relativas são exibidas em forma de percentual para indicar uma vantagem ou desvantagem em relação aos valores das métricas absolutas.

O conjunto de métricas adotado (absolutas e relativas) possibilitou a validação dos processos da abordagem "*SEnsembles*" por meio dos experimentos descritos no Capítulo 6, os quais foram executados no mínimo vinte vezes. Como os valores das métricas são valores médios obtidos, utilizou-se análise de variância (Teste F e ANOVA) e testes estatísticos (Teste T e Teste Tukey) para comparação das médias da métrica ASAM, ao nível de significância de 5%, conforme descrito em Triola (2013).

4.4 Conjuntos de dados

Nos experimentos foram utilizados dez conjuntos de dados, com destaque para o clássico conjunto de dados Lalonde (1986), um conjunto de dados com informações do Programa Bolsa Família que foi cedido por Martins (2013) e, por fim, os conjuntos de dados conforme os cenários descritos por Lee et al. (2010), sendo 6 cenários diferentes com variações em cada cenário referente à linearidade e à aditividade das covariáveis. Os conjuntos utilizados nos experimentos são descrito em maiores detalhes a seguir.

4.4.1 Conjunto de dados Lalonde (1986)

O clássico conjunto de dados Lalonde (1986) foi usado para investigação dos principais métodos para correspondência de instâncias e para a investigação dos métodos *ensembles*, uma vez que é disponibilizado na linguagem *R* (R FOUNDATION, 2016) em vários pacotes de correspondência de instâncias. A versão utilizada desse conjunto de dados possui 10 covariáveis (Tabela 4.1) e 614

observações, sendo 185 pertencentes ao grupo de tratamento e 429 observações pertencentes ao grupo de controle.

Tabela 4.1 – Covariáveis do conjunto de dados *Lalonde* (1986).

Covariáveis	Possíveis Valores	
Treat	1 (Tratado)	0 (Não tratado)
Age	Idade	
Educ	Anos de estudos	
Black	1 (negro)	0 (não)
Hispan	1 (hispânico)	0 (não)
Married	1 (casado)	0 (não)
Nodegree	1 (graduado)	0 (não)
re74	Renda em 1974 (pré-tratamento)	
re75	Renda em 1975 (pré-tratamento)	
re78	Renda em 1978 (resultado)	

Fonte: Elaborado pelo autor.

Já a Figura 4.1 ilustra parte do conjunto de dados *Lalonde*. Nota-se que a coluna *treat* possui somente valores 1 e 0, indicado se a instância pertence ao grupo de tratamento (1) ou ao grupo de controle (0), o que exemplifica a característica de disjunção intrínseca aos conjuntos de dados manipulados na presente pesquisa.

Figura 4.1 – Parte do conjunto de dados *Lalonde*.

Data: lalonde

	row.names	treat	age	educ	black	hispan	married	nodegree	re74	re75	re78
181	NSW181	1	33	12	1	0	1	0	20279.95	10941.35	15952.6
182	NSW182	1	25	14	1	0	1	0	35040.07	11536.57	36646.95
183	NSW183	1	35	9	1	0	1	1	13602.43	13830.64	12803.97
184	NSW184	1	35	8	1	0	1	1	13732.07	17976.15	3786.628
185	NSW185	1	33	11	1	0	1	1	14660.71	25142.24	4181.942
186	PSID1	0	30	12	0	0	1	0	20166.73	18347.23	25564.67
187	PSID2	0	26	12	0	0	1	0	25862.32	17806.55	25564.67
188	PSID3	0	25	16	0	0	1	0	25862.32	15316.21	25564.67
189	PSID4	0	42	11	0	0	1	1	21787.05	14265.29	15491.01

Fonte: Elaborado pelo autor.

A escolha inicial pelo conjunto de dados *Lalonde* se justifica por ser um dos conjuntos de dados mais utilizados pelos pacotes de *software* no contexto do processo de PSM.

4.4.2 Conjunto de Dados PBF (Martins, 2013)

Desde o início da pesquisa objetivou-se utilizar um conjunto de dados do PBF e, isso foi possível, pois Martins (2013) cedeu o conjunto de dados que manipulou em sua pesquisa de doutorado, que avaliou o impacto do PBF sobre a aquisição de alimentos. Esse conjunto de dados contém dados da Pesquisa de Orçamentos Familiares (POF) de 2008-09 de 55.970 domicílios brasileiros, dos quais 9.259 são beneficiários do PBF e 46711 não beneficiários.

O conjunto de dados PBF (Martins, 2013) possui 262 covariáveis, porém, foram utilizadas somente 15, que são aquelas também usadas por Martins (2013) em sua pesquisa de doutorado. Essas covariáveis são descritas na Tabela 4.2 a seguir.

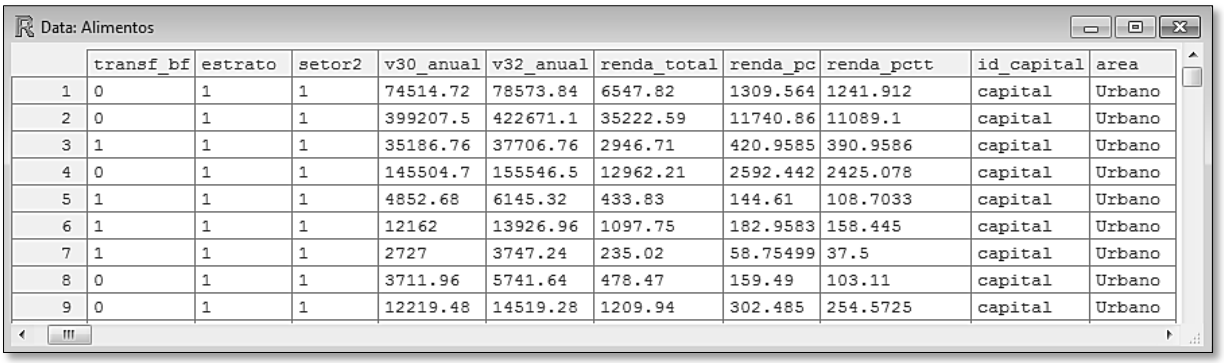
Tabela 4.2 – Covariáveis do conjunto de dados PBF (Martins, 2013).

Covariáveis	Descrição
transf_bf	1 (beneficiário) ou 0 (Não beneficiário)
renda_pc	Valores em Reais.
anoest_chefe	Anos de estudos
prop_pessoas_est1_masc	Qtd. Masculino de 0 a 9.
prop_pessoas_est1_fem	Qtd. Feminino de 0 a 9.
prop_pessoas_est2_masc	Qtd. Masculino de 10 a 15;
prop_pessoas_est2_fem	Qtd. Feminino de 10 a 15.
prop_pessoas_est3_masc	Qtd. Masculino de 16 a 20
prop_pessoas_est3_fem	Qtd. Feminino de 16 a 20.
prop_pessoas_est4_masc	Qtd. Masculino de 21 a 25
prop_pessoas_est4_fem	Qtd. Feminino de 21 a 25.
prop_pessoas_est5_masc	Qtd. Masculino de 21 a 65.
prop_pessoas_est5_fem	Qtd. Feminino de 21 a 65.
Nfam	Número de pessoas na família.
gasto_fora_est	Gasto com alimentação fora do domicílio

Fonte: Elaborado pelo autor.

Já a Figura 4.2 ilustra parte do conjunto de dados em questão, o qual possui a coluna *transf_bf* para indicar se o domicílio recebe ou não o benefício do PBF. Ao todo, esse conjunto de dados possui 55.970 instâncias, sendo 9.259 instâncias pertencentes ao grupo de beneficiários (tratamento) e 46.711 instâncias pertencentes ao grupo de não beneficiários (controle).

Figura 4.2 – Parte do conjunto de dados do PBF cedido por Martins (2013).



Data: Alimentos

	transf_bf	estrato	setor2	v30_anual	v32_anual	renda_total	renda_pc	renda_pctt	id_capital	area
1	0	1	1	74514.72	78573.84	6547.82	1309.564	1241.912	capital	Urbano
2	0	1	1	399207.5	422671.1	35222.59	11740.86	11089.1	capital	Urbano
3	1	1	1	35186.76	37706.76	2946.71	420.9585	390.9586	capital	Urbano
4	0	1	1	145504.7	155546.5	12962.21	2592.442	2425.078	capital	Urbano
5	1	1	1	4852.68	6145.32	433.83	144.61	108.7033	capital	Urbano
6	1	1	1	12162	13926.96	1097.75	182.9583	158.445	capital	Urbano
7	1	1	1	2727	3747.24	235.02	58.75499	37.5	capital	Urbano
8	0	1	1	3711.96	5741.64	478.47	159.49	103.11	capital	Urbano
9	0	1	1	12219.48	14519.28	1209.94	302.485	254.5725	capital	Urbano

Fonte: Elaborado pelo autor.

É importante ressaltar que o conjunto de dados PBF (Martins, 2013) também foi manipulado para se obter maior quantidade de instâncias duplicadas, uma vez que o conjunto de dados original, quando usado com quatro covariáveis para a correspondência de instâncias, apresentou 17 duplicatas e, quando usado quatorze covariáveis, apresentou apenas 2 duplicadas. Assim, além do conjunto original, duas outras variações foram utilizadas nos experimentos, sendo uma com 2.224 e outra com 2.280 duplicatas, conforme descrito na Tabela 4.3.

Tabela 4.3 – Manipulação do conjunto de dados PBF de Martins (2013).

Conjunto de Dados	Covariáveis utilizadas na correspondência	Instâncias Idênticas	Total de Instâncias
PBF 1 (Martins, 2013)	4	17 (8 Beneficiários e 9 não Beneficiários)	55.490, sendo: 9.259 Beneficiários e 46.711 Não Beneficiários
PBF 1 (Martins, 2013)	14	2 (1 par)	
PBF 2 Modificado a partir de Martins (2013)	14	2224 (1.112 pares)	
PBF 3 Modificado a partir de Martins (2013)	14	2.280 (1.185 Beneficiários e 1.095 não Beneficiários)	

Fonte: Elaborado pelo autor.

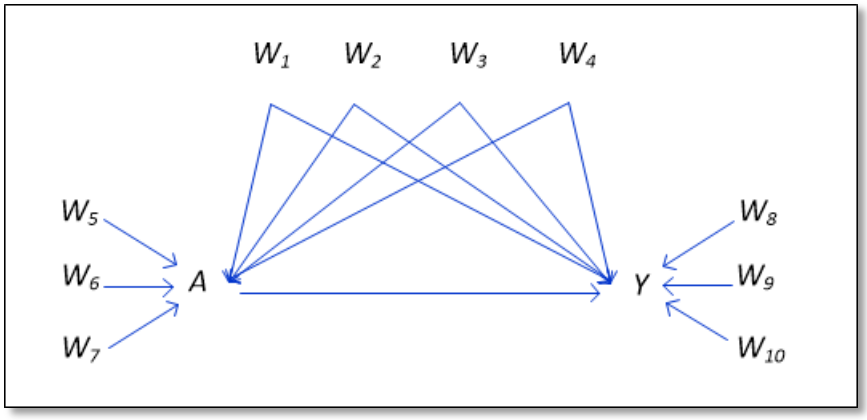
Contudo, as manipulações realizadas no PBF (Martins, 2013) não alteraram o número total de instâncias pertencentes ao grupo de beneficiários e não beneficiários, permanecendo o mesmo que o conjunto de dados original, ou seja, 9.259 são pertencentes ao grupo de beneficiários e 46.711 ao grupo de não beneficiários. E, ainda, as porcentagens de duplicatas introduzidas no conjunto original seguiram a mesma proporção do conjunto de dados Lalonde (1986), que é de aproximadamente 4%. Essa porcentagem foi introduzida para proporcionar comparações dos métodos de correspondências de instâncias.

4.4.3 Conjuntos de Dados Lee et al. (2010)

Os conjuntos de dados descritos por Lee et al. (2010) foram gerados a partir de uma codificação (Anexo A) que permite gerar até 7 cenários diferentes, com covariáveis com maiores e menores valores de linearidade e aditividade. Esses cenários se basearam na estrutura de simulação de dados sintéticos descrito por Setoguchi et al.(2018), mas com pequenas alterações introduzidas por Lee et al. (2010).

Para cada conjunto de dados são geradas 10 covariáveis, sendo quatro covariáveis (W_1, W_2, W_3 e W_4) associadas à exposição (A) e ao resultado (Y), três associadas à exposição (W_5, W_6 e W_7), e três associada ao resultado (W_8, W_9 e W_{10}), conforme se observa na Figura 4.3. Os valores dos W_{is} são gerados de maneira aleatória com base na distribuição normal com média zero e variancia um.

Figura 4.3 – Covariáveis do Conjunto de Dados utilizado por Lee et al. (2010).



Fonte: Adaptado de Lee et al. (2010).

Os cenários (A-G) utilizados por Lee et al. (2010) permitem similar conjuntos de dados contendo covariáveis com maiores e menores linearidade e aditividade, conforme observa-se na Tabela 4.4.

Tabela 4.4 – Cenários de conjuntos de dados do trabalho de Lee et al. (2010)

Cenário	Característica
A	Com linearidade e aditividade.
B	Leve (*mild*) não linearidade.
C	Moderada não linearidade.
D	Leve (*mild)* não aditividade.
E	Leve (mild) não aditividade e não linearidade.
F	Moderada não aditividade.
G	Moderada não aditividade e não linearidade.

Fonte: Elaborado pelo autor.

É importante mencionar que os conjuntos de dados podem ser gerados por meio de uma função que permite informar a quantidade de conjuntos de dados que serão gerados para um determinado cenário e, também, a quantidade de instâncias de cada conjunto de dados. Por exemplo, a Figura 4.4 apresenta a codificação para gerar 1000 conjuntos de dados para cada cenário, com 1000 instâncias em cada conjunto de dados gerado.

Figura 4.4 – Covariáveis do Conjunto de Dados utilizado por Lee et al. (2010).

```
cenarioA <- replicate(1000, F.generate(1000, "A"))
cenarioB <- replicate(1000, F.generate(1000, "B"))
cenarioC <- replicate(1000, F.generate(1000, "c"))
cenarioD <- replicate(1000, F.generate(1000, "D"))
cenarioE <- replicate(1000, F.generate(1000, "E"))
cenarioF <- replicate(1000, F.generate(1000, "F"))
cenarioG <- replicate(1000, F.generate(1000, "G"))
```

Fonte: Elaborado pelo autor.

Ressalta-se que na presente pesquisa os cenários descritos por Lee et al. (2010) foram utilizados para verificação da linearidade das covariáveis dos conjuntos de dados e, também, foram utilizados em experimentos pelos quais os escores de

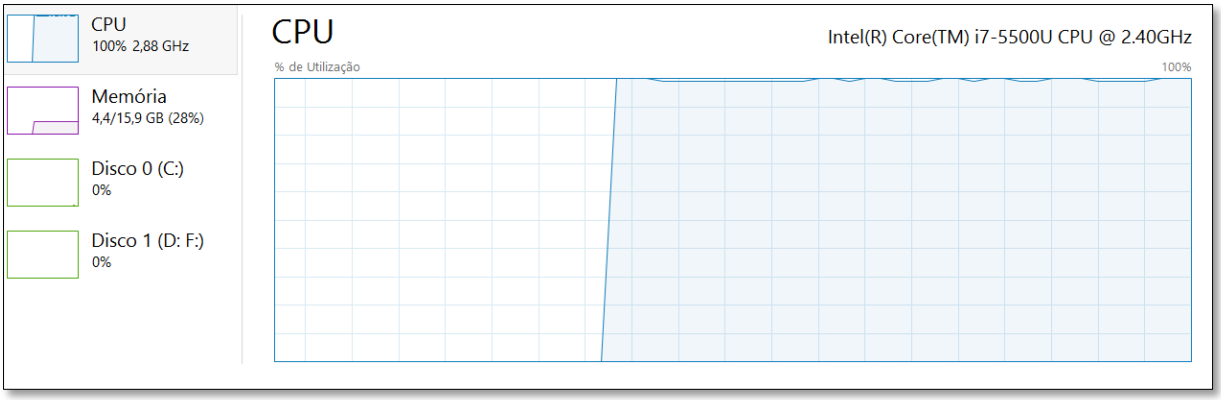
propensão foram estimados pelos *ensembles* de regressores. Dessa forma, objetivou-se verificar o comportamento dos *ensembles* em gerar os escores de propensão de acordo com cada cenário descrito por Lee et al. (2010), com covariáveis com maiores e menores valores de linearidade e aditividade.

4.5 Recursos

Os recursos utilizados na pesquisa limitam-se aos recursos de *software* e *hardware*. Com relação a *software,* merece destaque a linguagem *R* (R FOUNDATION, 2016), com licença (*General Public License*), o qual foi utilizado para codificação e validação da abordagem "*SEnsembles*". Adicionalmente, também se utilizou software para editoração de texto, planilha eletrônica e a ferramenta Bizagi (2016) para a modelagem dos processos da abordagem "*SEnsembles*" com base na notação BPMN (OMG, 2016).

Já em relação ao *hardware* foi utilizado somente um computador *notebook* da fabricante Dell, com processador i7 com 16 Gb de memória RAM e 1 TB de HD. Os experimentos para validação da abordagem "*SEnsembles*" exigiram consideravelmente desse computador, cujo processamento atingiu 100% do uso da CPU em alguns experimentos conforme pode-se visualizar na Figura 4.5.

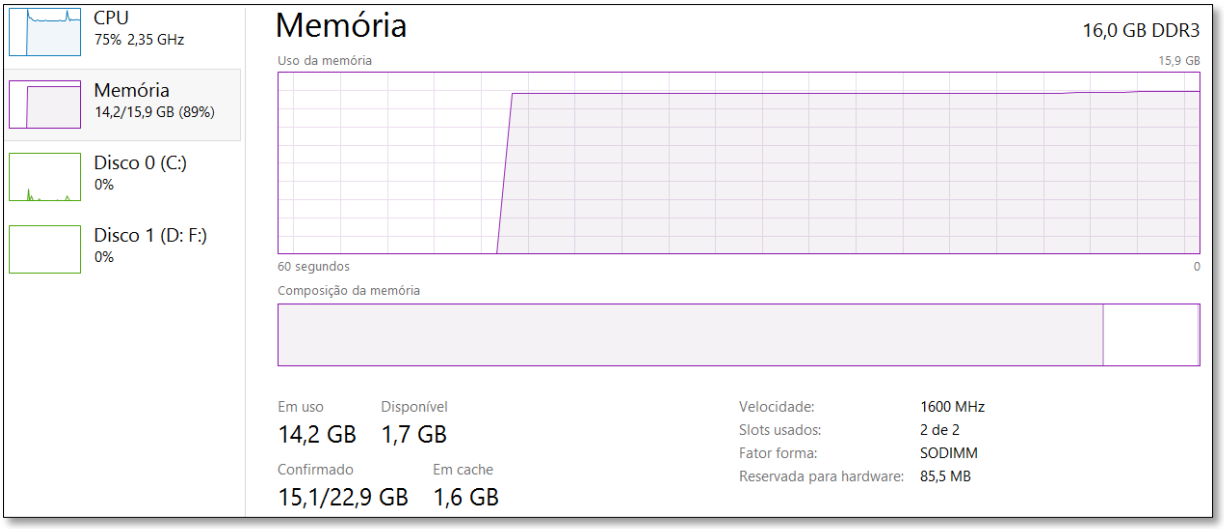
Figura 4.5 – Uso da CPU em alguns experimentos.



Fonte: Elaborado pelo autor.

Ressalta-se também que a memória foi muito exigida conforme ilustrado na Figura 4.6, a qual apresenta o uso de 89%, mas atingiu 97% de uso total em alguns experimentos.

Figura 4.6 – Uso da Memória em alguns experimentos.



Fonte: Elaborado pelo autor.

Por fim, a capacidade de processamento e memória primária do computador utilizado possibilitou a execução e validação dos experimentos sem restrições, com alocações, em alguns momentos, de até 100 conjuntos de dados em memória.

4.6 Acompanhamento das atividades

O acompanhamento das atividades da pesquisa foi realizado por meio de reuniões com o professor orientador, prof. Ricardo Rodrigues Ciferri, e com a profa. Marilde Terezinha Prado Santos, a qual possibilitou a idealização e início do projeto de pesquisa. Essas reuniões foram imprescindíveis para permitir o delineamento adequado do assunto abordado, delimitações da pesquisa, definição de diretrizes, correção de erros e, sem dúvida, foram importantíssimas para a conclusão da pesquisa.

Adicionalmente, a participação em reuniões do Grupo de Banco de Dados da UFSCar (GBD/UFSCar), bem com a apresentação de seminários nesse grupo, agregaram uma valor inestimável para a construção do conhecimento necessário ao desenvolvimento da pesquisa.

Destacam-se também, as reuniões para esclarecimento de dúvidas que foram realizadas com os professores da FEA-USP de Ribeirão Preto, prof. Dr. Luiz Guilherme Dácar da Silva Scorzafave e Prof. Dr. Walter Belluzzo Júnior, e também, aquelas que foram realizadas na fazenda da Embrapa São Carlos, com o pesquisador da área de estatística Me. Waldomiro Barioni.

Como visto, as dúvidas que foram surgindo ao longo da pesquisa foram esclarecidas por professores pesquisadores que dominam o tema sobre estudos observacionais e assuntos relacionados à estatística.

4.7 Considerações Finais

O presente capítulo apresentou a metodologia usada do desenvolvimento desta pesquisa, com a descrição dos principais métodos utilizados, das métricas adotadas para validar os resultados obtidos, dos conjuntos de dados utilizados nos experimentos, dos recursos necessários ao desenvolvimento e a forma de acompanhamento das atividades. A seguir será descrita a nova abordagem para correspondência de dados que é proposta nesta tese.

Capítulo 5

ABORDAGEM "*SENSEMBLES*"

5.1 Considerações Iniciais

Neste capítulo é descrita a proposta de uma nova abordagem para correspondência de instância, denominada "*SEnsembles*", cujo nome foi idealizado com base em seus processos principais, os quais possuem estratégias diferentes para gerar a correspondência das instâncias. Assim, o "*S*" teve origem no processo que se baseia na separação ("*Slice*") das instâncias com características idênticas (características iguais), enquanto que o termo "*Ensembles*" teve origem no processo que utiliza *ensembles* de regressores para estimar os escores de propensão das instâncias em substituição à técnica de regressão logística. Esses processos e as suas estratégias são detalhados nas próximas seções.

A seguir é apresentada uma visão geral da abordagem "*SEnsembles*", com a descrição dos seus processos principais e, o Apêndice A, apresenta a modelagem desses processos na notação BPMN (OMG, 2016).

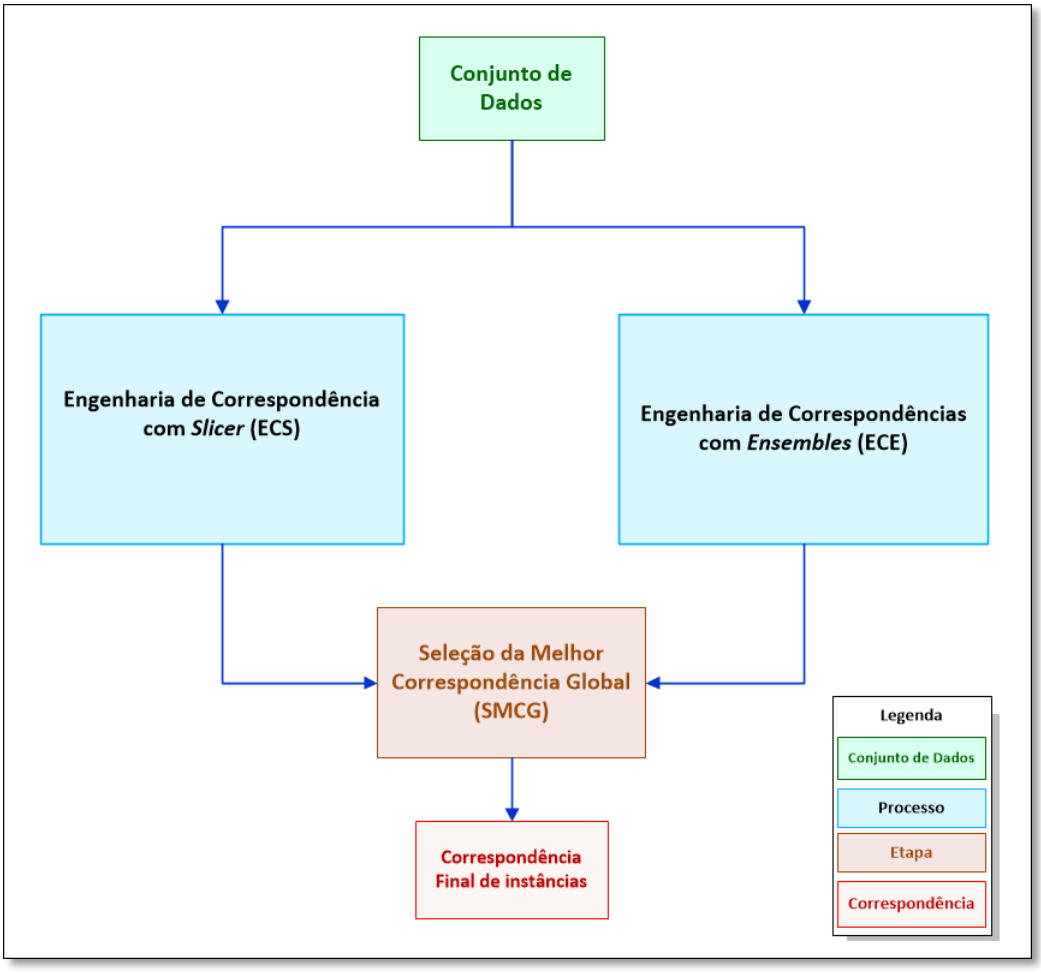
5.2 Visão Geral da Abordagem "*SEnsembles*"

A nova abordagem "*SEnsembles*" para correspondência de instâncias foi desenvolvida com o objetivo de melhorar a qualidade da correspondência de instâncias em conjuntos de dados que são manipulados em estudos observacionais,

os quais possuem como principal característica a disjunção de instâncias, ou seja, as instâncias estão alocados em dois grupos, que em geral recebem o nome de grupo de tratamento e de controle, e não há repetição de uma mesma instância nos dois grupos (i.e. não representam a mesma entidade no mundo real).

Conforme se observa na Figura 5.1, a abordagem "*SEnsembles*" possui dois processos que são executados de forma concomitante para prover a correspondência das instâncias, denominados de Engenharia de Correspondência com *Slicer* (ECS), Engenharia de Correspondência com *Ensembles* (ECE). Os processos ECS e ECE recebem como entrada o mesmo conjunto de dados particionado nos grupos de tratamento e de controle e, em seguida, cada um deles devolve uma saída, que é a correspondência das instâncias (instâncias pareadas). Em seguida, essas correspondências são recebidas como entrada pela etapa SMCG (Seleção da Melhor Correspondência Global) para serem avaliadas.

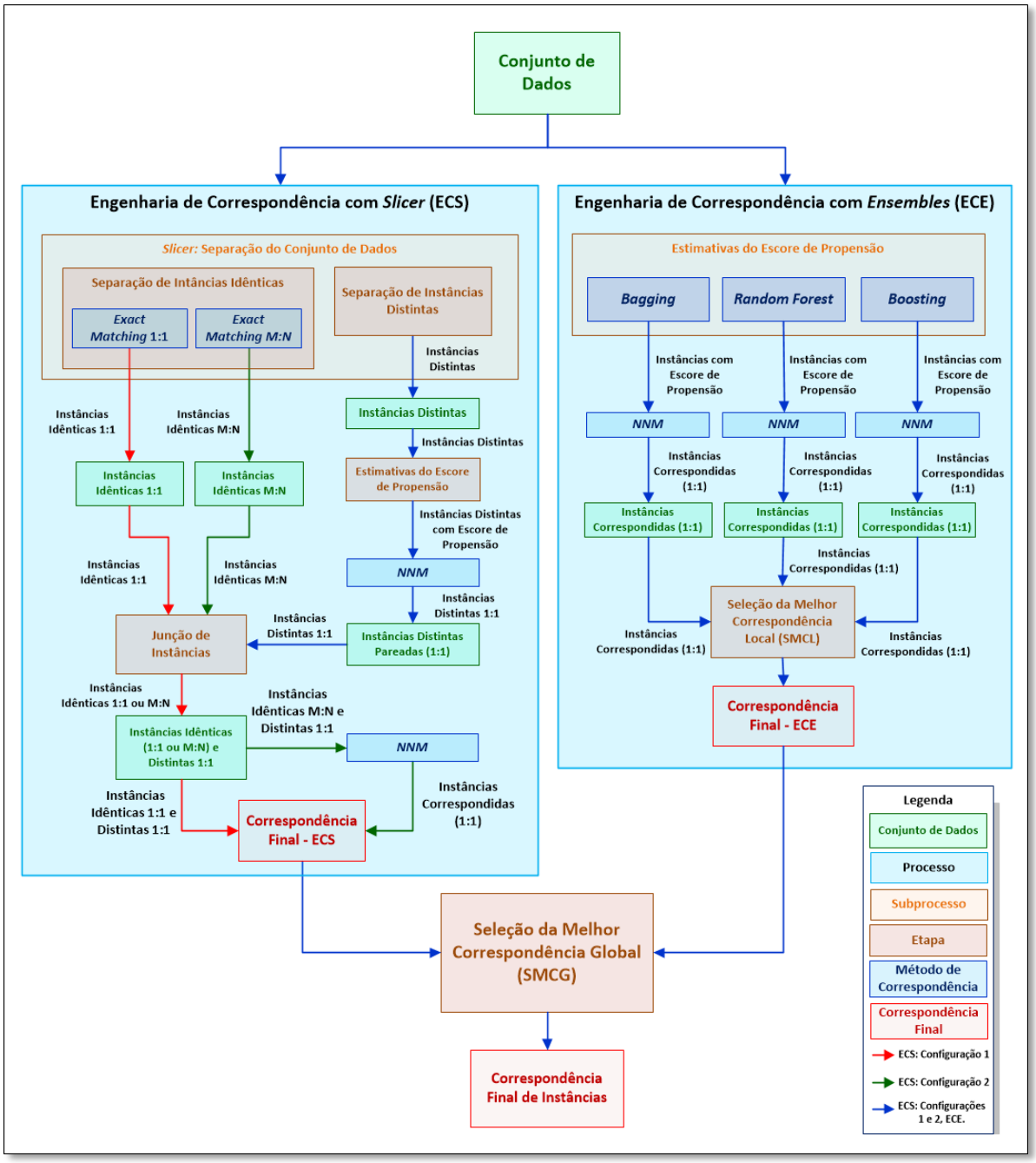
Figura 5.1 – Visão Geral da abordagem proposta "*SEnsembles*".



Fonte: Elaborado pelo autor.

A etapa SMCG utiliza a métrica ASAM para indicar qual correspondência deve ser escolhida como sendo a correspondência final da abordagem "*SEnsembles*". Ressalta-se que os processos ECS e ECE possuem estratégias diferenciadas entre si para gerar a correspondência das instâncias, conforme se observa na Figura 5.2, que ilustra uma visão detalhada desses processos.

Figura 5.2 – Visão detalhada das estratégias dos processos ECS e ECE da abordagem proposta "*SEnsembles*".



Fonte: Elaborado pelo autor.

O processo ECS possui uma estratégia baseada na separação das instâncias que possuem características exatamente idênticas considerando-se as covariáveis observadas, ou seja, considerando-se as covariáveis a serem utilizadas para se efetuar a correspondência das instâncias do grupo de tratamento e de controle. Já o processo ECE possui uma estratégia que substitui a regressão logística por *ensembles* de regressores para estimar os escores de propensão das instâncias. Esses processos (ECS e ECE) são detalhados a seguir.

5.3 Engenharia de Correspondência com *Slicer* (ECS)

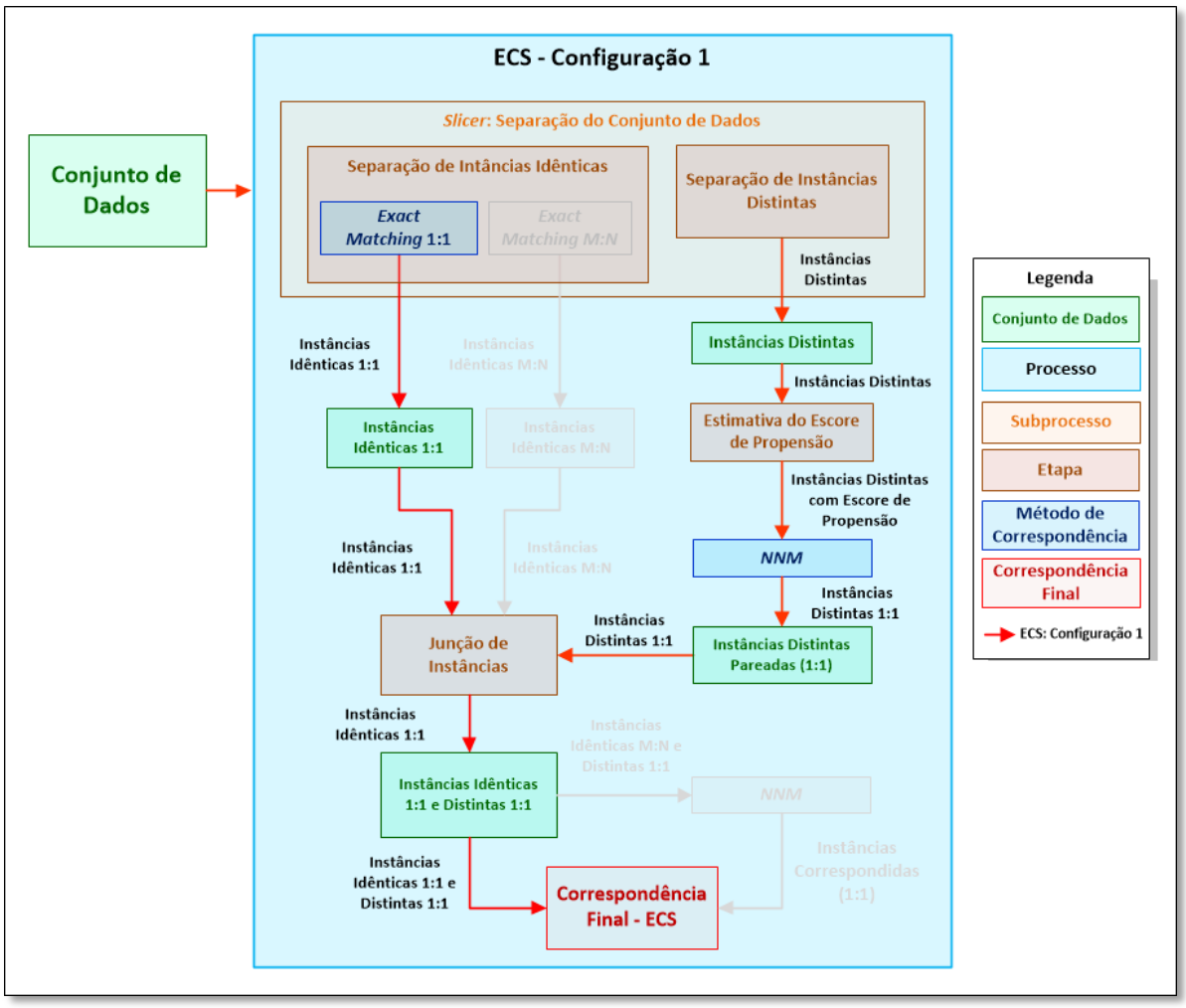
O processo de Engenharia de Correspondência com *Slicer* (ECS) possui uma estratégia que inicialmente separa o conjunto de dados em dois novos conjuntos de dados: conjunto de instâncias com características exatamente idênticas e conjunto de instâncias distintas.

