UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
DEPARTAMENTO DE COMPUTAÇÃO

Juliana Ferreira Alves

APLICAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA NA IDENTIFICAÇÃO DE MARCADORES GENÉTICOS PARA A DOENÇA DE ALZHEIMER

São Carlos, SP
2024

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
DEPARTAMENTO DE COMPUTAÇÃO

Juliana Ferreira Alves

APLICAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA NA IDENTIFICAÇÃO DE MARCADORES GENÉTICOS PARA A DOENÇA DE ALZHEIMER

Trabalho de Conclusão de Curso apresentado ao curso de Engenharia de Computação da Universidade Federal de São Carlos, como requisito parcial para a obtenção do título de Bacharel em Engenharia de Computação.

Orientador: Prof. Dr. Ricardo Cerri

São Carlos, SP
2024

Dedico este trabalho à minha mãe e a mim mesma.

Agradecimentos

Agradeço à minha mãe, Araci, que me criou com amor e me ensinou a determinação necessária para alcançar meus objetivos na vida. Também sou grata aos meus familiares, em especial minha avó Aliete, Tia Adilma, Tio Reginaldo, Tia Elizete, Tio Tarciano e Tia Viviane, que não apenas me proporcionaram apoio emocional, mas, ao saberem que eu seria admitida na UFSCar, me presentearam com um computador. Esse gesto foi essencial para meus primeiros anos na universidade. Não posso deixar de mencionar meu pai, João, cujo apoio foi fundamental durante todo o meu tempo na instituição.

Durante minha graduação, tive a oportunidade de conhecer pessoas incríveis que deixaram uma marca indelével na minha vida. Primeiramente, quero expressar minha gratidão ao meu orientador, Ricardo Cerri, que acreditou em mim em um momento de recomeço e me ofereceu um projeto que redirecionou minha trajetória acadêmica, influenciando minha escolha de carreira. Agradeço também pelo empenho na obtenção das bolsas FAPESP e BEPE, que, sem dúvida, foram cruciais em minha jornada. Além disso, quero estender meus agradecimentos a todos os amigos que fiz ao longo desse período, com destaque para Italo Ribeiro, meu grande companheiro durante toda a graduação, Daniel Sá Barreto, que me apoiou e orientou em diversos momentos, e meus amigos da Engenharia Química, Maria Vitória, Mariana Shiraishi, Paulo Henrique, Eder Ogeda, Leonardo Lima e Matheus Pinheiro. Juntos, desbravamos os primeiros anos da graduação e compartilhamos momentos inesquecíveis.

Durante minha estadia no exterior, conheci pessoas igualmente especiais que desempenharam um papel fundamental na conclusão deste projeto. Expresso minha gratidão ao Brito's Lab pelo conhecimento compartilhado, o acolhimento e as risadas nos momentos de discontração. Em especial, quero agradecer ao Henrique Alberto, Wen Hui, Felipe Eguti, Lirong Hu (também conhecida como Rose) e Luiz Brito. Também sou grata ao Sean Rowan, que sempre me deu apoio e me ensinou a enxergar o mundo com mais leveza, tanto durante meu perído exterior quanto no Brasil.

Não posso deixar de mencionar minha enorme gratidão ao PyLadies, especialmente ao PyLadies São Carlos, e à comunidade Python. Elas me introduziram ao mundo da tecnologia, proporcionaram oportunidades não apenas para aprender, mas também para ensinar. Vocês foram uma verdadeira inspiração!

Gostaria também de expressar minha gratidão a todos os colaboradores da universidade, aos professores que me ensinaram valiosas lições e me ajudaram a desenvolver resiliência, e a todas as pessoas que contribuem para o funcionamento da UFSCar. Seu esforço é fundamental para o sucesso da instituição.

Por fim gostaria de agradecer a mim mesma, por não ter desistido.

"– O mundo pertence aos corajosos

(Eva zu Beck)

Resumo

O polimorfismo de nucleotídeo único (SNP) é a variação em uma única posição na cadeia de nucleotídeos onde é formado o DNA. Visto que se trata de uma alteração genética, ele é de extrema importância para o estudo da saúde humana. Através dele, é possível prever as respostas de indivíduos a determinados medicamentos, buscar genes relacionados a doenças hereditárias em um grupo familiar e também pode ser associado a doenças mais complexas como doenças cardiovasculares, diabetes, câncer e a Doença de Alzheimer. Com a utilização de Aprendizado de Máquina supervisionado, é possível realizar estudos da relação dos SNPs com doenças complexas, sendo cada SNP uma variável de entrada para tais algoritmos. Desta forma, o objetivo deste projeto foi investigar a relação dos SNPs com a Doença de Alzheimer, através de algoritmos de Aprendizado de Máquina. Para isso, foram utilizados conjuntos de dados de indivíduos e seus respectivos SNPs e diagnósticos (Normal, Comprometimento Cognitivo Leve ou Doença de Alzheimer). Assim, este estudo apresenta uma abordagem inovadora para a identificação de marcadores genéticos associados à Doença de Alzheimer (DA), unindo técnicas de aprendizado de máquina e análise genômica em um conjunto de quatro etapas cruciais. A primeira etapa consiste no pré-processamento e normalização dos dados, seguida pela implementação de estudos de associação genômica ampla (GWAS) em todos os conjuntos de dados gerados na fase anterior. A terceira etapa emprega métodos avançados de aprendizado de máquina no conjunto de dados mais promissor identificado nas etapas anteriores. Finalmente, a quarta etapa envolve uma análise comparativa dos resultados alcançados nas etapas de GWAS e aprendizado de máquina. Os resultados deste estudo revelaram um conjunto abrangente de SNPs associados à DA, incluindo tanto aqueles previamente conhecidos na literatura científica quanto novas descobertas promissoras. Além disso, destacou-se a importância do tratamento de dados no controle de qualidade, o que teve um impacto significativo nos resultados encontrados. Os modelos de Aprendizado de Máquina utilizados neste estudo mostraram perfis distintos de SNPs mais significativos, enfatizando a complexidade e a heterogeneidade da DA. Essa variação nos coeficientes e importâncias das características sublinha a necessidade de uma abordagem integrada e multifacetada para a pesquisa em genética da DA. Em resumo, este estudo demonstra o potencial do aprendizado de máquina e da análise genômica no avanço do conhecimento sobre a DA. Os resultados fornecem novas percepções para o campo emergente da genômica da DA, abrindo novas perspectivas para estratégias terapêuticas e diagnósticas mais eficazes. As descobertas apresentadas representam um avanço significativo em direção a um entendimento mais profundo da genética da DA e seu impacto na saúde humana.

Palavras-chave: Predição de Doenças. Doença de Alzheimer. Aprendizado de máquina supervisionado. Classificação. Seleção de Atributos. Ciência de dados. GWAS.

Abstract

Single nucleotide polymorphism (SNP) is the variation at a single position in the nucleotide chain where DNA is formed. Since it is a genetic alteration, it is of utmost importance for the study of human health. Through it, it is possible to predict individuals' responses to certain medications, search for genes related to hereditary diseases in a family group, and it can also be associated with more complex diseases such as cardiovascular diseases, diabetes, cancer, and Alzheimer's Disease. With the use of supervised Machine Learning, it is possible to conduct studies on the relationship between SNPs and complex diseases, with each SNP being an input variable for such algorithms. Thus, the aim of this project was to investigate the relationship between SNPs and Alzheimer's Disease, through Machine Learning algorithms. For this purpose, datasets of individuals and their respective SNPs and diagnoses (Normal, Mild Cognitive Impairment, or Alzheimer's Disease) were used. Therefore, this study presents an innovative approach to identifying genetic markers associated with Alzheimer's Disease (AD), combining machine learning techniques and genomic analysis into a set of four crucial steps. The first step consists of data preprocessing and normalization, followed by the implementation of Genome-Wide Association Studies (GWAS) on all datasets generated in the previous phase. The third step employs advanced machine learning methods on the most promising dataset identified in the previous steps. Finally, the fourth step involves a comparative analysis of the results achieved in the GWAS and machine learning stages. The results of this study revealed a comprehensive set of SNPs associated with AD, including both those previously known in the scientific literature and promising new discoveries. Furthermore, the importance of data handling in quality control was highlighted, which had a significant impact on the results obtained. The machine learning models used in this study showed distinct profiles of the most significant SNPs, emphasizing the complexity and heterogeneity of AD. This variation in coefficients and feature importances underscores the need for an integrated and multifaceted approach to AD genetics research. In summary, this study demonstrates the potential of machine learning and genomic analysis in advancing knowledge about AD. The results provide new insights into the emerging field of AD genomics, opening up new perspectives for more effective therapeutic and diagnostic strategies. The findings presented represent a significant advancement towards a deeper understanding of AD genetics and its impact on human health.

Keywords: Disease Prediction. Alzheimer's Disease. Supervised Machine Learning. Classification. Feature Selection. Data Science. GWAS

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

A Doença de Alzheimer é uma patologia degenerativa com grande impacto tanto na vida do portador da doença (que pode ter eventual incapacitação) quanto na vida dos indivíduos próximos a ele. Além disso, ela também é uma doença que exige muitos investimentos em seu tratamento. Em 2015, o custo total estimado no tratamento da Alzheimer no mundo era de 818 bilhões de dólares. Exemplificando, se o tratamento dessa doença fosse um país, seria a 18^a maior economia do mundo em 2016 (World..., 2015).

Dada a importância da Doença de Alzheimer no mundo, pesquisas que tentam identificar marcadores moleculares associados à doença têm grande impacto na medicina preventiva e personalizada para pacientes. Esses estudos podem possibilitar a melhoria na vida de milhares de pessoas que tendem a ter esta patologia, conseguindo assim começar suas prevenções e tratamentos antecipadamente. A estimativa é que haja 131.5 milhões de pessoas com a DA em 2050 (World..., 2015).

1.1 Objetivos

1.1.1 Objetivo geral

O objetivo desta pesquisa é identificar a correlações entre os marcadores genéticos e a Doença de Alzheimer. Para alcançar esse propósito, são adotados algoritmos de aprendizado de máquina supervisionados em conjunto com técnicas de GWAS. De maneira mais específica, empregamos algoritmos Baseados em Árvores de Decisão e Regressão Logística, fazendo uso da Regularização Ridge e Lasso. Esses métodos, por meio de estratégias matemáticas avançadas, possibilitam a seleção das variáveis mais relevantes no modelo. Essa abordagem é vantajosa, uma vez que se mostra apropriada para resolver problemas com um grande número de variáveis, o que se encaixa perfeitamente nos conjuntos de dados que serão construídos para esta pesquisa, nos quais cada exemplo abarcará centenas ou mesmo milhares de SNPs (variáveis genéticas).

1.1.2 Objetivos específicos

Os objetivos específicos deste trabalho são:

- Aplicar técnicas de tratamento de dados visando otimizar os resultados dos algoritmos.

- Comparar técnicas de Aprendizado de Máquina com abordagens comuns na área de Genética e Bioinformática para a seleção de SNPs relacionados à doença de Alzheimer.
- Testar e encontrar a melhor combinação de hiperparâmetros tanto para a etapa de engenharia de dados quanto para a modelagem de dados.

1.2 Organização do trabalho

A estrutura do presente documento é delineada em cinco capítulos, complementados por um apêndice. O Capítulo 1 inaugura o texto com uma contextualização da investigação proposta, elucidação dos objetivos e a justificativa da pesquisa. Uma revisão abrangente da literatura e teorias pertinentes ao estudo é exposta no Capítulo 2. A metodologia adotada para a condução dos experimentos é articulada no Capítulo 3. O Capítulo 4 consolida os procedimentos experimentais e os achados resultantes. Conclusões derivadas da pesquisa, bem como recomendações para indagações futuras, são apresentadas no Capítulo 5.

2 Fundamentação Teórica

2.1 SNPs e sua Relação com a Doença de Alzheimer

Polimorfismos de nucleotídeo único (SNPs) são variações em uma determinada posição na sequência de DNA. SNPs são categorizados de acordo com suas variações nucleotídicas. Se as mudanças forem do tipo C $\rightarrow$ T ou G $\rightarrow$ A eles são considerados SNPs de transição, e caso as variações sejam to tipo C $\rightarrow$ A, A $\rightarrow$ T, T $\rightarrow$ C ou T $\rightarrow$ G, são considerados de transversão (Edwards et al., 2007). SNPs também são o tipo de polimorfismo genético mais frequente, e tais variações têm muito impacto em como os organismos se desenvolvem e se comportam no meio ambiente. Além disso os SNPs são evolutivamente estáveis, pois não sofrem mudanças significativas ao longo das gerações. Desta forma, eles podem ser associados à doenças hereditárias e também à doenças complexas tais como doenças cardiovasculares, diabetes, câncer e à doença de Alzheimer (Edwards et al., 2007), que é o foco desta pesquisa.

