Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Modelos de classificação combinados a redes bipartidas em dados da NFL

Felipe Baptistão Durante Molina

Trabalho de Conclusão de Curso

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Modelos de classificação combinados a redes bipartidas em dados da NFL

Felipe Baptistão Durante Molina
Orientadora: Prof^a Dr^a Andressa Cerqueira

Trabalho de Conclusão de Curso apresentado como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos
Fevereiro de 2025

Federal University of São Carlos
Exact and Technology Sciences Center
Department of Statistics

Classification models combined with bipartite networks in NFL data

Felipe Baptistão Durante Molina
Advisor: Prof[a] Dr[a] Andressa Cerqueira

Bachelors dissertation submitted to the Department of Statistics, Federal University of São Carlos - DEs-UFSCar, in partial fulfillment of the requirements for the degree of Bachelor in Statistics.

São Carlos
February 2025

Felipe Baptistão Durante Molina

Modelos de classificação combinados a redes bipartidas em dados da NFL

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por Felipe Baptistão Durante Molina e aprovado pela banca examinadora.

Aprovado em 5 de dezembro de 2024.

Banca Examinadora:

- Prof^a Dr^a Andressa Cerqueira (Orientadora)
- Prof. Dr. Thomas Peron
- Prof. Dr. Danilo Lourenço Lopes

Agradecimentos

Primeiramente, gostaria de agradecer a Deus por toda a força em todos esses anos de caminhada, por toda companhia nos momentos em que estive sozinho, por ser meu guia nos momentos em que estive perdido e por todos pequenos milagres realizados em minha vida para que eu pudesse chegar até aqui.

Também gostaria de agradecer minha mãe, Giovana, por ser minha base e acreditar em mim desde sempre. Gostaria de agradecer ao meu pai, Rubens, por todo esforço e ajuda na minha jornada, e aos meus irmãos Bruno e João Francisco pelos momentos de companhia. Também agradeço minhas avó Ana Maria por tantas coisas que seria impossível escrever aqui, agradeço minha avó Juraci por todas as orações e aos meus tios Antônio e Gisele por toda a ajuda desde o início.

Ainda, agradeço a todos os meus amigos, seja da época da escola quanto da faculdade, por deixar a caminhada muito mais leve, pelos momentos de alegria e companhia e por me lembrarem das coisas da vida que valem a pena. Também gostaria de agradecer minhas orientadora, Andressa, pelos 3 anos de orientação, por me ajudar a evoluir nessa caminhada, e pela motivação para o futuro.

"Ser brilhante não é o suficiente, você tem que trabalhar. A inteligência não é um privilégio, é um dom, e deve ser utilizado para o bem da humanidade."

(Dr. Otto Octavius)

Resumo

Redes complexas tem sido vastamente utilizadas para retratar dados reais, descrevendo a interação entre objetos por meios de grafos, possuindo uma grande aplicabilidade na análise de dados em diversas áreas, como por exemplo em aeroportos, relações sociais, internet (world wide web) e esportes. Neste estudo, utilizando dados provenientes da Liga Nacional de Futebol Americano (NFL), as redes são utilizadas para representar o jogo de um certo time, representando a dinâmica dos passes durante uma partida e, com isso, utilizar essas informações para extrair resultados e realizar previsões sobre os jogos, como, no caso desse estudo, a probabilidade de campanha positiva em uma temporada. Umas das ferramentas utilizadas para a análise de redes complexas são as medidas de centralidade, as quais transformam informações sobre toda a rede em um único valor de acordo com determinada característica e, dessa forma, podem ser incorporadas em modelos de classificação por meio de covariáveis. Logo, medidas de centralidade serão utilizadas em modelos de classificação para tentar explicar o evento de campanha positiva de uma equipe em uma temporada na NFL.

Palavras-chave: *Redes Complexas, Modelos de Classificação, NFL.*

Abstract

Complex networks have been widely used to depict real data, describing the interaction between objects through graphs, with great applicability in data analysis across various areas, such as airports, social relations, the internet *World Wide Web*, and sports. In this study, using data from the National Football League (NFL), networks are employed to represent the gameplay of a specific team, illustrating the dynamics of passes during a match and utilizing this information to extract outcomes and make predictions about games, such as in this study, the probability of one obtain a positive campaign in a season. One of the tools used for the analysis of complex networks is global measures, which resumes information about the entire network into a single value based on a particular characteristic and can in this way be incorporated into classification models as covariates. Therefore, global measures were used in classification models to attempt to explain the outcome of positive campaign in a NFL season.