O conjunto de instâncias com características idênticas pode ser obtido por meio de duas configurações. Com a Configuração 1 são obtidos pares de instâncias idênticas contendo uma instância do grupo de tratamento e uma do grupo de controle e, com a Configuração 2, obtém-se um conjunto de dados com M:N instâncias, sendo *M* instâncias do grupo de tratamento e *N* instâncias do grupo de controle, as quais são obtidas pelo método *exact matching* sem considerar a subclassificação, que é uma das características desse método.

Já o conjunto de dados com instâncias distintas obtém-se com a retirada das instâncias idênticas do conjunto de dados de entrada. Assim, são consideradas instâncias distintas aquelas que não estão no conjunto de instâncias idênticas.

A Figura 5.3 ilustra a estratégia do processo ECS da abordagem proposta "*SEnsembles*" quando escolhida a Configuração 1. Observa-se que as instâncias distintas são correspondidas (pareadas) pelo método *NNM* com escores de propensão estimados pela regressão logística. O uso de *calipers* poder ser considerado neste pareamento, mas com alguma eliminação de instâncias distintas menos similares.

Figura 5.3 – Configuração 1 do processo ECS da abordagem proposta “*SEnsembles*”.

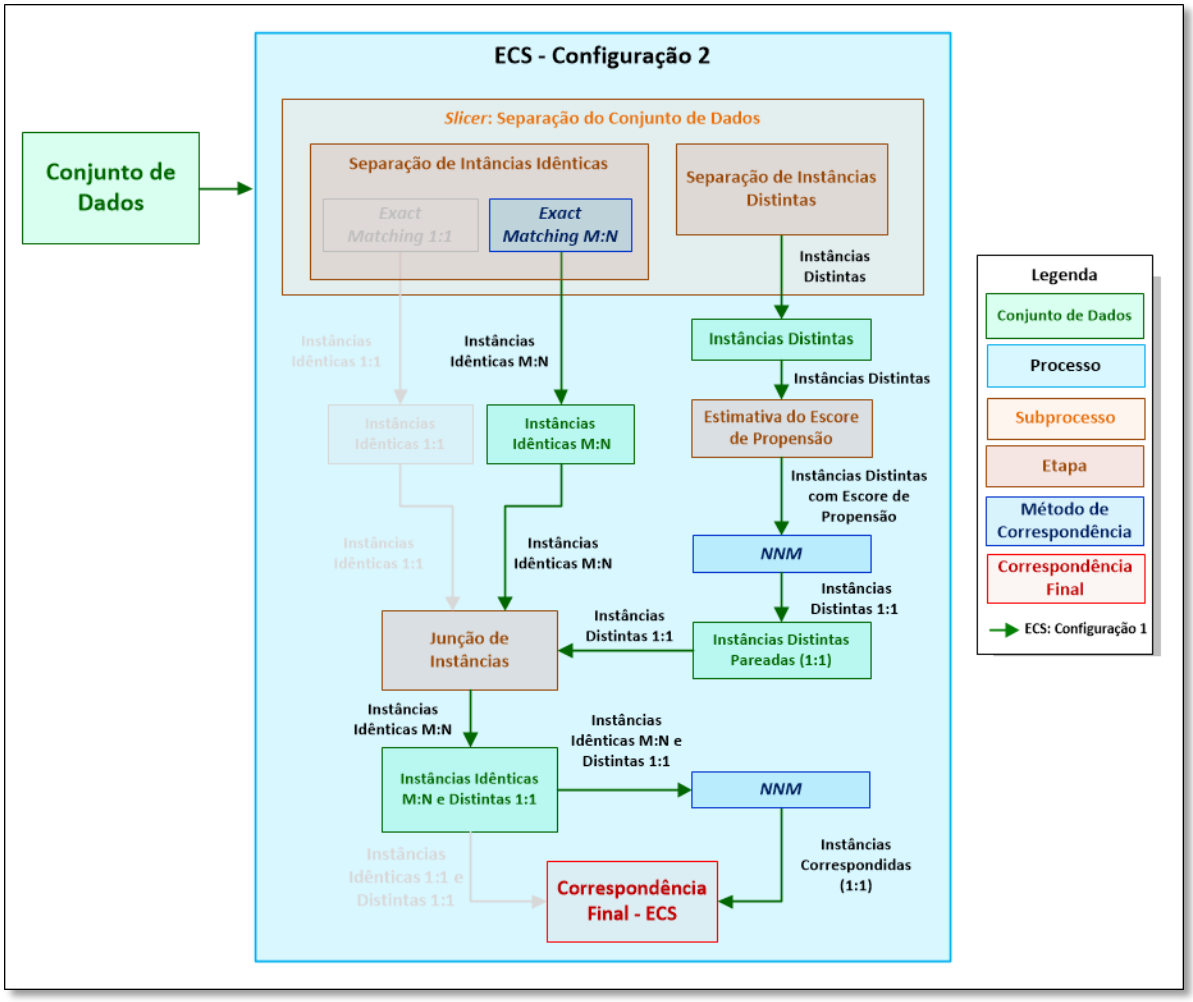


Fonte: Elaborado pelo autor.

Após o pareamento das instâncias distintas, o processo denominado Junção de Instâncias realiza a junção das instâncias idênticas com as distintas pareadas, formando assim, um novo conjunto de dados contendo somente instâncias pareadas, o qual considera-se como à correspondência final do processo ECS.

Entretanto, quando escolhida a Configuração 2 (Figura 5.4), após a junção das instâncias, realiza-se novo pareamento com o método *NNM*, uma vez que as instâncias idênticas nessa configuração ainda não estão pareadas, pois são obtidas na forma (M:N), ou seja, *M* instâncias do grupo de tratamento e *N* instâncias do grupo de controle. Assim, as instâncias devem ser novamente pareadas para gerar a correspondência final do processo ECS.

Figura 5.4 – Configuração 2 do processo ECS da abordagem proposta "*SEnsembles*".



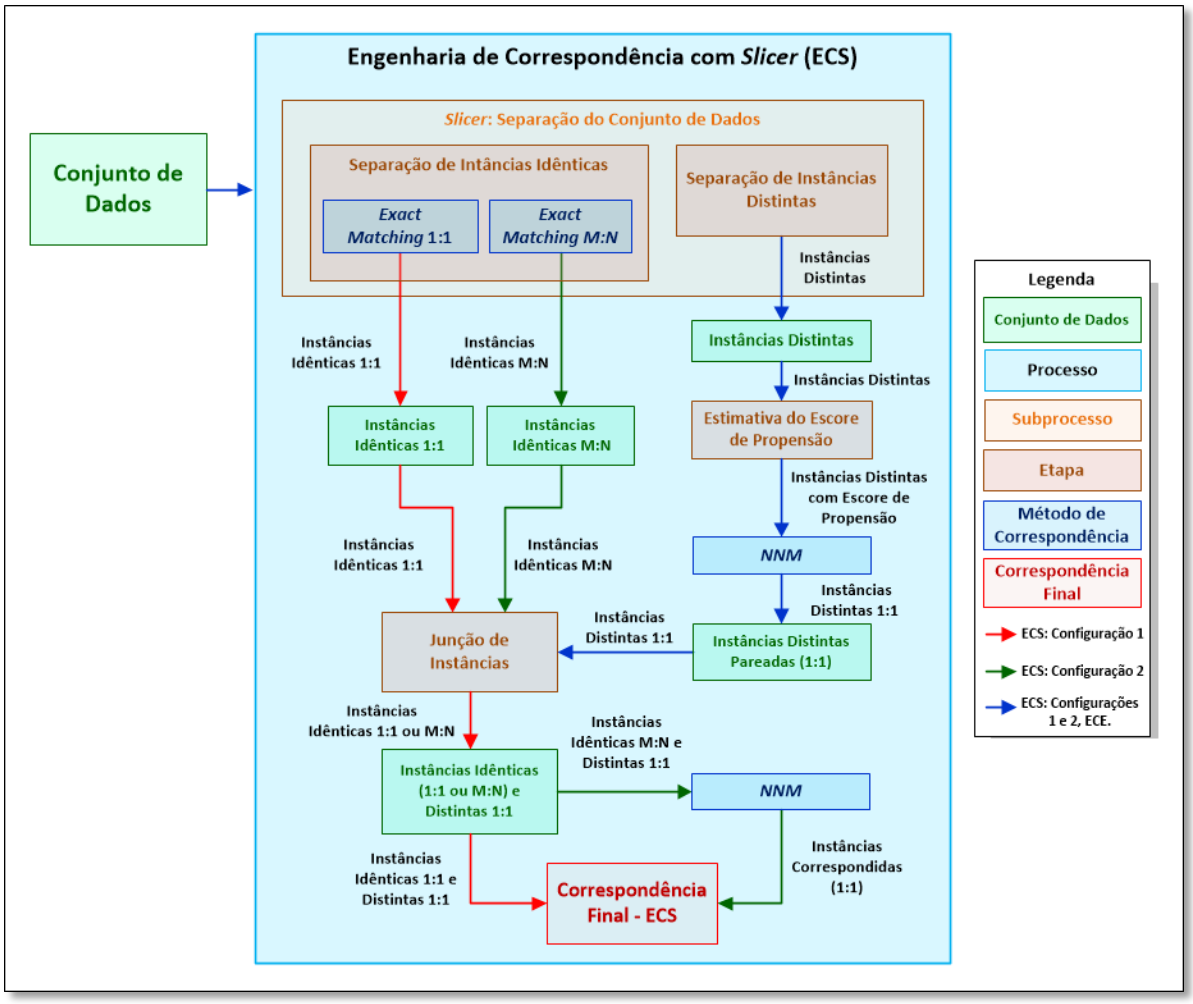
Fonte: Elaborado pelo autor.

É importante destacar que o uso de *calipers,* quando informados pelo usuário especialista, são utilizados em todos os pareamentos realizados pelo processo ECS, seja na Configuração 1 ou 2. Dessa forma, possibilita-se refinar a correspondência das instâncias com provável redução do número de pares de instâncias correspondidos.

Como visto, o processo ECS permite efetuar correspondência de instâncias por meio de duas configurações que se diferenciam pela forma de obtenção das instâncias com características exatamente idênticas. Porém, em ambas as configurações executa-se o método *NNM* com correspondência 1:1 sem substituição, para se efetuar o pareamento das instâncias distintas, conforme

ilustrado na Figura 5.5, a qual apresenta conjuntamente as duas configurações do processo ECS.

Figura 5.5 – Configurações 1 e 2 do processo ECS da abordagem proposta "*SEnsembles*".



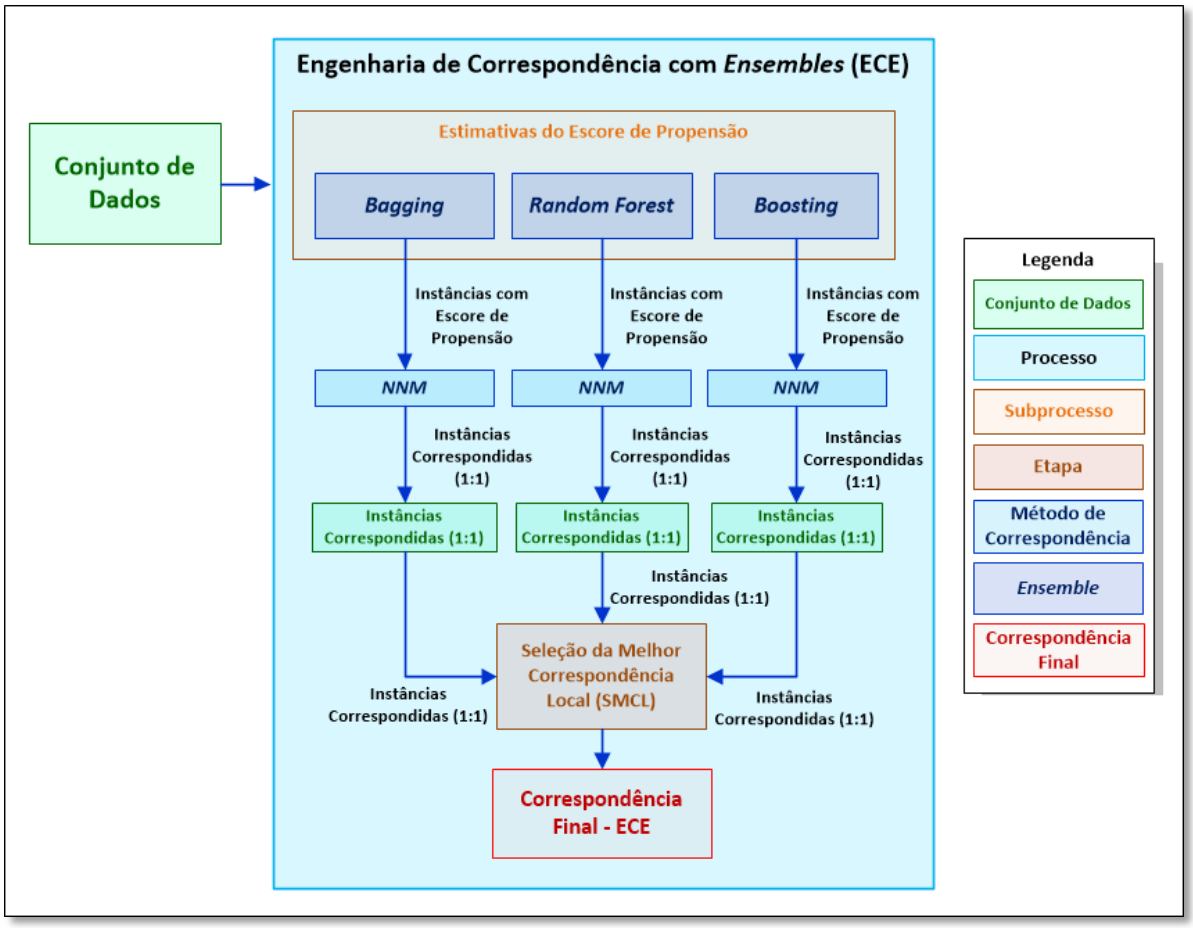
Fonte: Elaborado pelo autor.

Por fim, destaca-se que o processo ECS pode ser vantajoso quando há no conjunto de dados várias instâncias com características exatamente idênticas, considerando-se as covariáveis informadas pelo usuário especialista ao se solicitar que se efetuem a correspondência das instâncias.

5.4 Engenharia de Correspondência com *Ensembles* (ECE)

O processo de Engenharia de Correspondência com *Ensembles* (ECE) possui uma estratégia que utiliza e*nsembles* de regressores, mais precisamente, os *ensembles bagging*, *random forest* e *boosting*, para substituir a regressão logística ao estimar os escores de propensão das instâncias, conforme se observa na Figura 5.6.

Figura 5.6 – Engenharia de Correspondência com *Ensembles* (ECE) da abordagem proposta “*SEnsembles*”.



Fonte: Elaborado pelo autor.

A escolha *ensembles* foi baseada em trabalhos recentes na literatura que os utilizam em abordagens com ponderação pelo escore de propensão, diferentemente da abordagem "*SEnsembles*", que está inserida no contexto do processo de PSM. Observa-se na Figura 5.6 que a estratégia do processo ECE se inicia com a estimativa dos escores de propensão das instâncias pelos *ensembles bagging, random forest* e *boosting*. Em seguida, as instâncias são correspondidas (pareadas) utilizando-se o método *NNM* com os escores de propensão estimados pelos *ensembles.* O uso de *caliper,* caso seja informado pelo usuário especialista, também é considerado nesse pareamento.

O método *NNM* gera três correspondências de instâncias, sendo uma para cada estimativa de escore de propensão (uma de cada *ensemble*). Em seguida, essas correspondências são avaliadas na etapa de Seleção da Melhor Correspondência Local (SMCL), o qual seleciona a correspondência final do processo ECE com base no valor da métrica ASAM, ou seja, a etapa SMCL seleciona a correspondência que possui as instâncias mais similares com base nas covariáveis observadas.

5.5 Seleção da Melhor Correspondência Global (SMCG)

A etapa da Seleção da Melhor Correspondência Global busca definir a correspondência final de instâncias da abordagem "*SEnsembles*". Para isso, essa etapa recebe como entrada as correspondências dos processos ECS e ECE e, em seguida, as avalia para determinar qual delas possui as instâncias correspondidas de maneira mais similar, com base nas covariáveis utilizadas para efetuar a correspondência. Essa avaliação é feita também com base na métrica ASAM (i.e. quanto menor o valor dessa métrica, mais similares são as instâncias correspondidas). Assim, considera-se que a correspondência final da abordagem "*SEnsembles*" é aquela que apresenta o menor valor da métrica ASAM.

5.6 Considerações Finais

O presente capítulo apresentou a abordagem "*SEnsembles"* que possui dois processos, ECS e ECE, com estratégias específicas e diferenciadas entre si para gerar as correspondências das instâncias do grupo de tratamento e de controle. Ao final, a etapa SMCG avalia qual deles gerou a melhor correspondência de instância para que seja considerada a correspondência final da abordagem.

Ressalta-se que a abordagem "*SEnsembles*" está preparada para avaliar as correspondências de instâncias com base na métrica ASAM. Porém, permite-se, ao usuário especialista, definir a troca dessa métrica pelo número de pares de instâncias correspondidas. Essa alteração faz com que a abordagem "*SEnsembles*" priorize o maior número de pares de instâncias na avaliação da correspondência e, consequentemente, reduza o número de instâncias descartadas do grupo de tratamento com o uso de *calipers.*

O próximo capítulo apresenta a validação da abordagem "*SEnsembles*" com a descrição de um conjunto de experimentos e a discussão dos seus resultados.

Capítulo 6

VALIDAÇÕES

6.1 Considerações Iniciais

As validações da abordagem "*SEnsembles*" foram realizadas com experimentos que buscaram avaliar a qualidade das correspondências geradas por seus processos (ECS e ECE), comparando-as entre si e com as correspondências geradas pelo método *NNM*, utilizando-se a regressão logística para estimar os escores de propensão.

O método *NNM* foi utilizado como referencial (*baseline*) pode ser o mais utilizado no contexto do processo de PSM para se efetuar a correspondência de instâncias, principalmente, a correspondência pretendida pela abordagem "*SEnsembles*" (i.e. pareamento 1:1 sem substituição). Dessa forma, foi possível comparar e avaliar a qualidade das seguintes correspondências de instâncias:

- As correspondências finais das Configurações 1 e 2 do processo ECS da abordagem proposta "*SEnsembles*" (Seção 6.2);
- As correspondências geradas pelo processo ECE da abordagem proposta "*SEnsembles*" a partir das estimativas dos escores de propensão realizadas pelos *ensembles bagging, random forest e boosting* (Seção 6.3);
- As correspondências geradas pelo processo ECS da abordagem proposta "*SEnsembles*" com as correspondências obtidas com o método *baseline NNM* (Seção 6.4);
- As correspondências geradas pelo processo ECE da abordagem proposta "*SEnsembles*" com as correspondências obtidas com o método *baseline NNM* (Seção 6.5);

- As correspondências geradas pelo processo ECS da abordagem proposta "*SEnsembles*" com as correspondências obtidas com o processo ECE da abordagem proposta "SEnsembles" (Seção 6.6);
- As correspondências geradas pela abordagem proposta "*SEnsembles*" com as correspondências obtidas com o método *NNM* (Seção 6.7).

Para avaliar a qualidade da correspondência de instâncias utilizou-se a métrica ASAM. Além disso, em alguns experimentos utilizou-se como métrica o número de pares de instâncias correspondidas e o número de descartes de instâncias do grupo de tratamento. Já na Seção 6.8 adotou-se somente como métrica o tempo de execução em segundos, uma vez que os experimentos dessa seção buscaram comparar o desempenho de execução da abordagem proposta "*SEnsembles*" com o método *NNM* (*baseline*).

Em todos os experimentos realizados utilizou-se um intervalo de valores para o *caliper*, de zero a 0,30, ao se efetuar a correspondência de instância, o que permitiu avaliar a qualidade da correspondência ao se limitar o paramento de instâncias. Já em relação à configuração dos *ensembles* do processo ECE da abordagem proposta *"SEnsembles"*, adotou-se a configuração padrão disponibilizada pelos pacotes *ipred* e *randomforest* da linguagem R (R FOUNDATION, 2016), para os *ensembles bagging* e *random forest*, respectivamente, enquanto que para o *ensemble boosting* (pacote *gbm*) adotou-se 20.000 iterações e o valor 0,0005 para o parâmetro *shrinkage,* conforme utilizado por Lee et al. (2010).

Por fim, ressalta-se que todos os experimentos foram executados no mínimo 20 vezes e os valores das métricas apresentados são valores médios obtidos e, ainda, em todos os experimentos foi realizada análise de variância (ANOVA e Teste F) e testes estatísticos (Teste T e Teste Tukey) para comparação das médias obtidas para os valores da métrica ASAM ao nível de significância de 5%. Assim, foi possível efetuar comparações dos valores médio da métrica ASAM obtidos nos experimentos.

6.2 Comparação das correspondências geradas pelas Configurações 1 e 2 do processo ECS da abordagem proposta "*SEnsembles*"

Esta seção apresenta um comparativo da qualidade da correspondência gerada pelas Configurações 1 e 2 do processo ECS, uma vez que essas configurações se diferenciam entre si por suas estratégias, nas quais as instâncias com características exatamente idênticas são obtidas a partir do conjunto de dados.

Os conjuntos de dados utilizados nos experimentos são descritos na Tabela 6.1, na qual se observa o clássico conjunto de dados Lalonde (1986) e três conjuntos do Programa Bolsa Família, com destaque para o PBF 1, que foi utilizado por Martins (2013) em sua pesquisa de doutorado e, duas modificações desse conjunto (PBF 2 e 3), as quais foram realizadas para aumentar o número de instâncias com características idênticas e, o conjunto de dados do Cenário F (Lee et al., 2010), o qual não possui nenhum par de instâncias com característica idênticas. Assim, observa-se que o conjunto de dados PBF 1 possui 8 pares de instâncias quanto utilizado quatro covariáveis para efetuar a correspondência de instâncias e, apenas um, quando utilizado quatorze. Já o conjunto de dados PBF 2 possui 1.112 pares, enquanto que o PBF 3 possui 2.280 pares, em ambos, obtidos utilizando-se quatorze covariáveis para se efetuar a correspondência das instâncias. Porém, dos 2.280 pares, 1.185 instâncias são de beneficiários contendo 546 instâncias duplicadas e, 1.095 são de não beneficiários contendo 456 instâncias duplicadas.

Tabela 6.1 – Conjuntos de dados utilizados para comparar a qualidade da correspondência das configurações do processo ECS da abordagem proposta "*SEnsembles*", contendo a descrição do conjunto de dados, número do experimento, configuração, quantidade de atributos utilizada para efetuar a correspondência, quantidade de instâncias idênticas e total de instâncias.

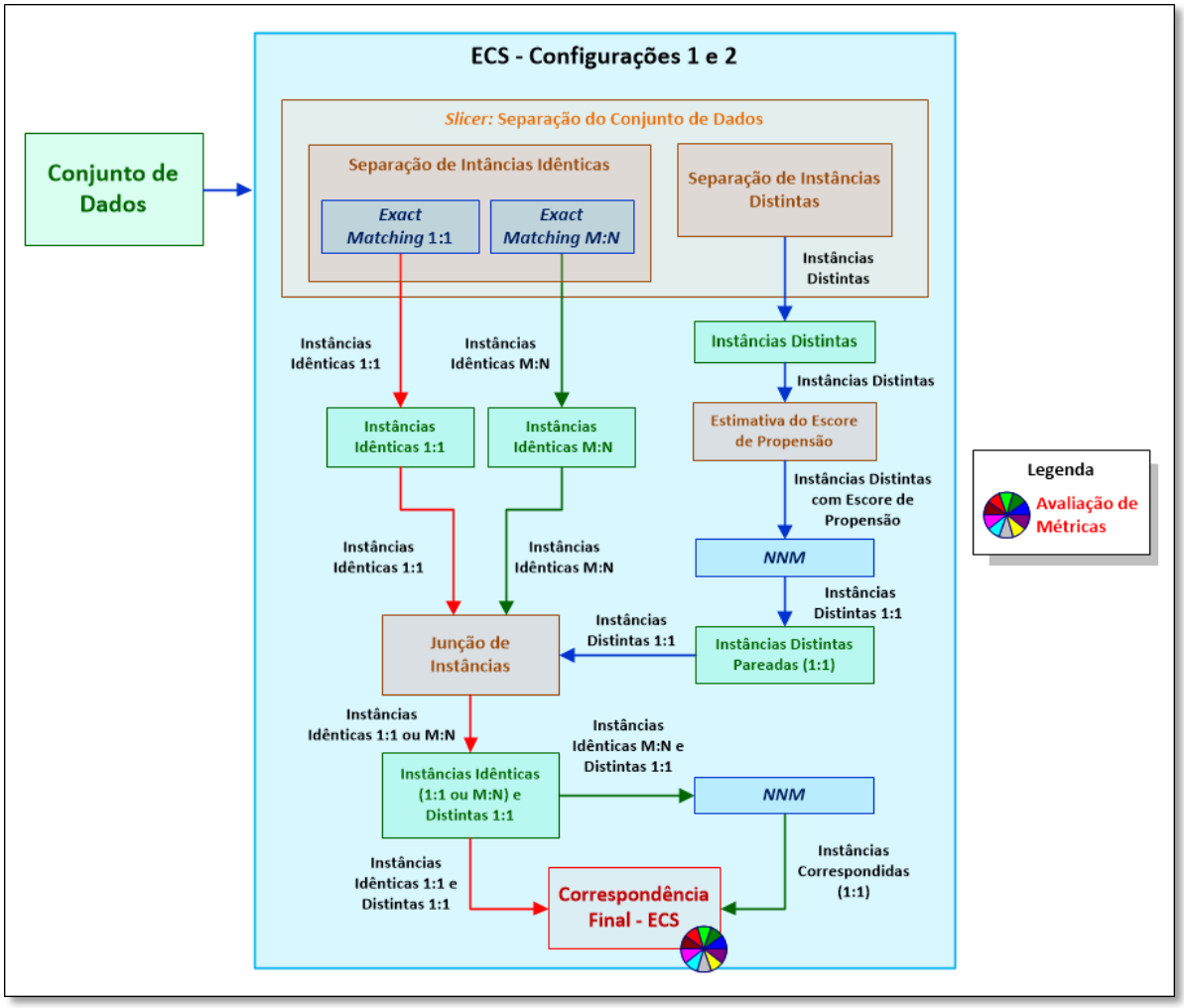
Conjunto de Dados	Exp.*	Conf.**	Atr.***	Instâncias Idênticas	Total de Instâncias
Lalonde (1986)	1	1	8	14 (7 Pares)	614, sendo: 185 Tratados e 429 Não Tratados
		2	8	25 (13 Tratados e 12 Não Tratados)	
PBF 1 (Martins, 2013)	2	1	4	16 (8 Pares)	55.490, sendo: 9.259 Beneficiários e 46.711 Não Beneficiários
		2	4	17 (8 Beneficiários e 9 não Beneficiários)	
	3	1 e 2	14	2 (1 par)	
PBF 2 Modificado a partir de Martins (2013)	4	1 e 2	14	2224 (1.112 pares)	
PBF 3 Modificado a partir de Martins (2013)	5	1	14	1.278 (639 pares)	
		2		2.280 (1.185 Beneficiários e 1.095 não beneficiários);	
Cenário F (Lee et al., 2010)	6	1	10	Nenhuma	1.000, sendo: 542 Tratados e 458 Não Tratados
		2			

*** Experimento; ** Configuração; ***Atributos.**

Fonte: Elaborado pelo autor.

As métricas (i.e. ASAM e número de pares de instâncias correspondidas) foram obtidas em apenas um único estágio de avaliação, que foi realizado depois de efetuado a correspondência por meio das configurações do processo ECS, conforme ilustrado na Figura 6.1.

Figura 6.1 – Estágio de avaliação de métricas para comparar as correspondências finais das Configurações 1 e 2 do processo ECS da abordagem proposta "*SEnsembles*".



Fonte: Elaborado pelo autor.

A Figura 6.2 ilustra o resultado do primeiro experimento no qual se utilizou o conjunto de dados Lalonde (1986) com 10 covariáveis para se efetuar a correspondências das instâncias. Observa-se que a Configuração 1 obteve os melhores valores da métrica ASAM (em verde) para os três menores *calipers* utilizados (zero, 0,05 e 0,10), enquanto que a Configuração 2 obteve para os quatro maiores *calipers* (0,15, 0,20, 0,25 e 0,30). Além disso, a Configuração 1 obteve o maior número de pares de instâncias correspondidas para todos *calipers.*

Figura 6.2 – Resultado das métricas ASAM e número de pares de instâncias correspondidas geradas pelas Configurações 1 e 2 do processo ECS da abordagem proposta "*SEnsembles*" no experimento 1, com variação de *caliper* de 0 a 0,30, e usando o conjunto de dados Lalonde (1986) com 10 covariáveis para efetuar a correspondência das instâncias.

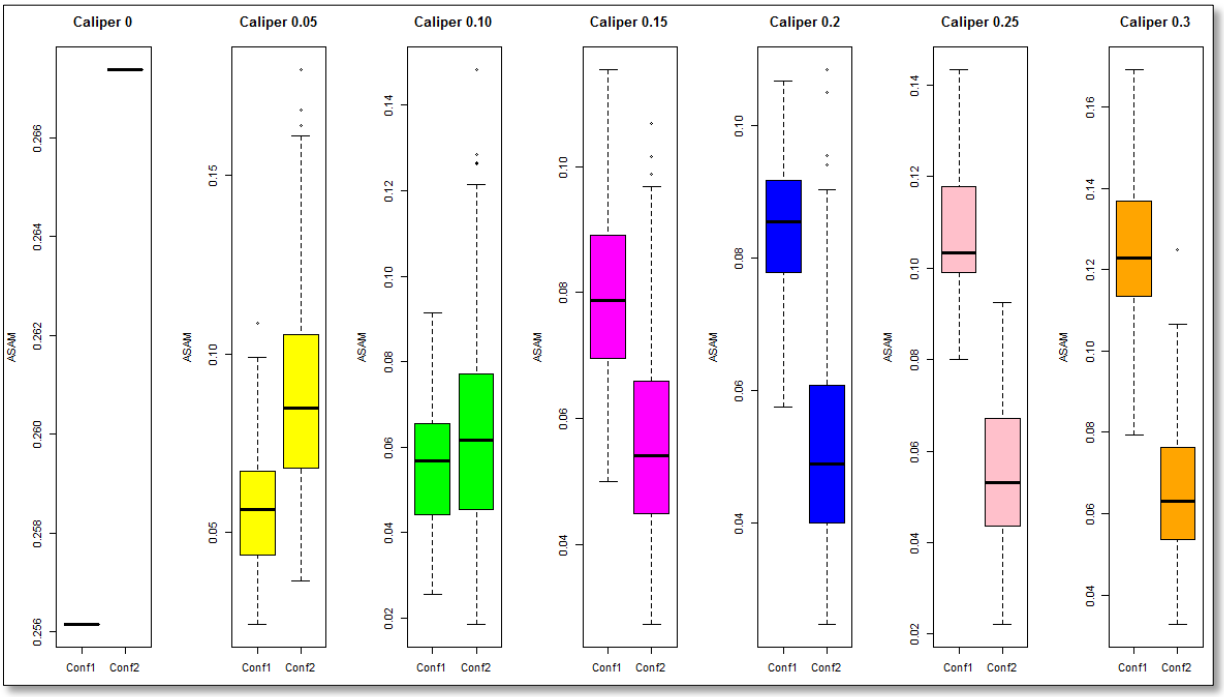
Experimento 1			
Lalonde - 10 Covariáveis			
Calipers	Métricas	Configuração 1	Configuração 2
0	ASAM	**0,2561**	0,2674
	Pares	185	184
0,05	ASAM	**0,0552**	0,0899
	Pares	108	73
0,10	ASAM	**0,0554**	0,0652
	Pares	109	86
0,15	ASAM	0,0804	**0,0567**
	Pares	112	91
0,20	ASAM	0,0849	**0,0530**
	Pares	113	96
0,25	ASAM	0,1077	**0,0550**
	Pares	115	95,53
0,30	ASAM	0,1252	**0,0654**
	Pares	116	98
Melhor ASAM			

Fonte: Elaborado pelo autor.

Os valores da métrica ASAM foram obtidos por meio de 100 execuções do Experimento 1, ou seja, os valores apresentados são valores médios obtidos. A Figura 6.3 ilustra a variação desses valores considerando-se as Configurações 1 e 2 do Processo ECS e os *calipers* utilizados (0 a 0,30). Observa-se que há uma semelhança das variações dos valores médios obtidos da métrica ASAM pelas Configurações 1 e 2 com os *calipers* 0,25 e 0,30, e de acordo com o resultado do Teste F para análise de variância, pode-se concluir que não há evidências de diferenças significativas das variâncias dos valores da métrica ASAM obtidos pelas Configurações 1 e 2, com os *calipers* 0,25 e 0,30, ao nível de significância de 5%, uma vez que o p-valor obtido foi de 0,9911 e 0.9844, respectivamente. Já de acordo com o Teste T e Teste Tukey, para comparação das médias da métrica ASAM, pode-se concluir que há evidências de diferenças significativas dos valores médios da métrica ASAM, ao nível de significância de 5%, obtidos pelas Configurações 1 e 2

em todos os *calipers* utilizados, pois os p-valores obtidos foram inferiores ao nível de significância adotado.

Figura 6.3 – Variação da métrica ASAM geradas pelas Configurações 1 e 2 do processo ECS da abordagem proposta "*SEnsembles*" no experimento 1, com *calipers* de 0 a 0,30, e usando o conjunto de dados Lalonde (1986) com 10 covariáveis para efetuar a correspondência das instâncias.



Fonte: Elaborado pelo autor.

Já nos experimentos 2 e 3 utilizou-se o mesmo conjunto de dados do PBF (PBF 1), mas com uma diferença em relação à quantidade de covariáveis utilizada para se efetuar as correspondências das instâncias, sendo 4 covariáveis no experimento 2, e 14 no experimento 3. A Figura 6.4 ilustra os resultados desses experimentos, na qual se observa que a Configuração 1, no experimento 2, obteve o melhor valor da métrica ASAM em seis *calipers* (exceto no *caliper* 0,05) e o maior número de pares de instâncias correspondidas, considerando-se todos os *calipers.* Porém, essa superioridade não foi encontrada no experimento 3, no qual a Configuração 1 somente obteve o melhor ASAM em dois *calipers* (zero e 0,15), enquanto que a Configuração 2 obteve em seis (*excetuando* o *caliper* 0,15), mas sempre resultando no menor número de pares de instâncias correspondidas. Além disso, nota-se, no experimento 3, que as duas configurações obtiveram o mesmo

valor da métrica ASAM com o *caliper* zero, uma vez que as duas obtiveram o mesmo número instâncias com características idênticas (um par).

Figura 6.4 – Resultado das métricas ASAM e número de pares de instâncias correspondidas geradas pelas Configurações 1 e 2 do processo ECS da abordagem proposta "*SEnsembles*" nos experimentos 2 e 3, com variação de *caliper* de 0 a 0,30, e usando o conjunto de dados PBF (Martins, 2013) com 4 e 14 covariáveis para efetuar a correspondência das instâncias.

		Experimento 2		Experimento 3	
		PBF 1 - 4 Covariáveis		PBF 1 - 14 Covariáveis	
Calipers	Métricas	Configuração 1	Configuração 2	Configuração 1	Configuração 2
0	ASAM	**0,0755**	0,0772	**0,0557**	**0,0557**
	Pares	9259	9259	9259	9259
0,05	ASAM	0,0304	**0,0189**	0,0178	**0,0082**
	Pares	8106	7709	7839	7499
0,10	ASAM	**0,0278**	0,0369	0,0169	**0,0142**
	Pares	8261	8075	7985	7817
0,15	ASAM	**0,0343**	0,0416	**0,0209**	0,0220
	Pares	8412	8251	8135	8031
0,20	ASAM	**0,0495**	0,0561	0,0263	**0,0259**
	Pares	8554	8312	8263	8130
0,25	ASAM	**0,0632**	0,0712	0,0340	**0,0328**
	Pares	8680	8360	8396	8221
0,30	ASAM	**0,0813**	0,0856	0,0449	**0,0418**
	Pares	8812	8461	8530	8311

Melhor ASAM

Fonte: Elaborado pelo autor.