2.1.1 Como os SNPs estão Relacionados a Doença de Alzheimer

A doença de Alzheimer (DA) é uma patologia degenerativa, e a mais frequentemente associada à idade, resultando em manifestações cognitivas e neuropsiquiátricas causadoras de uma eventual incapacitação (Sereniki; Vital, 2008). Portanto, trata-se não apenas de um problema de saúde pública, mas também representa um grande impacto social tanto para os pacientes quanto para as pessoas que convivem com eles. Em 2016 a estimativa era de 46.8 milhões de casos no mundo, e a projeção é de 74.7 milhões até 2030, e 131.5 milhões em 2050 (World..., 2015). Histopatologicamente a doença caracteriza-se por um grande dano sináptico e pela morte neuronal observada nas regiões cerebrais encarregadas pelas funções cognitivas, tais como o córtex cerebral, o hipocampo, o córtex entorrinal e o estriado ventral (Sereniki; Vital, 2008). Apesar da idade avançada (mais de 65 anos) ser o fator de risco mais conhecido associado à DA, indivíduos jovens também podem desenvolvê-la.

Existem quatro principais tipos de estudos a respeito da genética da DA. Dentre eles temos a análise de ligação genética, que foi a primeira a desvendar a base genética de estudos da DA. Ela visa identificar regiões cromossômicas associadas à doença, mas sem necessariamente identificar o gene ou a mutação associada (Tanzi; Bertram, 2005). Em seguida tem-se o estudo de genes candidatos, que compara as variações genéticas de pessoas com a doença e de pessoas saudáveis. Essa abordagem identificou que alelos de risco do gene APOE são fortes candidatos ao aparecimento da Doença de Alzheimer (Mohan;

Man; Yang, 2016). Também temos os estudos de associações genômicas em larga escala (GWAS), que graças ao desenvolvimento da microtecnologia, conseguem realizar uma avaliação simultânea de milhões de SNPs. Com eles foi possível identificar 20 loci genéticos associados ao aumento da suscetibilidade à DA em pessoas idosas (Karch; Cruchag; Goate, 2014). Por fim existem as tecnologias de *Next Generation Sequencing* (NGS), que fornecem estratégias de sequenciamento rápido e econômico, e têm implicações importantes no estudo de muitas doenças neurológicas. Recentemente essas tecnologias têm sido aplicadas para identificar fatores em pequenas famílias com inexplicáveis casos da DA em pessoas com menos 65 anos (early-onset Alzheimer Disease, ou EOAD) (Bertram, 2016).

Neste projeto trabalharemos principalmente com GWAS, pois o foco desta pesquisa é identificar quais SNPs estão associados à Doença de Alzheimer em pessoas com mais de 65 anos (LOAD). Trata-se de um problema geneticamente mais complexo do que o relacionado a EOAD, pois envolve diversos genes e fatores ambientais (Mohan; Man; Yang, 2016). Antes dos GWAS serem utilizados em larga escala, o gene APOE era o único com fator de risco bem estabelecido para a doença, mas com os avanços das tecnologias foi possível encontrar diversas regiões no genoma associadas a esse fator (Mohan; Man; Yang, 2016).

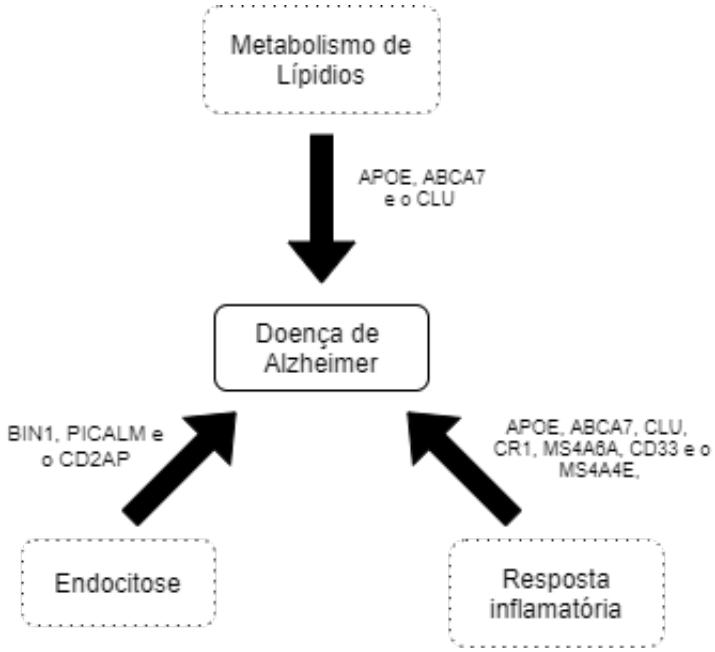
Devido aos GWAS foi possível perceber que grande parte dos genes que estão ligados à Doença de Alzheimer se agrupam em três grupos de funções, sendo resposta inflamatória, metabolismo de lipídios, e endocitose (Mohan; Man; Yang, 2016). Para esta pesquisa utilizaremos os 10 genes que foram classificados com maior relação de influência no surgimento da DA. Segundo o repositório Alzgene (Alzgene..., 2011), eles são: o APOE, responsável pelo metabolismo de lipídios; o ABCA7 e o CLU, responsáveis pelo metabolismo de lipídios e por respostas inflamatórias; o CR1, o CD33 e o MS4A, encarregados pelas respostas inflamatórias; e por fim o BIN1, o PICALM e o CD2AP, responsáveis pela endocitose (Mohan; Man; Yang, 2016; Alzgene..., 2011). Essas relações estão ilustradas na Figura 1.

2.1.2 GWAS

O foco deste estudo é identificar SNPs que estão associados à DA, utilizando estudos de associação genômica ampla. SNPs, ou Polimorfismos de Nucleotídeo Único, referem-se a variações em locais específicos na cadeia de DNA. Essas variações são categorizadas com base no tipo de substituição de nucleotídeo que ocorre. Por exemplo, substituições de C→T ou G→A são referidas como SNPs de transição, enquanto substituições de C→A, A→T, T→C ou T→G são referidas como SNPs de transversão (Edwards et al., 2007).

O fluxo de trabalho de um GWAS envolve várias etapas que são cuidadosamente planejadas para evitar viéses e erros. O primeiro passo é obter dados de um grupo abrangente e diversificado de indivíduos, que inclui informações genotípicas e fenotípicas para

Figura 1 – Esquema das principais funções dos genes que estão relacionados à Doença de Alzheimer.



Adaptado de (Mohan; Man; Yang, 2016).

as populações de caso e controle. Informações genotípicas correspondem ao DNA obtido usando arrays de SNPs ou estratégias de sequenciamento. Antes de prosseguir para o teste de associação, é recomendado aplicar um procedimento de controle de qualidade aos dados para aumentar sua confiabilidade. Isso envolve a realização de limpeza de dados usando abordagens estatísticas e conceitos biológicos para mitigar viéses que poderiam interferir nos resultados. Neste estudo, enfatizamos o impacto do Controle de Qualidade no GWAS, criando vários cenários, conforme explicado na Subseção 2.1.2.1. Após o teste de associação, várias análises pós-GWAS podem ser conduzidas para interpretar os resultados.

2.1.2.1 Controle de Qualidade no GWAS

Os dados de entrada para a análise de associação incluem números de identificação individual, estágio da doença, sexo e SNPs obtidos por meio de sequenciamento, juntamente com informações sobre o lote genotípico. Para minimizar erros e vieses, técnicas de Controle de Qualidade (QC) são necessárias para os dados de entrada. Esses métodos filtram SNPs e indivíduos usando equações estatísticas e matemáticas, com base em conceitos biológicos.

Para preservar a de qualidade dos dados e reduzir significativamente o número de SNPs candidatos, vários filtros estatísticos são aplicados em estudos de Associação de Genoma Ampla (GWAS) (Anderson et al., 2010). Essas técnicas evitam que os efeitos dos

SNPs sejam mascarados pela alta dimensionalidade do conjunto de dados. Neste estudo, apesar de também trabalharmos com modelos de Aprendizado de Máquina, utilizando métodos tradicionais usados em GWAS para filtrar SNPs e evitar as situações mencionadas acima. Os filtros usados em nosso estudo são os seguintes:

- **Frequência do Alelo Menor (MAF):** Na genética de populações, o termo frequência do alelo menor (MAF) refere-se à frequência de um alelo menos prevalente em um determinado locus (posição) em uma população. É descrito como a prevalência do alelo menos comum, ou segundo alelo mais frequente, em um locus específico em uma população. O alelo majoritário, que é o alelo mais prevalente naquele locus, é oposto ao alelo menor. Em GWAS, o valor calculado para MAF pode diminuir o número de SNPs, excluindo variantes raras do banco de dados. Portanto, MAF é a frequência do segundo alelo mais comum na população estudada. SNPs com MAF inferior a um limite especificado (por exemplo, 1%) são removidos (Zeng et al., 2015);

- **Desequilíbrio de Ligação (LD):** A associação não aleatória de alelos em vários loci em uma população é conhecida como desequilíbrio de ligação (LD). Em outras palavras, refere-se à tendência de alelos específicos serem herdados juntos com mais frequência do que seria previsto pelo acaso em vários loci. Em GWAS, os pesquisadores geralmente genotipam um número significativo de SNPs em todo o genoma em casos e controles para encontrar SNPs que estão relacionados à doença ou característica em estudo. Os pesquisadores podem encontrar grupos de SNPs que estão em LD e, portanto, provavelmente serão transmitidos juntos ao examinar os padrões de LD entre os SNPs. Esses grupos de SNP são conhecidos como haplótipos e podem ser usados para localizar partes do genoma que estão ligadas a uma doença ou característica específica. Portanto, o uso de LD como filtro serve para garantir o controle de qualidade e evitar a perda de informações (Malo; Libiger; Schork, 2008);

- **Equilíbrio de Hardy-Weinberg (HWE):** O equilíbrio de Hardy-Weinberg pressupõe que, em uma população mendeliana, as frequências alélicas permanecerão constantes ao longo das gerações. Portanto, é usado para remover variantes que não estejam em conformidade com suas expectativas;

- **Ausência de Genótipo (GENO) e Ausência de Amostra (SAMPLE):** Para ambos os parâmetros, o procedimento de filtragem é aplicado escolhendo um limite e calculando a ausência de dados de Genótipo ou Amostra. Valores de ausência acima do limite são excluídos do banco de dados.

Esses critérios, embora necessários para garantir a qualidade e relevância dos SNPs utilizados, introduzem um viés de seleção, pois excluem variantes que podem conter in-

formações úteis. Esta metodologia inclui também a conflitante escolha entre a redução da complexidade do modelo e a potencial perda de informações genéticas relevantes, destacando a importância de uma seleção criteriosa e informada dos dados. A referência a essa abordagem, evidenciando a complexidade e os desafios na seleção de características em dados genômicos para a predição de doenças complexas como a de Alzheimer.

2.1.3 Modelagem Computacional e Doença de Alzheimer

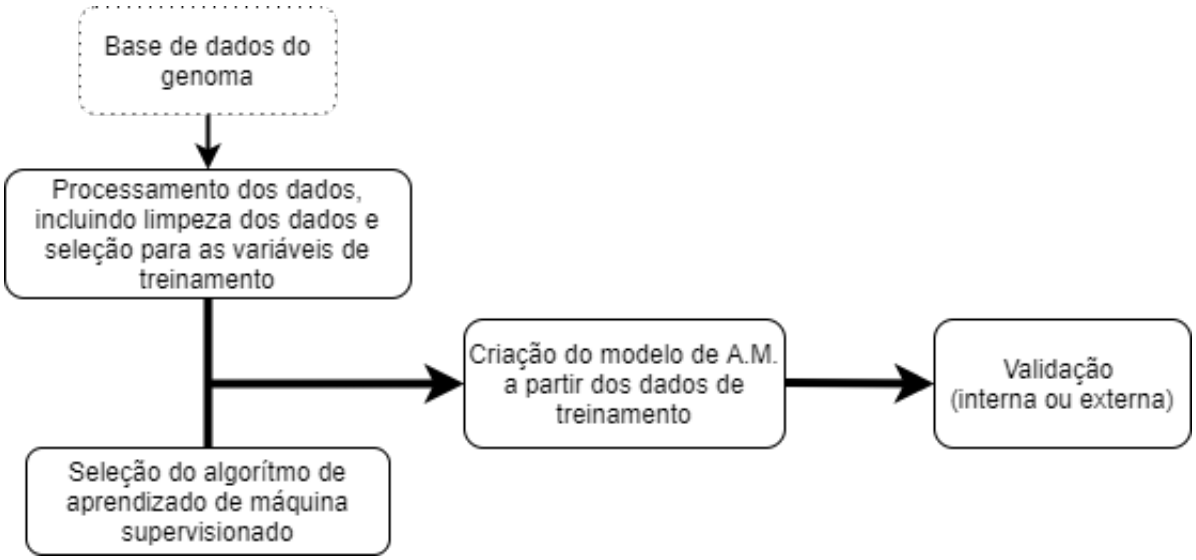
Desde de o Projeto Genoma Humano, que surgiu em meados de 1990 visando o mapeamento e sequenciamento do genoma humano (além da descoberta de conhecimentos dos genes associados ou não com doenças genéticas), as tecnologias a respeito do sequenciamento do DNA têm avançado drasticamente (Bueno, 2009). Dessa forma houve uma crescente abundância de dados genéticos e outras informações biológicas dos seres humanos (Laksman; Detsky, 2011; Spiegel; Hawkins, 2012). Esse crescimento exponencial na quantidade de dados teve como resultado uma evolução nos conceitos de medicina precisa e personalizada, que tem como objetivo trazer tratamentos médicos sob medida para os pacientes de acordo com suas características genéticas (Laksman; Detsky, 2011; Johnson, 2017). A recente inclusão de fatores de riscos genéticos, como os SNPs associados a fenótipos e doenças, melhorou a modelagem de predição para doenças individuais (Laksman; Detsky, 2011). Essas modelagens costumam ser feitas através da pontuação de risco poligênico ou através do Aprendizado de Máquina (Ho et al., 2019).