Keywords: *Complex Networks, Classification Models, NFL.*

Lista de Figuras

2.1 Representação gráfica das redes correspondentes às matrizes adjacência 1 e 2. 26
2.2 Representação gráfica das redes correspondentes à matriz adjacência 3. . . 27
2.3 Grafo bipartido com seus subconjuntos separados. 28
2.4 Representação de um ciclo par. 28

3.1 Rede passes completos - Minnesota Vikings - 2023 32
3.2 Rede passes para *touchdown* - Minnesota Vikings - 2023 32

5.1 Possíveis valores de $P(Y = 1|x, \beta)$ em função de η. 45
5.2 Esquema de uma árvore de regressão. 52
5.3 Regiões no espaço das covariáveis da árvore referente à Figura 5.2. 52

6.1 Frequência de valores (1 = Campanha Positiva, 0 = Caso Contrário). . . . 58
6.2 Boxplot para a medida de centralidade Força (1 = Campanha Positiva, 0 = Caso Contrário). 59
6.3 Boxplot para a medida de centralidade Assimetria (1 = Campanha Positiva, 0 = Caso Contrário). 59
6.4 Boxplot para a medida de centralidade Densidade (1 = Campanha Positiva, 0 = Caso Contrário). 60
6.5 Boxplot para a medida de centralidade Equidade (1 = Campanha Positiva, 0 = Caso Contrário). 60
6.6 Matriz de correlação para as diferentes medidas de centralidade. 62
6.7 Correlação das medidas dentro da rede de passes completos 63
6.8 Pontos de Alavanca e Distância de Cook. 67
6.9 Gráfico de índice por resíduo Componente do Desvio. 67
6.10 Gráfico de envelope simulado. 68

6.11 Árvore de Classificação para o cenário 1. 69
6.12 Árvore de Classificação para os cenários 2 e 3. 69
6.13 Árvore de Classificação para o cenário 3 (sem força td). 70
6.14 Medida de importância para o cenário 1. 71
6.15 Medida de importância para o cenário 2. 72
6.16 Medida de importância para o cenário 3. 72

Sumário

1 Introdução **19**

1.1 Objetivos 20

1.2 Organização do trabalho 21

2 Redes Complexas **23**

2.1 Definição 24

2.2 Representação Gráfica 25

2.3 Redes Bipartidas 27

3 Banco de Dados **29**

3.1 Construção da Rede 29

4 Medidas de Centralidade **35**

4.1 Força 36

4.2 Densidade de Ligação 36

4.3 Assimetria de força de interação 38

4.4 Equidade de interação 39

4.5 Outras medidas 40

4.6 Aplicação na Rede 41

5 Modelos de Classificação **43**

5.1 Regressão Logística 43

5.1.1 Definição Geral 43

5.1.2 Seleção de Variáveis 48

5.2 Árvores de Classificação 51

5.2.1 Definição Geral 51

5.2.2 Seleção de Variáveis 54

6 Aplicação **57**

6.1 Análise Descritiva 57

6.2 Regressão Logística 63

6.3 Árvores de Classificação 68

6.4 Floresta Aleatória 70

6.5 Conclusão 73

7 Conclusões Gerais **75**

Referências Bibliográficas **77**

Capítulo 1

Introdução

Redes complexas têm sido amplamente utilizadas recentemente para a modelagem de sistemas reais onde há a presença de componentes que se interligam de alguma maneira. Basicamente, as redes são conjuntos de objetos, chamados vértices, que se conectam devido à determinada característica ou evento, formando arestas entre si. Dessa forma, as redes são utilizadas para explorar padrões presentes nessas conexões, investigando como a estrutura da respectiva rede influencia em todo o sistema de estudo. Por exemplo, redes de aeroportos podem ser representadas por este sistema a fim de identificar sua evolução ao longo do tempo (Da Rocha, 2009), bem como redes de citações bibliográficas entre artigos científicos (Price, 2011), e até mesmo em rede de amizades entre alunos de uma escola de caratê (Zachary, 1977).