Os resultados dos experimentos 2 e 3 demonstram que a quantidade de covariáveis utilizada para se efetuar a correspondência das instâncias influenciou na quantidade de instâncias idênticas obtida pelas configurações, sendo 8 pares quando utilizado 4 covariáveis (experimento 2) e, apenas um único par, quando utilizado 14 covariáveis (experimento 3), o que também influenciou nas métricas obtidas.

Em se tratando do experimento 4, no qual se utilizou o conjunto de dados PBF 2 contendo 1.112 pares de instâncias com características idênticas, os resultados foram similares aos resultados do experimento 2, ou seja, a Configuração 1 obteve os melhores valores da métrica ASAM (exceto com o *caliper* 0,05) e o maior número de pares de instâncias correspondidas, conforme se observa na Figura 6.5. Nota-se, também, que ocorreu novamente um empate das configurações

com o *caliper* zero, pois em ambas as configurações obteve-se o mesmo número de instâncias com características idênticas. Porém, com os demais *calipers* a Configuração 1 foi superior a 2.

Figura 6.5 – Resultado das métricas ASAM e número de pares de instâncias correspondidas geradas pelas Configurações 1 e 2 do processo ECS da abordagem proposta "*SEnsembles*" no experimento 4, com variação de *caliper* de 0 a 0,30, e usando o conjunto de dados PBF 2 com 14 covariáveis para efetuar a correspondência das instâncias.

Experimento 4			
PBF 2 - 14 Covariáveis			
Calipers	Métricas	Configuração 1	Configuração 2
0	ASAM	**0,0380**	**0,0380**
	Pares	9259	9259
0,05	ASAM	0,0153	**0,0072**
	Pares	8225	7961
0,10	ASAM	**0,0145**	0,0146
	Pares	8331	8227
0,15	ASAM	**0,0175**	0,0213
	Pares	8442	8309
0,20	ASAM	**0,0204**	0,0248
	Pares	8550	8359
0,25	ASAM	**0,0275**	0,0316
	Pares	8662	8411
0,30	ASAM	**0,0357**	0,0401
	Pares	8768	8463
Melhor ASAM			

Fonte: Elaborado pelo autor.

Já no experimento 5, no qual se utilizou-se o conjunto de dados PBF 3 contendo instâncias duplicadas pertencentes a um mesmo grupo, seja de beneficiários (tratamento) ou de não beneficiários (controle), a Configuração 1 obteve os melhores valores da métrica ASAM em cinco *calipers* (exceto nos *calipers* 0,05 e 0,10) e o maior número de pares de instâncias correspondidas em todos, conforme ilustrado na Figura 6.6. Observa-se também que a Configuração 2 obteve o melhor ASAM com os *calipers* 0,05 e 0,10, mas com redução do número de pares de instâncias correspondidas. Essa situação foi recorrente nos experimentos, ou seja, quando a Configuração 2 obteve o melhor ASAM, sempre houve uma redução do número de pares de instâncias correspondidas.

Figura 6.6 – Resultado das métricas ASAM e número de pares de instâncias correspondidas geradas pelas Configurações 1 e 2 do processo ECS da abordagem proposta "*SEnsembles*" no experimento 5, com variação de *caliper* de 0 a 0,30, e usando o conjunto de dados PBF 3 com 14 covariáveis para efetuar a correspondência das instâncias.

Experimento 5			
PBF 3 - 14 Covariáveis			
Calipers	Métricas	Configuração 1	Configuração 2
0	ASAM	**0,0373**	0,0550
	Pares	9259	9169
0,05	ASAM	0,0153	**0,0071**
	Pares	8202	7903
0,10	ASAM	0,0155	**0,0154**
	Pares	8321	8143
0,15	ASAM	**0,0172**	0,0222
	Pares	8435	8236
0,20	ASAM	**0,0227**	0,0286
	Pares	8555	8285
0,25	ASAM	**0,0302**	0,0347
	Pares	8680	8342
0,30	ASAM	**0,0390**	0,0402
	Pares	8801	8408
Melhor ASAM			

Fonte: Elaborado pelo autor.

Por último, no experimento 6, no qual se utilizou o conjunto de dados do Cenário F (Lee et al, 2010), que não possui nenhum par de instâncias com características idênticas, a Configuração 1 obteve os melhores valores da métrica ASAM em seis *calipers* (exceto 0,10) e um empate com a Configuração 2 com o *caliper* 0,15, enquanto que a Configuração 2 somente obteve o melhor ASAM em dois *calipers* (0,10 e 0,15). Porém, os valores obtidos foram próximos, com destaque para o mesmo número de pares de instâncias correspondidas, conforme se observa na Figura 6.7.

Figura 6.7 – Resultado das métricas ASAM e número de pares de instâncias correspondidas geradas pelas Configurações 1 e 2 do processo ECS da abordagem proposta "*SEnsembles*" no experimento 5, com variação de *caliper* de 0 a 0,30, e usando o conjunto de dados Cenário F (Lee et al., 2010) com 10 covariáveis para efetuar a correspondência das instâncias.

Experimento 6			
Cenário F - 10 Covariáveis			
Calipers	Métricas	Configuração 1	Configuração 2
0	ASAM	**0,3847**	0,3847
	Pares	456	456
0,05	ASAM	**0,0302**	0,0304
	Pares	293	293
0,10	ASAM	0,0345	**0,0340**
	Pares	302	302
0,15	ASAM	**0,0415**	**0,0415**
	Pares	310	310
0,20	ASAM	**0,0526**	0,0528
	Pares	317	317
0,25	ASAM	**0,0680**	0,0681
	Pares	325	325
0,30	ASAM	**0,0800**	0,0803
	Pares	333	333
Melhor ASAM			

Fonte: Elaborado pelo autor.

A Tabela 6.2 apresenta um resumo dos resultados dos experimentos desta seção, com a descrição dos conjuntos de dados utilizados nos experimentos, configurações aplicadas do processo ECS e os melhores resultados obtidos da métrica ASAM. Observa-se que a Configuração 1 obteve as melhores correspondências com o menor valor de ASAM, excetuando, principalmente, o *caliper* 0,05. Nota-se também que a Configuração 2 obteve os melhores resultados quando o número de instâncias com características idênticas foi pequeno, no caso do experimento 3, quando esse número foi de um único par. Entretanto, quando o número de instâncias com características exatamente idênticas foi maior, como no experimento 4, a Configuração 1 foi superior a 2, a qual obteve os melhores valores de ASAM em seis *calipers*.

Tabela 6.2 – Resumo dos resultados dos experimentos pelos quais se comparou as correspondências das Configurações 1 e 2 do processo ECS da abordagem proposta "*SEnsembles*", com a descrição dos conjuntos de dados utilizados, número do experimento, configuração utilizada, quantidade de atributo utilizada para se efetuar a correspondência de instâncias, quantidade de instâncias idênticas dos conjuntos de dados e os melhores resultados da métrica ASAM.

Conjunto de Dados	Exp.*	Conf.**	Atr. ***	Instâncias Idênticas	Melhores Resultados (ASAM)
Lalonde (1986)	1	1	8	14 (7 Pares)	Configuração 1: em três *calipers* (zero, 0,05 e 0,10).
		2	8	25 (13 Tratados e 12 Não Tratados)	Configuração 2: em quatro *calipers* (0,15, 0,20, 0,25 e 0,30).
PBF 1 (Martins, 2013)	2	1	4	16 (8 Pares)	Configuração 1: em seis *calipers* (zero, 0,10, 0,15, 0,20, 0,25 e 0,30).
		2	4	17 (8 Beneficiários e 9 não Beneficiários)	Configuração 2: um *caliper* (0,05).
	3	1 e 2	14	2 (1 par)	Configuração 1: em dois *calipers* (zero e 0,15). Empate com a Configuração 2 com o *caliper* zero. Configuração 2: em cinco *calipers*(0,05, 0,10, 0,20, 0,25 e 0,30).
PBF 2 Modificado a partir de Martins (2013)	4	1 e 2	14	2.224 (1.112 pares)	Configuração 1: em seis *calipers* (zero, 0,10, 0,15, 0,20, 0,25 e 0,30). Empate com a Configuração 2 com o *caliper* zero. Configuração 2: um *caliper* (0,05).
PBF 3 Modificado a partir de Martins (2013)	5	1	14	1.278 (639 pares)	Configuração 1: em cinco *calipers* (zero, 0,15, 0,20, 0,25 e 0,30). Configuração 2: em dois *calipers* (0,05 e 0,10)
		2		2.280 (1.185 Beneficiários e 1.095 não beneficiários)	
Cenário F (Lee et al., 2010	6	1 e 2	10	Nenhuma	Configuração 1: em seis *calipers* (zero, 0,05, 0,15, 0,20, 0,25 e 0,30). Empate com a Configuração 2 em dois *calipers* (0,10 e 0,15). Configuração 2: em dois *calipers* (0,10 e 0,15)

***Experimento; **Configuração; ***Atributos.**

Fonte: Elaborado pelo autor.

Como visto, nesta seção buscou-se comparar a qualidade das correspondências geradas pelas configurações do processo ECS, com base na métrica ASAM e, de forma, adicional, utilizou-se também o número de pares de instâncias correspondidas. Os resultados demonstram que quando há uma quantidade expressiva do número de instâncias com características idênticas, como ocorreu nos experimentos 4 e 5, a Configuração 1 obtém os melhores valores para a métrica ASAM para a maioria dos *calipers* analisados, enquanto que a Configuração 2 obtém os melhores valores quando essa quantidade é pequena, como ocorreu no experimento 3 (um único par). Dessa forma, em geral, a Configuração 1 foi superior a 2, mas isso pode variar de acordo com o conjunto de dados e o número de instâncias com características idênticas desse conjunto.

6.3 Comparação das correspondências geradas pelo Processo ECE da abordagem proposta "*SEnsembles*"

Esta seção apresenta um comparativo da qualidade da correspondência de instâncias gerada pelo processo ECE, o qual possui uma estratégia que se baseia na substituição da regressão logística por *ensembles* de regressores para estimar os escores de propensão. Assim, ao comparar essas correspondências, buscou-se avaliar quais *ensembles* produziriam estimativas que propiciaram as melhores correspondências de instâncias.

Nos experimentos utilizou-se o clássico conjunto de dados Lalonde (1986), os conjuntos de dados descritos no trabalho de Lee et al. (2010), os quais foram construídos de forma sintética com a codificação do Anexo A e, por fim, o conjunto de dados PBF 3, o qual possui instâncias duplicados (iguais) pertencentes a um mesmo grupo, seja de beneficiários ou de não beneficiários.

Os cenários descritos por Lee et al. (2010) foram escolhidos por possuírem covariáveis com maiores e menores valores de linearidade e aditividade e, além disso, porque foram utilizados por esses autores para comparar os *ensembles* em uma abordagem que pondera as instâncias pelo escore de propensão, ou seja, em uma abordagem IPSW. Dessa forma, a escolha desses cenários possibilitou

comparar os mesmos *ensembles* de regressores em uma abordagem PSM, a qual utiliza os escores de propensão para gerar as correspondências das instâncias e, não, ponderá-los.

Ressalta-se que para cada cenário (B a G) foram construídos vinte conjuntos de dados, contendo em cada um 1.000 instâncias, conforme se observa na Tabela 6.3, a qual apresenta uma descrição dos conjuntos de dados utilizados nos experimentos desta seção.

Tabela 6.3 – Conjuntos de dados utilizados para comparar a qualidade da correspondência do processo ECS da abordagem proposta "*SEnsembles*", contendo a descrição do conjunto de dados, número do experimento, quantidade de atributos utilizada para efetuar correspondência, quantidade de conjuntos utilizados em cada experimento e o total de instâncias.

Conjunto de Dados		Exp.*	Atributos	Conjuntos Utilizados	Instâncias
Lalonde (1986)		1	8	1	614
Lee et al. (2010)	Cenário B: Leve (*mild*) não linearidade.	2	10	20	1.000
	Cenário C: Moderada não linearidade.	3	10	20	1.000
	Cenário D: Leve (*mild)* não aditividade.	4	10	20	1.000
	Cenário E: Leve (*mild)* não aditividade e não linearidade.	5	10	20	1.000
	Cenário F: Moderada não aditividade.	6	10	20	1.000
	Cenário G: Moderada não aditividade e não linearidade	7	10	20	1.000
PBF 3 Modificada a partir de Martins (2013)		8	14	1	55.970

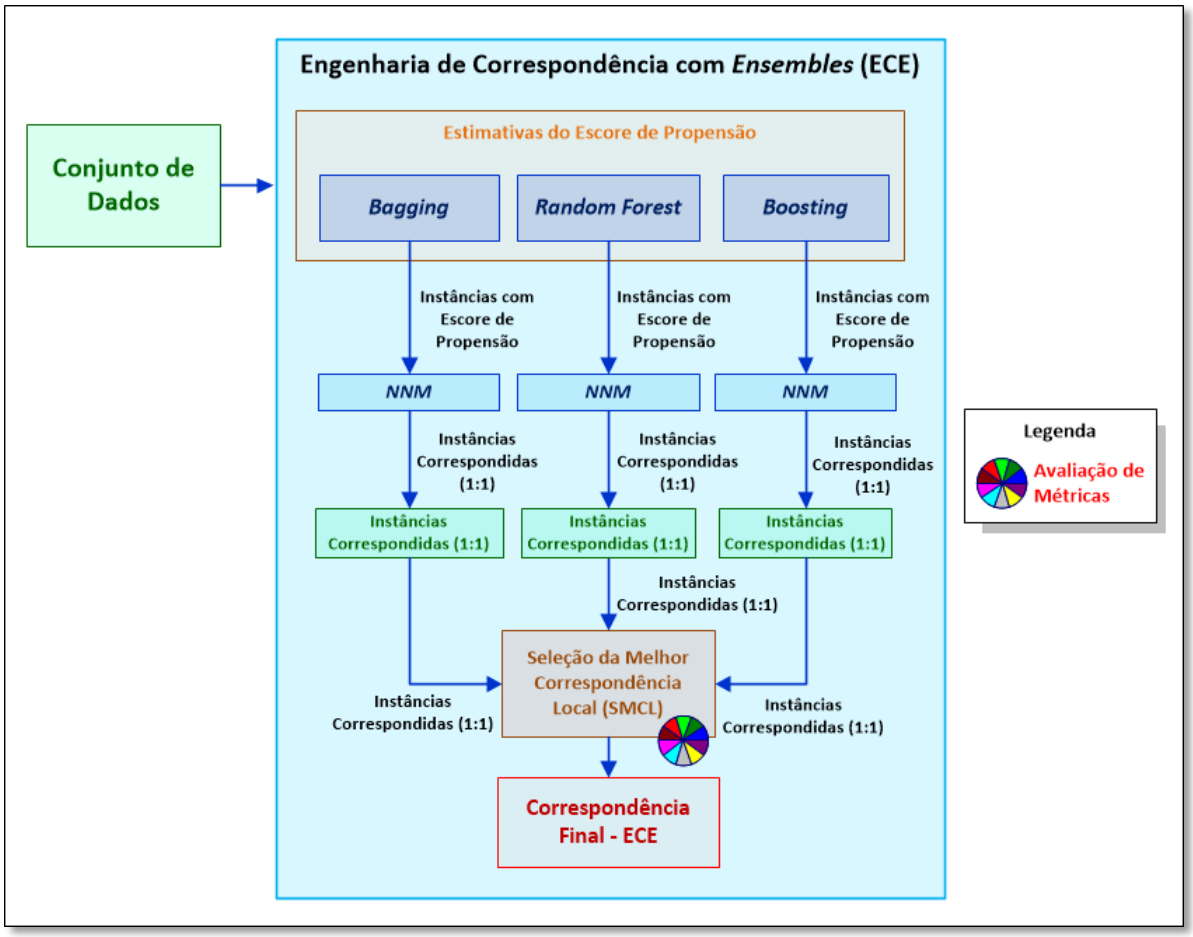
*** Experimento.**

Fonte: Elaborado pelo autor.

As métricas adotadas para comparar as correspondências do processo ECE são as mesmas usadas anteriormente para comparar as configurações do processo ECS na Seção 6.2 (i.e. ASAM e o numero de pares de instâncias correspondidas),

as quais foram obtidas em um único estágio, ao final de cada correspondência gerada, ou seja, na etapa SMCL conforme ilustrado na Figura 6.8.

Figura 6.8 – Estágio de avaliação de métricas para comparar as correspondências do processo ECE da abordagem proposta "*SEnsembles*".



Fonte: Elaborado pelo autor.

Já em relação aos resultados, no experimento 1, no qual se utilizou o conjunto de dados Lalonde (1986), o *ensemble random forest* ao substituir a regressão logística proporcionou os melhores resultados da métrica ASAM em todos os *calipers* e, também, o mesmo número de pares de instâncias correspondidas que o obtido quando utilizado o *ensembles boosting*, conforme ilustrado pela cor verde na Figura 6.9. Por outro lado, o maior número de pares de instâncias correspondidas foi obtido ao se utilizar o *ensemble bagging*, mas com aumento, talvez não desejável, da métrica ASAM.

Figura 6.9 – Resultado das métricas ASAM e número de pares de instâncias correspondidas geradas pelo processo ECE da abordagem proposta "*SEnsembles*" no experimento 1, com base em escores de propensão estimados por *ensembles* de regressores, *bagging*, *random forest* e *boosting*, com variação de *caliper* de 0 a 0,30 e, usando o conjunto de dados Lalonde (1986) com 10 covariáveis para efetuar a correspondência das instâncias.

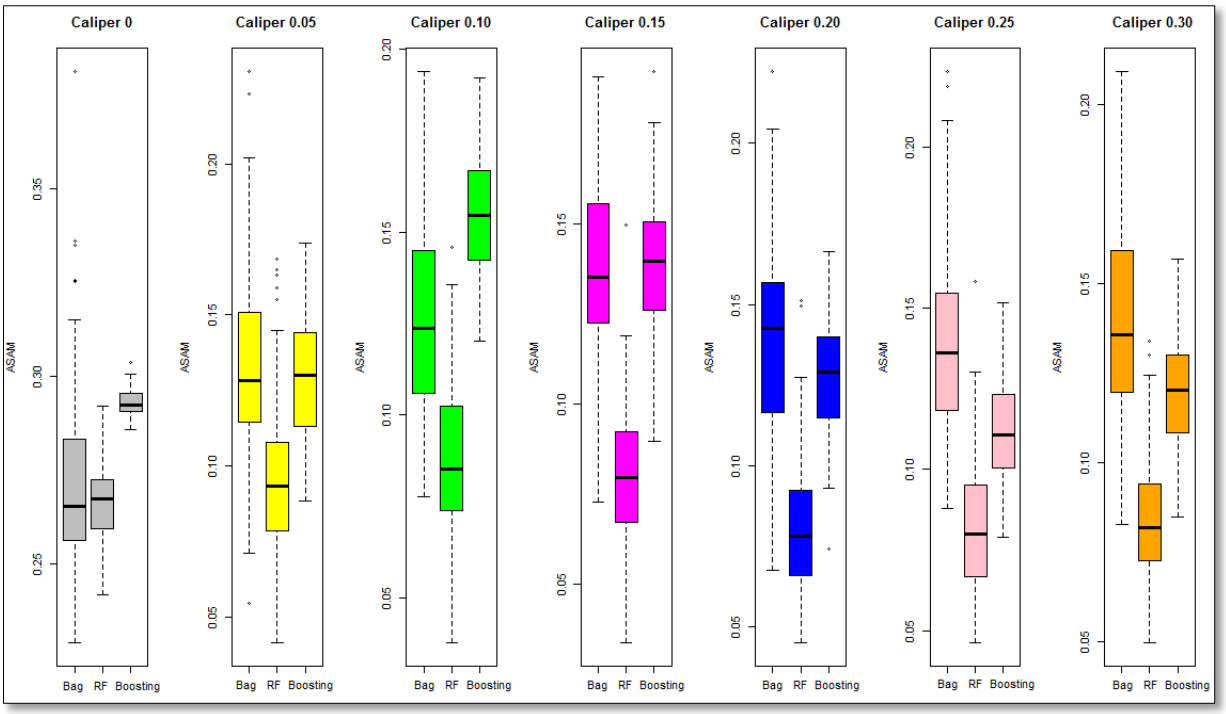
Experimento 1 - Lalonde				
Calipers	Métricas	*Bagging*	*Random Forest*	*Boosting*
0	ASAM	0,2714	**0,2664**	0,2929
	Pares	185	185	185
0,05	ASAM	0,1325	**0,0957**	0,1299
	Pares	98	85	84
0,10	ASAM	0,1276	**0,0871**	0,1541
	Pares	102	90	90
0,15	ASAM	0,1373	**0,0802**	0,1391
	Pares	105	92	92
0,20	ASAM	0,1380	**0,0815**	0,1274
	Pares	108	95	95
0,25	ASAM	0,1390	**0,0830**	0,1125
	Pares	111	98	98
0,30	ASAM	0,1392	**0,0840**	0,1206
	Pares	113	101	101
Melhor ASAM				

Fonte: Elaborado pelo autor.

A Figura 6.10 ilustra a variação dos valores da métrica ASAM do experimento 1, considerando-se os *ensembles bagging* (Bag), *random forest* (RF) e *boosting* do Processo ECE da abordagem proposta "*SEnsembles*" e os *calipers* utilizados (0 a 0,30) e, de acordo com o Teste F para análise de variância, pode-se concluir que não há evidencia de diferença significativa da variância dos valores obtidos da métrica ASAM, ao nível de significância de 5%, entre os *ensembles bagging* e *random forest* com o *caliper* 0,05 (p-valor: 0,06223) e, entre os *ensembles random forest* e *boosting* com o *caliper* 0,15 (p-valor: 0,5945) e 0,30 (p-valor: 0,3008).

Já em relação aos testes estatísticos para comparação da média, não há evidência de diferença significativa dos valores médios da métrica ASAM, ao nível de significância de 5%, entre os *ensembles bagging* e *random forest* com o *caliper* zero e, entre os *ensembles bagging* e *boosting* com os *calipers* 0,05 e 0,15, uma vez que o Teste T resultou em p-valores de 0,0625, 0,4867 e 0,5445, respectivamente e, p-valores de 0,0599, 0,7625 e 0,8052 no Teste Tukey.

Figura 6.10 – Variação da métrica ASAM geradas pelo processo ECE da abordagem proposta "*SEnsembles*" no experimento 1, considerando os *ensembles bagging* (*Bag*), *random forest* (RF) e *boosting*, com *calipers* de 0 a 0,30, e usando o conjunto de dados Lalonde (1986) com 10 covariáveis para efetuar a correspondência das instâncias.



Fonte: Elaborado pelo autor.

Já nos experimentos 2 e 5, nos quais se utilizou, respectivamente, os Cenários B e E (Lee et al., 2010), observou-se que os melhores valores para a métrica ASAM, em todos os *calipers*, foram obtidos quando utilizado o *ensemble boosting,* conforme observa-se na Figura 6.11. Além disso, excetuando o *caliper* zero, o melhor valor da métrica ASAM foi obtido sempre acompanhado de uma redução do número de pares de instâncias correspondidas. Por outro lado, o *ensemble bagging* apresentou, mais uma vez, o maior número de pares de instâncias correspondidas e os valores mais altos da métrica ASAM em todos os *calipers.*

Figura 6.11 – Resultado das métricas ASAM e número de pares de instâncias correspondidas geradas pelo processo ECE da abordagem proposta "*SEnsembles*" nos experimentos 2 e 5, com base em escores de propensão estimados por *ensembles* de regressores, *bagging*, *random forest* e *boosting*, com variação de *caliper* de 0 a 0,30 e, usando, respectivamente nesses experimentos, os conjuntos de dados dos Cenários B e E (Lee et al., 2010) com 10 covariáveis para efetuar a correspondência das instâncias.

Experimento 2 - Cenário B				
Calipers	Métricas	*Bagging*	*Random Forest*	*Boosting*
0	ASAM	0,2304	0,2229	**0,2110**
	Pares	455	455	455
0,05	ASAM	0,1286	0,1026	**0,0590**
	Pares	339	332	282
0,10	ASAM	0,1377	0,1091	**0,0556**
	Pares	347	343	291
0,15	ASAM	0,1428	0,1166	**0,0545**
	Pares	356	351	299
0,20	ASAM	0,1498	0,1236	**0,0538**
	Pares	363	359	307
0,25	ASAM	0,1566	0,1318	**0,0567**
	Pares	370	367	316
0,30	ASAM	0,1630	0,1415	**0,0620**
	Pares	375	374	324

Experimento 5 - Cenário E				
Calipers	Métricas	*Bagging*	*Random Forest*	*Boosting*
0	ASAM	0,2573	0,2517	**0,2399**
	Pares	459	459	459
0,05	ASAM	0,1317	0,0998	**0,0620**
	Pares	319	311	264
0,10	ASAM	0,1365	0,1065	**0,0604**
	Pares	328	321	273
0,15	ASAM	0,1464	0,1142	**0,0581**
	Pares	336	329	281
0,20	ASAM	0,1550	0,1222	**0,0578**
	Pares	345	336	289
0,25	ASAM	0,1633	0,1310	**0,0586**
	Pares	353	344	298
0,30	ASAM	0,1661	0,1412	**0,0639**
	Pares	357	352	306

Melhor ASAM

Fonte: Elaborado pelo autor.

Nos demais experimentos, o *ensemble bagging* resultou nos melhores valores da métrica ASAM com o *caliper* zero, conforme se observa nas Figuras 6.12 (experimentos 3 e 4), Figura 6.13 (experimentos 6 e 7) e Figura 6.14 (experimento 8). Porém, nesses experimentos o *ensemble boosting* propiciou os melhores resultados da métrica ASAM nos *calipers* superiores a zero (em verde nas figuras a seguir), ao mesmo tempo em que reduziu o número de pares de instâncias correspondidas.

Ressalta-se, que nos experimentos 3, 4, 6 e 7, nos quais se utilizou, respectivamente, os conjuntos de dados referentes aos cenários C, D, F e G (Lee et al., 2010), o *ensemble bagging* resultou no maior número de pares de instâncias correspondidas. Entretanto, no experimento 8, no qual se utilizou o conjunto de dados do PBF 3, o maior número de pares de instâncias foi obtido em todos os *calipers* quando utilizado o *ensemble random forest* (Figura 6.14).

Figura 6.12 – Resultado das métricas ASAM e número de pares de instâncias correspondidas geradas pelo processo ECE da abordagem proposta "*SEnsembles*" nos experimentos 3 e 4, com base em escores de propensão estimados por *ensembles* de regressores, *bagging*, *random forest* e *boosting*, com variação de *caliper* de 0 a 0,30 e, usando, respectivamente nesses experimentos, os conjuntos de dados dos Cenários C e D (Lee et al., 2010) com 10 covariáveis para efetuar a correspondência das instâncias.

Experimento 3 - Cenário C				
Calipers	Métricas	*Bagging*	*Random Forest*	*Boosting*
0	ASAM	**0,3260**	0,3289	0,3365
	Pares	446	446	446
0,05	ASAM	0,1000	0,0705	**0,0633**
	Pares	305	299	259
0,10	ASAM	0,1036	0,0753	**0,0599**
	Pares	316	308	268
0,15	ASAM	0,1100	0,0808	**0,0590**
	Pares	326	316	277
0,20	ASAM	0,1147	0,0883	**0,0567**
	Pares	332	324	285
0,25	ASAM	0,1234	0,0961	**0,0579**
	Pares	340	333	294
0,30	ASAM	0,1302	0,1058	**0,0611**
	Pares	347	341	302

Experimento 4 - Cenário D				
Calipers	Métricas	*Bagging*	*Random Forest*	*Boosting*
0	ASAM	**0,3520**	0,3543	0,3598
	Pares	473	473	473
0,05	ASAM	0,1377	0,1043	**0,0610**
	Pares	324	309	264
0,10	ASAM	0,1463	0,1124	**0,0573**
	Pares	334	319	273
0,15	ASAM	0,1534	0,1211	**0,0558**
	Pares	346	327	280
0,20	ASAM	0,1621	0,1303	**0,0561**
	Pares	354	335	288
0,25	ASAM	0,1709	0,1404	**0,0594**
	Pares	360	343	297
0,30	ASAM	0,1829	0,1494	**0,0679**
	Pares	366	350	305

Melhor ASAM

Fonte: Elaborado pelo autor.

Figura 6.13 – Resultado das métricas ASAM e número de pares de instâncias correspondidas geradas pelo processo ECE da abordagem proposta "*SEnsembles*" nos experimentos 6 e 7, com base em escores de propensão estimados por *ensembles* de regressores, *bagging*, *random forest* e *boosting*, com variação de *caliper* de 0 a 0,30 e, usando, respectivamente nesses experimentos, os conjuntos de dados dos Cenários F e G (Lee et al., 2010) com 10 covariáveis para efetuar a correspondência das instâncias.

Experimento 6 - Cenário F				
Calipers	Métricas	*Bagging*	*Random Forest*	*Boosting*
0	ASAM	**0,3682**	0,3698	0,3828
	Pares	456	456	456
0,05	ASAM	0,1187	0,1009	**0,0542**
	Pares	322	317	277
0,10	ASAM	0,1275	0,1089	**0,0500**
	Pares	333	327	286
0,15	ASAM	0,1365	0,1173	**0,0500**
	Pares	342	335	295
0,20	ASAM	0,1457	0,1277	**0,0500**
	Pares	347	343	303
0,25	ASAM	0,1552	0,1380	**0,0548**
	Pares	354	350	311
0,30	ASAM	0,1659	0,1479	**0,0631**
	Pares	365	358	320

Experimento 7 - Cenário G				
Calipers	Métricas	*Bagging*	*Random Forest*	*Boosting*
0	ASAM	**0,3064**	0,3104	0,3175
	Pares	458	458	458
0,05	ASAM	0,1067	0,0749	**0,0631**
	Pares	315	304	263
0,10	ASAM	0,1118	0,0795	**0,0620**
	Pares	322	314	272
0,15	ASAM	0,1158	0,0844	**0,0605**
	Pares	329	322	281
0,20	ASAM	0,1225	0,0915	**0,0587**
	Pares	335	329	289
0,25	ASAM	0,1284	0,0999	**0,0601**
	Pares	343	338	298
0,30	ASAM	0,1374	0,1095	**0,0609**
	Pares	357	346	307

Melhor ASAM

Fonte: Elaborado pelo autor.

Figura 6.14 – Resultados das métricas ASAM e número de pares de instâncias correspondidas geradas pelo processo ECE da abordagem proposta "*SEnsembles*" no experimento 8, com base em escores de propensão estimados por *ensembles* de regressores, *bagging*, *random forest* e *boosting*, com variação de *caliper* de 0 a 0,30 e, usando o conjunto de dados PBF 3 com 14 covariáveis para efetuar a correspondência das instâncias.

Experimento 8 - PBF 3				
Calipers	Métricas	*Bagging*	*Random Forest*	*Boosting*
0	ASAM	**0,0191**	0,0900	0,0880
	Pares	9259	9259	9259
0,05	ASAM	0,1883	0,0845	**0,0521**
	Pares	8580	8893	8037
0,1	ASAM	0,1941	0,0905	**0,0616**
	Pares	8640	8988	8180
0,15	ASAM	0,1965	0,0965	**0,0714**
	Pares	8668	9089	8327
0,20	ASAM	0,1982	0,1028	**0,0835**
	Pares	8670	9172	8472
0,25	ASAM	0,2018	0,1112	**0,0928**
	Pares	8694	9245	8619
0,30	ASAM	0,2014	0,1167	**0,1050**
	Pares	8699	9259	8762
Melhor ASAM				

Fonte: Elaborado pelo autor.

A Tabela 6.4 apresenta um resumo dos resultados dos experimentos desta seção. Observa-se que os melhores valores da métrica ASAM foram obtidos por *ensembles* diferentes ao substituírem a regressão logística. Assim, no experimento 1 obteve-se os melhores valores com o *ensemble random forest*, *boosting* nos experimentos 2 e 5, e nos demais experimentos (3, 4, 6, 7 e 8), com o *ensemble bagging* com o *caliper zero* e, com o *ensemble boosting*, com os demais *calipers.*

Tabela 6.4 – Resumo dos resultados dos experimentos pelos quais se comparou as correspondências geradas pelo processo ECE da abordagem proposta "*SEnsembles*", com a descrição dos conjuntos de dados, número do experimento e os melhores resultados obtidos da métrica ASAM.

Conjunto de Dados		Experimento	Melhores Resultados (ASAM)
Lalonde (1986)		1	*Random Forest* em todos os *calipers*.
Lee et al. (2010)	Cenário B: Leve (*mild*) não linearidade.	2	*Boosting* em todos os *calipers.*
	Cenário E: Leve (*mild)* não aditividade e não linearidade.	5	
	Cenário C: Moderada não linearidade	3	*Bagging* com o *caliper* zero. *Boosting* nos demais *calipers*.
	Cenário D: Leve (*mild)* não aditividade.	4	
	Cenário F: Moderada não aditividade.	6	
	Cenário G: Moderada não aditividade e não linearidade	7	
PBF 3 Modificado a partir de Martins (2013)		8	

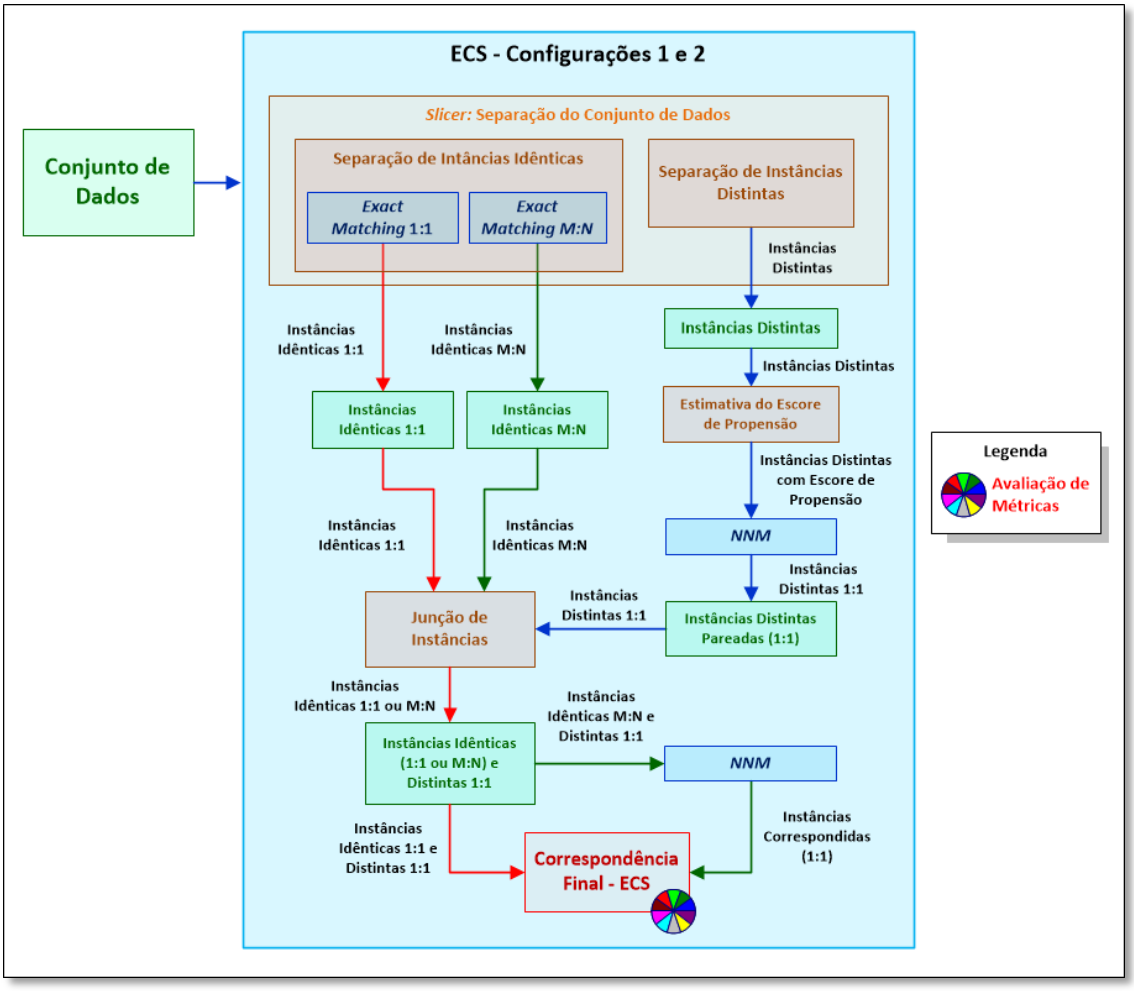
Fonte: Elaborado pelo autor.