Nesta pesquisa, os métodos de Aprendizado de Máquina supervisionados serão aplicados para realizar a predição da Doença de Alzheimer, com o objetivo principal de identificar quais os SNPs mais importantes para esse diagnóstico. Esses métodos utilizam algoritmos computacionais e estatísticos para realizarem predições médicas, ou seja, esses métodos realizam predições por meio da modelagem matemática sobre as associações complexas entre um conjunto de SNPs de risco e o fenótipo da doença estudada (Ho et al., 2019).

O processo de criação de um modelo de Aprendizado de Máquina supervisionado pode ser dividido em três etapas. Na primeira etapa, modelos são gerados por meio de um treinamento, em que se utiliza de um algoritmo responsável por criar as relações entre os dados genéticos do indivíduo e seu diagnóstico de saúde. Assim, o poder de predição para a doença estudada é alcançado por meio do mapeamento das variáveis encontradas dentro dos dados de treinamento do genoma (Ho et al., 2019). Em alguns modelos de aprendizado, utilizam-se estratégias como gradiente descendente e execuções interativas de estimação de parâmetros para maximização desempenho. Na segunda etapa, o processo de treinamento e otimização de parâmetros é repetido até que um bom modelo seja alcançado. Na penúltima etapa, após a etapa de treinamento, o melhor modelo obtido é avaliado em um conjunto separado (Ho et al., 2019), podendo ser utilizado para realizar

predições. Esse fluxo de etapas do aprendizado supervisionado é ilustrado na Figura 2.

Figura 2 – Ilustração do fluxo de etapas do Aprendizado de Máquina supervisionado.



Adaptado de (Ho et al., 2019).

2.2 Aprendizado de Máquina para Classificação da Doença de Alzheimer

O Aprendizado de Máquina é um campo da inteligência artificial no qual os algoritmos podem aprender com os dados que lhes são fornecidos. Essa área da ciência da computação e estatística é aplicada a problemas complexos nos quais não se pode encontrar boas soluções utilizando algoritmos convencionais. Ou seja, as soluções para os problemas não são codificadas explicitamente, mas encontradas por um algoritmo de aprendizado. Os algoritmos de Aprendizado de Máquina também são aplicados na análise de grandes quantidades de dados, e em problemas dinâmicos. Assim, sistemas utilizando Aprendizado de Máquina podem se adaptar a novos dados, e dessa forma também a novos parâmetros (Géron, 2002).

Existem diferentes paradigmas de Aprendizado de Máquina. Eles dividem-se em supervisionado, não supervisionado, semi supervisionado ou por reforço. Além disso podem funcionar comparando novos dados com dados já conhecidos, ou então detectar padrões em dados de treinamento para criar um modelo preditivo. Essas classificações são comparadas utilizando os termos aprendizado baseado em instância versus aprendizado baseado em modelo (Géron, 2002). Esses paradigmas podem ser combinados da melhor maneira para resolver um determinado problema. A seguinte seção apresenta os conceitos do aprendizado supervisionado, bem como dos algoritmos que serão utilizados neste trabalho.

2.3 Aprendizado de Máquina Supervisionado

No Aprendizado de Máquina supervisionado, são fornecidos além dos dados de treinamento, as soluções almejadas para resolver o problema, conhecidas como classes, rótulos ou classes (Géron, 2002), que podem ser numéricos ou categóricos. No caso desta pesquisa, os dados de treinamentos são os SNPs de cada indivíduo e os rótulos são os diagnósticos de saúde que podem variar entre Normal, Transtorno Cognitivo Leve, ou Doença de Alzheimer.

A classificação de dados é uma das tarefas mais comuns em aprendizado supervisionado. Nesse caso o algoritmo aprende a classificar novos dados por meio de um treinamento utilizando muitos dados de entrada, que no nosso caso são os SNPs, junto com suas classes, que são os estados de saúde dos indivíduos (Géron, 2002).

Outra tarefa muito comum dentro do aprendizado supervisionado é a regressão. Nela, ao invés de uma classe, a saída do algoritmo deve ser um valor numérico. Existem diferentes tipos de algoritmos de regressão, sendo alguns adequados para cenários com centenas ou milhares de variáveis, nos quais deseja-se identificar as variáveis mais importantes. Esse é justamente o cenário dos conjuntos de dados que serão utilizados nesta pesquisa, possuindo centenas ou milhares de SNPs.

Com intuito de também utilizar tais algoritmos de regressão neste trabalho, podemos usar técnicas de transformação de dados para faze-los resolver um problema de classificação. No caso do problema da predição do diagnóstico da Doença de Alzheimer, é possível mapear os valores categóricos de diagnóstico para valores numéricos probabilísticos.

A utilização de alguns algoritmos de regressão mostra-se adequada para a proposta desta pesquisa. Dado que os conjuntos de dados possuem centenas ou milhares de SNPs, são interessantes os algoritmos que utilizam estratégias de regularização, permitindo selecionar os SNPs mais importantes para o modelo.

Além da regressão, utilizaremos também modelos baseados em árvores de decisão. Estes, por sua vez, são por natureza algoritmos de classificação.

2.3.1 Regressão Linear

A regressão linear é um método que encontra a reta que melhor representa as relações entre as variáveis de entrada (preditoras) com as variáveis de saída (classes) de um problema. Ou seja, suponha que gostaria-se de criar um modelo de Aprendizado de Máquina para predizer o valor de um imóvel, tendo seu tamanho em metros quadrados como atributo preditor. Com esses parâmetros, o modelo é treinado com uma quantidade n de dados de entrada (contendo os atributos e seus respectivos classes) a fim de encontrar

a função que rege a relação. Essa relação é apresentada matematicamente na Equação2.1 (James et al., 2014).

$$Y \approx \hat{\beta}_0 + \hat{\beta}_1 X \tag{2.1}$$

O simbolo ≈ na Equação 2.1 pode ser lido como "é aproximadamente modelado como". Na Equação, os coeficientes $\hat{\beta}_0$ e $\hat{\beta}_1$ representam respectivamente o *"Intercept"* e o *"Slop"* quando se trata de modelos lineares. Desta forma temos o Intercept como coeficiente linear da reta, ou seja o ponto onde ela cruza o eixo y, e o Slop como coeficiente angular, que determina a inclinação da reta, representando assim um hiperplano do espaço. Como mencionado anteriormente, ao treinar os modelos com uma quantidade n de dados, busca-se encontrar quais são os valores para $\hat{\beta}_0$ e $\hat{\beta}_1$ (James et al., 2014). Dessa maneira o exemplo dado para predição de valores de imóveis pode ser descrito pela Equação 2.2 a seguir.

$$Valor\ do\ imóvel \approx \beta\_0 + \beta\_1 \times Area \tag{2.2}$$

Seja $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \times X_i$ a predição Y baseada no i-ésimo valor de X. Considere um valor $e_i = Y_i - \hat{Y}_i$, ou então $e_i = Y_i - \hat{\beta}_0 + \hat{\beta}_1 \times X_i$. Isso significa que e_i é a diferença entre o valor real (esperado) de Y_i e a sua predição (dada pelo algoritmo) $\hat{Y}_i$. Estatisticamente, definimos a Soma Quadrática Residual (RSS) para todas as predições $\hat{Y}$ como $RSS = e_1^2 + e_2^2 + \cdots + e_n^2 = \sum_{i=1}^{n} e_i^2$ (James et al., 2014).

Para encontrar os coeficientes β_0 e β_1 é utilizado o Método dos Mínimos Quadrados (MMQ), que propõe escolher os coeficientes de modo à minimizar o RSS. Algebricamente, esses coeficientes são encontrados por meio das Equações 2.3 e 2.4 (James et al., 2014).

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \tag{2.3}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} \tag{2.4}$$

Note que $\overline{X}$ e $\overline{Y}$ são as médias amostrais de Xe Y tendo n como a quantidade de amostras. Estes valores são obtidos com $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ e $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ (James et al., 2014).

No entanto na maioria dos problemas reais são utilizadas diversas variáveis independentes (atributos) para realizar uma regressão. Assim, suponha como no exemplo anterior, que além do tamanho em metros^2 do imóvel, o modelo precisaria também da quantidade de quartos, da localização e da especificação se existem escolas e supermercados próximos ao imóvel. Nesse sentido, todos esses fatores influenciam para a predição do

valor do imóvel. Portanto, o modelo é considerado uma regressão com múltiplas variáveis. Cada variável incluída pode ser interpretada como uma nova dimensão para o problema.

A Equação 2.5 representa um modelo de regressão linear aplicado a um exemplo X_i, em que Y é o valor predito, a é o número de atributos, X_{ij} é o valor da j-ésima característica de X_i, e β são os coeficientes de regressão, aprendidos durante o treinamento do modelo.

$$Y = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_n X_{ia} = \sum_{j=1}^{a} \beta_j X_{ij} + \beta_0 \tag{2.5}$$

Outro ponto a ser ressaltado nessa situação é em relação ao tipo da variável que esta sendo utilizada. Basicamente pode-se dividir as variáveis em quatro grupos: quantitativas, qualitativas, proxy, e dicotômicas. As variáveis quantitativas representam atributos que podem ser contados ou medidos (Matta, 2007). As variáveis qualitativas têm valores não mensuráveis (Matta, 2007). As varíaveis proxy substituem outras variáveis de difícil mensuração. Um exemplo de variável proxy é a renda per capita, que possibilita estimar a qualidade de vida (Proxy..., 2020), que é de difícil mensuração. Por fim, as variáveis dicotômicas são booleanas, pois assumem apenas dois valores. Elas são fortemente atreladas à ausência ou presença de um atributo no modelo (Matta, 2007). Para representação de SNPs, por exemplo, pode-se utilizar tanto atributos quantitativos quanto dicotômicos.

2.3.2 Regressão Ridge e Regressão Lasso

Para reduzir o sobreajuste de modelos (erros em modelos que funcionam bem nos dados de treinamento mas têm desempenho ruim em dados novos) são utilizadas técnicas chamadas de regularização, neste trabalho utilizamos os regularizadores Ridge, Lasso e Elastic-net. Os regularizadores ajudam a controlar a complexidade do modelo, impedindo que os coeficientes se tornem muito grandes, o que pode levar ao sobreajuste. A Regressão Ridge (Hoerl; Kennard, 1970) e a Regressão Lasso (Tibshirani, 1996) são dois algoritmos que possuem estratégias de regularização de seus modelos de regressão (Alkhateeb et al., 2022), já a Regressão Elastic-Net é uma combinação ponderada de ambas anteriores.

A Regressão Ridge, também conhecida como regularização L2, é uma versão da regressão linear em que é adicionado o termo $\alpha \frac{1}{2} \sum_{j=1}^{a} \beta_i^2$ à função. Esse termo é uma penalização, que força o algoritmo de aprendizado a manter os coeficientes do modelo o mais reduzido possíveis, além de ajustar os dados.

O α utilizado é o hiperparâmetro que controla quanto o modelo deve ser regularizado. Se $\alpha = 0$ não há regularização, mas se $\alpha \to \infty$ todos os coeficientes tenderão a 0 e tem-se uma linha plana que passa pela media dos dados como resultado. A Equação 2.6 mostra a função de custo da Regressão Ridge, sendo RSS sigla para *Residual Sum of Squares* ou Soma Quadrática Residual.

$$\mathrm{J}(\beta) = RSS(\beta) + \alpha\frac{1}{2}\sum_{i=1}^{a}\beta_i^2 \tag{2.6}$$

Na Regressão Lasso,também conhecida como regularização L1, assim como na regressão Ridge, é adicionado um termo de regularização na função. No entanto utiliza-se a norma dos coeficientes, ao invés do quadrado como acontece na Regressão Ridge. A Equação 2.7 apresenta a função de custo da Regressão Lasso.

$$\mathrm{J}(\beta) = RSS(\beta) + \alpha\sum_{i=1}^{a}|\beta_i| \tag{2.7}$$

A utilização da norma do coeficiente ao invés do quadrado no termo de regularização é uma diferença sutil entre as regressões Ridge e Lasso, porém muito importante. Ela faz com que os coeficientes da regressão Lasso de fato atinjam o valor 0, enquanto na regressão Ridge, os coeficientes se aproximam de 0, mas nunca chegam a 0 de fato. Isso faz com que a regressão Lasso seja adequada para a seleção de atributos.