Dentre os inúmeros campos de aplicação das redes complexas, pode-se destacar o campo esportivo. Com o avanço recente de técnicas de análise de dados, esportes de alto nível começam a aplicar essas técnicas a fim de melhorarem o desempenho de seus atletas e equipes para obterem os resultados desejados. Dentre os meios utilizados, as redes complexas são utilizadas para representar a dinâmica de certas partidas, jogadas, ou relações extra-campo, como relações entre transferências de jogadores entre clubes (Félix *et al.*, 2019). No futebol, por exemplo, estudos modelam a dinâmica de passes entre jogadores em uma determinada partida por meio de redes complexas, a fim de estudar sua estrutura e melhorar o desempenho da equipe na partida (Echegoyen Blanco *et al.*, 2018).

Além das redes complexas, na área esportiva existe a presença de modelos de classificação que são bastante explorados para a análise de desempenho de equipes em partidas. Tais modelos são úteis para identificar algumas características denominadas de covariáveis

que possam explicar alguma outra característica de interesse, chamada de variável resposta. Além disso, estes métodos permitem realizar predições a respeito da resposta escolhida, sendo possível prever resultados, número de pontos, entre outras características a respeito de partidas. Dentre os métodos de classificação, técnicas mais tradicionais são utilizadas, como a regressão logística (Prasetio *et al.*, 2016), além de técnicas mais recentes como Naive Bayes (Rahman *et al.*, 2018) e árvores de decisão (de Stefano *et al.*, 2020).

Ainda, é possível relacionar os modelos de classificação com as redes complexas. Isso se dá ao transformar informações sobre a rede em covariáveis, ou seja, extrair características a respeito das redes e incorporá-las nos modelos de interesse (Xavier, 2024). Essa transformação se dá pelas medidas de centralidade, onde de acordo com certa característica de interesse, toda a rede é resumida a um único valor, o qual dará um indicativo a respeito dessa certa característica. Inúmeras medidas podem ser aplicadas de acordo com o tipo de rede e o objetivo do estudo em questão (Newman, 2018).

Um dos esportes em que é possível aplicar essa metodologia é o futebol americano, mais precisamente na Liga Nacional de Futebol (NFL). A NFL é atualmente a liga mais lucrativa e assistida do mundo, e por conta da natureza altamente estratégica do esporte, a presença de análises de dados é amplamente explorada nesse cenário. Podem-se destacar os modelos de classificação (Gifford e Bayrak, 2023) e até mesmo de redes complexas (de Oliveira Salim e Brandao, 2018), porém apenas com a análise de dados extra-campo, sem o foco na partida em si.

Dessa forma, o foco deste estudo consiste em propor e aplicar conjuntamente a metodologia de redes complexas e modelos de classificação em dados da NFL, sendo possível identificar quais variáveis impactam no resultado de uma partida e iniciar o estudo conjunto entre redes complexas e modelos de classificação no âmbito do futebol americano.

1.1 Objetivos

Em muitos esportes, modelos de classificação desempenham um papel importante ao explicarem resultados, desempenho, dentre outras características de times e atletas, além de realizar previsões a respeito dessas respostas. Para isso, inúmeros valores de covariáveis pertencentes a uma equipe, ou até mesmo a uma partida, são coletados, valores estes que serão incorporados ao modelos de classificação para a realização de futuras predições.

No entanto, este trabalho possui o objetivo de verificar se características presentes em redes complexas, mais precisamente nas redes que representam o jogo passado de uma equipe, podem ser utilizadas como ferramenta para auxiliar nas predições de resultados de jogos da NFL. Dessa maneira, as covariáveis não serão coletadas diretamente do banco de dados das partidas, mas sim serão obtidas através de características presentes na construção de redes complexas de todos os jogos do time em uma temporada, obtidas por meio de medidas de centralidade.

Logo, o objetivo do trabalho é identificar quais medidas de centralidade relacionadas a redes complexas anuais de cada time da NFL podem explicar o evento da equipe possuir campanha positiva (mais vitórias do que derrotas) ou negativa, já que esta é uma característica importante nas análises das equipes ao longo dos anos. Assim, ao final desse trabalho, o uso de redes complexas a respeito do jogo passado das equipes em dados da NFL será analisado através da significância das covariáveis presentes nos modelos de classificação neste conjunto de dados, obtendo evidências se a metodologia aqui proposta é de fato útil ou não nestes dados.

1.2 Organização do trabalho

Para o organização deste estudo, primeiramente uma visão geral sobre redes complexas será dada no Capítulo 2, a fim de expor principais conceitos a respeito do objeto principal deste trabalho. Após isso, o Capítulo 3 trará todo o processo da construção do banco de dados utilizado além de aplicação das redes complexas neste conjunto de dados. A respeito das covariáveis a serem utilizadas, o Capítulo 4 exibirá a explicação detalhada de cada uma destas medidas, com a tradução delas em termos do problema e sua lógica matemática. Ainda, o Capítulo 5 possui um visão aprofundada em todos os modelos de classificação que serão utilizados na aplicação a qual se apresenta no Capítulo 6.