Como visto na Tabela 6.4 (acima), não houve um *ensemble* que resultou nos melhores resultados da métrica ASAM considerando-se todos os experimentos. Pelo contrário, nota-se que os três *ensembles* utilizados, *bagging*, *random forest* e *boosting*, resultaram em correspondências com os melhores valores dessa métrica com o *caliper* zero ao se efetuar a correspondência das instâncias. Dessa forma, justifica-se o uso desses *ensembles* na abordagem "*SEnsembles*".

6.4 Comparação do processo ECS da abordagem proposta "*SEnsembles*" com o *baseline NNM*

Esta seção apresenta um comparativo da qualidade da correspondência de instâncias gerada pelo processo ECS da abordagem proposta "*SEnsembles*", com as geradas pelo método *baseline NNM* com base em escores de propensão estimados com a regressão logística. Nesse comparativo foram realizados dois estágios de avaliação de métricas. O primeiro estágio foi realizado após o processo de correspondência do método *NNM* e, o segundo, após geradas as correspondências de instâncias pelas Configurações 1 e 2 do processo ECS da abordagem proposta "*SEnsembles*", conforme ilustrado na Figura 6.15.

Figura 6.15 – Estágios de avaliação de métricas para comparar as correspondências das Configurações 1 e 2 do processo ECS com as do método *NNM*.



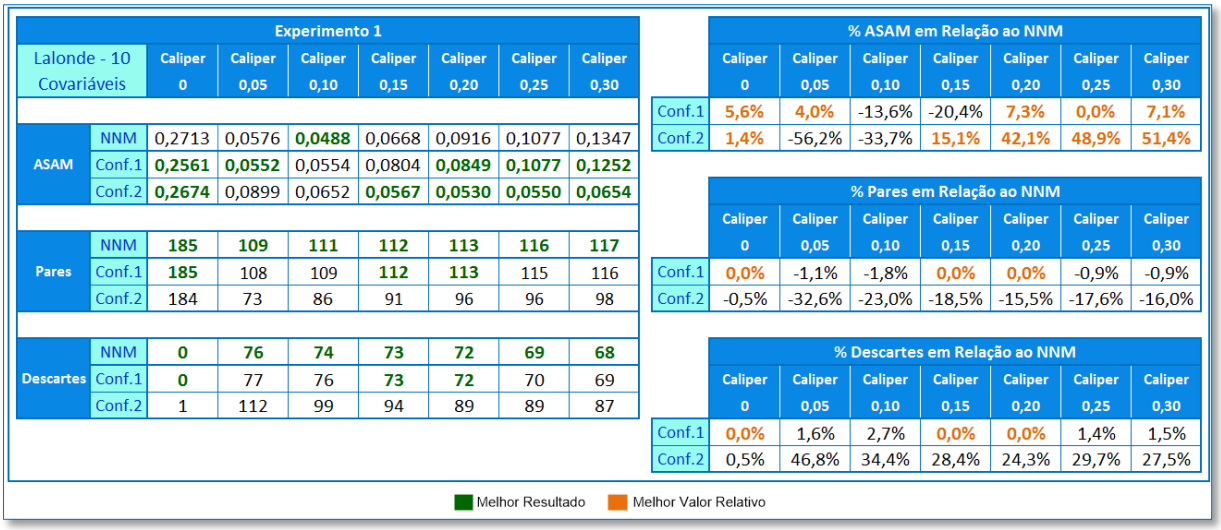
Fonte: Elaborado pelo autor.

Nos experimentos utilizou-se os conjuntos de dados descritos na Tabela 6.1, ou seja, o clássico conjunto de dados Lalonde (1986), o conjunto de dados PBF 1 (Martins, 2013) e duas variações desse último conjunto conforme já mencionadas na Seção 6.2 (PBF 2 e 3). Além disso, as métricas adotadas foram classificadas em absolutas, que são as utilizadas nas seções anteriores (i.e. ASAM e número de pares de instâncias correspondidas) e o número de instâncias descartadas do grupo de tratamento e, em métricas relativas a essas absolutas, que representam vantagem ou desvantagem do processo ECS em relação ao método *NNM*.

É importante destacar que as métricas absolutas são exibidas do lado esquerdo das figuras desta seção, enquanto que as relativas são exibidas do lado direito e em percentual. Adicionalmente, para uma melhor visualização dessas métricas, adotou-se um padrão de cores, no qual a cor verde indica um melhor valor absoluto da métrica ASAM em relação ao valor obtido pelo método *NNM* e, a cor laranja, indica uma vantagem da métrica relativa ASAM em comparação ao mesmo método.

Os resultados do experimento 1, no qual se utilizou o conjunto de dados Lalonde (1986), demonstram que o processo ECS (melhor valor dentre a Configuração 1 e 2) obteve os melhores valores absolutos da métrica ASAM em cinco *calipers*, excetuando, em ambas, o *caliper* 0,10, conforme se observa na Figura 6.16. Dessa forma, destaca-se que o processo ECS gerou ganhos de 1,4% a 51,4% em relação ao *NNM* (*baseline*) para 6 dos 7 *calipers* (exceto para o *caliper* 0,10), com a Configuração 1. Além disso, nota-se que a Configuração 1 obteve a menor redução do número de pares de instâncias correspondidas e os menores descartes do grupo de tratamento do que a Configuração 2, quando comparadas ao método *NNM*. Por sua vez, a Configuração 2 resultou nas maiores diferenças da métrica ASAM em relação ao método *NNM*, mas sempre acompanhadas do aumento do número de instâncias descartadas do grupo de tratamento.

Figura 6.16 – Resultado das métricas ASAM, número de pares de instâncias correspondidas e número de instâncias descartadas do grupo de tratamento geradas pelo processo ECS da abordagem proposta "*SEnsembles*" e pelo método *NNM* (*baseline* da comparação) no experimento 1, com variação de *caliper* de 0 a 0,30 e, usando o conjunto de dados Lalonde (1986) com 10 covariáveis para efetuar a correspondência das instâncias.



Experimento 1

Lalonde - 10 Covariáveis		Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
ASAM	NNM	0,2713	0,0576	**0,0488**	0,0668	0,0916	0,1077	0,1347
	Conf.1	**0,2561**	**0,0552**	0,0554	0,0804	**0,0849**	**0,1077**	**0,1252**
	Conf.2	**0,2674**	0,0899	0,0652	**0,0567**	**0,0530**	**0,0550**	**0,0654**
Pares	NNM	**185**	**109**	**111**	**112**	**113**	**116**	**117**
	Conf.1	**185**	108	109	**112**	**113**	115	116
	Conf.2	184	73	86	91	96	96	98
Descartes	NNM	**0**	**76**	**74**	**73**	**72**	**69**	**68**
	Conf.1	**0**	77	76	**73**	**72**	70	69
	Conf.2	1	112	99	94	89	89	87

% ASAM em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Conf.1	5,6%	4,0%	-13,6%	-20,4%	7,3%	0,0%	7,1%
Conf.2	1,4%	-56,2%	-33,7%	15,1%	42,1%	48,9%	51,4%

% Pares em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Conf.1	0,0%	-1,1%	-1,8%	0,0%	0,0%	-0,9%	-0,9%
Conf.2	-0,5%	-32,6%	-23,0%	-18,5%	-15,5%	-17,6%	-16,0%

% Descartes em Relação ao NNM

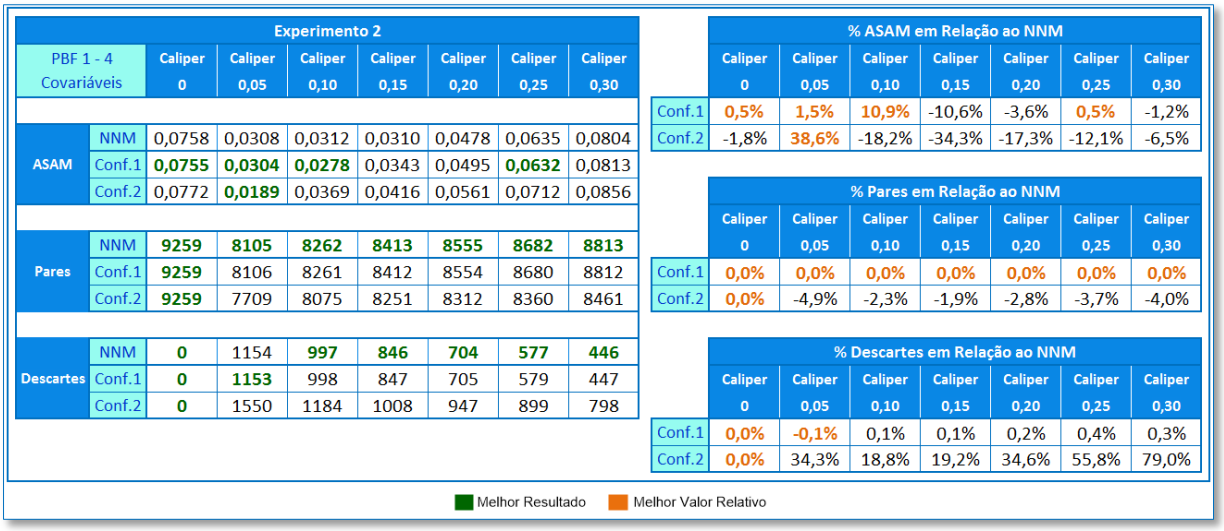
	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Conf.1	0,0%	1,6%	2,7%	0,0%	0,0%	1,4%	1,5%
Conf.2	0,5%	46,8%	34,4%	28,4%	24,3%	29,7%	27,5%

Fonte: Elaborado pelo autor.

No experimento 2, no qual se utilizou o conjunto de dados PBF 1 (Martins, 2013), com quatro covariáveis para efetuar a correspondência das instâncias, a Configuração 1 obteve os melhores valores absolutos da métrica ASAM em cinco *calipers* (zero, 0,05, 0,10 e, 0,25), enquanto que a Configuração 2 venceu no caliper 0,05. Dessa forma, o processo ECS da abordagem proposta "*SEnsembles*" obteve os melhores ASAM em 4 dos 7 *calipers* e, gerou ganhos de 0,5% a 38,6% em relação ao *baseline NNM.* Além disso, a Configuração 1 também não resultou na redução do número de pares de instâncias correspondidas e obteve um pequeno descarte de instâncias do grupo de tratamento (de 0,1% a 0,4%), quando comparada ao método *NNM*, conforme se observa na Figura 6.17.

Ressalta-se também que a Configuração 1 com o *caliper* 0,05 obteve uma instância descartada a menos que o método *NNM,* o que equivale a -0,1% do total de instâncias do grupo de tratamento. Já Configuração 2 obteve o maior descarte de instâncias do grupo de tratamento com o *caliper* 0,05 (34,3%), ao mesmo tempo em que gerou o ganho de 38,5% no valor do ASAM com esse *caliper.*

Figura 6.17 – Resultado da métricas ASAM, número de pares de instâncias correspondidas e número de instâncias descartadas do grupo de tratamento geradas pelo processo ECS da abordagem proposta "*SEnsembles*" e pelo método *NNM* (*baseline* da comparação) no experimento 2, com variação de *caliper* de 0 a 0,30 e, usando o conjunto de dados PBF 1 (Martins, 2013) com 4 covariáveis para efetuar a correspondência das instâncias.



Experimento 2

PBF 1 - 4 Covariáveis		Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
ASAM	NNM	0,0758	0,0308	0,0312	0,0310	0,0478	0,0635	0,0804
	Conf.1	**0,0755**	**0,0304**	**0,0278**	0,0343	0,0495	**0,0632**	0,0813
	Conf.2	0,0772	**0,0189**	0,0369	0,0416	0,0561	0,0712	0,0856
Pares	NNM	**9259**	**8105**	**8262**	**8413**	**8555**	**8682**	**8813**
	Conf.1	**9259**	8106	8261	8412	8554	8680	8812
	Conf.2	**9259**	7709	8075	8251	8312	8360	8461
Descartes	NNM	**0**	1154	**997**	**846**	**704**	**577**	**446**
	Conf.1	**0**	**1153**	998	847	705	579	447
	Conf.2	**0**	1550	1184	1008	947	899	798

% ASAM em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Conf.1	0,5%	1,5%	10,9%	-10,6%	-3,6%	0,5%	-1,2%
Conf.2	-1,8%	38,6%	-18,2%	-34,3%	-17,3%	-12,1%	-6,5%

% Pares em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Conf.1	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
Conf.2	0,0%	-4,9%	-2,3%	-1,9%	-2,8%	-3,7%	-4,0%

% Descartes em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Conf.1	0,0%	-0,1%	0,1%	0,1%	0,2%	0,4%	0,3%
Conf.2	0,0%	34,3%	18,8%	19,2%	34,6%	55,8%	79,0%

Fonte: Elaborado pelo autor.

Já no experimento 3, no qual se utilizou o mesmo conjunto do experimento 2, mas com 14 covariáveis, a Configuração 1 somente obteve o melhor valor absoluto da métrica ASAM em um único *caliper* (0,15), e a Configuração 2 obteve em cinco *calipers* (0,05, 0,10, 0,20, 0,25 e 0,30). Assim, o processo ECS da abordagem proposta "*SEnsembles*" não obteve o melhor valor da métrica ASAM apenas com o *caliper zero*, com ganhos de 0,7% a 53,8%, conforme ilustrado pela cor laranja na Figura 6.18. Além disso, as configurações também resultaram em menores ou nenhuma redução do número de pares de instâncias correspondidas. Porém, exceto com o *caliper* zero, a Configuração 2 gerou os maiores descartes de instâncias do grupo de tratamento, de 9,3 % a 30,2%.

O insucesso da Configuração 1 no experimento 3 pode ser resultante do único par de instâncias com características idênticas encontrado no conjunto de dados PBF1 (Martins, 2013), quando utilizado 14 covariáveis para efetuar a correspondência de instâncias. Dessa forma, observa-se que a Configuração 1 é mais vantajosa quando o número de pares de instâncias idênticas dos conjuntos de dados não é pequeno.

Figura 6.18 – Resultado das métricas ASAM, número de pares de instâncias correspondidas e número de instâncias descartadas do grupo de tratamento geradas pelo processo ECS da abordagem proposta "*SEnsembles*" e pelo método *NNM* (*baseline* da comparação) no experimento 3, com variação de *caliper* de 0 a 0,30 e, usando o conjunto de dados PBF 1 (Martins, 2013) com 14 covariáveis para efetuar a correspondência das instâncias.

Experimento 3

PBF 1 - 14 Covariáveis		Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
ASAM	NNM	**0,0555**	0,0177	0,0167	0,0211	0,0261	0,0333	0,0445
	Conf.1	0,0557	0,0178	0,0169	**0,0209**	0,0263	0,0340	0,0449
	Conf.2	0,0557	**0,0082**	**0,0142**	0,0220	**0,0259**	**0,0328**	**0,0418**
Pares	NNM	**9259**	**7838**	**7985**	**8135**	**8264**	**8397**	**8531**
	Conf.1	**9259**	7839	7985	8135	8263	8396	8530
	Conf.2	**9259**	7499	7817	8031	8130	8221	8311
Descartes	NNM	**0**	1421	**1274**	**1124**	**995**	**862**	**728**
	Conf.1	**0**	**1420**	**1274**	**1124**	996	863	729
	Conf.2	**0**	1760	1442	1228	1129	1038	948

% ASAM em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Conf.1	-0,3%	-0,7%	-0,8%	**1,0%**	-0,9%	-2,2%	-0,9%
Conf.2	-0,3%	**53,8%**	**14,8%**	-4,5%	**0,7%**	**1,5%**	**6,0%**

% Pares em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Conf.1	**0,0%**	**0,0%**	**0,0%**	**0,0%**	**0,0%**	**0,0%**	**0,0%**
Conf.2	0,0%	-4,3%	-2,1%	-1,3%	-1,6%	-2,1%	-2,6%

% Descartes em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Conf.1	**0,0%**	**-0,1%**	**0,0%**	**0,0%**	0,1%	0,1%	0,1%
Conf.2	**0,0%**	23,9%	13,2%	9,3%	13,4%	20,4%	30,2%

Melhor Resultado Melhor Valor Relativo

Fonte: Elaborado pelo autor.

Já no experimento 4, no qual se utilizou o conjunto de dados PBF 2, considerando-se 14 covariáveis para efetuar a correspondência das instâncias, as Configuração 1 e 2 obtiveram os melhores valores absolutos da métrica ASAM em cinco *calipers* (cada uma), conforme se observa na Figura 6.19. Nota-se também que a Configuração 1 obteve os melhores valores nos *calipers* 0,10, 0,15, 0,20, 0,25 e 0,30, e a Configuração 2 obteve os melhores valores excetuando os *calipers* zero e 0,15. Assim, considerando-as em conjunto, as configurações do processo ECS da abordagem "*SEnsembles*" somente não obtiveram o melhor valor da métrica ASAM com o *caliper* zero, com ganhos de 3,9% a 48,4%. Além disso, ambas as configurações obtiveram pequenas reduções do número de pares de instâncias correspondidas, com a maior desvantagem para a Configuração 2 com o *caliper* 0,30 (-5,0%).

Figura 6.19 – Resultado das métricas ASAM, número de pares de instâncias correspondidas e número de instâncias descartadas do grupo de tratamento geradas pelo processo ECS da abordagem proposta "*SEnsembles*" e pelo método *NNM* (*baseline* da comparação) no experimento 4, com variação de *caliper* de 0 a 0,30 e, usando o conjunto de dados PBF 2 com 14 covariáveis para efetuar a correspondência das instâncias.

Experimento 4

PBF 2 - 14 Covariáveis		Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
ASAM	NNM	**0,0375**	0,0139	0,0163	0,0191	0,0258	0,0336	0,0430
	Conf.1	0,0380	0,0153	**0,0145**	**0,0175**	**0,0204**	**0,0275**	**0,0357**
	Conf.2	0,0380	**0,0072**	**0,0146**	0,0213	**0,0248**	**0,0316**	**0,0401**
Pares	NNM	**9259**	**8241**	**8374**	**8505**	**8651**	**8779**	**8906**
	Conf.1	**9259**	8225	8331	8442	8550	8662	8768
	Conf.2	**9259**	7961	8227	8309	8359	8411	8463
Descartes	NNM	**0**	**1018**	**885**	**754**	**608**	**480**	**353**
	Conf.1	**0**	1034	928	817	709	597	491
	Conf.2	**0**	1298	1032	950	900	848	796

% ASAM em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Conf.1	-1,1%	-10,2%	**11,5%**	**7,9%**	**20,8%**	**18,0%**	**17,1%**
Conf.2	-1,1%	**48,4%**	**10,4%**	-11,6%	**3,9%**	**6,0%**	**6,8%**

% Pares em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Conf.1	**0,0%**	-0,2%	-0,5%	-0,7%	-1,2%	-1,3%	-1,6%
Conf.2	**0,0%**	-3,4%	-1,8%	-2,3%	-3,4%	-4,2%	-5,0%

% Descartes em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Conf.1	**0,0%**	1,6%	4,9%	8,4%	16,7%	24,5%	39,3%
Conf.2	**0,0%**	27,4%	16,7%	26,0%	48,2%	76,7%	125,6%

Melhor Resultado Melhor Valor Relativo

Fonte: Elaborado pelo autor.

Já no experimento 5, a Configuração 1 obteve novamente os melhores valores absolutos da métrica ASAM em seis *calipers* (exceto no *caliper* 0,05). Por sua vez, a Configuração 2 obteve os melhores valores em três *calipers* (0,05, 0,10 e 0,30). Assim, o processo ECS da abordagem "*SEnsembles*" obteve os melhores valores da métrica ASAM em 7 *calipers* de 7 possíveis, com ganhos de 2,2% a 50,6%, conforme se observa-se na Figura 6.20. Além disso, as duas configurações obtiveram pequenas reduções do número de pares de instâncias correspondidas quando comparadas ao método *NNM*, com destaque para a Configuração 1, na qual a perda foi de no máximo -0,9% com o *caliper* 0,30. Já a Configuração 2 obteve os maiores descartes de instâncias pertencentes aos grupo de tratamento.

Figura 6.20 – Resultado das métricas ASAM, número de pares de instâncias correspondidas e número de instâncias descartadas do grupo de tratamento geradas pelo processo ECS da abordagem proposta "*SEnsembles*" e pelo método *NNM* (*baseline* da comparação) no experimento 5, com variação de *caliper* de 0 a 0,30 e, usando o conjunto de dados PBF 3 com 14 covariáveis para efetuar a correspondência das instâncias.

Experimento 5

PBF 3 - 14 Covariáveis		Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
ASAM	NNM	0,0382	0,0144	0,0159	0,0188	0,0257	0,0333	0,0426
	Conf.1	**0,0373**	0,0153	**0,0155**	**0,0172**	**0,0227**	**0,0302**	**0,0390**
	Conf.2	0,0550	**0,0071**	**0,0154**	0,0222	0,0286	0,0347	**0,0402**
Pares	NNM	**9259**	**8217**	**8348**	**8474**	**8610**	**8746**	**8881**
	Conf.1	**9259**	8202	8321	8435	8555	8680	8801
	Conf.2	9169	7903	8143	8236	8285	8342	8408
Descartes	NNM	**0**	**1042**	**911**	**785**	**649**	**513**	**378**
	Conf.1	**0**	1057	938	824	704	579	458
	Conf.2	90	1356	1116	1023	974	917	851

% ASAM em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Conf.1	2,2%	-6,7%	2,5%	8,4%	11,7%	9,3%	8,4%
Conf.2	-44,0%	50,6%	3,2%	-18,5%	-11,0%	-4,4%	5,7%

% Pares em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Conf.1	0,0%	-0,2%	-0,3%	-0,5%	-0,6%	-0,8%	-0,9%
Conf.2	-1,0%	-3,8%	-2,5%	-2,8%	-3,8%	-4,6%	-5,3%

% Descartes em Relação ao NNM

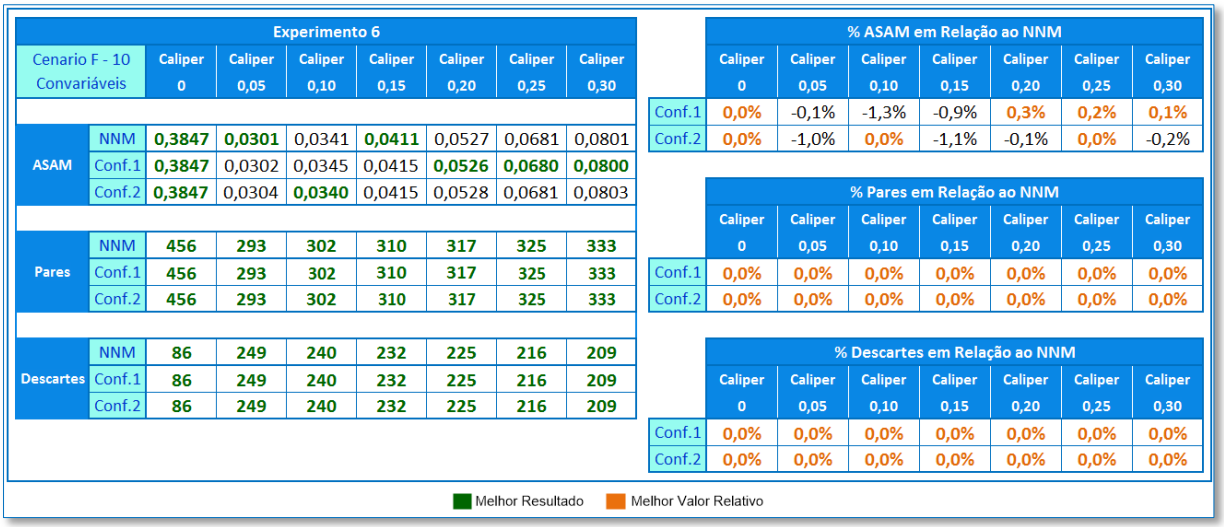
	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Conf.1	0,0%	1,4%	3,0%	5,0%	8,5%	13,0%	21,3%
Conf.2	1,0%	30,1%	22,5%	30,4%	50,0%	78,9%	125,3%

Melhor Resultado Melhor Valor Relativo

Fonte: Elaborado pelo autor.

Por último, no experimento 6, no qual se utilizou o conjunto de dados do Cenário F (Lee et al., 2010), que não possui nenhum par de instâncias com características idênticas, os resultados demonstraram que o processo ECS da abordagem "*SEnsembles*" obteve correspondências similares ao método *NNM,* com pequenas diferenças em relação ao valor da métrica ASAM. Além disso, o número de pares de instâncias correspondidas e o número de descartes de instâncias do grupo de tratamento foram exatamente os mesmos, conforme se observa pela cor laranja na Figura 6.21. Dessa forma, para conjuntos de dados sem instâncias com características idênticas, as correspondências do processo ECS da abordagem "*SEnsembles*" são similares às correspondências do método *NNM*.

Figura 6.21 – Resultado das métricas ASAM, número de pares de instâncias correspondidas e número de instâncias descartadas do grupo de tratamento das correspondências geradas pelo processo ECS da abordagem proposta "*SEnsembles*" e pelo método *NNM* (*baseline* da comparação) no experimento 5, com variação de *caliper* de 0 a 0,30 e, usando o conjunto de dados Cenário F (Lee et al., 2010) com 10 covariáveis para efetuar a correspondência das instâncias.



Experimento 6

Cenario F - 10 Convariáveis		Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
ASAM	NNM	0,3847	0,0301	0,0341	0,0411	0,0527	0,0681	0,0801
	Conf.1	0,3847	0,0302	0,0345	0,0415	0,0526	0,0680	0,0800
	Conf.2	0,3847	0,0304	0,0340	0,0415	0,0528	0,0681	0,0803
Pares	NNM	456	293	302	310	317	325	333
	Conf.1	456	293	302	310	317	325	333
	Conf.2	456	293	302	310	317	325	333
Descartes	NNM	86	249	240	232	225	216	209
	Conf.1	86	249	240	232	225	216	209
	Conf.2	86	249	240	232	225	216	209

% ASAM em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Conf.1	0,0%	-0,1%	-1,3%	-0,9%	0,3%	0,2%	0,1%
Conf.2	0,0%	-1,0%	0,0%	-1,1%	-0,1%	0,0%	-0,2%

% Pares em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Conf.1	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
Conf.2	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%

% Descartes em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Conf.1	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
Conf.2	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%

Fonte: Elaborado pelo autor.

A Tabela 6.5 apresenta um resumo dos resultados obtidos nos experimentos desta seção, bem como uma observação sobre a métrica ASAM. Observa-se que no experimento 1, o processo ECS da abordagem proposta "*SEnsembles*" obteve os melhores valores da métrica ASAM em seis *calipers* (exceto com o *caliper* 0,10), com ganhos de 1,4% a 51,4%. Já no experimento 2, o processo ECS obteve os melhores valores em quatro *calipers* (zero, 0,05, 0,10 e 0,25), com ganhos de 0,5% a 38,6%.

No experimento 3 e 4, o processo ECS obteve os melhores valores da métrica ASAM em seis *calipers* (exceto no zero), com ganhos de 0,7% a 53,8% no experimento 3 e, de 3,9% a 48,4% no experimento 4. Já no experimento 5 obteve os ganhos de 2,2% a 30,6% em todos os *calipers* e, por último, no experimento 6, o processo ECS gerou correspondências similares ao *NNM*, com pequenos ganhos de 0,1% e 0,3%.

Tabela 6.5 – Resumo dos resultados dos experimentos pelos quais se comparou as correspondências geradas pelo processo ECS da abordagem proposta "*SEnsembles*" com as geradas pelo método *NNM* (*baseline* da comparação), com a descrição dos conjuntos de dados, número do experimento, configuração, quantidade de atributos, quantidade de instâncias idênticas, melhores resultados e os ganhos obtidos.

Conjunto de Dados	Exp.*	Conf.**	Atr.***	Instâncias Idênticas	Melhores Resultados	Ganhos
Lalonde (1986)	1	1	8	14 (7 Pares)	Processo ECS com 6 *calipers* (exceto no 0,10).	De 1,4% a 51,4%.
		2	8	25 (13 Tratados e 12 Não Tratados)		
PBF 1 (Martins, 2013)	2	1	4	16 (8 Pares)	Processo ECS com 4 *calipers* (zero, 0,05, 0,10 e 0,25)	De 0,5% a 38,6%.
		2	4	17 (8 Beneficiários e 9 não Beneficiários)		
	3	1 e 2	14	2 (1 par)	Processo ECS com 6 *calipers* (exceto no zero).	De 0,7% a 53,8%.
PBF 2 Modificado a partir de Martins (2013)	4	1 e 2	14	2.224 (1.112 pares)	Processo ECS com 6 *calipers* (exceto no zero).	De 3,9% a 48,4%.
PBF 3 Modificado a partir de Martins (2013)	5	1	14	1.278 (639 pares)	Processo ECS com todos os *calipers.*	De 2,2% a 50,6%.
		2		2.280 (1185 Beneficiários e 1.095 não beneficiários);		
Cenário F (Lee et al., 2010)	6	1	10	Nenhuma	Processo ECS com 4 *calipers* (0,10, 0,20, 0,25 e 0,30).	De 0,1% a 0,3%.
		2				

***Experimento; **Configuração; ***Atributos.**

Fonte: Elaborado pelo autor.

Como visto, o processo ECS da abordagem proposta "*SEnsembles*" venceu o *baseline NNM* em todos os experimentos, com ganhos que variaram de 0,1% a 53,8%, conforme se observa na Tabela 6.6, a qual apresenta um mapeamento de

calipers que proporcionaram uma melhora da qualidade da correspondência pelo processo ECS em relação ao *NNM*.

Tabela 6.6 – Mapeamento de *calipers* nos quais se obteve melhoria da qualidade da correspondência quando comparado as configurações do processo ECS da abordagem proposta "*SEnsembles*" com o método *NNM* (*baseline* de comparação).

Conjunto de Dados	Atr.*	Conf.**	Idênticos Obtidos	*Calipers*						
				Zero	0,05	0,10	0,15	0,20	0,25	0,30
Lalonde (1986)	8	1	14 (7 Pares)	5,6%	4,0%			7,3%	0,0%	7,1%
		2	25 (13 Tratados e 12 Não Tratados)	1,4%			15,1%	42,1%	48,9%	51,4%
PBF 1 (Martins, 2013)	4	1	16 (8 Pares)	0,5%	1,5%	10,9%			0,5%	
		2	17 (8 Beneficiários e 9 não Beneficiários)		38,6%					
	14	1	2 (1 par)				1,0%			
		2	2 (1 par)		53,8%	14,8%		0,7%	1,5%	6,0%
PBF 2 Modificado a partir de Martins (2013)	14	1	2224 (1.112 pares)			11,5%	7,9%	20,8%	18,0%	17,1%
		2	2224 (1.112 pares)		48,4%	10,4%		3,9%	6,0%	6,8%
PBF 3 Modificado a partir de Martins (2013)	14	1	1.278 (639 pares)	2,2%		2,5%	8,4%	11,7%	9,3%	8,4%
		2	2.280 (1.185 Beneficiários e 1.095 não beneficiários);		50,6%	3,2%				5,7%
Cenário F (Lee et al., 2010)	10	1	Nenhum	0%				0,3%	0,2%	0,1%
		2	Nenhum	0%		0%			0%	

***Atributos; **Configuração do Processo ECS.**

Fonte: Elaborado pelo autor.

Ressalta-se, entretanto, com exceção do *caliper* zero, o ganho obtido pelo processo ECS sempre foi acompanhado de uma redução do número de pares de instâncias correspondidas conforme se observa pela cor em vermelho na Tabela 6.7 e, um aumento talvez não desejável do número de instâncias do grupo de

tratamento. Nota-se que a maior redução de pares de instâncias correspondidas ocorreu no experimento 1 com os *calipers* mais altos (0,15, 0,20, 0,25 e 0,30). Nos demais experimentos essa redução não atingiu 5%, com exceção dos experimento 2 e 3 com o *caliper* 0,30.

Tabela 6.7 – Mapeamento de *calipers* com a descrição das porcentagens de pares de instâncias obtidas pelas Configurações 1 e 2 do processo ECS da abordagem proposta "*SEnsembles*" em relação ao método *NNM* (*baseline* da comparação), somente nas situações que houve melhoria da qualidade da correspondência.

Conjunto de Dados	Atr.*	Conf.**	Instâncias Idênticas	*Calipers*						
				Zero	0,05	0,10	0,15	0,20	0,25	0,30
Lalonde (1986)	8	1	14 (7 Pares)	0%	-1,1%			0%	-0,9%	-0,9%
		2	25 (13 Tratados e 12 Não Tratados)	-0,5%			-18,5%	-15,5%	-17,6%	-16,0%
PBF 1 (Martins, 2013)	4	1	16 (8 Pares)	0%	0%	0%			0%	
		2	17 (8 Beneficiários e 9 não Beneficiários)		-4,9%					
	14	1	2 (1 par)				0%			
		2	2 (1 par)		-4,3%	-2,1%		-1,6%	-2,1%	-2,6%
PBF 2 Modificado a partir de Martins (2013)	14	1	2224 (1.112 pares)			-0,5%	-0,7%	-1,2%	-1,3%	-1,6%
		2	2224 (1.112 pares)		-3,4%	-1,8%		-3,4%	-4,2%	-5,0%
PBF 3 Modificado a partir de Martins (2013)	14	1	1.278 (639 pares)	0%		-0,3%	-0,5%	-0,6%	-0,8%	-0,9%
		2	2.280 (1.185 Beneficiários e 1.095 não beneficiários);		-3,8%	-2,5%				-5,3%
Cenário F (Lee et al., 2010)	10	1	Nenhum	0%				0%	0%	0%
		2	Nenhum	0%		0%			0%	

***Atributos; **Configuração do Processo ECS.**

Fonte: Elaborado pelo autor.

Como visto, o processo ECS melhorou as correspondências de instâncias com vários *calipers* nos experimentos realizados, com uma redução de pares de instâncias correspondidas que não atingiu 5% (exceto em dois experimento com o *caliper* 0,30). Além disso, no conjunto de dados sem instâncias com características idênticas (experimento 6), os resultados demonstram que o processo ECS da abordagem proposta "*SEnsembles*" é similar ao método *NNM*.