2.3.3 Regressão Logística

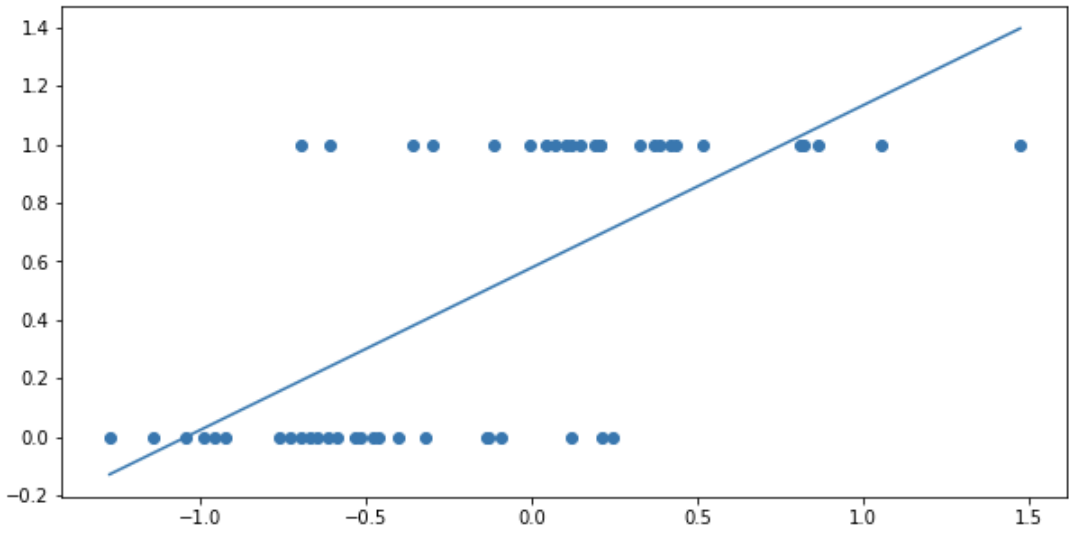
Em alguns casos quando a variável dependente é dicotômica, ou seja ela só tem dois possíveis resultados, a Regressão Logística é utilizada para realizar previsões, apenas adaptando essencialmente a Regressão Linear para a tarefa de classificação. De maneira ilustrativa, a Figura 3 apresenta um gráfico em que a variável X é contínua e a variável Y é dicotômica. Dessa maneira, se simplesmente tentarmos encontrar uma reta que representa a regressão linear, a mesma não obterá resultados satisfatórios na tarefa de predição, como pode ser notado na figura (James et al., 2014).

Para contornar este problema, ao invés de predizer o valor de $\hat{Y}$ pode-se predizer a Probabilidade de $\hat{Y}$, obtendo assim valores de predição que variam entre 0 e 1. A curva que rege esta predição é dada pela função sigmoide, dada pela Figura 4, na qual os pontos em cima desta curva refletem a probabilidade do problema em questão ser verdadeiro (James et al., 2014). Estatisticamente, a probabilidade do resultado esperado pode ser escrita como:

$$Pr(\text{Resultado padrão} = \text{Verdadeiro}|\text{Entrada}) \tag{2.8}$$

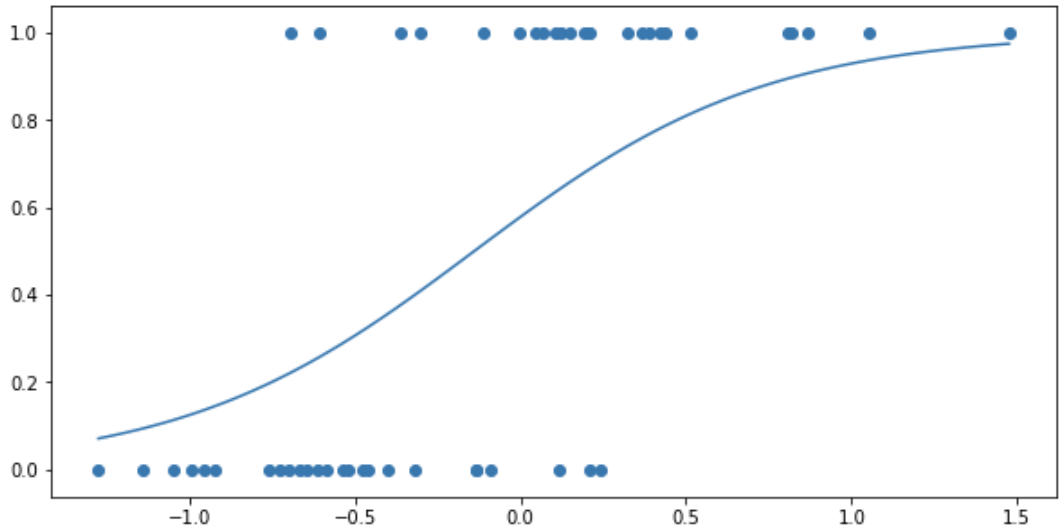
A Regressão Logística é uma generalização de um modelo linear. Assim, se quisermos encontrar uma relação entre a curva de probabilidade com uma reta que a representa, é preciso realizar transformação das mesmas através de $\hat{Y} = log(\frac{P(x)}{1-P(x)})$, também conhecida como $log(odds)$. Assim quando temos um ponto que está exatamente em 0.5 no eixo Y da Função Logística e o transformamos na Função Linear, ele se encontrará em cima

Figura 3 – Exemplificação da regressão linear aplicada em um problema de categorização.



Retirada de (James et al., 2014).

Figura 4 – Exemplo de Curva Logística que representa uma classificação simples.



Retirada de (James et al., 2014).

do eixo X (ou seja, no ponto 0 em Y). Quando tem-se um ponto contido em 1 ou 0 na Função Logística, ao realizar a transformação linear, este ponto estará em $+\infty$ ou $-\infty$ respectivamente.

Percebe-se então que é possível encontrar uma reta para realizar analises de modelos lineares. Porém quando se trata de encontrar a reta que melhor resolve o problema, não é possível utilizar o Método dos Mínimos Quadrados para encontrar os coeficientes β_0 e β_1. Para encontrar estes coeficientes utilizamos a função da Máxima Verossimilhança (ou maximum-likelihood function), que é dada pela Equação 2.9 (James et al., 2014)

$$Likelihood = \prod_{i=1}^{n}(P(x_i)) \prod_{j=1}^{n}(1 - P(x_j)) \tag{2.9}$$

Por outra perspectiva, a função de custo também pode ser obtida através de uma versão otimizada da função apresentada anteriormente, na qual é aplicada propriedades logarítmicas, a fim de reduzir a complexidade computacional da equação, se tornando a Equação 2.10, na qual $\hat{y} = P(x_i)$ (Li, 2015)

$$Likelihood = \sum_{i=1}^{n}(-y \cdot log(\hat{y}) - (1 - y) \cdot log(1 - \hat{y})) \tag{2.10}$$

Conseguimos, assim, obter uma Função Linear que melhor representa o problema em questão, podendo sofrer regularização através dos modelos Ridge (L2) e Lasso (L1), que serão detalhados a seguir, para diminuir os erros de generalização.

Aplicando a regularização L2 os coeficientes são "encolhidos" impondo uma restrição nos mesmo. Esta regularização ocorre como representado na Equação 2.11 (UFPR, xx; Aditya, 2018)

$$Likelihood = \sum_{i=1}^{n}(-y \cdot log(\hat{y}) - (1 - y) \cdot log(1 - \hat{y})) + \frac{1}{2}\sum_{j=1}^{a}\beta_j^2 \tag{2.11}$$

No entanto quando necessita-se que determinados coeficientes atinjam o valor 0 é usado a regularização L1, pois a regularização L2 apesar de se aproximar não atinge o valor 0. A função regularizada com o termo L2 é exemplificada na Equação 2.12. Embora possa parecer uma pequena modificação, as implicações práticas são significantes (UFPR, xx; Aditya, 2018)

$$Likelihood = \sum_{i=1}^{n}(-y \cdot log(\hat{y}) - (1 - y) \cdot log(1 - \hat{y})) + \sum_{j=1}^{n}|\beta_j| \tag{2.12}$$

2.3.4 Árvores de Decisão

O conceito a cerca de árvores de decisão é fundamental em alguns algoritmos de Aprendizado de Máquina para problemas de classificação e Regressão. Elas funcionam dividindo o espaço de entrada em regiões distintas, seguindo uma abordagem de *dividir para conquistar*. Em cada nó da árvore, uma decisão é tomada com base em um único atributo, dividindo o conjunto de dados em subconjuntos mais homogêneos (Breiman et al., 1984).

O processo pra construir uma árvore de decisão começa com o nó raiz, onde toda a amostra de treinamento é considerada. A cada passo, um atributo é selecionado para dividir o conjunto de dados, visando maximizar a pureza dos nós filho resultantes. Esse processo continua recursivamente até que um critério de parada seja atingido, como uma profundidade máxima da árvore ou um número mínimo de amostras em um nó (Quinlan, 1986).

Os critérios de divisão, como o índice Gini ou a entropia, são utilizados para avaliar a qualidade de uma divisão. O índice Gini mede a impureza de um nó, enquanto a entropia é uma medida da desordem ou incerteza. Estes critérios ajudam a determinar a melhor característica para dividir o conjunto de dados em cada etapa da construção da árvore (Raileanu; Stoffel, 2004).

2.3.5 Random Forest

Random Forest, introduzido por Breiman em 2001, é um método de Aprendizado de Máquina baseado em conjunto que melhora a performance e a precisão de modelos de decisão através da combinação de várias árvores de decisão (Breiman, 2001). Esta abordagem mitiga o problema de sobreajuste, comum em árvores de decisão individuais, especialmente em conjuntos de dados com alta variabilidade. O princípio fundamental do Random Forest é que um grupo de árvores fracas, cada uma contribuindo com suas previsões, pode formar um modelo forte mais robusto e preciso.

para a construção de uma Random Forest começamos com a criação de múltiplas árvores de decisão através de um processo chamado de bootstrap aggregating, ou bagging. Cada árvore é treinada em um subconjunto aleatório do conjunto de dados, selecionado com substituição, proporcionando diferentes perspectivas dos dados para cada árvore (Ho, 1998). Durante a construção da árvore, uma amostra aleatória das características é considerada para cada divisão, o que contribui para a diversidade entre as árvores. Esta diversidade é crucial para a robustez do modelo e para a redução do risco de sobreajuste.

Uma vez treinado, o Random Forest faz previsões agregando as saídas de todas as árvores individuais. No caso de problemas de classificação, isso geralmente envolve um sistema de ”votação por maioria”, onde cada árvore ”vota”para uma classe e a classe com

a maioria das votações é escolhida como a previsão final. Para problemas de regressão, a média ou a mediana das previsões de todas as árvores é calculada. Este processo de agregação contribui significativamente para a precisão e estabilidade do modelo, especialmente em conjuntos de dados com ruído (Breiman, 2001).

Random Forest é valorizado por sua versatilidade, sendo eficaz em uma ampla gama de problemas de classificação e regressão. Uma vantagem notável é a sua capacidade de lidar com conjuntos de dados com um grande número de características e com dados faltantes. Além disso, fornece uma medida de importância de características, útil para entender os fatores que mais influenciam as previsões. Esta metodologia tem sido aplicada com sucesso em domínios como bioinformática, reconhecimento de padrões e análise financeira (Liaw; Wiener, 2002).

2.3.6 XGBoost

XGBoost, que significa eXtreme Gradient Boosting, é uma implementação avançada e eficiente do algoritmo de boosting baseado em gradientes (Chen; Guestrin, 2016). Ele tem ganhado popularidade devido ao seu desempenho superior em várias competições de ciência de dados e é amplamente utilizado em problemas de Aprendizado de Máquina por sua velocidade e precisão. XGBoost melhora os métodos tradicionais de boosting por meio de otimizações de engenharia de software.

O XGBoost é baseado no princípio do boosting de gradiente, que é um método de aprendizado de conjunto. O objetivo é construir sequencialmente um conjunto de modelos fracos (geralmente árvores de decisão) e combiná-los para formar um modelo forte. A função objetivo do XGBoost pode ser descrita como:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{2.13}$$

onde $\mathcal{L}$ é a função de perda total, l é uma função de perda convexa que mede a diferença entre a previsão $\hat{y}_i$ e a observação real y_i, f_k são as funções que representam as árvores de decisão, K é o número de árvores, e Ω é uma função de regularização que penaliza a complexidade do modelo (Friedman, 2001).

Uma característica chave do XGBoost é sua capacidade de fazer uso eficiente dos recursos computacionais. Ele implementa várias otimizações, como paralelização da construção da árvore, poda de árvore eficiente e tratamento eficiente de valores faltantes. Além disso, XGBoost introduz um algoritmo regularizado para controlar o sobreajuste, o que é particularmente útil em conjuntos de dados com um alto grau de variabilidade (Chen; Guestrin, 2016)

2.3.7 Otimização por hiperparâmetros

Otimização por hiperparâmetros é a técnica que visa otimizar o resultados de um modelo de Aprendizado de Máquina selecionando os hiperparâmetros mais adequados para a modelagem. Neste trabalho, para realizar esta tarefa foi empregado o algoritmo da *Random Search*. Ao contrário do *Grid Search*, que avalia todas as possíveis combinações de hiperparâmetros, o *Random Search* seleciona aleatoriamente combinações dentro de um espaço de hiperparâmetros definido. Esta metodologia é particularmente benéfica em contextos com amplos espaços de hiperparâmetros, pois reduz significativamente o tempo computacional necessário para a identificação de uma combinação eficaz. Bergstra e Bengio demonstraram que o *Random Search* pode ser mais eficiente que o Grid Search, especialmente quando poucos hiperparâmetros são cruciais para o desempenho do modelo (Bergstra; Bengio, 2012). No, no Scikit-Learn, biblioteca utilizada neste trabalho, a implementação do *Random Search* permite a integração com Cross-Validation, facilitando uma avaliação mais robusta e confiável dos modelos selecionados.

A Cross-Validation é uma técnica crítica para avaliar a capacidade de generalização de um modelo preditivo. Ela envolve dividir o conjunto de dados em várias partes ou ”folds”, treinando o modelo em algumas dessas partes e testando nas demais. Esta abordagem é particularmente relevante quando combinada com a *Random Search*, como oferecido pelo Scikit-Learn. Ao aplicar Cross-Validation durante o processo de *Random Search*, é possível obter uma avaliação mais abrangente e imparcial do desempenho do modelo. A K-Fold Cross-Validation, onde ’K’ representa o número de grupos nos quais o conjunto de dados é dividido, é uma das variantes mais comuns. Conforme discutido por James et al., a Cross-Validation proporciona uma estimativa mais precisa do desempenho do modelo em novos dados, comparativamente à divisão simples em conjuntos de treino e teste, sendo particularmente vantajosa em conjuntos de dados de tamanho limitado (James et al., 2013).

3 Metodologia

Este capítulo tem como objetivo descrever a metodologia de experimentação utilizada neste trabalho, assim como os conjuntos de dados e algoritmos utilizados.

3.1 Arquitetura da metodologia

A metodologia adotada para o desenvolvimento deste estudo estrutura-se em quatro etapas cruciais, delineadas da seguinte forma: (1) Pré-processamento e normalização dos dados; (2) Implementação de estudos de associação genômica ampla (GWAS) em todos os conjuntos de dados derivados da primeira etapa, seguido pela seleção criteriosa do conjunto de dados que revelou os resultados mais promissores em termos de significância dos polimorfismos de nucleotídeo único (SNPs); (3) Emprego de métodos avançados de aprendizado de máquina no conjunto de dados identificado como ótimo na fase anterior; (4) Análise comparativa dos resultados alcançados nas etapas de GWAS e aprendizado de máquina.

A Figura 5 esquematiza as etapas mencionadas e suas respectivas sub-tarefas, as quais serão detalhadas nas seções subsequentes deste artigo.

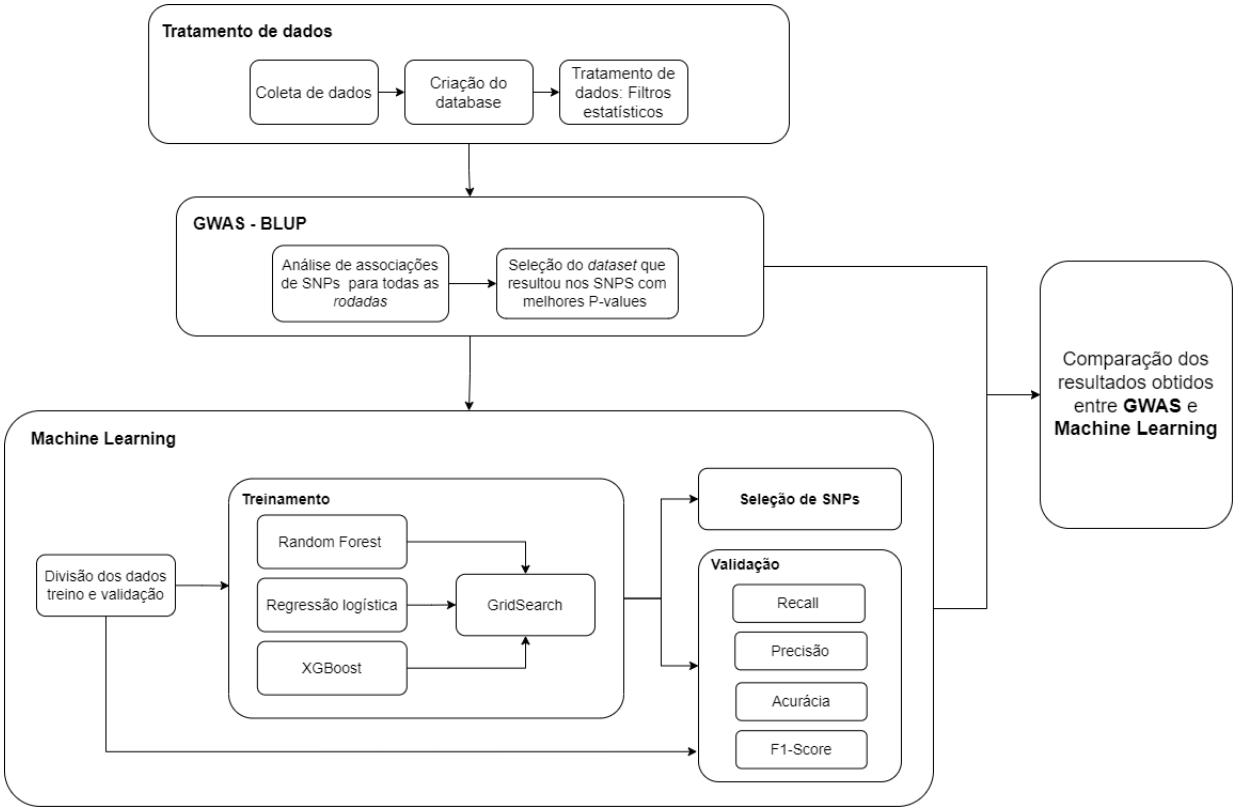
3.2 Obtenção dos dados

Os dados utilizados neste estudo foram obtidos a partir do banco de dados da Iniciativa de Neuroimagem da Doença de Alzheimer (ADNI, na sigla em inglês)[1]. Lançada em 2004 como uma parceria público-privada, a ADNI tem como objetivo identificar se imagens cerebrais, marcadores biológicos, avaliações clínicas e neuropsicológicas podem ser combinados para medir a progressão do comprometimento cognitivo leve (CCL) e da doença de Alzheimer (DA) em estágios iniciais.

Os dados genotípicos inicialmente continham 620.901 variantes, com uma média de 30.785 variantes ausentes. O arquivo de entrada incluiu 757 indivíduos, sendo 449 homens e 308 mulheres. O fenótipo foi dividido em três categorias: Normal (CN), com 214 amostras; Doença de Alzheimer (DA) com 175; e Comprometimento Cognitivo Leve (CCL), com 367. Essas categorias foram estimadas pela ADNI usando vários biomarcadores, que são substâncias, medições ou indicadores de um estado biológico que podem ser identificados antes do aparecimento de sintomas clínicos.

[1] http://adni.loni.usc.edu

Figura 5 – Arquitetura da metodologia



Elaborado pela própria autora.

Além dos dados genotípicos, também foi obtido um conjunto de dados de fenótipos de cada indivíduo. Esta base contem informações sobre sexo, idade, etnia e gênero.

3.3 Tratamento dos dados

3.3.1 Preprocessamento de dados

O conjunto de dados genotípicos inicial consistia em três arquivos separados: um arquivo de genótipo (Tabela 2), um arquivo de mapa para identificar os IDs de SNP no arquivo de genótipo (Tabela 3), e um arquivo de fenótipo (Tabela 4). Para consolidar essas informações em um único arquivo, utilizou-se o software PLINK, um conjunto de ferramentas de análise de associação de genoma completo. Com o PLINK, foi mesclado com sucesso os arquivos separados e foram aplicados os filtros mencionados na Subseção 2.1.2.1 para criar o arquivo final do conjunto de dados (Tabela 1). Este arquivo consolidado é adequado para uso em análises de GWAS e modelos de Aprendizado de Máquina.

Tabela 1 – Layout de modelo de conjunto de dados: arquivo de saída do PLINK.

ID Individual	Fenótipo	rs3094315	rs12563034	...	MitoC16272T
sujeito1	CN	0	1	...	0
sujeito2	DA	0	0	...	1
sujeito3	CN	2	2	...	2
...	...	...	...	...	...
sujeito_n	CCL	1	1	...	2

Nota: Este arquivo mapeia genótipos e diagnósticos de saúde para cada indivíduo. As informações de SNP são apresentadas em colunas, com cada coluna representando um alelo. Os nomes dos SNP são obtidos do arquivo MAP.

Tabela 2 – Arquivo de entrada de genótipo do PLINK

ID Individual	Genótipo
sujeito1	0102122102120
sujeito2	0112120212012
...	...
sujeito_n	0212120211011

Nota: Coluna 1: ID Individual; Coluna 2: Dados de genotipagem, onde 0 representa homozigoto, e 1 e 2 representam diferentes tipos heterozigotos.

Tabela 3 – Arquivo de entrada de mapa do PLINK

ID de SNP	Cromossomo	Posição
rs3094315	1	742429
rs12563034	1	758311
...	...	...
MitoC16272T	26	16272

Nota: O objetivo do arquivo MAP é identificar a localização de cada SNP dentro do cromossomo, bem como seu ID de SNP correspondente. A ordem dos IDs de SNP no arquivo MAP corresponde à ordem dos dados de genótipo na Tabela 1, coluna 2.

Tabela 4 – Arquivo de entrada de fenótipo do PLINK

ID Individual	Fenótipo
sujeito1	1
sujeito2	3
sujeito3	1
...	...
sujeito_n	2

Nota: Este arquivo fornece um mapeamento dos diagnósticos de saúde para cada indivíduo, onde o valor 1 indica normalidade cognitiva, o valor 2 indica comprometimento cognitivo leve e o valor 3 indica doença de Alzheimer.

3.3.2 Processamento de Dados para Controle de Qualidade

Nesta seção, será discutido como filtros estatísticos foram aplicados para garantir o controle de qualidade no conjunto de dados. O controle de qualidade (CQ) é uma etapa essencial no pré-processamento de dados, e melhora a consistência dos dados e a validade dos resultados. Assim, foram criados múltiplos conjuntos de dados variando os valores para cada hiperparâmetro de CQ durante a etapa de CQ, conforme mostrado na Tabela 5.

Cada combinação de parâmetros de controle de qualidade resultou em um conjunto de dados distinto que foi então usado para gerar um modelo estatístico. Como cada conjunto de parâmetros de controle de qualidade produziu um conjunto de dados único e subsequente modelo, cada iteração é referida como uma "rodada". Em outras palavras, uma rodada é uma combinação específica de hiperparâmetros de CQ que resultou em um conjunto de dados particular, que foi então usado para construir um modelo. Ao conduzir múltiplas *rodadas* com combinações variadas de parâmetros de controle de qualidade, foi explorado o impacto desses parâmetros no desempenho do modelo final e identificar o conjunto ótimo de parâmetros para nossa análise.

Para ilustrar, considere a 1ª rodada, que consiste em um conjunto de dados gerado combinando hiperparâmetros de CQ. Especificamente, o conjunto de dados tem uma taxa de ausência de genótipos de 0,02, uma taxa de ausência de amostras de 0,05, uma frequência alélica menor (MAF) de 0,01, um limiar de equilíbrio de Hardy-Weinberg (HWE) de 5e-6 e um corte de desequilíbrio de ligação (LD) de 80.

Neste estudo, foi realizado o CQ usando várias combinações de parâmetros, e os conjuntos de dados resultantes foram comparados para avaliar o impacto de cada parâmetro na qualidade dos dados. A qualidade dos conjuntos de dados foi avaliada usando diferentes métricas, como a porcentagem de dados ausentes, o número de amostras e SNPs após o CQ, e o desvio do equilíbrio de Hardy-Weinberg esperado. Posteriormente, o impacto do CQ em análises subsequentes foi avaliado, como análise de componentes principais e análise de associação genômica ampla.

Os filtros listados na Tabela 5 foram aplicados com sucesso aos dados usando o conjunto de ferramentas PLINK e um script em Python que automatizou o processo, resultando na criação de 144 arquivos de dados filtrados. Cada um desses arquivos de dados foi submetido a análises GWAS neste estudo.

Para usar o PLINK, os dados devem estar formatados em um dos vários formatos de entrada padrão aceitáveis, incluindo arquivos .vcf, binários ou de texto. O conjunto de dados ADNI foi pré-formatado em formato binário.

Tabela 5 – Parâmetros para garantir o controle de qualidade

Hiperparâmetros de Controle de Qualidade	
Nomes dos Hiperparâmetros	Valores
MAF	0,01, 0,05, 0,1
LD	0,9, 0,85, 0,8
HWE	$0,5e-6$, $10e-8$
GENO	0,02, 0,05, 0,1, 0,2
AMOSTRA	0,05, 0,1

Nota: Esta tabela apresenta o dicionário de hiperparâmetros que foram usados para filtrar os SNPs. Cada chave no dicionário corresponde a um hiperparâmetro específico, e o valor associado é uma lista de valores de filtro potenciais para esse hiperparâmetro. A coluna 'Valores' lista os valores de filtro potenciais para cada hiperparâmetro na forma de uma lista.

3.4 Análise exploratória dos dados

3.4.1 ANOVA

Adicionalmente, durante este estudo, os efeitos das características demográficas sobre o fenótipo foram analisados para verificar se seriam usados como parâmetros nos modelos. As características demográficas incluem informações sobre os indivíduos, como raça, gênero, etnia e idade. Se relevantes, esses dados seriam usados para treinar os modelos e melhorar a seleção de SNPs. Para analisar o impacto dessas características, um modelo linear foi utilizado para calcular a Análise de Variância (ANOVA) e obter a importância de cada variável demográfica.