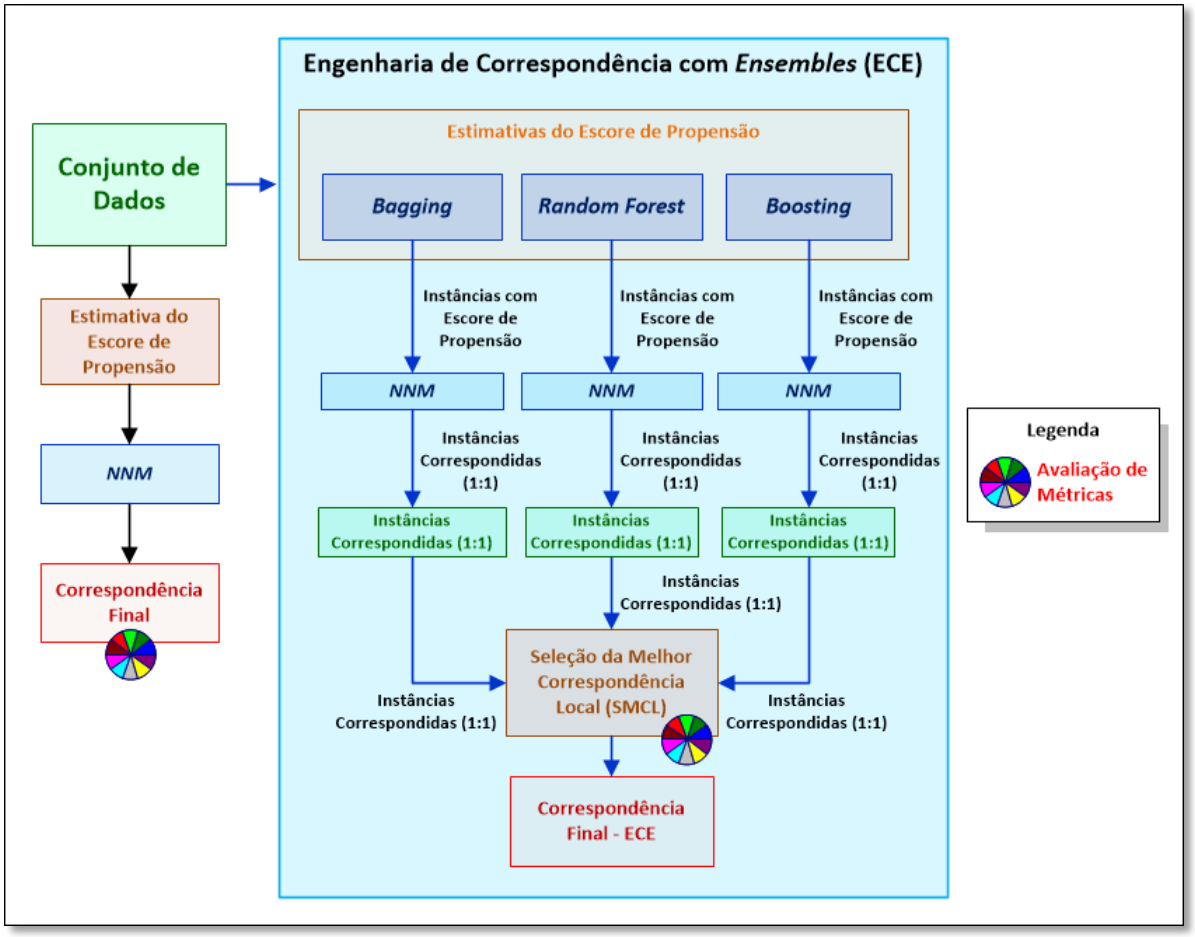
6.5 Comparação do processo ECE da abordagem proposta "*SEnsembles*" com o *baseline NNM*

Esta seção realiza uma análise comparativa da qualidade das correspondências geradas pelo processo ECE da abordagem proposta "*SEnsembles*" com as correspondências geradas pelo método *NNM* (*baseline* de comparação) utilizando as métricas ASAM, número de pares de instâncias correspondidas e número de instâncias descartadas do grupo de tratamento. Nesse comparativo foram realizados dois estágios de avaliação das métricas, conforme ilustrado na Figura 6.22. O primeiro estágio foi realizado após o processo de correspondência do método *NNM* e, o segundo, após geradas as correspondências pelo processo ECE, ou seja, na etapa SMCL.

Nos experimentos desta seção adotou-se as mesmas métricas absolutas e relativas da seção anterior (Seção 6.4). Assim, os valores das métricas absolutas para as correspondências efetuadas pelo método *NNM* são considerados como referencias para se obter as métricas relativas. Além disso, adotou-se também o mesmo padrão de cores também da seção anterior.

As métricas foram obtidas com base em oito experimentos, nos quais se utilizou os conjuntos de dados descritos na Tabela 6.4, ou seja, o clássico conjunto de dados Lalonde (1986), os conjuntos de dados dos cenários (B a G) (Lee et al., 2010) e o conjunto de dados PBF 3.

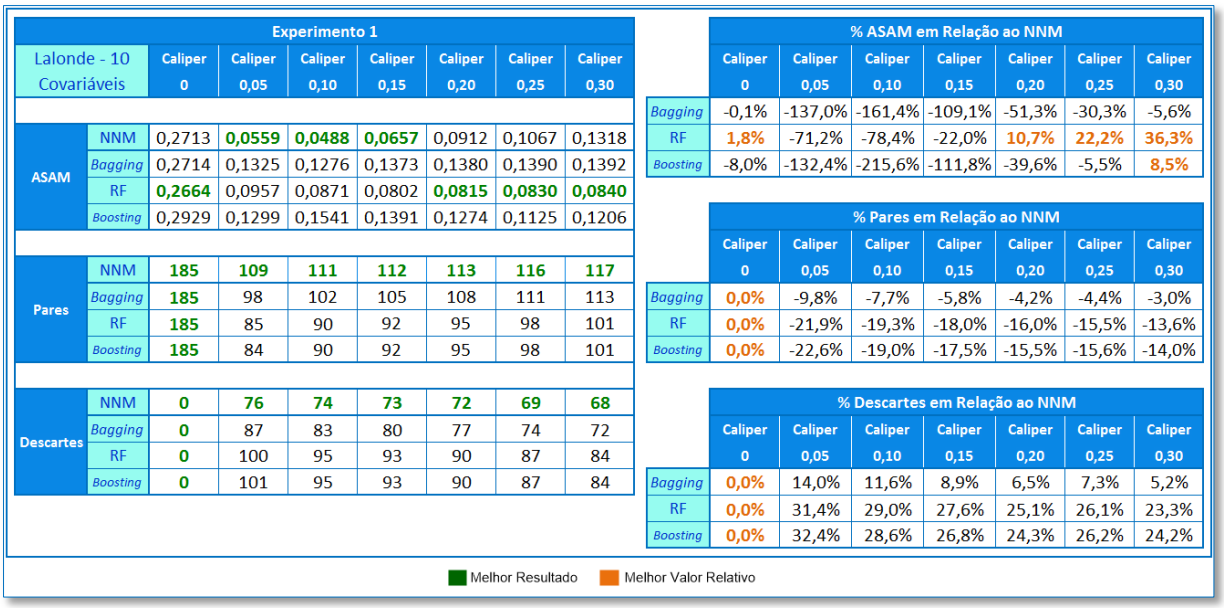
Figura 6.22 – Estágios de avaliação de métricas para comparar as correspondências geradas pelo processo ECE abordagem proposta "*SEnsembles*" com as do método *NNM* (*baseline* da comparação).



Fonte: Elaborado pelo autor.

No experimento 1, no qual utilizou-se o conjunto de dados Lalonde (1986), o *ensemble random forest* ao substituir a regressão logística proporcionou os melhores valores da métrica ASAM em quatro *calipers* (zero, 0,20, 0,25 e 0,30), com ganhos de 1,8% a 36,3% e, especificamente com o *caliper* zero, não reduziu o número de pares de instâncias correspondidas, se comparado ao método *NNM,* conforme ilustrado na Figura 6.23.

Figura 6.23 – Resultado das métricas ASAM, número de pares de instâncias correspondidas e número de instâncias descartadas do grupo de tratamento geradas pelo processo ECE da abordagem proposta "*SEnsembles*" e pelo método *NNM* (*baseline* de comparação) no experimento 1, com variação de *caliper* de 0 a 0,30 e, usando o conjunto de dados Lalonde (1986) com 10 covariáveis para efetuar a correspondência das instâncias.



Experimento 1

Lalonde - 10 Covariáveis		Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
ASAM	NNM	0,2713	**0,0559**	**0,0488**	**0,0657**	0,0912	0,1067	0,1318
	Bagging	0,2714	0,1325	0,1276	0,1373	0,1380	0,1390	0,1392
	RF	**0,2664**	0,0957	0,0871	0,0802	**0,0815**	**0,0830**	**0,0840**
	Boosting	0,2929	0,1299	0,1541	0,1391	0,1274	0,1125	0,1206
Pares	NNM	**185**	**109**	**111**	**112**	**113**	**116**	**117**
	Bagging	**185**	98	102	105	108	111	113
	RF	**185**	85	90	92	95	98	101
	Boosting	**185**	84	90	92	95	98	101
Descartes	NNM	**0**	**76**	**74**	**73**	**72**	**69**	**68**
	Bagging	**0**	87	83	80	77	74	72
	RF	**0**	100	95	93	90	87	84
	Boosting	**0**	101	95	93	90	87	84

% ASAM em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Bagging	-0,1%	-137,0%	-161,4%	-109,1%	-51,3%	-30,3%	-5,6%
RF	**1,8%**	-71,2%	-78,4%	-22,0%	**10,7%**	**22,2%**	**36,3%**
Boosting	-8,0%	-132,4%	-215,6%	-111,8%	-39,6%	-5,5%	**8,5%**

% Pares em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Bagging	**0,0%**	-9,8%	-7,7%	-5,8%	-4,2%	-4,4%	-3,0%
RF	**0,0%**	-21,9%	-19,3%	-18,0%	-16,0%	-15,5%	-13,6%
Boosting	**0,0%**	-22,6%	-19,0%	-17,5%	-15,5%	-15,6%	-14,0%

% Descartes em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Bagging	**0,0%**	14,0%	11,6%	8,9%	6,5%	7,3%	5,2%
RF	**0,0%**	31,4%	29,0%	27,6%	25,1%	26,1%	23,3%
Boosting	**0,0%**	32,4%	28,6%	26,8%	24,3%	26,2%	24,2%

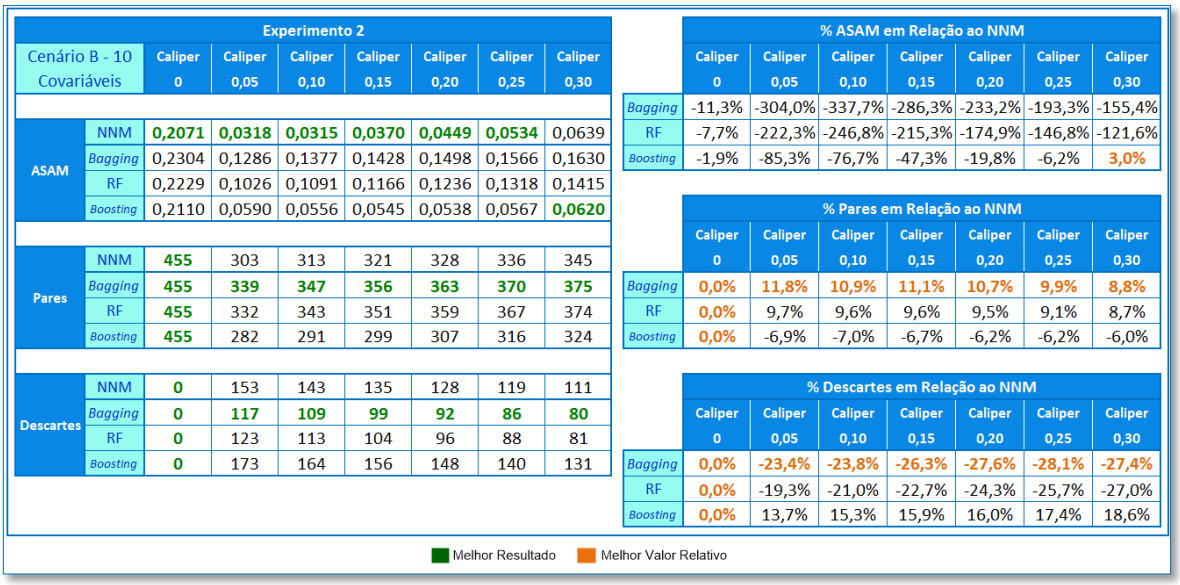
Melhor Resultado Melhor Valor Relativo

Fonte: Elaborado pelo autor.

Já no experimento 2, no qual se utilizou o conjunto de dados do Cenário B (Lee et al., 2010), apenas o *ensemble boosting* com o *caliper* 0,30 resultou no melhor valor da métrica ASAM, com ganho de 3,0% e uma redução do número de pares de instâncias correspondidas de -6,0%. O processo ECE, para os demais valores de *caliper*, não produziu as melhores correspondências quando comparado ao *baseline NNM,* conforme ilustrado na Figura 6.24.

Já no experimento 3, no qual se utilizou o conjunto de dados do Cenário B (Lee et al., 2010), todos os *ensembles* utilizados para substituir a regressão logística obtiveram os melhores valores da métrica ASAM com o *caliper* zero e, o *ensemble boosting* resultou no melhor valor de ASAM em três *calipers* (zero, 0,25 e 0,30), conforme se observa na Figura 6.25. Para estes valores de *caliper*, ao analisar os *ensembles* conjuntamente, o processo ECE da abordagem proposta "*SEnsembles*" resultou em ganhos de 7,5% a 24,8% quando comparado ao método *NNM* e, com uma redução do número de pares de instâncias correspondidas com o *ensemble boosting*, com os *calipers* 0,25 e 0,30, de -15,7% e -15,4%, respectivamente.

Figura 6.24 – Resultado das métricas ASAM, número de pares de instâncias correspondidas e número de instâncias descartadas do grupo de tratamento geradas pelo processo ECE da abordagem proposta "*SEnsembles*" e pelo método *NNM* (*baseline* de comparação) no experimento 2, com variação de *caliper* de 0 a 0,30 e, usando o conjunto de dados do Cenário B (Lee et al., 2010) com 10 covariáveis para efetuar a correspondência das instâncias.



Experimento 2

Cenário B - 10 Covariáveis		Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
ASAM	NNM	0,2071	0,0318	0,0315	0,0370	0,0449	0,0534	0,0639
	Bagging	0,2304	0,1286	0,1377	0,1428	0,1498	0,1566	0,1630
	RF	0,2229	0,1026	0,1091	0,1166	0,1236	0,1318	0,1415
	Boosting	0,2110	0,0590	0,0556	0,0545	0,0538	0,0567	0,0620
Pares	NNM	455	303	313	321	328	336	345
	Bagging	455	339	347	356	363	370	375
	RF	455	332	343	351	359	367	374
	Boosting	455	282	291	299	307	316	324
Descartes	NNM	0	153	143	135	128	119	111
	Bagging	0	117	109	99	92	86	80
	RF	0	123	113	104	96	88	81
	Boosting	0	173	164	156	148	140	131

% ASAM em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Bagging	-11,3%	-304,0%	-337,7%	-286,3%	-233,2%	-193,3%	-155,4%
RF	-7,7%	-222,3%	-246,8%	-215,3%	-174,9%	-146,8%	-121,6%
Boosting	-1,9%	-85,3%	-76,7%	-47,3%	-19,8%	-6,2%	3,0%

% Pares em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Bagging	0,0%	11,8%	10,9%	11,1%	10,7%	9,9%	8,8%
RF	0,0%	9,7%	9,6%	9,6%	9,5%	9,1%	8,7%
Boosting	0,0%	-6,9%	-7,0%	-6,7%	-6,2%	-6,2%	-6,0%

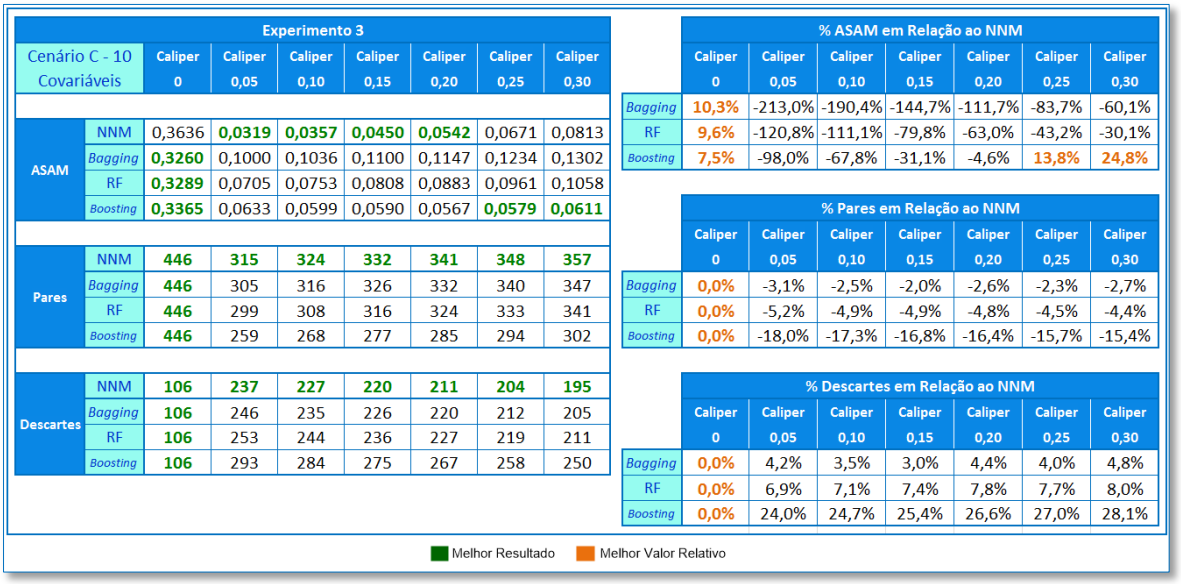
% Descartes em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Bagging	0,0%	-23,4%	-23,8%	-26,3%	-27,6%	-28,1%	-27,4%
RF	0,0%	-19,3%	-21,0%	-22,7%	-24,3%	-25,7%	-27,0%
Boosting	0,0%	13,7%	15,3%	15,9%	16,0%	17,4%	18,6%

Melhor Resultado Melhor Valor Relativo

Fonte: Elaborado pelo autor.

Figura 6.25 – Resultado das métricas ASAM, número de pares de instâncias correspondidas e número de instâncias descartadas do grupo de tratamento geradas pelo processo ECE da abordagem proposta "*SEnsembles*" e pelo método *NNM* (*baseline* de comparação) no experimento 3, com variação de *caliper* de 0 a 0,30 e, usando o conjunto de dados do Cenário C (Lee et al., 2010) com 10 covariáveis para efetuar a correspondência das instâncias.



Experimento 3

Cenário C - 10 Covariáveis		Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
ASAM	NNM	0,3636	0,0319	0,0357	0,0450	0,0542	0,0671	0,0813
	Bagging	0,3260	0,1000	0,1036	0,1100	0,1147	0,1234	0,1302
	RF	0,3289	0,0705	0,0753	0,0808	0,0883	0,0961	0,1058
	Boosting	0,3365	0,0633	0,0599	0,0590	0,0567	0,0579	0,0611
Pares	NNM	446	315	324	332	341	348	357
	Bagging	446	305	316	326	332	340	347
	RF	446	299	308	316	324	333	341
	Boosting	446	259	268	277	285	294	302
Descartes	NNM	106	237	227	220	211	204	195
	Bagging	106	246	235	226	220	212	205
	RF	106	253	244	236	227	219	211
	Boosting	106	293	284	275	267	258	250

% ASAM em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Bagging	10,3%	-213,0%	-190,4%	-144,7%	-111,7%	-83,7%	-60,1%
RF	9,6%	-120,8%	-111,1%	-79,8%	-63,0%	-43,2%	-30,1%
Boosting	7,5%	-98,0%	-67,8%	-31,1%	-4,6%	13,8%	24,8%

% Pares em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Bagging	0,0%	-3,1%	-2,5%	-2,0%	-2,6%	-2,3%	-2,7%
RF	0,0%	-5,2%	-4,9%	-4,9%	-4,8%	-4,5%	-4,4%
Boosting	0,0%	-18,0%	-17,3%	-16,8%	-16,4%	-15,7%	-15,4%

% Descartes em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Bagging	0,0%	4,2%	3,5%	3,0%	4,4%	4,0%	4,8%
RF	0,0%	6,9%	7,1%	7,4%	7,8%	7,7%	8,0%
Boosting	0,0%	24,0%	24,7%	25,4%	26,6%	27,0%	28,1%

Melhor Resultado Melhor Valor Relativo

Fonte: Elaborado pelo autor.

Já os resultados do experimento 4 acompanharam os resultados do experimento 3, ou seja, os *ensembles* proporcionaram os melhores valores da métrica ASAM com o *caliper zero*, e, mais uma vez, o *ensemble boosting* apresentou os melhores valores da métrica ASAM em três *caliper* (zero, 0,25 e 0,30), conforme se observa na Figura 6.26. Para estes valores de *caliper*, ao analisar os *ensembles* conjuntamente, o processo ECE da abordagem proposta "*SEnsembles*" resultou em ganhos de 0,4% a 8% se comparado ao *NNM e*, com redução do numero de pares de instâncias correspondidas de -4,1% e -4,0%, com os *calipers* 0,25 e 0,30, respectivamente.

Figura 6.26 – Resultado das métricas ASAM, número de pares de instâncias correspondidas e número de instâncias descartadas do grupo de tratamento geradas pelo processo ECE da abordagem proposta "*SEnsembles*" e pelo método *NNM* (*baseline* de comparação) no experimento 4, com variação de *caliper* de 0 a 0,30 e, usando o conjunto de dados do Cenário D (Lee et al., 2010) com 10 covariáveis para efetuar a correspondência das instâncias.

Experimento 4

Cenário D - 10 Covariáveis		Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
ASAM	NNM	0,3612	**0,0324**	**0,0346**	**0,0410**	**0,0503**	0,0621	0,0738
	Bagging	**0,3520**	0,1377	0,1463	0,1534	0,1621	0,1709	0,1829
	RF	**0,3543**	0,1043	0,1124	0,1211	0,1303	0,1404	0,1494
	Boosting	**0,3598**	0,0610	0,0573	0,0558	0,0561	**0,0594**	**0,0679**
Pares	NNM	**473**	279	288	294	302	309	318
	Bagging	**473**	**324**	**334**	**346**	**354**	**360**	**366**
	RF	**473**	309	319	327	335	343	350
	Boosting	**473**	264	273	280	288	297	305
Descartes	NNM	**48**	242	234	227	219	212	204
	Bagging	**48**	**197**	**187**	**175**	**167**	**161**	**155**
	RF	**48**	212	202	194	186	178	171
	Boosting	**48**	258	248	241	233	224	216

% ASAM em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Bagging	2,5%	-324,6%	-323,1%	-274,4%	-222,5%	-175,3%	-148,0%
RF	1,9%	-221,5%	-225,1%	-195,5%	-159,1%	-126,2%	-102,6%
Boosting	0,4%	-87,9%	-65,7%	-36,2%	-11,7%	4,3%	8,0%

% Pares em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Bagging	0,0%	16,1%	16,1%	17,5%	17,2%	16,4%	15,2%
RF	0,0%	10,8%	10,9%	11,0%	10,9%	10,8%	10,4%
Boosting	0,0%	-5,6%	-5,2%	-4,7%	-4,4%	-4,1%	-4,0%

% Descartes em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Bagging	0,0%	-18,6%	-19,8%	-22,7%	-23,7%	-24,0%	-23,7%
RF	0,0%	-12,4%	-13,4%	-14,3%	-15,0%	-15,7%	-16,2%
Boosting	0,0%	6,4%	6,3%	6,1%	6,1%	6,0%	6,2%

Melhor Resultado Melhor Valor Relativo

Fonte: Elaborado pelo autor.

Já os resultados do experimento 5 foram muito similares aos resultados do experimento 2, no qual o processo ECE da abordagem proposta "*SEnsembles*" somente proporcionou o melhor valor da métrica ASAM com o *ensemble boosting*

com o *caliper* 0,30, com ganho de 6,4% e com uma redução de -4,3% no número de pares de instâncias correspondidas, conforme se observa na Figura 6.27.

Figura 6.27 – Resultado das métricas ASAM, número de pares de instâncias correspondidas e número de instâncias descartadas do grupo de tratamento das correspondências geradas pelo processo ECE da abordagem proposta "*SEnsembles*" e pelo método *NNM* (*baseline* de comparação) no experimento 5, com variação de *caliper* de 0 a 0,30 e, usando o conjunto de dados do Cenário E (Lee et al., 2010) com 10 covariáveis para efetuar a correspondência das instâncias.

Experimento 5

Cenário E - 10 Covariáveis		Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
ASAM	NNM	0,2381	0,0337	0,0334	0,0399	0,0474	0,0564	0,0682
	Bagging	0,2573	0,1317	0,1365	0,1464	0,1550	0,1633	0,1661
	RF	0,2517	0,0998	0,1065	0,1142	0,1222	0,1310	0,1412
	Boosting	0,2399	0,0620	0,0604	0,0581	0,0578	0,0586	0,0639
Pares	NNM	459	281	290	298	305	312	320
	Bagging	459	319	328	336	345	353	357
	RF	459	311	321	329	336	344	352
	Boosting	459	264	273	281	289	298	306
Descartes	NNM	0	177	169	161	154	147	139
	Bagging	0	140	131	123	114	106	102
	RF	0	148	138	130	123	115	107
	Boosting	0	194	186	178	169	161	152

% ASAM em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Bagging	-8,1%	-291,3%	-308,2%	-266,6%	-227,3%	-189,7%	-143,5%
RF	-5,7%	-196,4%	-218,5%	-185,8%	-157,9%	-132,4%	-106,9%
Boosting	-0,7%	-84,1%	-80,5%	-45,6%	-22,0%	-3,9%	6,4%

% Pares em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Bagging	0,0%	13,3%	13,1%	12,8%	13,0%	12,9%	11,4%
RF	0,0%	10,5%	10,6%	10,3%	10,2%	10,2%	10,0%
Boosting	0,0%	-6,0%	-5,9%	-5,6%	-5,1%	-4,5%	-4,3%

% Descartes em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Bagging	0,0%	-21,1%	-22,5%	-23,6%	-25,7%	-27,5%	-26,3%
RF	0,0%	-16,6%	-18,2%	-19,1%	-20,2%	-21,7%	-23,0%
Boosting	0,0%	9,6%	10,1%	10,3%	10,2%	9,6%	10,0%

Melhor Resultado Melhor Valor Relativo

Fonte: Elaborado pelo autor.

É importante destacar que nos experimentos 2 e 5 utilizou-se os conjuntos de dados dos Cenários B e E (Lee et al., 2010), respectivamente, o quais possuem uma covariável de ordem quadrática. Assim, percebe-se aqui que uma análise prévia dos conjuntos de dados para verificação da linearidade das covariáveis poderia auxiliar a execução dos processos ECS e ECE da abordagem "*SEnsembles*". Alguns esforços foram direcionados neste sentido, mas os resultados não foram concretos ou definitivos e, por isso, não foram incluídos nesta tese e foram considerados como um trabalho futuro.

Já os experimentos 6 e 7, nos quais utilizou-se os conjuntos de dados do Cenário F e G, respectivamente, apresentaram resultados semelhantes aos experimentos 3 e 4, ou seja, o processo ECE da abordagem proposta "*SEnsembles*" proporcionou os melhores valores da métrica ASAM com o *caliper* zero. No experimento 6, o *ensemble boosting* obteve os melhores valores da métrica ASAM em quatro *calipers* (zero, 0,20, 0,25 a 0,30), com redução do número de pares de instâncias abaixo de 5%, se comparado ao método *NNM* e, aumento do número de descartes de instâncias do grupo de tratamento ligeiramente superior a 6%, conforme se observa na Figura 6.28. Nota-se, que o processo ECE obteve ganhos de 0,5% a 21,2% do valor da métrica ASAM, com destaque para o *caliper* zero, no qual não reduziu o número de pares de instâncias correspondidas e não aumentou o número de descartes de instâncias do grupo de tratamento.

Figura 6.28 – Resultado das métricas ASAM, número de pares de instâncias correspondidas e número de instâncias descartadas do grupo de tratamento geradas pelo processo ECE da abordagem proposta "*SEnsembles*" e pelo método *NNM* (*baseline* de comparação) no experimento 6, com variação de *caliper* de 0 a 0,30 e, usando o conjunto de dados do Cenário F (Lee et al., 2010) com 10 covariáveis para efetuar a correspondência das instâncias.

Experimento 6

Cenário F - 10 Covariáveis		Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
ASAM	NNM	0,3847	**0,0301**	**0,0341**	**0,0411**	0,0527	0,0681	0,0801
	Bagging	**0,3682**	0,1187	0,1275	0,1365	0,1457	0,1552	0,1659
	RF	**0,3698**	0,1009	0,1089	0,1173	0,1277	0,1380	0,1479
	Boosting	**0,3828**	0,0542	0,0500	0,0500	**0,0500**	**0,0548**	**0,0631**
Pares	NNM	**456**	293	302	310	317	325	333
	Bagging	**456**	**322**	**333**	**342**	**347**	**354**	**365**
	RF	**456**	317	327	335	343	350	358
	Boosting	**456**	277	286	295	303	311	320
Descartes	NNM	**86**	249	240	232	225	216	209
	Bagging	**86**	**219**	**208**	**199**	**195**	**188**	**177**
	RF	**86**	224	215	207	199	191	184
	Boosting	**86**	265	256	247	239	230	222

% ASAM em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Bagging	4,3%	-294,2%	-274,5%	-232,0%	-176,6%	-127,8%	-107,2%
RF	3,9%	-235,1%	-219,8%	-185,5%	-142,4%	-102,6%	-84,7%
Boosting	0,5%	-80,1%	-46,7%	-21,7%	5,1%	19,6%	21,2%

% Pares em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Bagging	0,0%	10,1%	10,4%	10,6%	9,4%	8,8%	9,6%
RF	0,0%	8,4%	8,3%	8,0%	8,0%	7,7%	7,5%
Boosting	0,0%	-5,4%	-5,3%	-4,8%	-4,4%	-4,3%	-3,9%

% Descartes em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Bagging	0,0%	-11,9%	-13,1%	-14,1%	-13,3%	-13,2%	-15,3%
RF	0,0%	-9,8%	-10,4%	-10,7%	-11,3%	-11,6%	-11,9%
Boosting	0,0%	6,4%	6,6%	6,5%	6,2%	6,5%	6,3%

Melhor Resultado Melhor Valor Relativo

Fonte: Elaborado pelo autor.

Já no experimento 7, o processo ECE da abordagem proposta “SEnsembles” proporcionou os melhores valores da métrica ASAM em três *calipers* (zero, 0,25 e 0,30). Porém, com os *calipers* 0,25 e 0,30 resultou em uma redução do número de pares de instâncias correspondidas superior a 15%, conforme se observa na Figura 6.29. Porém, com o *caliper* zero isso não ocorreu.

Figura 6.29 – Resultado das métricas ASAM, número de pares de instâncias correspondidas e número de instâncias descartadas do grupo de tratamento geradas pelo processo ECE da abordagem proposta "*SEnsembles*" e pelo método *NNM* (*baseline* de comparação) no experimento 7, com variação de *caliper* de 0 a 0,30 e, usando o conjunto de dados do Cenário G (Lee et al., 2010) com 10 covariáveis para efetuar a correspondência das instâncias.

Experimento 7

Cenário G - 10 Covariáveis		Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
ASAM	NNM	0,3405	**0,0305**	**0,0355**	**0,0425**	**0,0537**	0,0672	0,0812
	Bagging	**0,3064**	0,1067	0,1118	0,1158	0,1225	0,1284	0,1374
	RF	**0,3104**	0,0749	0,0795	0,0844	0,0915	0,0999	0,1095
	Boosting	**0,3175**	0,0631	0,0620	0,0605	0,0587	**0,0601**	**0,0609**
Pares	NNM	**458**	**321**	**330**	**339**	**347**	**355**	**363**
	Bagging	**458**	315	322	329	335	343	357
	RF	**458**	304	314	322	329	338	346
	Boosting	**458**	263	272	281	289	298	307
Descartes	NNM	**83**	**220**	**211**	**203**	**195**	**187**	**178**
	Bagging	**83**	226	220	212	209	198	185
	RF	**83**	238	228	220	212	204	195
	Boosting	**83**	279	270	261	252	243	234

% ASAM em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Bagging	10,0%	-249,4%	-215,1%	-172,6%	-128,2%	-91,0%	-69,1%
RF	8,9%	-145,4%	-124,2%	-98,8%	-70,4%	-48,6%	-34,8%
Boosting	6,8%	-106,6%	-74,7%	-42,5%	-9,4%	10,5%	25,0%

% Pares em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Bagging	0,0%	-1,9%	-2,6%	-2,8%	-3,3%	-3,3%	-1,8%
RF	0,0%	-5,5%	-5,1%	-5,0%	-4,9%	-4,9%	-4,7%
Boosting	0,0%	-18,2%	-17,7%	-17,1%	-16,5%	-16,0%	-15,4%

% Descartes em Relação ao NNM

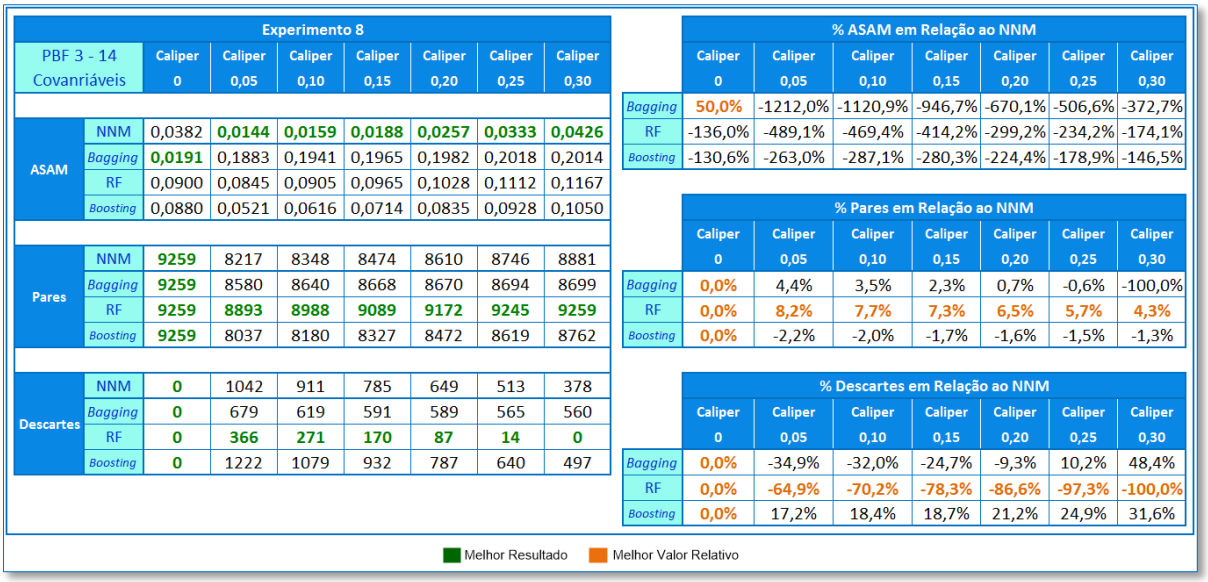
	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Bagging	0,0%	2,8%	4,1%	4,6%	7,3%	6,2%	3,6%
RF	0,0%	8,0%	8,0%	8,3%	8,7%	9,3%	9,6%
Boosting	0,0%	26,5%	27,7%	28,6%	29,4%	30,4%	31,5%

Melhor Resultado Melhor Valor Relativo

Fonte: Elaborado pelo autor.

Por fim, no experimento 8, no qual se utilizou o conjunto de dados PBF 3, o processo ECE somente resultou no melhor valor da métrica ASAM com o *ensemble bagging* com o *caliper* zero, com ganho de 50% se comparado ao método *NNM*, sem reduzir o número de pares de instâncias correspondidas e aumentar o número de descartes de instâncias do grupo de tratamento, conforme se observa na Figura 6.30.

Figura 6.30 – Resultado das métricas ASAM, número de pares de instâncias correspondidas e número de instâncias descartadas do grupo de tratamento geradas pelo processo ECE da abordagem proposta "*SEnsembles*" e pelo método *NNM* (*baseline* de comparação) no experimento 8, com variação de *caliper* de 0 a 0,30 e, usando o conjunto de dados do PBF 3 com 14 covariáveis para efetuar a correspondência das instâncias.