A ANOVA é um teste estatístico que determina se as correlações observadas em uma amostra podem ser generalizadas para toda a população. Portanto, os resultados deste teste indicam se é razoável tirar conclusões com base na amostra e se é válido usar determinadas informações demográficas para treinar o modelo. Diante disso, antes de incorporar as variáveis demográficas nos modelos, foi examinado sua significância por meio da ANOVA.

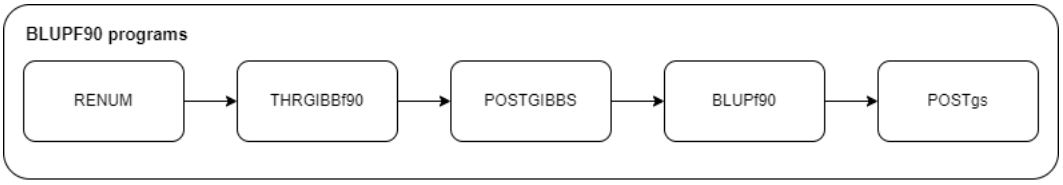
3.4.2 GWAS utilizando a família de programas BLUPF90

O BLUPF90 é uma família de programas projetada para o cálculo de modelos lineares mistos com foco em aplicações de melhoramento. Este software, que é comumente utilizado por bioinformatas, foi utilizado neste estudo para realização de GWAS e para comparação com os modelo de Aprendizado de Máquina. O BLUPF90 oferece um conjunto de funcionalidades, incluindo estimativa de variância usando vários métodos, a estimativa dos Melhores Estimadores Linearmente Não Viesados (BLUEs) e Preditores (BLUPs) para grandes conjuntos de dados, cálculo da acurácia em nível individual, permite a

utilização de informações de pedigree, pois é comumente usado na área de melhoramento genético, para estimar o mérito genético e conduzir estudos de associação genômica ampla (GWAS) (Misztal et al., 2015). O programa BLUPF90 foi desenvolvido para analisar grandes conjuntos de dados para aplicações de melhoramento com alta performance e sem necessidade de programação.

Neste estudo foram usados um subconjunto de programas dentro da família de programs BLUP, sendo eles: RUNUM, THRGIBBS1F90, POSTGIBBf90, BLUPf90 e POSTgs. Construímos um pipeline utilizando esses programas, conforme mostrado na Figura 6. A seguir, uma descrição de cada uma das aplicações que foram utilizadas.

Figura 6 – Pipeline de execução dos programas BLUPF90



Elaborado pela própria autora.

1. RUNUM é usado para gerar o arquivo de parâmetros (runf90.par) conforme mostrado na Figura 7. Um arquivo de parâmetros inclui o modelo estatístico, parâmetros e opções para executar os programas no pipeline. Os parâmetros e opções são descritos no manual do BLUPF90[2].

2. THRGIBBf90 executa as amostras da cadeia de Markov Monte Carlo para a estimativa dos componentes de variância dos Modelos Lineares Mistos (MLM) de um modelo de limiar, utilizado para variáveis de resposta categóricas. Os MLM diferem dos modelos lineares gerais (GLM) ao modelar termos de efeito aleatório além do termo residual. Assim, o componente aleatório ε, com variância $V[\varepsilon] = R$, do GLM (Equação 3.1) é estendido para $\varepsilon = Zu + e$, com variância $V[\varepsilon] = ZGZ' + R$. Portanto, a estimativa do componente de variância é uma operação chave dos MLMs.

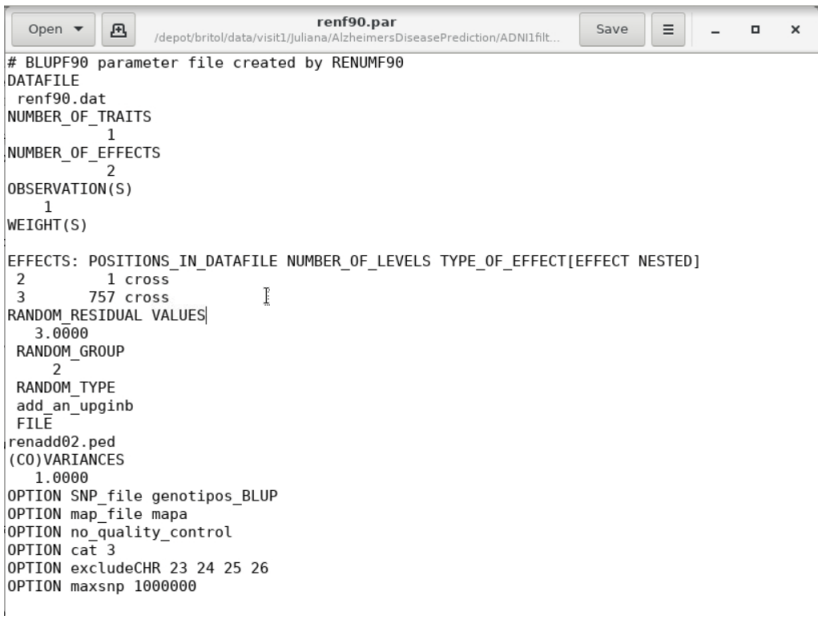
$$y = X\beta + \varepsilon \tag{3.1}$$

$$y = X\beta + Zu + e \tag{3.2}$$

Este método utiliza a Amostragem de Gibbs (GS) para a amostragem MCMC dos componentes de variância (Casella; George, 1992).

[2] http://nce.ads.uga.edu/wiki/lib/exe/fetch.phpmedia=blupf90_all8.pdf

Figura 7 – Cartão de parâmetros (arquivo par) usado para especificar os parâmetros e opções para executar a família de programas BLUP



Elaborado pela própria autora.

3. POSTGIBBSF90 resume as amostras MCMC para gerar os estimadores da média posterior dos componentes de variância ($\hat{\sigma_u^2}$ e $\hat{\sigma_e^2}$).

4. BLUPF90 utiliza os componentes de variância de POSTGIBBSF90 para ajustar o modelo genômico (3.2) para inferir o mérito genético dos indivíduos. Aqui, isso corresponde à probabilidade dos indivíduos caírem em uma das três categorias fenotípicas (ou seja, CN, CCL e DA).

5. POSTgs é um procedimento pós-hoc para extrair soluções de SNP dos méritos genéticos, conforme mostrado em 3.3. Subsequentemente, ele calcula os valores-p para todos os SNPs. A significância dessas associações permite comparar o impacto do controle de qualidade dos dados para os vários cenários considerados.

$$\hat{\beta} = MK^{-1}\hat{u} \tag{3.3}$$

O valor-p é a probabilidade do efeito de SNP deregressado estimado (α) dado que a hipótese nula (H_0) é verdadeira, assim $p(\alpha_j|H_0)$. Para o j^{simo} SNP, isso é computado de acordo com a Equação 3.4 (Aguilar, 2019).

$$p(\alpha_j|H_0) = 2(1 - \Phi\frac{|\alpha_j|}{\sigma_{\alpha_j}}) \tag{3.4}$$

Os efeitos de SNP deregressados (α) e seu desvio padrão são estimados a partir de β reescalado pelos elementos diagonais da equação do lado esquerdo utilizada para

resolver a equação do modelo misto. As equações são descritas por Aguilar et al. (2019).

3.5 Algoritmos utilizados

Nesta seção será discutido como os o treinamento dos modelos de Aprendizado de Máquina foram implementados.

Vale salientar que todos os modelos foram treinados utilizando um conjunto de treinamento (`X_train, y_train`) com 80% dos dados rotulados como DA e CN, e internamente validado usando a técnica de validação cruzada com 5 folds. O RandomizedSearchCV avaliou diferentes combinações de hiperparâmetros, utilizando o scorer F1 para a otimização. Para evitar advertências relacionadas à divisão por zero na pontuação F1, o valor de `zero_division` foi definido como 1. Em seguida o modelo foi avalido externamente no conjunto de teste (`X_test, y_test`).

3.5.1 Regressão Logística

Para encontrar o modelo de Regressão logística com maior robustez, os hiperparâmetros da `RandomizedSearchCV` incluíram o tipo de penalidade (`penalty`) com opções 'l1', 'l2' e 'elasticnet'; o parâmetro de regularização (`C`) variando de $1e-5$ a $1e2$; e a proporção de 'l1' em 'elasticnet' (`l1_ratio`) com valores 0, 0.5 e 1. O algoritmo 'saga' foi escolhido como o solver devido à sua compatibilidade com todos os tipos de penalidades. A otimização focou principalmente na pontuação F1, devido à sua importância na avaliação do equilíbrio entre precisão e recall.

Os hiperparâmetros ajustados na Regressão Logística incluíram:

- `penalty`: tipos de regularização para evitar o sobreajuste, incluindo 'l1', 'l2' e 'elasticnet'.
- `C`: intervalo do parâmetro de regularização, explorado de $1e^{-5}$ a $1e^{2}$.
- `solver`: definido como 'saga', que é compatível com todos os tipos de penalidades.
- `l1_ratio`: usado somente com 'elasticnet', com valores testados de 0, 0.5 e 1.

```
log_reg_rs = RandomizedSearchCV(LogisticRegression(random_state=42, max_iter=1000),
    param_grid, scoring=f1_scorer, n_jobs=-1, cv=5, verbose=10, random_state=42)
```

3.5.2 Random Forest

Para a Random Forest os hiperparâmetros ajustados incluíram o número de árvores (`n_estimators`) com valores de 250, 500 e 1000; a profundidade máxima das árvores

(`max_depth`) com valores 7, 9, 11 e 13; e o número máximo de recursos (`max_features`) considerados em cada divisão, testando valores relativos à 1%, a raiz quadrada e 10% e 50% do número total de SNPs.

```
rf_rs = RandomizedSearchCV(RandomForestClassifier(random_state=42),
    param_grid, scoring=f1_scorer, n_jobs=-1, cv=5, verbose=10, random_state=42)
```

3.5.3 XGBoost

Para o ajuste dos hiperparâmetros do XGBoost foram considerados:

- `n_estimators`: Número de árvores a serem construídas (100, 250, 500).
- `learning_rate`: Taxa de aprendizado para ajustar os pesos de cada árvore (0.01, 0.05, 0.1).
- `max_depth`: Profundidade máxima de cada árvore (3, 5, 7, 9).
- `subsample`: Fração dos dados a serem usados em cada árvore (0.5, 0.7, 0.9).
- `colsample_bytree`: Fração das colunas usadas em cada árvore (0.5, 0.7, 0.9).
- `gamma`: Parâmetro de regularização para controle de overfitting (0, 0.1, 0.2).

```
xgb_rs = RandomizedSearchCV(XGBClassifier(use_label_encoder=False,
    eval_metric='logloss', random_state=42), param_grid, scoring=f1_scorer,
    n_jobs=-1, cv=5, verbose=10, random_state=42)
```

3.5.4 Abordagem de Data Augmentation

A técnica de Data Augmentation foi realizada como uma tentativa complementar a análise, sendo feita através da geração de dados sintéticos. Seu objetivo é aumentar o número de amostras, no contexto deste trablho, em 400 unidades, no conjunto de dados, para diminuir a problema de alta dimensionalidade. Este processo consistiu em selecionar aleatoriamente pares de amostras do conjunto de treinamento e criar novas amostras pela média dos SNPs (Single Nucleotide Polymorphisms) desses pares, arredondando os resultados para obter valores discretos de 0, 1 ou 2. O número de amostras sintéticas geradas foi igual ao número de amostras no conjunto de treinamento original.

3.6 Métricas de avaliação

Em um problema de classificação binária, as classes são frequentemente designadas como classe positiva e classe negativa. Durante o processo de predição, é possível quantificar o número de amostras positivas e negativas que foram corretamente identificadas

(respectivamente, TP, ou *True Positives*, e TN, ou *True Negatives*), bem como aquelas que foram incorretamente classificadas (respectivamente, FP, ou *False Positives*, e FN, ou *False Negatives*).

Baseando-se nestes valores, duas métricas cruciais podem ser calculadas: precisão e revocação. A precisão (Equação 3.5) reflete a proporção de amostras corretamente identificadas como positivas em relação ao total de amostras classificadas como positivas. A revocação (Equação 3.6), por outro lado, indica a fração de amostras verdadeiramente positivas que foram corretamente classificadas como tal (Faceli et al., 2011, p. 164-165).

$$\text{Precisão} = \frac{TP}{TP + FP} \tag{3.5}$$

$$\text{Revocação} = \frac{TP}{TP + FN} \tag{3.6}$$

No contexto deste trabalho, a métrica principal utilizada é o F_1 *score* (Equação 3.7), definido como a média harmônica entre precisão e revocação. Este valor permite a integração dessas duas métricas importantes em um único indicador.

$$F_1 = 2 \cdot \frac{\text{Precisão} \cdot \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{3.7}$$

4 Experimentos e resultados

Este capítulo apresenta os resultados obtidos com os experimentos realizados seguindo a metodologia descrita no Capítulo 3.

4.1 ANOVA

O método ANOVA foi aplicado para investigar a relevância estatística da inclusão de variáveis demográficas como Idade, Raça, Etnia e Gênero nos modelos preditivos para a doença de Alzheimer. A hipótese inicial postulava uma associação significativa, pelo menos da variável Idade, com a incidência da doença. Os resultados da ANOVA são apresentados na Tabela 6. Com base nos resultados obtidos, observa-se que os p_value são consistentemente superiores ao limiar de 0,05, indicando que essas variáveis demográficas não constituem indicadores significativos para a previsão da probabilidade de ocorrência da doença de Alzheimer.

Tabela 6 – ANOVA: Dados Demográficos

	Graus de Liberdade	**Valor-p**
Gênero	1.0	0.87390
Raça	4.0	0.35020
Etnia	2.0	0.60523
Idade	1.0	0.47378

4.2 GWAS

4.2.1 Hiperparâmetros de controle de qualidade

Esta seção destaca o impacto da variação dos filtros estatísticos no controle de qualidade. Neste contexto, os conjuntos de dados sofreram cortes que serão expostos juntamente com variações nos valores de significância em cada *rodadas*.

Nos conjuntos de dados gerados, o resultado mais favorável foi observado no conjunto caracterizado por uma frequência alélica menor (MAF) de 0,01, um limiar de desequilíbrio de ligação (LD) de 85, uma significância de equilíbrio de Hardy-Weinberg (HWE) de 5e-6, uma taxa de ausência de amostra de 0,01 e uma taxa de ausência genética de 0,1. Notavelmente, o resultado mais significativo foi obtido para o SNP rs7918269, com um valor-p de 1.600000e-09, sendo este mais de 10 vezes maior que o segundo melhor valor encontrado. As figuras subsequentes ilustram o impacto da variação dos hiperparâmetros de QC e como o melhor resultado encontrado pode representar um ponto ótimo.

Figura 8 – Impacto da variação dos hiperparâmetros de QC: MAF e LD.

(a) Variação dos log(p_value) por Frequência Alélica Menor

(b) Variação dos log(p_value) por Desequilíbrio de Ligação

Figura 9 – Impacto da variação dos hiperparâmetros de QC: HWE e Ausência de Amostra.

(a) Variação dos log(p_value) por Equilíbrio de Hardy-Weinberg

(b) Variação dos log(p_value) por Ausência de Amostra

Figura 10 – Variação dos log(p_value) por Ausência Genética

*Todas as figuras de impacto da variação dos hiperparâmetro foram elaboradas pela própria autora

4.2.2 Análise de significância dos SNPs

Para analisar e avaliar cada *rodadas*, criou-se uma tabela contendo os 1000 SNPs mais significativos de cada *rodada*. Conforme mostrado na Tabela 7, para cada *rodadas*existem duas colunas. A primeira coluna representa o Nome do SNP e a segunda coluna representa seu valor-p. Desta forma, cada *rodadas*é ordenada do menor para o maior valor-p.

Essa representação de tabela para as *rodadas* simplifica o acesso a cada conjunto de dados das *rodadas*. Além disso, facilita a analisar qual conjunto de dados tem o maior valor de P_Value, a comparar se o SNP de maior significância que aparece em uma rodada aparecerá em alta posição nas outras *rodadas*, e quais SNPs aparecem mais vezes entre os SNPs mais bem classificados com valor de significância em todas as *rodadas*.

4.3 Modelos de Aprendizado de Máquina

4.3.1 Seleção de Hiperparâmetros

4.3.1.1 Random Forest

Os hiperparâmetros selecionados para o Random Forest foram: número de estimadores (n_estimators) = 250, número máximo de características (max_features) = 1273 e profundidade máxima (max_depth) = 11.

4.3.1.2 Regressão Logística

Para a Regressão Logística, os hiperparâmetros selecionados foram: força da regularização (C) = 0.19185373703841915, proporção L1 (l1_ratio) = 0.5, penalidade (penalty) = elasticnet e solucionador (solver) = saga.

4.3.1.3 XGBoost

No caso do XGBoost, os hiperparâmetros escolhidos foram: taxa de subsample = 0.9, número de estimadores (n_estimators) = 250, profundidade máxima (max_depth) = 9, taxa de aprendizado (learning_rate) = 0.1, gamma = 0.2 e colsample_bytree = 0.7.

4.3.2 Análise de Desempenho

4.3.2.1 Random Forest

O modelo Random Forest apresentou uma Acurácia de 57.69%, com uma Precisão de 100%, Recall de 5.71% e F1 Score de 10.81%. Apesar da alta Precisão, o baixo Recall sugere uma tendência do modelo em prever a classe negativa, limitando sua capacidade de identificar corretamente a classe positiva.

Tabela 7 – Ilustração dos resultados: p_value dos SNPs classificados para cada conjunto de parâmetros

	rodada 1		**rodada 2**		...	**rodada 144**	
	rodada1 - Nome do SNP	rodada1 - Valor-p do SNP	rodada2 - Nome do SNP	rodada2 - Valor-p do SNP	...	rodada144 - Nome do SNP	rodada144 - Valor-p do SNP
0	rs7918269	1.600000e-09	rs969329	1.130000e-08	...	rs2901028	4.730000e-08
1	rs7568837	7.485000e-07	rs7948482	1.911300e-06	...	rs7109945	1.587000e-06
2	rs17086056	1.421700e-06	rs4723295	2.079300e-06	...	rs2041610	1.698900e-06
3	rs2313048	1.605300e-06	rs13035707	2.792800e-06	...	rs1947582	3.836100e-06
4	rs3847701	1.605300e-06	rs6827843	3.339900e-06	...	rs7599284	3.949600e-06
...	...	...	...	...	...	...	...
996	rs3914966	1.041643e-03	rs11902906	1.195012e-03	...	rs2245197	1.007786e-03
997	rs12451379	1.044560e-03	rs2724157	1.195477e-03	...	rs9365499	1.007786e-03
998	rs12451379	1.044578e-03	rs3759	1.196857e-03	...	rs1035496	1.007786e-03
999	rs753002	1.045522e-03	rs2515032	1.197187e-03	...	rs2622927	1.009046e-03

4.3.2.2 Regressão Logística

A Regressão Logística também alcançou uma Acurácia de 57.69%, com Precisão de 60%, Recall de 17.14% e F1 Score de 26.67%. Este modelo demonstrou um equilíbrio um pouco melhor entre Precisão e Recall em comparação com o Random Forest.

4.3.2.3 XGBoost

O modelo XGBoost teve uma Acurácia de 53.85%, com Precisão de 47.62%, Recall de 28.57% e F1 Score de 35.71%. Este modelo apresentou um desempenho mais equilibrado entre as métricas, mas com uma Acurácia geral menor.

4.3.3 Comparação dos modelos

Tabela 8 – Comparação de Desempenho nos Dados de Teste

Modelo	Acurácia	Precisão	Recall	F1 Score
Random Forest	57.69%	100%	5.71%	10.81%
Regressão Logística	57.69%	60%	17.14%	26.67%
XGBoost	53.85%	47.62%	28.57%	35.71%

Desta forma os resultados indicam que, embora a Acurácia dos modelos seja similar, existem diferenças significativas em termos de Precisão, Recall e F1 Score. O modelo Random Forest, apesar de sua alta Precisão, apresenta um baixo Recall, sugerindo uma tendência a prever predominantemente a classe negativa (CN). A Regressão Logística e o XGBoost mostram um equilíbrio melhor entre Precisão e Recall, com o XGBoost apresentando um desempenho ligeiramente mais balanceado. No entanto, todos os modelos mostram espaço para melhorias, particularmente no que diz respeito ao equilíbrio entre as métricas de desempenho.

4.3.4 Abordagem de Data Augmentation

Após a aplicação de Data Augmentation, a melhor combinação de hiperparâmetros para a Regressão Logística encontrada foi: 'C' = 0.015577217702693031, 'l1_ratio' = 0.7, 'penalty' = 'l2', 'solver' = 'saga'. Os resultados de desempenho nos dados de teste foram os seguintes:

- Acurácia: 58.97%
- Precisão: 66.67%
- Recall: 17.14%
- F1 Score: 27.27%

Comparando estes resultados com aqueles obtidos sem Data Augmentation (Acurácia de 57.69%, Precisão de 60%, Recall de 17.14% e F1 Score de 26.67%), observa-se uma leve melhoria na Acurácia e na Precisão. No entanto, o Recall e o F1 Score permaneceram praticamente inalterados.

A aplicação de Data Augmentation resultou em uma pequena melhoria no desempenho do modelo de Regressão Logística, especialmente em termos de Acurácia e Precisão. Esta melhoria pode ser atribuída ao aumento na variabilidade e no volume dos dados de treinamento, permitindo ao modelo aprender com um conjunto de dados mais rico e potencialmente reduzir o overfitting.

No entanto, o impacto no Recall e no F1 Score foi limitado, sugerindo que, enquanto a Data Augmentation pode ajudar o modelo a identificar melhor a classe positiva, não houve uma melhora significativa na sua capacidade de classificar corretamente todas as instâncias positivas. Este resultado indica que, embora a Data Augmentation seja uma técnica promissora, pode ser necessário explorar outras abordagens ou combinações de técnicas para alcançar melhorias substanciais em todas as métricas de desempenho.

Apesar da inclusão da Data Augmentation no pipeline ter demonstrado um impacto positivo, na performance do modelo, não é possível afirmar que a melhoria foi resposta somente a esse fato. Portanto, por ser um processo muito mais custoso no treinamento, essa técnica não entrou para o pipeline principal.

4.3.5 Análise Comparativa dos SNPs Associados à Doença de Alzheimer

A investigação de marcadores genéticos associados à Doença de Alzheimer foi realizada mediante a aplicação de diversas técnicas de aprendizado de máquina e um estudo de associação genômica ampla (GWAS). Os SNPs mais significativos foram identificados para cada modelo, e suas importâncias foram contrastadas para aferir consistências e discrepâncias entre os métodos. Esta análise foi conduzida sobre os resultados de cada modelo que podem ser consultados na Tabela 9.

A Regressão Logística destacou SNPs com coeficientes negativos e positivos, sugerindo a presença de alelos protetores e de risco para a DA. Já o modelo Random Forest atribuiu maior importância a um conjunto diferente de SNPs, com *feature importances* variando na ordem de 10^{-4}. O modelo XGBoost identificou uma terceira coleção de SNPs, com importâncias na faixa de 10^{-2}, enfatizando a relevância de cada SNP de maneira mais uniforme. Por fim, o GWAS utilizando o BLUPF90 forneceu um conjunto de SNPs baseados em log(p_value) extremamente significativos, alguns dos quais não foram detectados pelos outros modelos.

A comparação dos resultados com a literatura evidencia que alguns dos SNPs identificados estão em loci já associados à DA, enquanto outros representam descobertas

potencialmente novas, demandando validação adicional (Smith et al., 2019). Por exemplo, o SNP rs7535533, que demonstrou um valor de importância extremamente significativo no modelo XGBoost, e seu gene, RGSL1, foi previamente associado à DA em múltiplos estudos (article, 2019a; article, 2007; article, 2019b) , sugerindo um papel potencial em vias biológicas relacionadas à patogênese da doença. Outros SNPs, como rs13026208, em que seu gene, GALNT13, também mostrou associações em pesquisas anteriores(MAP-AD, 2023), reforçando sua relevância para o fenótipo de interesse

4.3.5.1 Tabela de SNPs Significativos

A Tabela 9 compara os SNPs mais significativos identificados em cada modelo. A sobreposição e as diferenças entre os modelos são críticas para futuras investigações, podendo revelar novas vias biológicas envolvidas na patogênese da DA ou confirmar a relevância de caminhos já conhecidos.