Experimento 8

PBF 3 - 14 Covanriáveis		Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
ASAM	NNM	0,0382	**0,0144**	**0,0159**	**0,0188**	**0,0257**	**0,0333**	**0,0426**
	Bagging	**0,0191**	0,1883	0,1941	0,1965	0,1982	0,2018	0,2014
	RF	0,0900	0,0845	0,0905	0,0965	0,1028	0,1112	0,1167
	Boosting	0,0880	0,0521	0,0616	0,0714	0,0835	0,0928	0,1050
Pares	NNM	**9259**	8217	8348	8474	8610	8746	8881
	Bagging	**9259**	8580	8640	8668	8670	8694	8699
	RF	**9259**	**8893**	**8988**	**9089**	**9172**	**9245**	**9259**
	Boosting	**9259**	8037	8180	8327	8472	8619	8762
Descartes	NNM	**0**	1042	911	785	649	513	378
	Bagging	**0**	679	619	591	589	565	560
	RF	**0**	**366**	**271**	**170**	**87**	**14**	**0**
	Boosting	**0**	1222	1079	932	787	640	497

% ASAM em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Bagging	**50,0%**	-1212,0%	-1120,9%	-946,7%	-670,1%	-506,6%	-372,7%
RF	-136,0%	-489,1%	-469,4%	-414,2%	-299,2%	-234,2%	-174,1%
Boosting	-130,6%	-263,0%	-287,1%	-280,3%	-224,4%	-178,9%	-146,5%

% Pares em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Bagging	**0,0%**	4,4%	3,5%	2,3%	0,7%	-0,6%	-100,0%
RF	**0,0%**	**8,2%**	**7,7%**	**7,3%**	**6,5%**	**5,7%**	**4,3%**
Boosting	**0,0%**	-2,2%	-2,0%	-1,7%	-1,6%	-1,5%	-1,3%

% Descartes em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
Bagging	**0,0%**	-34,9%	-32,0%	-24,7%	-9,3%	10,2%	48,4%
RF	**0,0%**	**-64,9%**	**-70,2%**	**-78,3%**	**-86,6%**	**-97,3%**	**-100,0%**
Boosting	**0,0%**	17,2%	18,4%	18,7%	21,2%	24,9%	31,6%

Melhor Resultado Melhor Valor Relativo

Fonte: Elaborado pelo autor.

A Tabela 6.8 apresenta um resumo dos resultados dos experimentos desta seção, com descrição dos conjuntos de dados, número do experimento, melhores resultados obtidos e uma observação sobre eles. Observa-se, que no experimento 1, o processo ECE da abordagem proposta "*SEnsembles*" obteve os melhores ASAM em quatro *calipers*, com ganhos de 1,8% a 36,5%. Já nos experimentos 2 e 5, esse ganho foi obtido somente com o *caliper* zero, sendo de 3,0% e 6,4%, respectivamente. Entretanto, o processo ECE obteve ganhos de 7,5% a 24,8%, no experimento 3, de 0,4% a 8% no experimento 4, de 0,5% a 21,2% no experimento 6 de 6,8% a 25% no experimento 7 e, por último, no experimento 8, de 50% somente com o *caliper* zero.

Tabela 6.8 – Resumo dos resultados dos experimentos pelos quais se comparou as correspondências geradas pelo processo ECE da abordagem proposta "*SEnsembles*" e pelo método *NNM* (*baseline da comparação*), com a descrição dos conjuntos de dados, número do experimento, melhores resultados e os ganhos obtidos.

Conjunto de Dados		Exp.*	Melhores Resultados	Ganho
Lalonde (1986)		1	Processo ECE com quatro *calipers* (zero, 0,20, 0,25 e 0,30).	De 1,8% a 36,5%.
Lee et al. (2010)	Cenário B: Leve (*mild*) não linearidade.	2	Processo ECE com um *caliper* (0,30).	3,0%.
	Cenário C: Moderada não linearidade.	3	Processo ECE com três *calipers* (zero, 0,25 e 0,30).	De 10,3% a 24,8%
	Cenário D: Leve (*mild)* não aditividade.	4	Processo ECE com três *calipers* (zero, 0,25 e 0,30).	De 2,5% a 8%
	Cenário E: Leve (*mild)* não aditividade e não linearidade	5	Processo ECE com um *caliper* (0,30).	6,4%.
	Cenário F: Moderada não aditividade.	6	Processo ECE com quatro *calipers* (zero, 0,20, 0,25 e 0,30).	De 4,3% a 21,2%
	Cenário G: Moderada não aditividade e não linearidade	7	Processo ECE com três *calipers* (zero, 0,25 e 0,30).	De 10% a 25%
PBF 3 – 14 Covariáveis Modificada a partir de Martins (2013)		8	Processo ECE com um *caliper* (zero)	50%.

***Experimentos.**

Fonte: Elaborado pelo autor.

É importante destacar que o processo ECE da abordagem "*SEnsembles*" obteve os melhores valores da métrica ASAM quando utilizado o menor *caliper* (zero) e os maiores (0,20, 0,25 e 0,30). Assim, em uma faixa que varia de 0,05 a 0,15, ou seja, com três *calipers*, não obteve os melhores resultados, conforme se observa na Tabela 6.9, a qual apresenta um mapeamento de *calipers* nos quais se obteve melhoria da qualidade da correspondência de instâncias quando comparado o processo ECE ao *baseline NNM*.

Tabela 6.9 – Mapeamento de *calipers* nos quais se obteve melhoria da qualidade da correspondência quando comparado o processo ECE da abordagem proposta "*SEnsembles*" com o método *NNM* (*baseline* de comparação).

Conjunto de Dados		**Exp.***	***Calipers***						
			Zero	**0,05**	**0,10**	**0,15**	**0,20**	**0,25**	**0,30**
Lalonde (1986)		1	1,8%				10,7%	22,2%	36,3%
Lee et al. (2010)	Cenário B: Leve (*mild*) não linearidade.	*2*							3,0%
	Cenário C: Moderada não linearidade	3	10,3%					3,8%	24,8%
	Cenário D: Leve (mild) não aditividade.	4	2,5%					4,3%	8,0%
	Cenário E: Leve (*mild)* não aditividade e não linearidade.	5							6,4%
	Cenário F: Moderada não aditividade.	6	4,3%				5,1%	19,6%	21,2%
	Cenário G: Moderada não aditividade e não linearidade	7	10,0%					10,5%	25,0%
PBF 3 – 14 Covariáveis Modificada a partir de Martins (2013)		8	50%						

*** Experimento.**

Fonte: Elaborado pelo autor.

Entretanto, com exceção do *caliper* zero, o ganho obtido pelo processo ECE sempre foi acompanhado de uma redução do número de pares de instâncias correspondidas, conforme se observa pela cor em vermelho na Tabela 6.10 e, um aumento talvez não desejável do número de instâncias descartadas do grupo de tratamento. Nota-se que a redução do número de pares de instâncias correspondidas ficou abaixo de 5% nos experimentos 4, 5 e 6 e, em 6% no experimento 2. Nos demais experimentos, 1 e 8, esse número foi superior 15%.

Tabela 6.10 – Mapeamento de *calipers* com a descrição das porcentagens de pares de instâncias obtidas pelo processo ECE da abordagem proposta "*SEnsembles*" em relação ao método *NNM* (*baseline* da comparação), somente nas situações que houve melhoria da qualidade da correspondência.

Conjunto de Dados		Exp.*	*Calipers*						
			Zero	0,05	0,10	0,15	0,20	0,25	0,30
Lalonde (1986)		1	0%				-16%	-15,5%	-13,6%
Lee et al. (2010)	Cenário B: Leve (*mild*) não linearidade.	2							-6%
	Cenário C: Moderada não linearidade	3	0%					-15,7%	-15,4%
	Cenário D: Leve (*mild)* não aditividade.	4	0%					-4,1%	-4,0%
	Cenário E: Leve (*mild)* não aditividade e não linearidade.	5							-4,3%
	Cenário F: Moderada não aditividade.	6	0%				-4,4%	-4,3%	-3,9%
	Cenário G: Moderada não aditividade e não linearidade	7	0%					-16,0%	-15,4%
PBF 3 – 14 Covariáveis Modificada a partir de Martins (2013)		8	0%						

***Experimento.**

Fonte: Elaborado pelo autor.

Como visto, o processo ECE resultou nas melhores correspondências com o *caliper* zero, no qual não reduziu o número de pares de instâncias correspondidas e, com os *calipers* 0,20, 0,25 e 0,30, nos quais houve redução do número de pares de instâncias correspondidas de -3,9% a -16%, dependendo do *caliper* e do conjunto de dados utilizados. Ressalta-se que o ganho obtido está vinculado aos tipos de *ensembles* utilizados (*bagging*, *random forest* e *boosting)*, os quais foram selecionados por possuírem características diferentes, uma vez que os *ensembles bagging* e *boosting* manipulam o conjunto de dados de entrada, enquanto que o *ensemble random forest* manipula as covariáveis de entrada. Além disso, o uso de *ensembles* com características diferentes também permitiu ganhos considerando-se diferentes aspectos dos dados manipulados. Por exemplo, uma pequena variação do conjunto de dados pode gerar alterações significativas no resultado do *ensemble bagging*. Já o *ensemble boosting* é mais sensível a *outlines*, pois em cada iteração os pondera com maior peso, enquanto que o *ensemble random forest* utiliza as covariáveis de maneira aleatória para a construção dos modelos regressores.

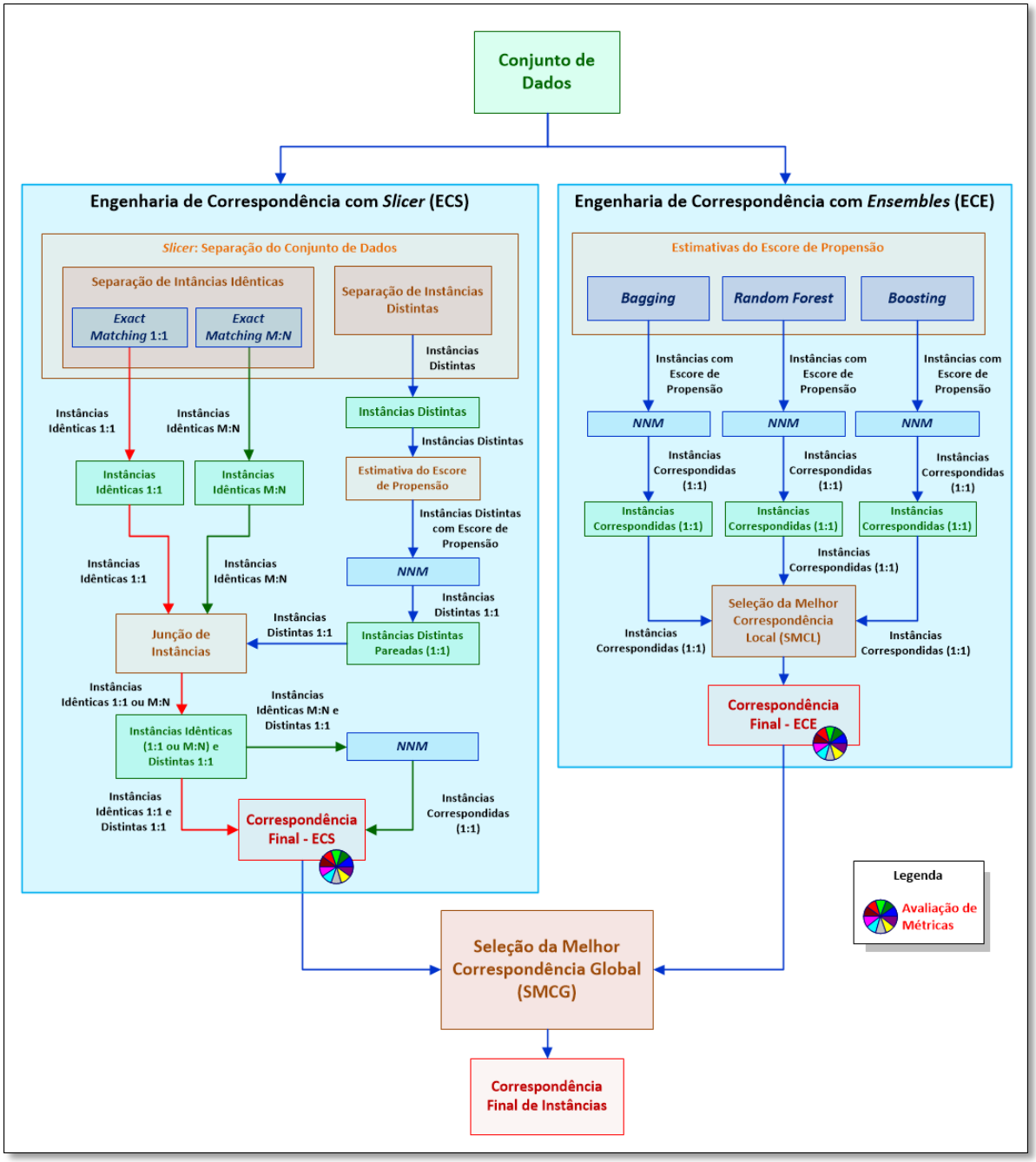
6.6 Comparação das correspondências geradas pelos processos ECS e ECE da abordagem proposta "*SEnsembles*"

Esta seção realiza uma análise comparativa entre as correspondências geradas pelos processos ECS e ECE da abordagem proposta "*SEnsembles"*. Para isso, foram realizados dois estágios de avaliação de métricas, um no final do processo ECS e outro no final do processo ECE, conforme ilustrada na Figura 6.31. Nesse comparativo foram utilizadas somente as métricas absolutas (i.e. ASAM, número de pares de instâncias correspondidas e número de descarte de instância do grupo de tratamento), uma vez que não há valores referenciais para comparação.

Nos experimentos desta seção utilizou-se o clássico conjunto de dados Lalonde (1986), o conjunto de dados PBF 3, que foi modificado a partir do conjunto de dados cedido por Martins (2013) para contemplar instâncias com características idênticas nos grupos de tratamento e de controle e, por fim, o conjunto de dados do

Cenário F descrito por Lee et al. (2010), o qual não possui nenhum par de instâncias idênticas.

Figura 6.31 – Estágios de avaliação das métricas utilizadas para comparar as correspondências geradas pelo processo ECS x ECE da abordagem proposta "*SEnsembles*".



Fonte: Elaborado pelo autor.

No experimento que usou o conjunto de dados Lalonde (1986), o processo ECE, considerando-se as suas configurações, obteve os melhores valores para métrica ASAM em todos *calipers*, conforme se observa na Figura 6.32.

Figura 6.32 – Resultado das métricas ASAM, número de pares de instâncias correspondidas e número de instâncias descartadas do grupo de tratamento geradas pelos processos ECS e ECE da abordagem proposta "*SEnsembles*", com variação de *caliper* de 0 a 0,30 e, usando o conjunto de dados do Lalonde (1986) com 10 covariáveis para efetuar a correspondência das instâncias.

ECS x ECE								
Lalonde - 10 Covariáveis			Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
ASAM	ECE		0,2664	0,0957	0,0871	0,0802	0,0815	0,0830	0,0840
	ECS	Conf.1	**0,2561**	**0,0552**	**0,0554**	0,0804	0,0849	0,1077	0,1252
		Conf.2	0,2674	0,0899	0,0652	**0,0567**	**0,0530**	**0,0550**	**0,0654**
Pares	ECE		**185**	85	90	92	95	98	101
	ECS	Conf.1	**185**	**108**	**109**	**112**	**113**	**115**	**116**
		Conf.2	184	73	86	91	96	96	98
Descartes	ECE		**0**	100	95	93	90	87	84
	ECS	Conf.1	**0**	**77**	**76**	**73**	**72**	**70**	**69**
		Conf.2	1	112	99	94	89	89	87

Melhor Resultado

Fonte: Elaborado pelo autor.

Já no experimento que usou o conjunto de dados do PBF 3 com 14 covariáveis para efetuar a correspondência das instâncias, o processo ECS obteve os melhores valores da métrica ASAM em seis *calipers* (exceto o *caliper* zero), enquanto que o processo ECE obteve em apenas um *caliper* (zero), conforme se observa-se na Figura 6.33.

Figura 6.33 – Resultado das métricas ASAM, número de pares de instâncias correspondidas e número de instâncias descartadas do grupo de tratamento geradas pelos processos ECS e ECE da abordagem proposta "*SEnsembles*", com variação de *caliper* de 0 a 0,30 e, usando o conjunto de dados do PBF 3 com 14 covariáveis para efetuar a correspondência das instâncias.

ECS x ECE								
PBF 3 - 14 Covariáveis			Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
ASAM	ECE		**0,0191**	0,0521	0,0616	0,0714	0,0835	0,0928	0,1050
	ECS	Conf.1	0,0373	0,0153	0,0155	**0,0172**	**0,0227**	**0,0302**	**0,0390**
		Conf.2	0,0550	**0,0071**	**0,0154**	0,0222	0,0286	0,0347	0,0402
Pares	ECE		**9259**	8037	8180	8327	8472	8619	8762
	ECS	Conf.1	**9259**	**8202**	**8321**	**8435**	**8555**	**8680**	**8801**
		Conf.2	9169	7903	8143	8236	8285	8342	8408
Descartes	ECE		**0**	1222	1079	932	787	640	497
	ECS	Conf.1	**0**	**1057**	**938**	**824**	**704**	**579**	**458**
		Conf.2	90	1356	1116	1023	974	917	851

Melhor Resultado

Fonte: Elaborado pelo autor.

Por fim, no experimento no qual se utilizou o Cenário F (Lee et al., 2010), o processo ECE obteve os melhores valores da métrica ASAM em quatro *calipers* (zero, 0,20, 0,25 e 0,30), enquanto que o processo ECS obteve os melhores valores da métrica ASAM apenas em três calipers (0,05, 0,10 e 0,15), conforme se observa na Figura 6.34.

Figura 6.34 – Resultado das métricas ASAM, número de pares de instâncias correspondidas e número de instâncias descartadas do grupo de tratamento geradas pelos processos ECS e ECE da abordagem proposta "*SEnsembles*", com variação de *caliper* de 0 a 0,30 e, usando o conjunto de dados do Cenário F (Lee et al., 2010) com 10 covariáveis para efetuar a correspondência das instâncias.

ECS x ECE									
Cenário F - 10 Covariáveis			Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
ASAM	ECE		**0,3682**	0,0542	0,0500	0,0500	**0,0500**	**0,0548**	**0,0631**
	ECS	Conf.1	0,3847	**0,0302**	0,0345	**0,0415**	0,0526	0,0680	0,0800
		Conf.2	0,3847	0,0304	**0,0340**	**0,0415**	0,0681	0,0681	0,0803
Pares	ECE		**456**	277	286	295	303	311	320
	ECS	Conf.1	**456**	**293**	**302**	**310**	**317**	**325**	**333**
		Conf.2	**456**	**293**	**302**	**310**	**317**	**325**	**333**
Descartes	ECE		**86**	265	256	247	239	230	222
	ECS	Conf.1	**86**	**249**	**240**	**232**	**225**	**216**	**209**
		Conf.2	**86**	**249**	**240**	**232**	**225**	**216**	**209**

Melhor Resultado

Fonte: Elaborado pelo autor.

A Tabela 6.11 apresenta os resumos dos experimentos desta seção. Nota-se que há uma variação, dependendo do conjunto de dados utilizados, do processo da abordagem proposta "*SEnsembles*" que obteve os melhores valores da métrica ASAM. Nota-se que o processo ECS proporcionou os melhores resultados em todos os *calipers* quando utilizado o conjunto de dados Lalonde. Já, com o conjunto de dados PBF 3, que possui a mesma proporção de instâncias com características idênticas ao conjunto de dados Lalonde, o processo ECE obteve o melhor ASAM com o *caliper* zero e, o processo ECS com os demais. E, por último, com o conjunto de dados do Cenário F (Lee et. al., 2010), que não possui nenhum par de instâncias idênticas, o processo ECS obteve os melhores valores para a métrica ASAM com três *calipers* (0,05, 0,10 e 0,15), enquanto o processo ECE nos demais *calipers* (zero, 0,20, 0,25 e 0,30).

Tabela 6.11 – Resumo dos resultados dos experimentos pelos quais se comparou as correspondências geradas pelo processo ECS e ECE da abordagem proposta "*SEnsembles*", com a descrição dos conjuntos de dados, número de instâncias com características idênticas, melhores resultados obtidos da métrica ASAM e observações sobre esses resultados.

Conjunto de Dados	Instâncias Idênticas	Melhores Resultados (ASAM)
Lalonde (1986)	13 Tratados e 2 Não Tratados	ECS em todos os *calipers*.
PBF 3 Modificado a partir de Martins (2013).	1.185 Beneficiários e 1.095 não beneficiários	ECS com os demais *calipers* (0,05, 0,10, 0,15, 0,20, 0,25 e 0,30). ECE com o *caliper* zero.
Cenário F (Lee et al., 2010).	Nenhuma	ECS: com três *calipers* (0,05, 0,10 e 0,15). ECE com quatro *calipers* (zero, 0,20, 0,25 e 0,30).

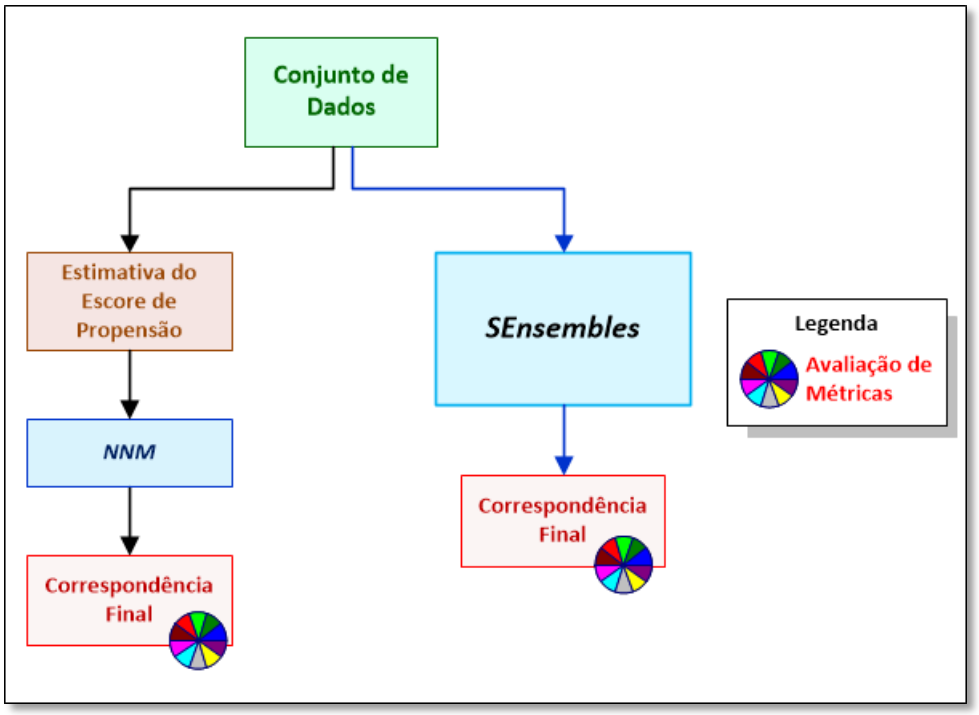
Fonte: Elaborado pelo autor.

O processo ECS obteve os melhores valores de ASAM para a grande maioria dos valores de *caliper* nos 3 experimentos. Porém, o processo ECE superou o processo ECS em 5 valores de *caliper* em 2 dos experimentos. Dessa forma, verifica-se aqui a necessidade de manter os dois processos na abordagem proposta "*SEnsembles*", uma vez que as melhores correspondências variam de acordo com o conjunto de dados e com o *caliper* utilizado.

6.7 Comparação da abordagem proposta "*SEnsembles*" com o *baseline NNM*

Esta seção realiza uma análise comparativa da qualidade das correspondências geradas pela abordagem proposta "*SEnsembles*", considerando a melhor correspondência gerada por seus processos, seja ECS ou ECE, com as correspondências geradas pelo método *NNM* (*baseline* de comparação) utilizando as métricas ASAM, número de pares de instâncias correspondidas e número de instâncias descartadas do grupo de tratamento. Nesse comparativo, foram realizados dois estágios de avaliação dessas métricas, conforme se observa na Figura 6.35.

Figura 6.35 – Estágios de avaliação das métricas utilizadas para comparar as correspondências geradas pela abordagem proposta "SEnsembles" com o método *NNM* (*baseline* da comparação).



Fonte: Elaborado pelo autor.

Nos experimentos desta seção utilizou-se os mesmos conjuntos de dados da seção anterior (Seção 6.6), ou seja, o clássico conjunto de dados Lalonde (1986), o conjunto de dados PBF 3 e o conjunto de dados do Cenário F descrito por Lee et al. (2010).

No experimento que foi utilizado o conjunto de dados Lalonde (1986), a abordagem "*SEnsembles*" proporcionou os melhores valores da métrica ASAM em seis *calipers* (exceto 0,10), com ganhos de 5,6% a até 51,4%. Além disso, nota-se que a redução do número de pares de instâncias correspondidas com os *calipers* zero e 0,05 ficou abaixo de -1,1% e, o número de instâncias descartadas do grupo de tratamento com esses *calipers* não foi superior a 1,6%, conforme se observa na Figura 6.36.

Figura 6.36 – Resultado das métricas ASAM, número de pares de instâncias correspondidas e número de instâncias descartadas do grupo de tratamento geradas pela abordagem proposta "*SEnsembles*" e *NNM* (*baseline* da comparação), com variação de *caliper* de 0 a 0,30 e, usando o conjunto de dados Lalonde (1986) com 10 covariáveis para efetuar a correspondência das instâncias.

SEnsembles x NNM

Lalonde		Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
ASAM	NM	0,2713	0,0576	**0,0488**	0,0668	0,0916	0,1077	0,1347
	SEnsembles	**0,2561**	**0,0552**	0,0554	**0,0567**	**0,0530**	**0,0550**	**0,0654**
Pares	NM	**185**	**109**	**111**	**112**	**113**	**116**	**117**
	SEnsembles	**185**	108	109	91	96	96	98
Descartes	NM	**0**	**76**	**74**	**73**	**72**	**69**	**68**
	SEnsembles	**0**	77	76	94	89	89	87

% ASAM em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
SEnsembles	5,6%	4,0%	-13,6%	15,1%	42,1%	48,9%	51,4%

% Pares em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
SEnsembles	0,0%	-1,1%	-1,8%	-18,5%	-15,5%	-17,6%	-16,0%

% Descartes em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
SEnsembles	0,0%	1,6%	2,7%	28,4%	24,3%	29,7%	27,5%

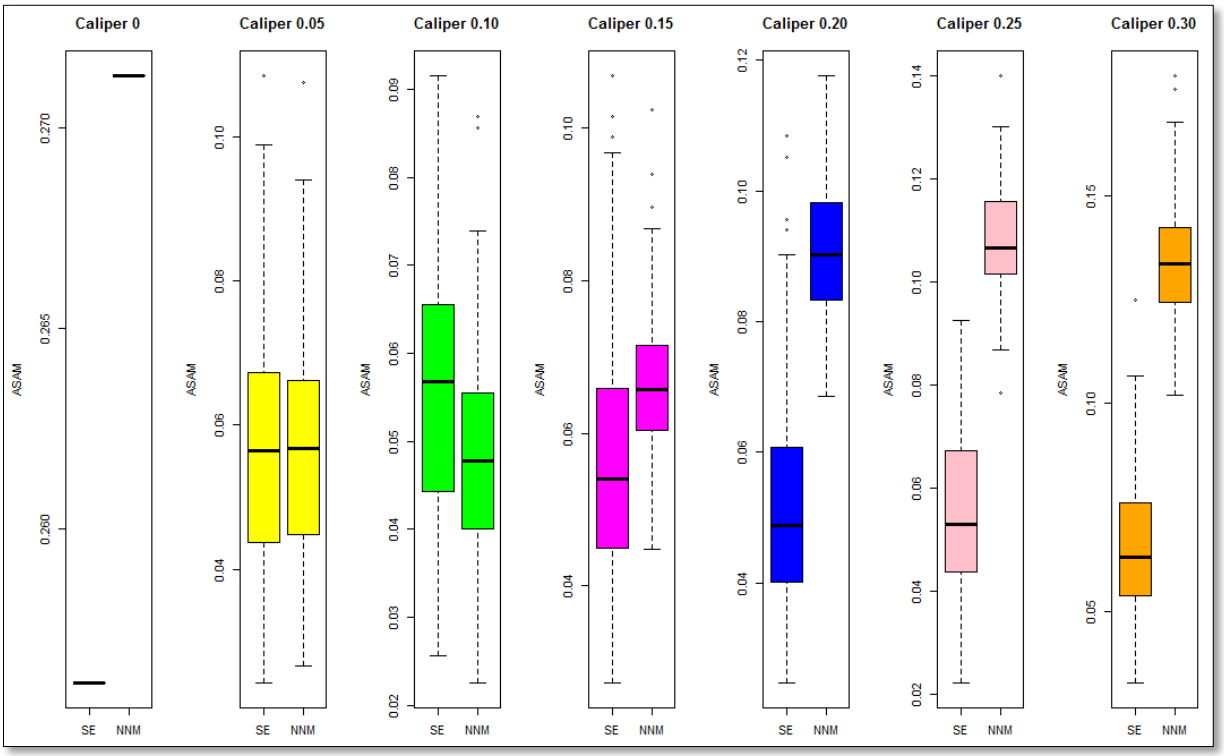
Melhor Resultado Melhor Valor Relativo

Fonte: Elaborado pelo autor.

Em média, a abordagem proposta "*SEnsembles*" proporcionou uma melhora de 27,9% do valor da métrica ASAM, excetuando o *caliper* 0,10, com uma redução de -11,4% no número de pares de instâncias correspondidas e, um aumento de 18,6% do número de instâncias descartadas do grupo de tratamento.

A Figura 6.37 ilustra a variação dos valores da métrica ASAM gerados pela abordagem proposta "*SEnsembles*" e o *NNM* (*baseline*), considerando-se os *calipers* utilizados (0 a 0,30) e, de acordo com o Test F para análise de variância, pode-se concluir que não há evidencia de diferença significativa da variância dos valores obtidos da métrica ASAM, ao nível de significância de 5%, entre a abordagem proposta "*SEnsembles*" e o *NNM* com os *calipers*: 0,05, 0,10 e 0,30, obtendo-se p-valores de 0,6288, 0,2004 e 0,1358, respectivamente. Já em relação aos testes estatísticos para comparação da média, não há evidência de diferença significativa dos valores médios obtidos da métrica ASAM, ao nível de significância de 5%, entre a abordagem proposta "*SEnsembles*" e o *NNM* com o *caliper* 0,05, uma vez que o Teste T e Teste Tukey resultaram em um p-valor de 0,3115.

Figura 6.37 – Variação da métrica ASAM gerada pela abordagem proposta "*SEnsembles*" (SE) e *NNM* (*baseline* da comparação), com *calipers* de 0 a 0,30 e, usando o conjunto de dados Lalonde (1986) com 10 covariáveis para efetuar a correspondência das instâncias.



Fonte: Elaborado pelo autor.

Já no experimento que usou o conjunto de dados PBF 3, a abordagem "SEnsembles" obteve os melhores valores da métrica ASAM em todos os *calipers*, com ganhos de 2,5% até 50,0%, com pequena redução do número de pares de instâncias correspondidas, o qual não ultrapassou, no pior resultado, uma redução de aproximadamente -3,8%. Porém, o maior descarte de instâncias do grupo de tratamento ocorreu em quatro calipers, 0,05, 0,20, 0,25 e 0,30, sendo de 30,1%, 8,5%, 13% e 21,3% respectivamente, conforme se observa pela Figura 6.38.

Figura 6.38 – Resultado das métricas ASAM, número de pares de instâncias correspondidas e número de instâncias descartadas do grupo de tratamento geradas pela abordagem proposta "*SEnsembles*" e *NNM* (*baseline* da comparação), com variação de *caliper* de 0 a 0,30 e, usando o conjunto de dados PBF 3 com 14 covariáveis para efetuar a correspondência das instâncias.

SEnsembles x NNM

PBF 3 - 14 Covariáveis		Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
ASAM	NM	0,0382	0,0144	0,0159	0,0188	0,0257	0,0333	0,0426
	SEnsembles	**0,0191**	**0,0071**	**0,0155**	**0,0172**	**0,0227**	**0,0302**	**0,0390**
Pares	NM	9259	8217	8348	8474	8610	8746	8881
	SEnsembles	9259	7903	8321	8435	8555	8680	8801
Descartes	NM	0	1042	911	785	649	513	378
	SEnsembles	0	1356	938	824	704	579	458

% ASAM em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
SEnsembles	50,0%	50,6%	2,5%	8,4%	11,7%	9,3%	8,4%

% Pares em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
SEnsembles	0,0%	-3,8%	-0,3%	-0,5%	-0,6%	-0,8%	-0,9%

% Descartes em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
SEnsembles	0,0%	30,1%	3,0%	5,0%	8,5%	13,0%	21,3%

Melhor Resultado Melhor Valor Relativo

Fonte: Elaborado pelo autor.

Em média, a abordagem proposta "*SEnsembles*" proporcionou uma melhora de 20,1% do valor da métrica ASAM, excetuando o *caliper* 0,10, com uma redução de -1,0% no número de pares de instâncias correspondidas e, um aumento de 11,6% do número de instâncias descartadas do grupo de tratamento. Ressalta-se também que o *caliper* zero houve uma melhora de 50% da métrica ASAM, sem redução do número de pares de instâncias correspondidas ou aumento do número de instâncias descartadas do grupo de tratamento.

Por fim, no experimento que usou o conjunto de dados do Cenário F (Lee et al., 2010), a abordagem proposta "*SEnsembles*" proporcionou os melhores valores da métrica ASAM em cinco *calipers* (zero, 0,20, 0,25 e 0,30), com ganhos de 4,3% a 21,2%, com pequenas reduções do número de pares de instâncias correspondidas (de até -4,4%), conforme se observa na Figura 6.39. Além disso, com os *calipers* 0,20, 0,25 e 0,30 resultou nos maiores descartes de instâncias do grupo de tratamento, com valores superiores a 6,2%.

Figura 6.39 – Resultado das métricas ASAM, número de pares de instâncias correspondidas e número de instâncias descartadas do grupo de tratamento geradas pela abordagem proposta "*SEnsembles*" e *NNM* (*baseline* da comparação), com variação de *caliper* de 0 a 0,30 e, usando o conjunto de dados do Cenário F (Lee et al., 2010) com 10 covariáveis para efetuar a correspondência das instâncias.

SEnsembles x NNM

Cenário F - 10 Covariáveis		Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
ASAM	NM	0,3847	0,0301	0,0341	**0,0411**	0,0527	0,0681	0,0801
	SEnsembles	**0,3682**	0,0302	**0,0340**	0,0415	**0,0500**	**0,0548**	**0,0631**
Pares	NM	**456**	**293**	**302**	**310**	317	325	333
	SEnsembles	**456**	**293**	**302**	**310**	303	311	320
Descartes	NM	**86**	**249**	**240**	**232**	225	216	209
	SEnsembles	**86**	**249**	**240**	**232**	239	230	222

% ASAM em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
SEnsembles	4,3%	-0,1%	0,0%	-0,9%	5,1%	19,6%	21,2%

% Pares em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
SEnsembles	0,0%	0,0%	0,0%	0,1%	-4,4%	-4,3%	-3,9%

% Descartes em Relação ao NNM

	Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
SEnsembles	0,0%	0,2%	0,0%	0,1%	6,2%	6,5%	6,3%

Melhor Resultado Melhor Valor Relativo

Fonte: Elaborado pelo autor.