Tabela 9 – Comparação dos 20 SNPs mais significativos entre os modelos de Regressão Logística, Random Forest, XGBoost e GWAS (BLUPF90)

Rank	Regressão Logística		Random Forest		XGBoost		GWAS (BLUPF90)	
	SNP	Coeficiente	SNP	Importância	SNP	Importância	SNP	P-Value
1	rs3008922	-2.15e-02	rs861750	9.35e-04	rs13026208	1.05e-02	rs7918269	1.6e-09
2	rs1936770	-1.82e-02	rs11709715	8.04e-04	rs7752155	1.03e-02	rs7568837	7.485e-07
3	rs5931572	-1.71e-02	rs7810927	7.70e-04	rs7535533	1.00e-02	rs17086056	1.4217e-06
4	rs4074535	1.67e-02	rs1980449	7.16e-04	rs2045325	9.32e-03	rs2313048	1.6053e-06
5	rs3923971	1.64e-02	rs409430	6.79e-04	rs10126412	8.89e-03	rs3847701	1.7123e-06
6	rs11071021	-1.56e-02	rs9573808	6.78e-04	rs6507641	8.73e-03	rs3763619	2.1707e-06
7	rs6636902	-1.49e-02	rs629081	6.69e-04	rs6037744	8.40e-03	rs501596	5.0406e-06
8	rs4517051	1.48e-02	rs3781556	6.67e-04	rs10769990	8.25e-03	rs2454568	5.3074e-06
9	rs573015	1.45e-02	rs255125	6.36e-04	rs17631450	8.13e-03	rs12899040	9.3628e-06
10	rs10847292	-1.43e-02	rs13421115	6.16e-04	rs17042385	8.11e-03	rs2084148	9.4484e-06
11	rs1893163	-1.42e-02	rs2123381	6.15e-04	rs7523907	8.09e-03	rs10445975	1.01714e-05
12	rs343494	-1.42e-02	rs11042834	6.12e-04	rs10870468	7.47e-03	rs17061864	1.23179e-05
13	rs5932828	1.41e-02	rs4910752	6.09e-04	rs10870468	7.47e-03	rs7959451	1.26233e-05
14	rs9573808	1.41e-02	rs1569660	5.86e-04	rs5908533	6.98e-03	rs2418495	1.28885e-05
15	rs10095724	1.37e-02	rs3920209	5.58e-04	rs4429270	6.94e-03	rs17690887	1.58476e-05
16	rs1517955	1.37e-02	rs12507354	5.50e-04	rs10933234	6.94e-03	rs10145908	1.64487e-05
17	rs497511	-1.34e-02	rs2293336	5.44e-04	rs17842504	6.86e-03	rs10048757	1.71341e-05
18	rs5908533	-1.33e-02	rs2242160	5.36e-04	rs693754	6.76e-03	rs7026908	1.78216e-05
19	rs7880350	1.33e-02	rs1993116	5.26e-04	rs2235396	6.65e-03	rs2749907	1.84709e-05
20	rs12147012	-1.28e-02	rs7610060	5.26e-04	rs343043	6.20e-03	rs4147990	1.87379e-05

5 Conclusão

Este estudo apresentou uma abordagem inovadora e multidisciplinar para a identificação de marcadores genéticos associados à Doença de Alzheimer (DA), empregando diversas técnicas de aprendizado de máquina e análise genômica. Através da combinação de métodos como Random Forest, Regressão Logística, XGBoost, e GWAS via BLUPF90, foi possível identificar um conjunto abrangente de SNPs associados à DA, alguns dos quais já conhecidos na literatura científica e outros representando possíveis novas descobertas.

Além disso, este estudo também contou com desenvolvimento de um tratamento de dados que foi crucial para os resultados encontrados e mostraram a relêvância da escolha de bons filtros para o Controle de Qualidade dos dados.

Os resultados obtidos pelos modelos de Aprendizado de Máquina refletiram nuances importantes na caracterização genética da DA. Cada modelo apresentou um perfil distinto de SNPs mais significativos, evidenciando a complexidade e a heterogeneidade da doença. A variação nos coeficientes e importâncias de características, conforme identificado pelos modelos, ressalta a importância de uma abordagem integrada e multifacetada para a pesquisa em genética da DA.

Os achados deste estudo não apenas corroboram associações genéticas previamente estabelecidas, mas também abrem caminhos para novas investigações. A identificação de SNPs não previamente associados à DA sugere novas áreas de investigação, potencialmente revelando novos mecanismos biológicos envolvidos na patogênese da doença. Estes resultados destacam a importância de continuar a explorar o genoma humano com abordagens analíticas avançadas para melhor compreender doenças complexas como a DA.

Investigar a Doença de Alzheimer é complexo devido à sua natureza multifatorial, envolvendo variações genéticas, fatores ambientais e estilo de vida. Além disso, a heterogeneidade dos sintomas e a progressão variável da doença entre indivíduos dificultam a identificação de padrões claros apenas usando SNPs. Essa complexidade implica desafios significativos para o desenvolvimento de modelos preditivos precisos com Aprendizado de Máquina, limitando a capacidade de prever a doença de maneira abrangente e específica para cada paciente e encontrar a significância de cada SNP. Assim, embora este estudo tenha demonstrado avanços significativos, é importante reconhecer tais limitações, e considera-las em interpretações e aplicações futuras.

Pesquisas futuras poderiam se beneficiar da integração de conjuntos de dados mais amplos, abordagens computacionais mais avançadas e técnicas de validação rigorosas para aprimorar a compreensão dos mecanismos genéticos da DA.

Em resumo, este estudo destaca o poder do aprendizado de máquina e da análise genômica no avanço do conhecimento sobre a Doença de Alzheimer. Os resultados obtidos fornecem contribuições para o campo emergente da genômica da DA, oferecendo novas perspectivas para o desenvolvimento de estratégias terapêuticas e diagnósticas mais eficazes. As descobertas aqui apresentadas representam um passo significativo no caminho para um entendimento mais profundo da genética da DA e seu impacto na saúde humana.

Referências

ADITYA, P. *L1 and L2 Regularization.* 2018. https://medium.com/@aditya97p/l1-and-l2-regularization-237438a9caa6. Accessed: 2020-09-02. Citado na página 25.

AGUILAR, I. e. o. Título do artigo. *Genetics Selection Evolution*, BioMed Central, v. 51, n. 1, p. 34, 2019. Disponível em: https://gsejournal.biomedcentral.com/articles/10.1186/s12711-019-0469-3. Citado na página 35.

ALKHATEEB, A. et al. Hybrid blockchain platforms for the internet of things (iot): A systematic literature review. *Sensors*, v. 22, p. 1304, 02 2022. Citado na página 22.

ALZGENE - TOP RESULTS. 2011. http://www.alzgene.org/TopResults.asp. Accessed: 2020-04-21. Citado na página 15.

ANDERSON, C. A. et al. Data quality control in genetic case-control association studies. *Nature Protocols*, v. 5, p. 1564–1573, 09 2010. Citado na página 16.

ARTICLE, A. of the. Candidate genes for the late onset alzheimer disease in human chromosome 10. *ResearchGate*, 2007. Disponível em: https://www.researchgate.net/publication/244092784_Candidate_genes_for_the_late_onset_Alzheimer_disease_in_human_chromosome_10. Citado na página 45.

ARTICLE, A. of the. Early-onset molecular derangements in the olfactory bulb of tg2576 mice: Novel insights into the stress-responsive olfactory kinase dynamics in alzheimer's disease. *Frontiers in Neuroscience*, 2019. Disponível em: https://www.frontiersin.org/articles/10.3389/fnagi.2019.00141/full. Citado na página 45.

ARTICLE, A. of the. Genetic insights into alzheimer's disease. *Annual Review of Pathology: Mechanisms of Disease*, 2019. Disponível em: https://www.annualreviews.org/doi/full/10.1146/annurev-pathmechdis-012419-032551. Citado na página 45.

BERGSTRA, J.; BENGIO, Y. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, v. 13, p. 281–305, 2012. Citado na página 28.

BERTRAM, L. Next generation sequencing in alzheimer's disease. In: _____. *Systems Biology of Alzheimer's Disease*. New York, NY: Springer New York, 2016. p. 281–297. ISBN 978-1-4939-2627-5. Disponível em: https://doi.org/10.1007/978-1-4939-2627-5_17. Citado na página 15.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Citado 2 vezes nas páginas 26 e 27.

BREIMAN, L. et al. *Classification and regression trees*. [S.l.]: Wadsworth International Group, 1984. Citado na página 26.

BUENO, M. R. P. O PROJETO GENOMA HUMANO. *Bioética*, v. 5, p. 1–10, 2009. Citado na página 18.

CASELLA, G.; GEORGE, E. I. Explaining the gibbs sampler. *The American Statistician*, Taylor Francis, v. 46, n. 3, p. 167–174, 1992. Citado na página 34.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, ACM, p. 785–794, 2016. Citado na página 27.

EDWARDS, D. et al. What are snps? In: _____. *Association Mapping in Plants*. New York, NY: Springer New York, 2007. p. 41–52. ISBN 978-0-387-36011-9. Disponível em: https://doi.org/10.1007/978-0-387-36011-9_3. Citado 2 vezes nas páginas 14 e 15.

FACELI, K. et al. *Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina*. 1. ed. Rio de Janeiro: LTC, 2011. ISBN 978-85-216-1880-5. Citado na página 38.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, Institute of Mathematical Statistics, p. 1189–1232, 2001. Citado na página 27.

GéRON, A. *Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow*. [S.l.]: Alta books editora, O'REILLY, 2002. Citado 2 vezes nas páginas 19 e 20.

HO, D. et al. Machine learning snp based prediction for precision medicine. *Frontiers in Genetics*, v. 10, p. 267, 2019. ISSN 1664-8021. Disponível em: https://www.frontiersin.org/article/10.3389/fgene.2019.00267. Citado 2 vezes nas páginas 18 e 19.

HO, T. K. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 20, n. 8, p. 832–844, 1998. Citado na página 26.

HOERL, A. E.; KENNARD, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, v. 12, p. 55–67, 1970. Citado na página 22.

JAMES, G. et al. *An Introduction to Statistical Learning with Applications in R*. [S.l.]: Springer, 2013. Citado na página 28.

JAMES, G. et al. *An Introduction to Statistical Learning: With Applications in R*. [S.l.]: Springer Publishing Company, Incorporated, 2014. ISBN 1461471370. Citado 4 vezes nas páginas 21, 23, 24 e 25.

JOHNSON, S. G. Genomic Medicine in Primary Care," in Genomic and Precision Medicine. *Amsterdam: Elsevier Inc.*, v. 5, p. 1–18, 2017. Citado na página 18.

KARCH, C. M.; CRUCHAG, C.; GOATE, A. M. Alzheimer's disease genetics: from the bench to the clinic. *Neuron*, v. 83, p. 11–26, 2014. Citado na página 15.

LAKSMAN, Z.; DETSKY, A. S. Personalized medicine: understanding probabilities and managing expectations. *J. Gen. Intern. Med.*, v. 26, p. 204–206, 2011. Citado na página 18.

LI, P. *Logistic Regression*. 2015. https://www.stat.rutgers.edu/home/pingli/papers/Logit.pdf. Accessed: 2020-09-02. Citado na página 25.

LIAW, A.; WIENER, M. Classification and regression by randomforest. *R news*, v. 2, n. 3, p. 18–22, 2002. Citado na página 27.

MALO, N.; LIBIGER, O.; SCHORK, N. Malo n, libiger o, schork nj. accommodating linkage disequilibrium in genetic-association analyses via ridge regression. am j hum genet 82: 375-385. *American journal of human genetics*, v. 82, p. 375–85, 03 2008. Citado na página 17.

MAP-AD. *Gene GALNT13 - MAP-AD*. 2023. https://map-ad.org/gene/galnt13. Acessado em 25 de janeiro de 2024. Citado na página 45.

MATTA, T. A. "AVALIAÇÃO DO VALOR DE IMÓVEIS POR ANALISE DE REGRESSÃO: UM ESTUDO DE CASO PARA A CIDADE DE JUIZ DE FORA. *ufjf*, p. 06–15, 2007. Citado na página 22.

MISZTAL, I. et al. *Manual for BLUPF90 family of programs*. 2015. http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90_all2.pdf. Accessed: 2022-07-28. Citado na página 34.

MOHAN, G.; MAN, Z.; YANG, L. Genes associated with Alzheimer's disease:an overview and current status. *Clinical Interventions in Aging*, v. 11, p. 665–681, 2016. Citado 2 vezes nas páginas 15 e 16.

PROXY variable. 2020. https://www.oxfordreference.com/view/10.1093/oi/authority.20110803100351624. Accessed: 2020-04-23. Citado na página 22.

QUINLAN, J. R. Induction of decision trees. *Machine learning*, Kluwer Academic Publishers, v. 1, n. 1, p. 81–106, 1986. Citado na página 26.

RAILEANU, L. E.; STOFFEL, K. Theoretical comparison between the gini index and information gain criteria. *Annals of Mathematics and Artificial Intelligence*, Springer, v. 41, n. 1, p. 77–93, 2004. Citado na página 26.

SERENIKI, A.; VITAL, M. A. B. F. A doença de Alzheimer: aspectos fisiopatológicos e farmacológicos. *Revista de Psiquiatria do Rio Grande do Sul*, scielo, v. 30, p. 0 – 0, 00 2008. ISSN 0101-8108. Citado na página 14.

SMITH, R. et al. Genetic insights into the mechanisms of alzheimer's disease. *Journal of Neurogenetics*, Taylor Francis, v. 33, n. 3, p. 73–83, 2019. Citado na página 45.

SPIEGEL, A. M.; HAWKINS, M. "Personalized medicine" to identify genetic risks for type 2 diabetes and focus prevention: can it fulfill its promise?. *Health Aff.*, v. 31, p. 43–49, 2012. Citado na página 18.

TANZI, R. E.; BERTRAM, L. Twenty years of the Alzheimer's disease amyloid hypothesis: a genetic perspective. *Cell*, v. 7, p. 545–555, 2005. Citado na página 14.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, v. 58, p. 267–288, 1996. Citado na página 22.

UFPR. *Regularização - Laboratório de Estatística e Geoinformação - LEG/UFPR*. xx. http://cursos.leg.ufpr.br/ML4all/apoio/Regularizacao.html#. Accessed: 2020-09-02. Citado na página 25.

WORLD Alzheimer Report. 2015. https://www.alz.co.uk/research/WorldAlzheimerReport2015.pdf. Accessed: 2020-04-21. Citado 2 vezes nas páginas 12 e 14.

ZENG, P. et al. Statistical analysis for genome-wide association study. *Journal of biomedical research*, v. 29, p. 285–97, 07 2015. Citado na página 17.