Em média, neste último experimento, a abordagem proposta "*SEnsembles*" proporcionou uma melhora de 10,1% do valor da métrica ASAM, excetuando os *calipers* 0,05 e 0,15, com uma redução de -2,5% no número de pares de instâncias correspondidas e, um aumento de 3,8% do número de instâncias descartadas do grupo de tratamento. Ressalta-se também que o *caliper* zero houve uma melhora de 50% da métrica ASAM, sem reduzir o número de pares de instâncias correspondidas ou aumento do número de descartes de instâncias do grupo de tratamento.

A Tabela 6.12 apresenta um resumo do comparativo entre a abordagem proposta "*SEnsembles*" e o método *NNM* (*baseline* da comparação), com a descrição do conjunto de dados utilizado, melhores resultados obtidos e uma

observação desses resultados. Observa-se que a abordagem "*SEnsembles*" para o conjunto de dados Lalonde resultou em ganhos de 5,6% a 51,4%, com média de melhoria do valor da métrica ASAM de 27,9%, com redução média do número de pares de instâncias correspondidas de -11,4% e, média de descartes de instâncias do grupo de tratamento de 18,6%.

Tabela 6.12 – Resumo dos resultados dos experimentos pelos quais se comparou as correspondências geradas pela abordagem proposta "*SEnsembles*" e *NNM* (*baseline* da comparação), com a descrição dos conjuntos de dados, melhores resultados obtidos, médias das métricas utilizadas e observações sobre esses resultados.

Conjunto de Dados	Melhores Resultados	Médias	Ganhos
Lalonde (1986)	*SEnsembles* em seis *calipers* (exceto 0,05).	**ASAM:** 27,9% de melhoria. **Pares:** -11,4%. **Descartes:** 18,6%	De 5,6% a 51,4%.
PBF 3 – 14 Covariáveis Modificada a partir de Martins (2013)	*SEnsembles* em todos os *calipers.*	**ASAM:** 20,1% de melhoria. **Pares:** -1,0%. **Descartes:** 11,4%	De 2,5% a 50%.
Cenário F: Moderada não aditividade Lee et al. (2010).	*SEnsembles* em cinco os *calipers* (exceto 0,05 e 0,15)	**ASAM:** 10,0% de melhoria. **Pares:** -2,5 **Descartes:** 3,8%	De 4,3% a 21,2%

Fonte: Elaborado pelo autor.

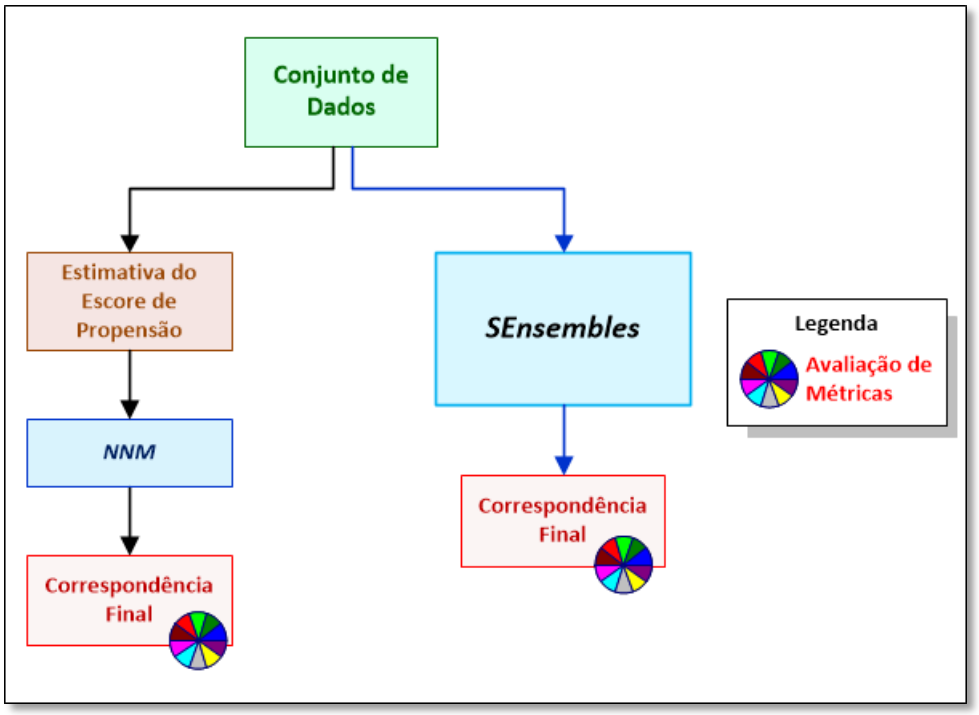
Já com o conjunto de dados PBF 3, a abordagem "*SEnsembles*" resultou em ganhos de 2,5% a 50%, com média de melhoria do valor da métrica ASAM de 20,1%, com redução média do número de pares de instâncias correspondidas de apenas -1,0% e, média de descartes de instâncias do grupo de tratamento de 11,4%. E por fim, para o conjunto de dados do Cenário F, a abordagem proposta "*SEnsembles*" resultou em ganhos de 4,3% a 21,2%, com redução média do número de pares de instâncias correspondidas de -2,5% e, média de descartes de instâncias do grupo de tratamento de 3,8%.

Como visto, a abordagem "*SEnsembles*" se apresenta como um alternativa viável para melhorar a qualidade da correspondência de instâncias em conjuntos de dados com características de disjunção, uma vez que proporcionou correspondências mais similares de instâncias nos experimentos com a maioria dos *calipers* utilizados.

6.8 Comparação de desempenho de execução da abordagem proposta "*SEnsembles*" com o *baseline NNM*

Esta seção apresenta um comparativo de desempenho de execução da abordagem proposta "*SEnsembles*" com o método *NNM* (*baseline*). Neste comparativo foi utilizado como métrica o tempo total de execução em segundos, considerando-se como o tempo total da abordagem *"SEnsembles"* a somatória do tempo necessário para a execução dos processos ECS e ECE. Dessa forma, foram obtidos dois tempos totais finais em cada experimento, ou seja, dois estágios de comparação de métricas, sendo um para a abordagem proposta "*SEnsembles*" e o outro para o método *NNM*, conforme se observa-se na Figura 6.40.

Figura 6.40 – Estágios de avaliação da métrica tempo total de execução (em segundos) utilizada para comparar o tempo necessário para a execução da abordagem proposta "*SEnsembles*" com o método *NNM* (*baseline* da comparação).



Fonte: Elaborado pelo autor.

Nos experimentos utilizou-se o conjunto de dados Lalonde (1986) e o conjunto de dados do Cenário F descrito por Lee et al. (2010), contendo, respectivamente, 614 e 1000 instâncias. Os tempos foram obtidos por meio de 100 execuções para cada *caliper* utilizado (zero, 0,10, 0,15, 0,20, 0,25 e 0,30). É importante ressaltar que o tempo do processo ECE da abordagem "*SEnsembles*" foi obtido considerando-se a somatória dos tempos para obter a correspondência das instâncias, quando utilizado os *ensembles* (*bagging, random forest e boosting*) em substituição à regressão logística ao estimar os escores de propensão, ou seja, considerou-se o pior cenário possível sem a execução em paralelo (simultânea) desses métodos. Além disso, também não foi considerado a execução em paralelo dos processos ECS e ECE da abordagem proposta "*SEnsembles*", isto é, também foi considerado o pior cenário possível nos experimentos realizados.

No experimento que foi utilizado o conjunto de dados Lalonde (1986), a abordagem "*SEnsembles*" proporcionou os tempos totais de execução próximos de cinco segundos para todos os calipers utilizados, enquanto o método *NNM* gerou tempos de execução inferiores a um segundo, conforme se observa na Figura 6.41, a qual apresenta o tempo dos processos ECS e ECE e, em seguida, o tempo total da abordagem proposta "*SEnsembles*" e do método *NNM*. Observa-se que a maior parte do tempo total de execução da abordagem proposta "*SEnsembles*" foi gerado pelo processo ECE (em destaque na cor amarela), mais precisamente, pela correspondência que utilizou o *ensemble boosting* para estimar os escores de propensão das instâncias. Por outro lado, observa-se também que o tempo de execução do processo ECS, seja utilizado a Configuração 1 ou 2, são bem próximos do método *NNM* (*baseline*), o que evidencia que o processo ECE foi o responsável por elevar o tempo total da abordagem proposta "*SEnsembles*" para em torno de cinco segundos.

Figura 6.41 – Resultados dos tempos de execução (em segundos) dos processos ECE e ECS da abordagem proposta "*SEnsembles*", Tempo Total da abordagem proposta "*SEnsembles*" considerando-se a somatória dos tempos de seus processos (ECE e ECS) e, o Tempo do método *NNM* (*baseline* da comparação), com variação de *caliper* de 0 a 0,30 e, usando o conjunto de dados Lalonde (1986) com 10 covariáveis para efetuar a correspondência das instâncias.

Tempo de Execução (em segundos) dos Processos ECS e ECE								
Lalonde - 10 Covariáveis		Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
ECE	*Bagging*	0,1854	0,1880	0,1832	0,1869	0,1811	0,1830	0,1859
	RF	0,3968	0,3933	0,3883	0,3823	0,3872	0,3838	0,3846
	Boosting	5,0747	5,2121	5,1193	5,1707	5,0853	5,0269	5,0703
	Total ECE:	**5,6569**	**5,7934**	**5,6907**	**5,7400**	**5,6536**	**5,5936**	**5,6408**
ECS	Conf.1	0,0707	0,0719	0,0714	0,0718	0,0722	0,0716	0,0719
	Conf.2	0,0994	0,0911	0,0882	0,0890	0,0888	0,0889	0,0897
SEnsembles (SE) - Tempo Total (em segundos)								
SE	ECS Conf.1 + Total ECE	5,7276	5,8653	5,7621	5,8118	5,7258	5,6652	5,7127
	ECS Conf.2 + Total ECE	5,7466	5,8845	5,7789	5,8289	5,7424	5,6825	5,7305
NNM (*Baseline* da comparação) - Tempo Total (em segundos)								
NNM	Regressão Logistica	0,0593	0,0623	0,0625	0,0655	0,0646	0,0651	0,0632

Fonte: Elaborado pelo autor.

Já no experimento que foi utilizado o conjunto de dados do Cenário F, a abordagem "*SEnsembles*" proporcionou os tempos totais de execução próximos de onze segundos, enquanto o método *NNM* gerou tempos de execução, novamente, inferiores a um segundo para todos os *calipers,* conforme se observa na Figura 6.42. Mais uma vez nota-se que os tempos das Configurações 1 e 2 são bem próximos do método *NNM* em todos os *calipers* utilizados e, ainda, o *ensembles boosting* foi responsável por elevar o tempo total da abordagem proposta "*SEnsembles*" para em torno de onze segundos em todos os *calipers*. Assim, nos dois experimentos realizados os tempos obtidos tiveram a mesma tendência, com o ensemble *boosting* elevando o tempo total da abordagem proposta "*SEnsembles*", enquanto que as Configurações 1 e 2 do processo ECS gerou tempos de execução próximos aos do método *NNM* (*baseline* da comparação).

Figura 6.42 – Resultados dos tempos de execução (em segundos) dos processos ECE e ECS da abordagem proposta "*SEnsembles*", Tempo Total da abordagem proposta "*SEnsembles*" considerando-se a somatória dos tempos de seus processos (ECE e ECS) e, o Tempo do método *NNM* (*baseline* da comparação), com variação de *caliper* de 0 a 0,30 e, usando o conjunto de dados do Cenário F (Lee et al., 2010) com 10 covariáveis para efetuar a correspondência das instâncias.

Tempo de Execução (em segundos) dos Processos ECS e ECE								
Cenário F - 10 Covariáveis		Caliper 0	Caliper 0,05	Caliper 0,10	Caliper 0,15	Caliper 0,20	Caliper 0,25	Caliper 0,30
ECE	*Bagging*	0,3314	0,3636	0,3617	0,3670	0,3606	0,3645	0,3513
	RF	1,2446	1,2452	1,2640	1,2526	1,2367	1,2459	1,2375
	Boosting	9,2682	9,4231	9,4046	9,4707	9,7012	9,5975	9,1508
	Total ECE:	**10,8442**	**11,0319**	**11,0303**	**11,0903**	**11,2985**	**11,2079**	**10,7397**
ECS	Conf.1	0,1865	0,2230	0,2194	0,2232	0,2225	0,2229	0,2211
	Conf.2	0,3181	0,2921	0,2953	0,2975	0,3003	0,3096	0,3965
SEnsembles (SE) - Tempo Total (em segundos)								
SE	ECS Conf.1 + Total ECE	11,0306	11,2549	11,2498	11,3135	11,5209	11,4307	10,9608
	ECS Conf.2 + Total ECE	11,2407	11,3239	11,3257	11,3878	11,5987	11,5175	11,1362
NNM (*Baseline* da comparação) - Tempo Total (em segundos)								
NNM	Regressão Logistica	0,1521	0,1845	0,1842	0,1828	0,1840	0,1902	0,1826

Fonte: Elaborado pelo autor.

Ressalta-se, contudo, mesmo que a abordagem proposta "*SEnsembles*" tenha gerado um tempo total de execução superior ao *baseline* da comparação (*NNM*), em torno de cinco segundos no primeiro experimento, considerando-se todos os *calipers* utilizados e, em torno de 11 segundos, no segundo experimento, o tempo total de execução não deverá ser levado em consideração pelo usuário especialista, uma vez que a abordagem proposta "*SEnsembles*" não gerou tempos elevados para se obter melhor qualidade na correspondência de instância.

6.9 Considerações Finais

O presente capítulo apresentou diversas comparações das correspondências geradas pelos processos ECS e ECE da abordagem "*SEnsembles",* bem como a comparação desses processos com o método *NNM*. Observou-se nos experimentos apresentados que a abordagem proposta "*SEnsembles"* produz correspondência de instâncias mais similares com diversos *calipers.* Porém, ressalta-se que a prática de executar várias vezes um método de correspondência de instâncias, também se aplica à abordagem "*SEnsembles*", uma vez que se permite alterar alguns parâmetros para a sua execução: Configuração 1 ou 2 da ECS, *calipers*, métrica de avaliação das correspondências e os *ensembles* a serem utilizados. Além disso, observou-se que o desempenho de execução da abordagem proposta "*SEnsembles*", com base nos experimentos realizados, não é um fator que impossibilitará o seu uso pelo usuário especialista, uma vez que o tempo de execução não é elevado para se obter melhor qualidade na correspondência de instância. Por fim, o próximo capítulo apresenta às conclusões da presente pesquisa.

Capítulo 7

CONCLUSÃO

7.1 Comprovação das Hipóteses

Esta seção apresenta uma discussão sobre a comprovação das hipóteses da pesquisa, as quais são descritas novamente abaixo apenas para facilitar a discussão a seguir.

Hipótese 1: “Uma abordagem para correspondência de instâncias, no contexto do processo de PSM, que considere *ensembles* de regressores, mais especificamente, *bagging, boosting e random forest*, pode permitir correspondências de instâncias mais similares, a partir de conjuntos de dados com características de disjunção de instâncias”.

Hipótese 2: “Uma estratégia para correspondência de instâncias que trate separadamente as características idênticas das instâncias pode permitir correspondências mais similares de instâncias e melhorar a qualidade final do processo de correspondência”.

É importante ressaltar que as hipóteses acima estão relacionadas com os dois processos da abordagem “*SEnsembles*” responsáveis em gerar a correspondência das instâncias. Assim, a Hipótese 1 está relacionado ao processo ECE e, a Hipótese 2, com o processo ECS, e suas comprovações estão discutidas a seguir.

A Hipótese 1 foi comprovada quando utilizou-se o *caliper* zero (menor valor) e os valores mais altos, 0,20, 0,25 e 0,30, para efetuar a correspondência das instâncias com o processo ECE da abordagem proposta "*SEnsembles*", com ganhos de até 36,3%, quando comparado ao método *NNM* (*baseline* da comparação), conforme se observa na Tabela 7.1, a qual apresenta os conjuntos de dados utilizados nos experimentos, número do experimento, *calipers* com os quais houve melhoria e os ganhos obtidos. Nota-se que com os *calipers* 0,05, 0,10 e 0,15 a substituição da regressão logística pelos *ensembles* não permitiu correspondências com instâncias mais similares.

Tabela 7.1 – Ganhos obtidos na melhoria da qualidade da correspondência pelo processo ECE da abordagem proposta "*SEnsembles"* em relação ao método *NNM* (*baseline* da comparação).

Conjunto de Dados		**Experimento**	***Caliper***	**Ganhos**
Lalonde (1986)		1	Zero, 0,20, 0,25 e 0,30	De 1,8% a 36,5%.
Lee et al. (2010)	Cenário B: Leve (*mild*) não linearidade.	2	0,30	3,0%.
	Cenário C: Moderada não linearidade	3	Zero, 0,25 e 0,30	De 7,5% a 24,8%
	Cenário D: Leve (*mild)* não aditividade.	4	Zero, 0,25 e 0,30	De 0,4% a 8%
	Cenário E: Leve (*mild)* não aditividade e não linearidade.	5	0,30	6,4%.
	Cenário F: Moderada não aditividade.	6	Zero, 0,20, 0,25 e 0,30	De 0,5% a 21,2%
	Cenário G: Moderada não aditividade e não linearidade	7	Zero, 0,25 e 0,30	De 6,8% a 25%
PBF 3 – 14 Covariáveis Modificada a partir de Martins (2013)		8	Zero	50%.

Fonte: Elaborado pelo autor.

Em média, com o *caliper* zero obteve-se 13,2% de melhoria da qualidade da correspondência, 7,9% com o *caliper* 0,20, 12,1% com *caliper* 0,25 e, 17,8% com o *caliper* 0,30, o que resulta em uma melhoria média da qualidade da correspondência de aproximadamente 12,7%. Em contrapartida, ao se analisar o número de pares de instâncias correspondidas, observa-se que com o *caliper* zero não houve redução desse número, mas com o *caliper* 0,20 resultou em redução média de 10,2%, 11,1% com o *caliper* 0,25 e, por último, 8,9% com o *caliper* 0,30, o que reduziu em média o número de pares de instâncias correspondidas em aproximadamente 7,6%, quando comparado ao método *NNM* (*baseline* da comparação).

Em se tratando da Hipótese 2, pode-se afirmar que foi comprovada para a maioria dos *caliper* utilizados para efetuar a correspondência com o processo ECS da abordagem proposta "*SEnsembles*", com ganhos de até 53,8%, quando comparado ao método *NNM* (*baseline* da comparação), conforme se observa na Tabela 7.2, a qual apresenta os conjuntos de dados utilizados nos experimentos, número do experimento, quantidade de covariáveis (atributos) utilizadas para efetuar a correspondência, *calipers* com os quais houve melhoria da qualidade e os ganhos obtidos.

Tabela 7.2 – Ganhos obtidos na melhoria da qualidade da correspondência pelo processo ECS da abordagem proposta "*SEnsembles*" em relação ao método *NNM* (*baseline* da comparação).

Conjunto de Dados	Atr.*	Exp.**	*Caliper*	Ganhos
Lalonde (1986)	10	1	Exceto 0,10	De 1,4% a 51,4%.
PBF 1 (Martins, 2013)	4	2	Zero, 0,05, 0,10 e 0,25	De 0,5% a 38,6%.
	14	3	Exceto zero	De 0,7% a 53,8%.
PBF 2 Modificado a partir de Martins (2013)	14	4	Exceto zero	De 3,9% a 48,4%.
PBF 3 Modificado a partir de Martins (2013)	14	5	Todos	De 2,2% a 50,6%.
Cenário F (Lee et al., 2010)	10	6	Exceto 0,05 e 0,15	De 0,1% a 0,3%.

***Atributo; **Experimento.**

Fonte: Elaborado pelo autor.

Em média, com o *caliper* zero obteve-se 1,6% de melhoria da qualidade da correspondência, 32,8% com o *caliper* 0,05, 7,6% com o *caliper* 0,10, 8,1% com o *caliper* 0,15, 12,4% com o *caliper* 0,20, 9,4% com o *caliper* 0,25 e, por último, 12,8% com o *caliper* 0,30, o que resulta em uma melhoria média da qualidade da correspondência de aproximadamente 12,1%. Em contrapartida, ao se analisar o número de pares de instâncias correspondidas, houve pequena redução (de 0,1%) com o *caliper* zero em apenas um experimento, mas com o *caliper* 0,05 resultou em uma redução média de 2,9%, de 1% com o *caliper* 0,10, de 4,9% com o *caliper* 0,15, de 3,2% com o *caliper* 0,20, de 3% com o *caliper* 0,25 e, por último, de 4% com o *caliper* 0,30, o que reduziu em média o número de pares de instâncias correspondidas em aproximadamente 2,7%, quando comparado ao método *NNM* (*baseline* da comparação).

Como visto, as hipóteses levantadas para a pesquisa, foram comprovadas, no caso da Hipótese 1, quando utilizado o *caliper* zero e os valores mais altos, de 0,20 a 0,30, e no caso da Hipótese 2, na maioria dos *calipers* utilizados (zero a 0,30).

7.2 Análise dos objetivos

Esta pesquisa de doutorado atingiu seu objetivo geral proposto, uma vez que possibilitou investigar alternativas computacionais para melhorar a qualidade das correspondências de instâncias em conjuntos de dados que são manipulados em estudos observacionais e, a partir dessa investigação, possibilitou o desenvolvimento de uma nova abordagem para correspondências de instâncias, utilizando-se de técnicas que considerem as características idênticas e *ensembles* de regressores em substituição à regressão logística.

A Abordagem "*SEnsembles*" permitiu obter grupos de tratamento e de controle mais balanceados, com correspondências de instâncias mais similares para a maioria dos *calipers* utilizados nos experimentos com base na medida ASAM.

Para exemplificar a melhoria da qualidade da correspondência obtida, na Tabela 7.3 apresenta-se um resumo das métricas ASAM e número de pares de instâncias obtidas para três conjuntos de dados, quando comparadas às obtidas

pelo método *NNM,* ao se efetuar a correspondência das instâncias com os *calipers* variando de zero a 0,30. Observa-se que a abordagem "*SEnsembles*" proporcionou as melhores correspondências (em verde) para a maioria dos *caliper* ao efetuar a correspondência das instâncias dos conjuntos de dados Lalonde (1986), PBF 3 e Cenário F (Lee et al., 2010). Além disso, nota-se que somente em três *calipers* (em vermelho) não houve uma melhora da correspondência das instâncias.

Tabela 7.3 – Melhoria da qualidade das correspondências de instâncias proporciona pela abordagem proposta "*SEnsembles*" se comparada ao método *NNM* , com a descrição dos conjuntos de dados, métricas ASAM e numero de pares de instâncias obtidas em cada *caliper.*

Conjunto de Dados	Métrica	*Calipers*							Média
		Zero	0,05	0,10	0,15	0,20	0,25	0,30	
Lalonde (1986)	ASAM	**5,6%**	**4,0%**	**-13,6%**	**15,1%**	**42,1%**	**48,9%**	**51,4%**	**27,9%**
	Pares	0,0%	-1,1%	-1,8%	-18,5%	-15,5%	-17,6%	-16,0%	-11,4%
PBF 3 Modificado a partir de Martins (2013)	ASAM	**50,0%**	**50,6%**	**2,5%**	**8,4%**	**11,7%**	**9,3%**	**8,4%**	**20,1%**
	Pares	0,0%	-3,8%	-0,3%	-0,5%	-0,6%	-0,8%	-0,9%	-1,0%
Cenário F: Moderada não aditividade. Lee et al. (2010)	ASAM	**4,3%**	**-0,1%**	**0%**	**-0,9%**	**5,1%**	**19,6%**	**21,2%**	**10,0%**
	Pares	0%	0%	0%	0,1%	-4,4%	-4,3%	-3,9%	-2,5%

Fonte: Elaborado pelo autor.

Já se tratando do número de pares de instâncias correspondidas, nota-se que as correspondências com o conjunto de dados Lalonde (1986), com *calipers* até 0,15, apresentaram pequena redução do número de pares. Porém, a partir do *caliper* 0,15, essa redução foi superior a 15%. Já as correspondências dos conjuntos de dados PBF 3 e Cenário F (Lee et al., 2010), apresentaram pequena redução do número de pares de instância correspondidas, com média de -1,0% e -1,8%, respectivamente. Ou seja, a abordagem "*SEnsembles*" proporcionou as melhores correspondências de instâncias para a maioria dos *calipers* utilizados e, o ganho obtido na qualidade das correspondências compensa a pequena perda do número

de pares de instância correspondidas. Porém, não se pode descartar a importância do usuário especialista para avaliar as correspondências obtidas.

Já em relação aos ganhos de melhoria proporcionados pela abordagem "*SEnsembles*", obteve-se de 5,6% a 51,4% com o conjunto de dados Lalonde, de 2,5% a 50% com o conjunto de dados PBF 3 e, de 4,3% a 21,2%, com o conjunto de dados do Cenário F (Lee et al., 2010), conforme se observa na Tabela 7.4. Em média, considerando-se os *calipers* em que o houve melhoria do valor da métrica ASAM nesses conjuntos, obteve-se 27,9%, 20,1% e 10,0%, respectivamente.

Tabela 7.4 – Melhoria da qualidade das correspondências de instâncias proporciona pela abordagem proposta "*SEnsembles*" se comparada ao método *NNM*, com a descrição dos conjuntos de dados, médias das métricas utilizadas o intervalo de ganho em relação do valor da métrica ASAM.

Conjunto de Dados	Médias	Ganhos
Lalonde (1986)	**ASAM:** 27,9% de melhoria. **Pares:** -11,4%. **Descartes:** 18,6%	De 5,6% a 51,4%.
PBF 3 – 14 Covariáveis Modificada a partir de Martins (2013)	**ASAM:** 20,1% de melhoria. **Pares:** -1,0%. **Descartes:** 11,4%	De 2,5% a 50%.
Cenário F: Moderada não aditividade Lee et al. (2010).	**ASAM:** 10,0% de melhoria. **Pares:** -2,5% **Descartes:** 3,8%	De 4,3% a 21,2%

Fonte: Elaborado pelo autor.

Ressalta-se também que todos os objetivos específicos foram atingidos e, foram reunidos aqui em três tópicos para facilitar a discussão, conforme a seguir:

- Os *ensembles* de regressores, mais precisamente, *bagging*, *boosting* e *random forest* foram investigados, o que permitiu a elaboração da estratégia do processo ECE, a qual os utiliza para substituir a regressão logística ao estimar os escores de propensão das instâncias no contexto do processo PSM;
- As características idênticas das instâncias também foram investigadas e isso propiciou a elaboração da estratégia do processo ECS, a qual permite ao usuário especialista definir qual configuração a ser aplicada para separar as instâncias com características idênticas, seja por meio de

um pareamento 1:1 (Configuração 1) ou M:N (Configuração 2). Com isso, investigou-se também se tais configurações poderiam resultar em melhores correspondências de instâncias e em quais situações;

- Em seguida, as estratégias do processo ECS e ECE foram codificadas e validadas por meio dos experimentos já demonstrados neste trabalho.

Por fim, ressalta-se que os próximos passos estarão direcionados para a criação de um pacote na linguagem *R* (R FOUNDATION, 2016) para distribuição da abordagem "*SEnsembles*". Além disso, pretende-se elaborar uma página para divulgação contendo informações sobre essa abordagem, pacote de distribuição e exemplos de aplicações.

7.3 Contribuições

A principal contribuição da presente pesquisa é a definição de uma nova abordagem para correspondência de instâncias, denominada "*SEnsembles*", a qual apresenta alternativas computacionais para melhorar a qualidade das correspondências em conjuntos de dados que são manipulados em estudos observacionais. Para isso, essa abordagem possui dois processos para gerar as correspondências, os quais utilizam estratégias que consideram as características idênticas das instâncias e *ensembles* de regressores para substituição à regressão logística.

O processo ECS possui uma estratégia que se baseia no tratamento em separado das instâncias com características idênticas. Assim, essa estratégia busca separar as instâncias idênticas das demais (distintas) para efetuar as correspondências somente das instâncias distintas. Ao final, junta-se as instâncias com características idênticas com as distintas já pareadas. Com isso, essa estratégia apresentou ganhos em relação à qualidade final das correspondências em situações em que o conjunto de dados possui aproximadamente 4% de instâncias com características idênticas (de acordo com os resultados experimentais descritos no Capítulo 6). E, além disso, em conjuntos de dados que não há instâncias com

tais características, essa estratégia apresentou resultados similares, com o mesmo número de pares de instâncias correspondidas e nenhum descarte de indivíduos do grupo de tratamento em relação ao método *NNM,* utilizando-se a regressão logística para estimar os escores de propensão, com os valores da métrica ASAM muito próximos

Já o processo ECE possui uma estratégia que utiliza *ensembles* de regressores, mais precisamente, *bagging, random forest e boosting*, para substituir a regressão logística ao estimar os escores de propensão. Inicialmente, os escores de propensão são estimados pelos *ensembles* e essas estimativas são utilizadas para gerar a correspondência das instâncias pelo método *NNM.* Ao final, avalia-se qual das estimativas permitiu gerar as melhores correspondências. Essa estratégia apresentou ganhos em relação à qualidade final da correspondência em situações em que se utilizou o *caliper* zero e os valores acima de 0,20 (até 0,30).

Como os processos ECS e ECE são executados de forma concomitante, avalia-se também, ao final, qual desses processos apresentou a melhor correspondência de instâncias, a qual é definida como sendo a correspondência final da abordagem "*SEnsembles*".

Para permitir a geração de correspondências de instâncias de maneira mais flexível, a abordagem "*SEnsembles*" foi projetada também para permitir configurações de alguns parâmetros pelo usuário especialista, tais como:

- Definição da configuração do processo ECS pela qual as instâncias com características idênticas serão obtidas, seja por meio de um pareamento 1:1 sem substituição, no caso da Configuração 1, ou M:N, no caso da Configuração 2;
- Definição dos valores dos *calipers* a serem utilizado para gerar correspondências de instâncias pelos processos ECS e ECE da abordagem "*SEnsembles*";
- Definição da métrica que será utilizada na avaliação das correspondências pelos processos ECE e ECS, permitindo assim, que se priorize a melhor qualidade das correspondências com o uso da métrica ASAM ou o maior número de pares de instâncias correspondidas;

- Definição dos ensembles de regressores que serão utilizados para estimar os escores de propensão no processo ECE. A definição padrão adotada é o uso dos três *ensembles* (*bagging*, *random forest* e *boosting*).

É importante ressaltar que a abordagem "*SEnsembles*" está preparada para avaliar as correspondências com base na métrica ASAM. Porém, conforme mencionado, permite-se ao usuário especialista intervir nessa métrica, alterando-a para o número de pares de instâncias correspondidas. Entretanto, quando isso ocorre, com exceção do uso do *caliper* zero, há um aumento do valor da métrica ASAM e, consequentemente, são geradas correspondências com instâncias menos similares. Mas, sobretudo, prevalece a opção do usuário especialista na busca por um maior número de pares de instâncias correspondidas.

Outra contribuição trata-se da utilização de um conjunto de dados real, o qual contém informações do PBF e que foi utilizado em outra pesquisa de doutorado (Martins, 2013). Neste sentido, ressalta-se que muitos trabalhos da literatura que utilizam *ensemble* no âmbito dos estudos observacionais, com destaque para os trabalhos de Setoguchi et al. (2008) e Lee et al. (2010), somente utilizaram conjuntos de dados simulados, ou seja, construídos de forma sintética e, ainda, em abordagens que ponderam as instâncias pelo escore de propensão e, não as correspondem, como acontece na presente abordagem que está inserida no contexto do processo PSM. Assim, destaca-se a capacidade da abordagem "*SEnsembles*" em gerar as melhores correspondências em conjuntos de dados reais, como o PBF, que contém 55.970 instâncias, sendo 9.259 pertencentes ao grupo de beneficiários (tratamento) e 46.711 ao grupo de não beneficiários (controle).

Destaca-se também como contribuição, a investigação que foi conduzida para comparar as correspondências geradas, tanto pelos processos da abordagem "*SEnsembles*" quanto pelo método *NNM*, quando aplicado *calipers* para impor um limite para seleção das instâncias do grupo de controle. Essa investigação permitiu verificar as melhores correspondências geradas em cada uma dos *calipers* utilizados (zero, 0,05, 0,10, 0,15, 0,20, 0,25 e 0,30).

Por fim, é importante ressaltar que a abordagem "*SEnsembles*" é inovadora pois alia estratégias para geração de correspondências que não encontradas na literatura. Essas estratégias são relacionadas aos processos pelos quais a

correspondência das instâncias é efetuada, seja considerando as características idênticas das instâncias (processo ECS) ou com o uso de *ensembles* de regressores para estimar os escores de propensão (processo ECE).

7.4 Trabalhos Futuros

Os trabalhos futuros descritos na presente seção estão relacionados com a continuidade da pesquisa e, consequentemente, com a evolução da abordagem "*SEnsembles*". Nesta perspectiva, sugerem-se os seguintes direcionamentos:

- Adicionar à abordagem proposta "*SEnsembles*" outros *ensembles* e/ou métodos que permitam estimar os escores de propensão das instâncias. Com isso, objetiva-se aumentar a precisão dessa estimativa e gerar melhores correspondências de instâncias;
- Adicionar à abordagem proposta "*SEnsembles*" a capacidade de efetuar correspondência M:N de indivíduos e com substituição;
- Investigar se uma integração dos processos ECS e ECE seria viável, permitindo gerar correspondências mais similares. Por exemplo, as instâncias distintas do processo ECS poderiam ter seus escores de propensão estimados pelo processo ECE antes de serem pareadas;
- Investigar como a linearidade das covariáveis poderia auxiliar os processos ECS e ECE da abordagem proposta "*SEnsembles*";
- Codificar o pacote conforme o padrão da linguagem R (R FOUNDATION, 2016) para permitir a distribuição da abordagem proposta "*SEnsembles*";
- Criar uma página *web* sobre o projeto, contendo o pacote de distribuição e exemplos de aplicações.

Por fim, as sugestões acima descrevem algumas direções para permitir a evolução da abordagem proposta "*SEnsembles*" em curto prazo.

REFERÊNCIAS

AL-GHANIM, M.; NOAH, S. A.; SEMBOK, T. M. Automating XML schema matching: a composite approach. In: INTERNATIONAL CONFERENCE ON ELETRICAL ENGINEERING AND INFORMATICS. **Proceedings**...Bandung, Indonesia: 2011, p. 1-6.

ANDERSON, D. R; SWEENEY, D. J.; WILLIAMS, T. A. Estatística aplicada à administração. São Paulo: Thomson, 2003.

ARCENEAUX, K. et al. Comparing experimental and matching methods using a large-scale voter mobilization experiment. **Political Analysis**, v. 14, n. 1, p. 37–62, 2006.

AUSTIN, P. C. A comparison of 12 algorithms for matching on the propensity score. **Statistics in Medicine**, v. 33, n. 6, p. 1057–1069, 2014.

AUSTIN, P. C. An introduction to propensity score methods for reducing the effects of confounding in observational studies. **Multivariate Behavioral Research**, v. 46, n. 3, p. 399–424, 2011.

AUSTIN, P. C. et al. Regression trees for predicting mortality in patients with cardiovascular disease: what improvement is achieved by using ensemble-based methods? **Biometrical Journal**, v. 54, n. 5, p. 657–73, 2012.

AUSTIN, P. C.; SMALL, D. S. The use of bootstrapping when using propensity-score matching without replacement: a simulation study. **Statistics in Medicine**, v. 33, n. 24, p. 4306– 4319, 2014.

BERNSTEIN, P. A. et al. Generic schema matching, ten years later. **PVLDB**, v. 4, n. 11, p. 695–701, 2011.

BIZAGI. Bizagi PMN Modeler. Disponível em: <http://www.bizagi.com/pt/produtos/bpm-suite/ modeler>. Acesso em: 09 out. 2016.

BREIMAN, L. Bagging predictors. **Machine Learning**, v. 24, n. 2, p. 123–140, 1996.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.

CAMPELLO, M.; GRAHAM, J. R.; HARVEY, C. R. The real effects of financial constraints: Evidence from a financial crisis. **Journal of Financial Economics**, v. 97, n. 3, p. 470–487, 2010.

CHRISTEN, P. Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. Springer, 2012.

CIFERRI, C. D. A. **Distribuição dos dados em ambientes de data warehousing: o Sistema WebD^{2}W e algoritmos voltados à fragmentação horizontal dos dados**. 2002. Tese (Doutorado em Ciência da Computação) – Centro de Informática, Universidade Federal de Pernambuco, Pernambuco, 2002.

COCHRAN, W. G. The effectiveness of adjustment by subclassification in removing bias in observational studies. **Biometrics**, v. 24, n. 2, p. 295–313, 1968.

COCHRAN, W. G. The planning of observational studies of human populations. v. 128, n. 2, p. 234–266, 1965.

COCHRAN, W. G.; RUBIN, D. R. Controlling bias in observational studies : a review. **Sankhyā: The Indian Journal of Statistics**, v. 35, n. 4, p. 417– 446, 1973.

D'AGOSTINO, R. B. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. **Statistics in Medicine**, v. 17, p. 2265–2281, 1998.

DEHEJIA, R. H.; WAHBA, S. Causal effects in nonexperimental studies : Reevaluating the evaluation of training programs. **Journal of the American Statistical**, v. 94, n. 448, p. 1053–1062, 1999.

DIETTERICH, T. G. Ensemble methods in machine learning. **Multiple Classifier Systems Lecture Notes in Computer Science**, v. 1857, p. 1–15, 2000.

DOAN, A.; HALEVY, A. Y. Semantic integration research in the database community: a brief survey. **AI Magazine**, v. 26, n. 1, p. 83–94, 2005.

DOAN, A; HALEVY, A. Y.; IVES, Z. G. Principles of data integration. Morgan Kaufamann, 2012.

DORNELES, C. F.; GONÇALVES, R.; MELLO, R. S. Approximate data instance matching: a survey. **Knowledge and Information Systems**, v. 27, n. 1, p. 1–21, 2010.

EFRON, B.; TIBSHIRANI, R. J. An introduction to the bootstrap. New York: Chapman & Hall, 1993.

ELLIS, A. R. et al. Confounding control in a non-experimental study of STAR*D data: Logistic regression balanced covariates better than boosted CART. **Ann Epidemiol**, v. 23, n. 4, p. 204–209, 2013.

FREUND, Y. Boosting a week learning algorithm by majority. **Information and Computation**, v. 121, n. 2, p. 256–285, 1995.

FREUND, Y.; SCHAPIRE, R. E. A decision-theoretic generalization of on-line learning and application to boosting. **Journal of Computer and System Sciences**, v. 55, n. 1, p. 119–139, 1997.

FREUND, Y.; SCHAPIRE, R. E. A Decision-theoretic generalization of on-line learning and an application to boosting. In: SECOND EUROPEAN CONFERENCE ON COMPUTATION LEARNING THEORY. **Proceedings**... London: 1995. p. 23-37.

FREUND, Y.; SCHAPIRE, R. E. Experiments with a new boosting algorithm. In: THIRTEENTH INTERNATIONAL CONFERENCE ON MACHINE LEARNING. **Proceedings**... Bari: 1996. p. 148-156.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. **The Annal of Statistcs.** V. 29, n. 5, 1189-1232, 2001.

GANGL, M. Scar Effects of Unemployment : An assessment of institutional complementarities. **American Sociological Review**, v. 71, n. 6, p. 986–1013, 2006.

GONG, J. et al. Efficient management of uncertainty in XML schema matching. **The VLDB Journal**, v. 21, n. 3, p. 385–409, 2012.

GRODSKY, E. Compensatory sponsorship in higher education. **American Journal of Sociology**, v. 112, n. 6, p. 1662–1712, 2007.

GU, X. S.; ROSENBAUM, P. R. Comparison of multivariate matching methods: structures, distances, and algorithms. **Journal of Computation and Graphics Statistics**, v. 2, n. 4, p. 405–420, 1993.

HO, D. E. et al. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. **Political Analysis**, v. 15, n. 3, p. 199–236, 2007.

HO, D. E. et al. MatchIt : nonparametric preprocessing for parametric causal inference. **Journal of Statistical Software**, v. 42, n. 8, p. 1–28, 2011.

HONG, G.; YU, B. Effects of kindergarten retention on children's social-emotional development: an application of propensity score method to multivariate, multilevel data. **Developmental Psychology**, v. 44, n. 2, p. 407–21, 2008.

IBRAHIM, H. et al. An Automatic domain independent schema matching in integrating schemas of heterogeneous relational databases. **Journal of Information Science and Engineering**, v. 30, n. 4, p. 1505–1536, 2014.

IMBENS, G. Nonparametric estimation of average treatment effects under exogeneity: a review. **The Review of Economics and Statistics**, v. 86, n. 1, p. 4–29, 2004.

INEP. Observatório da Educação. Disponível em: <http://observatorio.inep.gov.br/o-que-e>. Acesso em: 05 nov. 2016a.

INEP. Uma Análise da Evolução e dos Determinantes do Desempenho Escolar no Brasil. Disponível em: <http://observatorio.inep.gov.br/visualizar>. Acesso em: 05 nov. 2016b.

KARASNEH, Y. et al. A Model for matching and integrating heterogeneous relational biomedical databases schemas. In: IDEAS 2009. **Proceeding**... Cetraro, Italy: 2009.

KOU, Y. Improving the accuracy of entity identification through refinement. In: 2008 EDBT Ph.D. WORKSHOP (Ph.D. '08). **Proceedings…**New York: ACM, 2008. p. 39-48.

KUNCHEVA, L. I. Combining pattern classifiers: methods and algorithms. New Jersey: John Wiley & Sons, 2004. 350 p.

LAKATOS, E. M.; MARCONI, M. A. **Metodologia Científica**. São Paulo: Atlas, 2008.

LALONDE, R. J. Evaluation the econometric evaluations of training programs with experimental. **The American Economic Review**, v. 76, n. 4, p. 604–620, 1986.

LEE, B. K.; LESSLER, J.; STUART, E. A. Improving propensity score weighting using machine learning. **Statistics in Medicine**, v. 29, n. 3, p. 337–346, 2010.

LI, M. Using the propensity score method to estimate causal effects: a review and practical guide. **Organizational Research Methods**, v. 16, n. 2, p. 188–226, 13 jun. 2012.

LIAW, A.; WIENER, M. Classification and Regression by randomforest. **R News**, v. 2, n. 3, p. 18-22, 2002.

LITTNEROVA, S. et al. Why to use propensity score in observational studies? Case study based on data from the Czech clinical database AHEAD 2006–09. **Cor et Vasa**, v. 55, n. 4, p. 383–390, 2013.

LUNCEFORD, J. K.; DAVIDIAN, M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. **Statistics in Medicine**, v. 23, n. 19, p. 2937–60, 15 out. 2004.

MARTINS, A. P. B. **Impacto do Programa Bolsa Família sobre a aquisição de alimentos em famílias brasileiras de baixa renda**. 2014. Tese (Doutorado em Ciências) – Faculdade de Saúde Pública, Universidade de São Paulo, São Paulo 2013.

MCCAFFREY, D. F. et al. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. **Statistics in Medicine**, v. 32, n. 19, p. 3388–3414, 2013.

MCCAFFREY, D. F. et al. Propensity score estimation with boosted regression for evaluating causal effects in observational studies (Boosted). **Psychological Methods**, v. 9, n. 4, p. 403–425, 2004.

MDS, CEDEPLAR. **Avaliação de impacto do Programa Bolsa Família**. Disponível em:<http://aplicacoes.mds.gov.br/sagi/PainelPEI/Publicacoes/Avalia%C3%A7%C3%A3o%20de%20Impacto%20do%20Programa%20Bolsa%20Fam%C3%ADlia.pdf>. Acesso em: 22 nov. 2014.

MDS. **Avaliação de impacto do Programa Bolsa Família – 2ª Rodada (AIBF II)**. Disponível em: <http://www.mds.gov.br/biblioteca/secretaria-de-avaliacao-e-gestao-de-informacao-sagi/cadernos-de-estudos/avaliacao-de-impacto-do-programa-bolsa-familia/avaliacao-de-impacto-do-programa-bolsa-familia>. Acesso em: 22 nov. 2014.

OMG. Documents Associated With Business Process Model And Notation™ (BPMN™) Version 2.0. Disponível em: <http://www.omg.org/spec/BPMN/2.0/>. Acesso em: 09 out. 2016.

PARENT, C. SPACCAPIETRA, S. ERC+: an object based entity relationship approach. In: LOUCOPOULOS, P.; ZICARI, R. **Conceptual modelling, databases and Case:** an Integrated view of information systems development. New York: John Wiley & Sons, 1992.

PFEFFERMANN, D.; LANDSMAN, V. Are private schools better than public schools? Appraisal for Ireland by methods for observational studies. **Ann Appl Stat.**, v. 5, n. 3, p. 1726–1751, 2011.

R Foundation for Statistical Computing. The R project for statistical computing. Disponível em: <http://www.r-project.org/>. Acesso em: 09 out. 2016.

RAHM, E. Towards large-scale schema and ontology matching. In: BELLAHSENE, Z.; BONIFATI, A.; RAHM, E. (Eds.). **Schema Matching and Mapping**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. p. 3–27.

RAHM, E.; BERNSTEIN, P. A. A survey of approaches to automatic schema matching. **The VLDB Journal**, v. 10, n. 4, p. 334–350, 2001.

ROSENBAUM, P. R. A Characterization of optimal designs for observation studies. **Journal of the Royal Statistical Society**, v. 53, n. 3, p. 597–610, 1991.

ROSENBAUM, P. R. Optimal matching for observational studies. **Journal of the American Statistical Association**, v. 84, n. 408, p. 1024–1032, 1989.

ROSENBAUM, P. R.; Design of observational studies. New York: Springer Science+Business Media, 2010.

ROSENBAUM, P. R.; RUBIN, D. B. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. **The American Statistician**, v. 39, n. 1, p. 33–38, 1985.

ROSENBAUM, P. R.; RUBIN, D. B. The central role of the propensity score in observational studies for causal effects. **Biometrika**, v. 70, n 1, p. 41-55, 1983.

RUBIN, D. B. Matching to remove bias in observational studies. **Biometrics**, v. 29, n. 1, p. 159–183, 1973.

SAGI, T.; GAL, A. Schema matching prediction with applications to data source discovery and dynamic ensembling. **The VLDB Journal**, v. 22, n. 5, p. 689–710, 2013.

SCHAPIRE, R. E. The strength of weak learnability. **Machine Learning**, v. 5, n. 2, p. 197–227, 1990.

SETOGUCHI, S. et al. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. **Pharmacoepidemiol Drug Safety**, v. 17, n. 6, p. 546–555, 2008.

SHADISH, W. R.; STEINER, P. M. A Primer on propensity score analysis. **Newborn and Infant Nursing Reviews**, v. 10, n. 1, p. 19–26, 2010.

SMITH, J. A.; TODD, P. E. Does matching overcome LaLonde's critique of nonexperimental estimators? **Journal of Econometrics**, v. 125, n. 1-2, p. 305–353, 2005.

SPACCAPIETRA, S.; PARENT, C.; DUPONT, Y. Model independent assertions for integration of heterogeneous schemas. **The VLDB Journal**, v. 1, n. 1, p. 81–126, 1992.

SPACCAPIETRA, S.; PARENT C. View integration: a step forward in solving structural conflicts. **IEEE Transaction on Knowledge and Data Engineering**, v. 6, n. 2, p. 258–274, 1994.

STAFF, J. et al. Teenage alcohol use and educational attainment. **Journal of Studies on Alcohol and Drugs**, v. 69, p. 848–858, 2008.

STEINER, P. M.; COOK, D. L. Matching and propensity scores. In: LITTLE, T.D. (Ed). **The Oxford handbook of quantitative methods**. **Volume 1**: foundations. Oxford: Oxford Library of Psychology, 2013. p. 237-259.

STRAUSS, D. The many faces of logistic regression. **The American Statistician**, v. 46, n. 4, p. 321–327, 1992.

STUART, E. A. Matching methods for causal inference: A review and a look forward. **Statistical Science**, v. 25, n. 1, p. 1–21, 2010.

TRIOLA, M. F. Introdução à estatística: atualização da tecnologia. LTC: Rio de Janeiro, 2013.

WATKINS, S. et al. An empirical comparison of tree-based methods for propensity score estimation. **Health Services Research**, v. 48, n. 5, p. 1798–1817, 2013.

WESTREICH, D.; LESSLER, J.; FUNK, M. J. Propensity score estimation: machine learning and classification methods as alternatives to logistic regression. **Journal of Clinical Epidemiology**, v. 63, n. 8, p. 826–833, 2010.

WOLFE, F.; MICHAUD, K. Heart failure in rheumatoid arthritis: rates, predictors, and the effect of anti-tumor necrosis factor therapy. **The American Journal of Medicine**, v. 116, n. 5, p. 305–11, 2004.

GLOSSÁRIO

Estudos Clínicos Randomizados

É um tipo de estudo realizado principalmente na medicina no qual se observa os efeitos de um tratamento ou droga, e os indivíduos em estudo são alocados de maneira aleatória em dois grupos distintos.

Indivíduos

Termo utilizado em estudos observacionais para representar as pessoas participantes, os quais são separados, na maioria dos estudos, em dois grupos distintos: grupo de tratamento e de controle.

Instância

Representa uma entidade no mundo real e, nesta tese, foi utilizada para denominar indivíduos participantes de estudos observacionais.

Disjunção de Instâncias

Disjunção de instâncias ocorre quando as instâncias são separadas em dois grupos distintos e cada instância somente é alocada em um dos grupos. E, ainda, instâncias com as mesmas características (instâncias idênticas) alocadas em um mesmo grupo e instâncias idênticas alocadas em grupos diferentes, em ambas as situações, não representam a mesma instância no mundo real.

Apêndice A

Este apêndice apresenta os processos ECS e ECE da abordagem proposta "*SEnsembles*" representados na notação para modelagem de processo *Business Process Model and Notation*[1], ou simplesmente, BPMN, a qual é mantida pelo *Object Management Group*[2] *e* provê uma gama de conceitos para modelagem de processos e fluxo de trabalho (*workflows*).

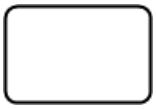
É importante lembrar que um processo descreve uma sequencia de atividades com um trabalho específico[1]. Uma atividade é um termo genérico utilizado para representar partes de um processo e pode ser classificada em subprocesso e tarefas. Um subprocesso é uma atividade que possui seu fluxo de execução detalhado, enquanto que uma tarefa é considerada atômica, ou seja, não possui seu fluxo detalhado. Por fim, o fluxo de execução pode ser divergido ou convergido por meio de *gateways*. Existem vários tipos de gateways, mas os utilizados nesta seção indicam uma execução em paralelo ou exclusivo (apenas um dos caminhos) do fluxo de execução.

A Tabela 1 ilustra os elementos básicos utilizados da notação BPMN para modelar os processos da abordagem proposta "*SEnsembles*". Sugere-se a leitura da especificação BPMN[1] para maiores detalhes.

[1] Object Management Group. Documents associated with Business Process Model And Notation™ (BPMN™) Version 2.0. Disponível em: <http://www.omg.org/spec/BPMN/2.0/>. Acesso em: 09 out. 2016.

[2] Object Management Group. Disponível em: <http://www.omg.org>. Acesso em: 09 out. 2016.

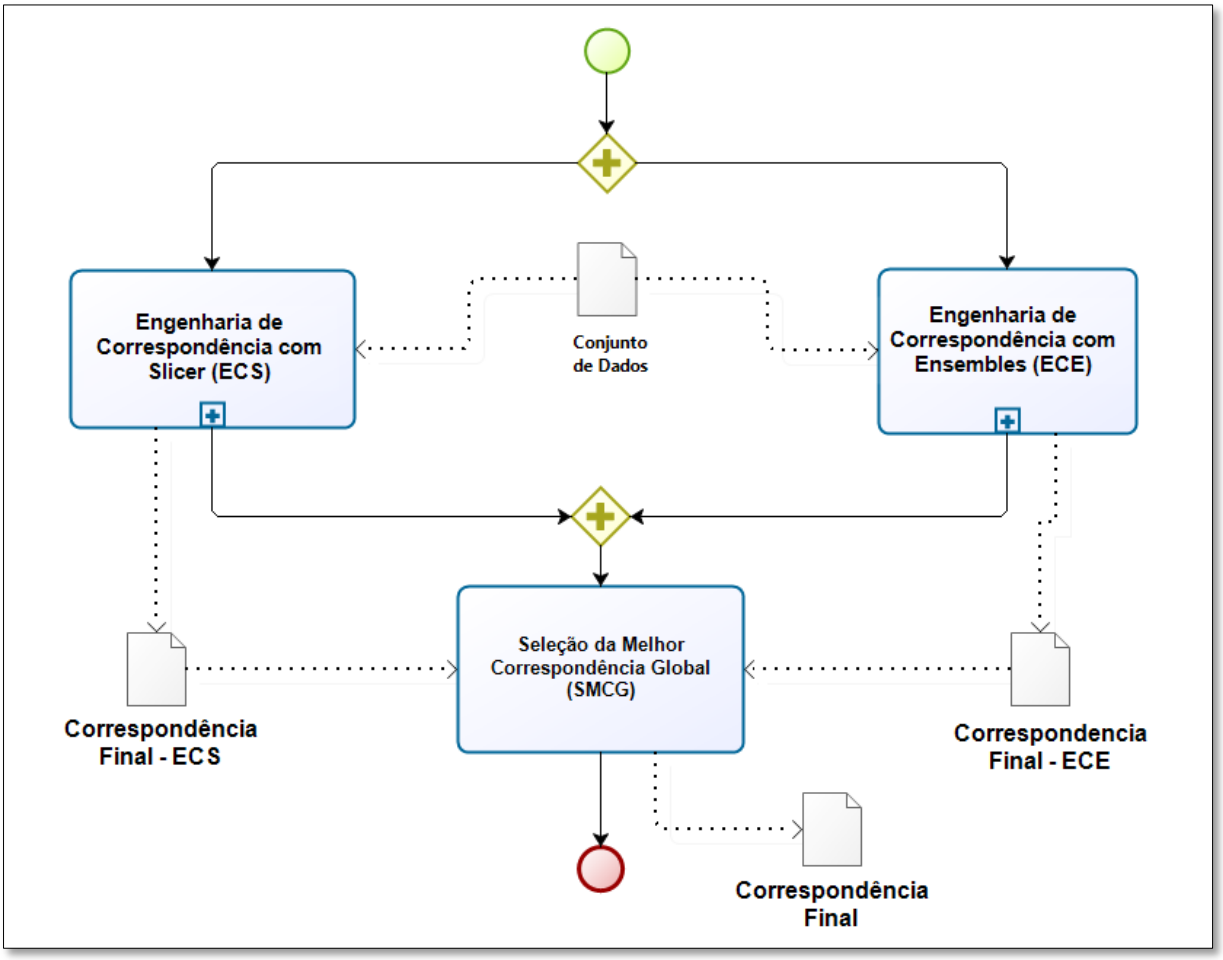
Tabela 1 – Elementos Básicos de Modelagem de Processos.

Elemento	Descrição	Representação Gráfica
Evento	Um evento é algo que acontece que afeta o curso de um processo. Os eventos são classificados em: inicial, intermediário e final.	
Atividade	Atividade é o termo genérico para representar partes de um processo e pode ser classificada em subprocesso ou tarefa.	
Subprocesso	Um subprocesso é uma atividade que possui seu fluxo de execução detalhado.	
Tarefa	Uma tarefa é uma atividade atômica, ou seja, não possui fluxo de execução detalhado.	
Gateway Paralelo	*Gateway* utilizado para indicar uma divergência ou convergência em paralelo do fluxo de execução.	
Gateway Exclusivo	*Gateway* utilizado para indicar uma divergência ou convergência exclusiva, ou seja, apenas um dos caminhos do fluxo de execução é executado.	
Objeto de Dados	Objetos de dados fornecem informações para as atividades ou armazenam o que elas produzem.	
Fluxo de Sequencia	Um fluxo de sequencia é usado para representar a ordem em que as atividades são executadas em um processo.	
Associação	Associação é usada para vincular informações com elementos gráficos.	

Fonte: Adaptado de OMG (2016).

A Figura 1 ilustra a modelagem na notação BPMN da visão geral da abordagem proposta “*SEnsembles*”, a qual é equivalente à Figura 5.1 da tese.

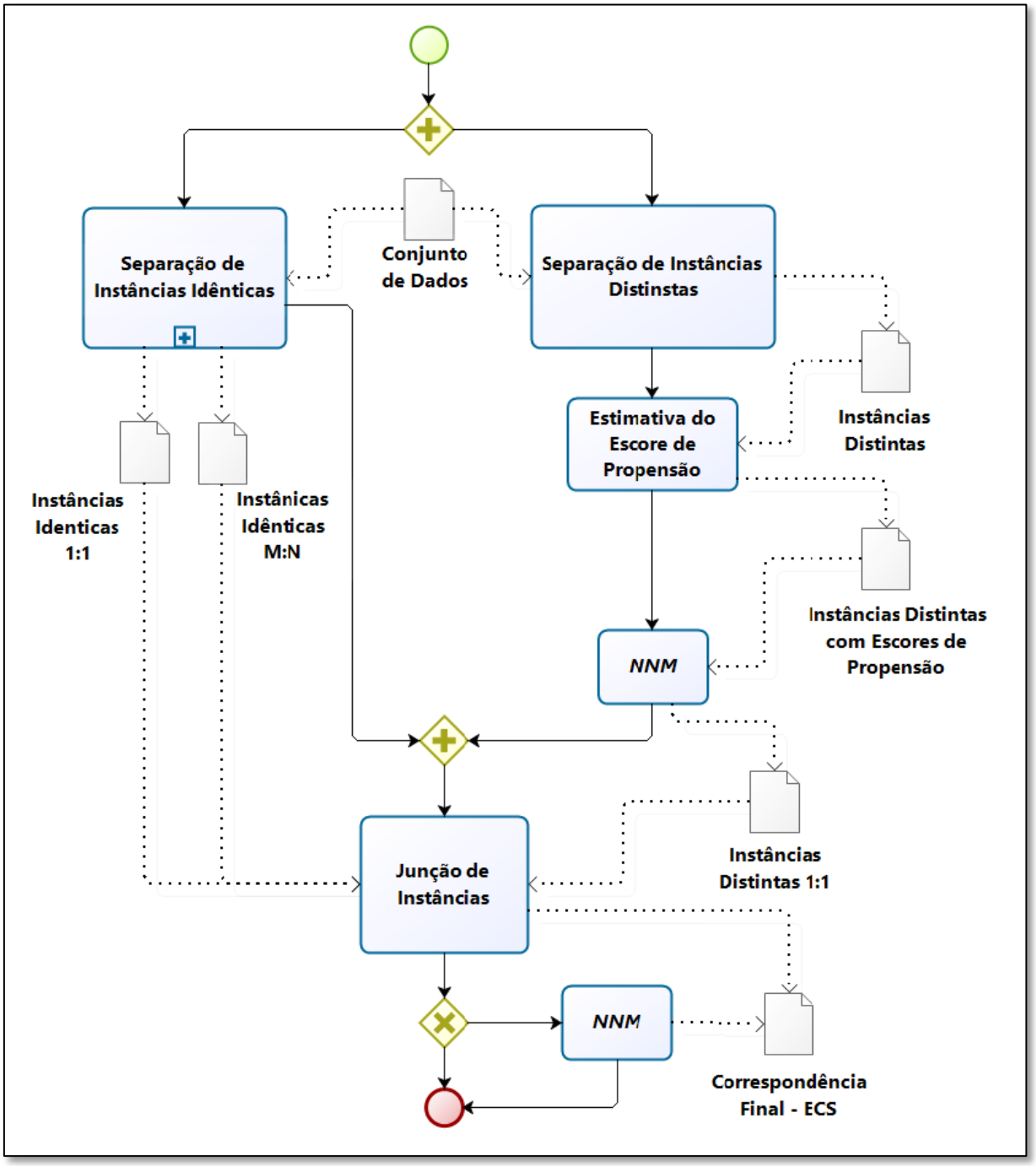
Figura 1 – Visão Geral da abordagem proposta "*SEnsembles*" equivalente à Figura 5.1.



Fonte: Elaborado pelo autor.

A Figura 2 ilustra a modelagem na notação BPMP do processo ECS da abordagem proposta "*SEnsembles*", a qual é equivalente à Figura 5.5 da tese.

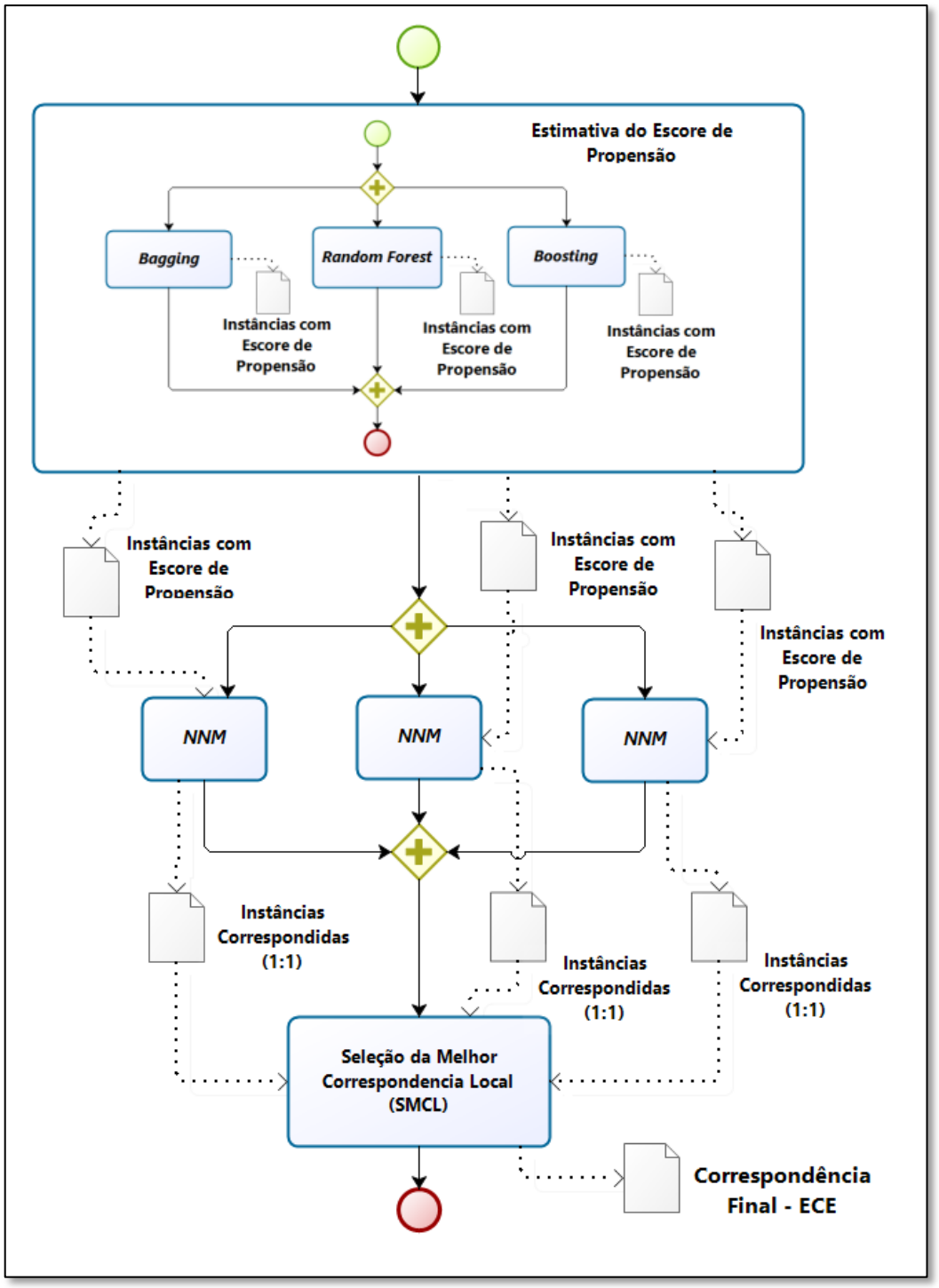
Figura 2 – Engenharia de Correspondência com *Slicer* (ECS) da abordagem proposta "*SEnsembles*" equivalente à Figura 5.5 da tese.



Fonte: Elaborado pelo autor.

A Figura 3 ilustra a modelagem na notação BPMP do processo ECE da abordagem proposta "*SEnsembles*", a qual é equivalente à Figura 5.6 da tese.

Figura 3 – Engenharia de Correspondência com *Ensembles* (ECE) da abordagem proposta "*SEnsembles*" equivalente à Figura 5.6 da tese.



Fonte: Elaborado pelo autor.

Como visto, a notação BPMN possibilitou representar os processos da abordagem proposta "*SEnsembles*" com maior formalismo.

Anexo A

GERAÇÃO DE DADOS SINTÉTICOS

```
#### Início ####


#------------------ Global Variables -----------------

# coefficients for data generation models
b0 <- 0
b1 <- 0.8
b2 <- -0.25
b3 <- 0.6
b4 <- -0.4
b5 <- -0.8
b6 <- -0.5
b7 <- 0.7
a0 <- -3.85
a1 <- 0.3
a2 <- -0.36
a3 <- -0.73
a4 <- -0.2
a5 <- 0.71
a6 <- -0.19
a7 <- 0.26
g1 <- -0.4 # effect of exposure


#------------------ Functions------------------

# function:
# generate continuous random variable correlated to variable x by rho
# invoked by the "F.generate" function
# Parameters:
# x - data vector
# rho - correlation coefficient
# Returns: a correlated data vector of the same length as x

F.sample.cor <- function(x, rho) {
  y <- (rho * (x - mean(x)))/sqrt(var(x)) + sqrt(1 - rho^2) *
rnorm(length(x))
  #cat("Sample corr = ", cor(x, y), "\n")
  return(y)
}
```

```
# function: generate simulation datasets
# inputs: sample size N, scenario
# outputs: 1 dataset of size N
# binary variables: w1, w3, w5, w6, w8, w9
# continous variables: w2, w4, w7, w10
# confounders: w1, w2, w3, w4
# exposure predictors only: w5, w6, w7
# outcome predictors only: w8, w9, w10
# correlations: (w1,w5)=0.2, (w2,w6)=0.9, (w3,w8)=0.2, (w4,w9)=0.9


F.generate <- function(size, scenario) {
  ## begin
  w1 <- rnorm(10, mean=0, sd=1)
  w2 <- rnorm(size, mean=0, sd=1)
  w3 <- rnorm(size, mean=0, sd=1)
  w4 <- rnorm(size, mean=0, sd=1)
  w5 <- F.sample.cor(w1, 0.2)
  w6 <- F.sample.cor(w2, 0.9)
  w7 <- rnorm(size, mean=0, sd=1)
  w8 <- F.sample.cor(w3, 0.2)
  w9 <- F.sample.cor(w4, 0.9)
  w10 <- rnorm(size, mean=0, sd=1)


  # dichotomize variables (will attenuate correlations above)
  w1 <- ifelse(w1 > mean(w1), 1, 0)
  w3 <- ifelse(w3 > mean(w3), 1, 0)
  w5 <- ifelse(w5 > mean(w5), 1, 0)
  w6 <- ifelse(w6 > mean(w6), 1, 0)
  w8 <- ifelse(w8 > mean(w8), 1, 0)
  w9 <- ifelse(w9 > mean(w9), 1, 0)


  # scenarios for data generation models
  # A: model with additivity and linearity
  # B: mild non-linearity
  # C: moderate non-linearity
  # D: mild non-additivity
  # E: mild non-additivity and non-linearity
  # F: moderate non-additivity
  # G: moderate non-additivity and non-linearity


  # binary exposure modeling
  if (scenario == "A")
      z.a_trueps <- (1 + exp( -(0 + b1*w1 + b2*w2 + b3*w3 + b4*w4 + b5*w5 +
      b6*w6 + b7*w7) ) )
  else
     if (scenario == "B")
          z.a_trueps <- (1 + exp( -(0 + b1*w1 + b2*w2 + b3*w3 + b4*w4 +
          b5*w5 + b6*w6 + b7*w7 + b2*w2*w2) ) )^-1
    else
          if (scenario == "C")
              z.a_trueps <- (1 + exp( -(0 + b1*w1 + b2*w2 + b3*w3 + b4*w4 +
              b5*w5 + b6*w6 + b7*w7 + b2*w2*w2 +b4*w4*w4 + b7*w7*w7) ) )^-1
          else
```

```
            if (scenario == "D")
                z.a_trueps <- (1 + exp( -(0 + b1*w1 + b2*w2 + b3*w3 + b4*w4
                + b5*w5 + b6*w6 + b7*w7 + b1*0.5*w1*w3 + b2*0.7*w2*w4 +
                b4*0.5*w4*w5 + b5*0.5*w5*w6) ) )^-1
            else
                if (scenario == "E")
                    z.a_trueps <- (1 + exp( -(0 + b1*w1 + b2*w2 + b3*w3 +
                    b4*w4 + b5*w5 + b6*w6 + b7*w7 + b2*w2*w2 +
                    b1*0.5*w1*w3 + b2*0.7*w2*w4 + b4*0.5*w4*w5 +
                    b5*0.5*w5*w6) ) )^-1
                else
                    if (scenario == "F")
                        z.a_trueps <- (1 + exp( -(0 + b1*w1 + b2*w2 +
                        b3*w3 + b4*w4 + b5*w5 + b6*w6 + b7*w7 +
                        b1*0.5*w1*w3 + b2*0.7*w2*w4 + b3*0.5*w3*w5 +
                        b4*0.7*w4*w6 + b5*0.5*w5*w7 + b1*0.5*w1*w6 +
                        b2*0.7*w2*w3 + b3*0.5*w3*w4 + b4*0.5*w4*w5 +
                        b5*0.5*w5*w6) ) )^-1
                    else
                        { # scenario G
                        z.a_trueps <- (1 + exp( -(0 + b1*w1 + b2*w2 +
                        b3*w3 + b4*w4 + b5*w5 + b6*w6 + b7*w7
                        + b2*w2*w2 + b4*w4*w4 + b7*w7*w7 + b1*0.5*w1*w3 +
                        b2*0.7*w2*w4+b3*0.5*w3*w5 + b4*0.7*w4*w6 +
                        b5*0.5*w5*w7 + b1*0.5*w1*w6 + b2*0.7*w2*w3 +
                        b3*0.5*w3*w4 + b4*0.5*w4*w5 + b5*0.5*w5*w6) ) )^-1
                        }


  # probability of exposure: random number betw 0 and 1

  # if estimated true ps > prob.exposure, than received exposure (z.a=1)
  prob.exposure <- runif(size)
  z.a <- ifelse(z.a_trueps > prob.exposure, 1, 0)

  # continuous outcome modeling
  y.a <- a0 + a1*w1 + a2*w2 + a3*w3 + a4*w4 +a5*w8 + a6*w9 + a7*w10 + g1*z.a

  # create simulation dataset
  sim <- as.data.frame(cbind(w1, w2, w3 ,w4, w5, w6, w7, w8, w9, w10, z.a, y.a))
  return(sim)

} ## end function


#------------------ Call ------------------

# Example: Generate 1000 datasets of N=1000:

cenarioA <- replicate(1000, F.generate(1000, "A"))
cenarioB <- replicate(1000, F.generate(1000, "B"))
cenarioC <- replicate(1000, F.generate(1000, "c"))
cenarioD <- replicate(1000, F.generate(1000, "D"))
cenarioE <- replicate(1000, F.generate(1000, "E"))
cenarioF <- replicate(1000, F.generate(1000, "F"))
cenarioG <- replicate(1000, F.generate(1000, "G"))


#### Fim. ####
```