Capítulo 2

Redes Complexas

Redes complexas compreendem uma sistema que consegue retratar interações entre objetos das mais variadas áreas disciplinares, como por exemplo da computação, biologia, ciências sociais, esportes, entre outras. Tais redes são construídas por meio de grafos nos quais seus vértices (também denominados nós) são os objetos de estudo e as arestas representam as conexões presentes entre os vértices de acordo com certa forma de interação pré definida entre eles. É de interesse, por meio da análise de redes complexas buscar compreender o sistema que está sendo modelado, estudando a estrutura e dinâmica presente no ambiente, bem como observar padrões de interesse, nós com certo grau de importância, formação de comunidades, dentre outras características de interesse (Newman, 2018).

A teoria dos grafos deu-se início no Século XVIII com um artigo publicado por Leonard Euller, matemático e físico suíço. Euller tratava do problema das Sete pontes de Königsberg, onde se discutia a possibilidade de atravessar todas as pontes sem repetir a passagem duas vezes sobre a mesma ponte. A prova, então, que tal feito não seria de fato possível originou a primeira ideia da teoria dos grafos, de tal modo que o matemático transformou as pontes em arestas e as quatro localidades em pontos, criando o primeiro traço de um grafo (Mata, 2020).

Com isso, ao longo dos anos, a utilização de redes complexas para representarem sistemas reais se tornou cada vez mais frequente. No campo das comunicações, PRICE (1965) desenvolveu uma rede de citações entre artigos publicados na época, onde cada interação correspondia à uma citação que um determinado artigo realizava de outro, e o estudo da estrutura desta rede tinha a finalidade de descobrir tópicos de interesse relacionados entre os autores. Ainda com relação ao estudo da estrutura das redes, há um destaque no campo das comunicações para a rede da World Wide Web (WWW), onde os

sites são os objetos de estudo com os *hyperlinks* entre eles sendo as conexões presentes na rede.

Além disso, uma importante rede no campo das relações sociais, a qual serviu de passo inicial para os estudos de agrupamento em redes complexas é o clube de caratê de Zachary (Zachary, 1977), onde procurava-se compreender as relações de amizade entre os alunos do clube, formando grupos daqueles que apresentavam um ciclo de amigos em comum. A biologia, também, é um campo o qual utiliza as redes complexas para realizar o estudo de cadeias alimentares entre espécies, identificando a relação entre presa e predador como uma conexão entre os seres presentes no ambiente, observando a estrutura e formação de comunidades presentes no sistema (Rejmanek e Starỳ, 1979).

Logo, as redes complexas podem ser utilizadas nos mais diversos campos de estudo, relacionando quaisquer tipos de objetos presentes em um sistema de acordo com seu tipo de conexão. Suas propriedades e medidas são de extrema importância para a identificação de padrões no ambiente, estudando a importância de seus objetos individualmente ou até mesmo em grupo.

2.1 Definição

Uma rede é representada por um grafo G pode ser definida por meio de um conjunto V contendo os seus vértices e um conjunto E contendo as arestas conectando os elementos do conjunto V, assim denotamos $G = (V, E)$. As conexões realizadas são representadas por meio de uma matriz $\mathbf{A}$ de dimensão $n \times n$ chamada matriz adjacência , em que n são o número de vértices presentes na rede.

Primeiramente, para redes **sem peso** temos uma matriz adjacência binária, apresentando apenas entradas 0 ou 1, indicando apenas a presença de conexão entre dois vértices. Logo, as entradas da matriz adjacência são definidas da seguinte maneira:

$$A_{ij} = \begin{cases} 1\,, & \text{se existe conexão entre os vértices } i \text{ e } j \\ 0\,, & \text{caso contrário} \end{cases} . \tag{2.1}$$

Para o caso **com peso**, no entanto, a matriz adjacência não possui apenas entradas binárias, mas sim valores de acordo com seu peso. O peso é uma quantificação da força de ligação entre dois vértices, podendo ser determinado de acordo com certa característica presente no ambiente representado pela rede, por exemplo, em uma rede aérea, o peso

entre dois aeroportos pode ser o número de voos realizados ou até mesmo o número de passageiros. Agora, a matriz adjacência será denotada por **W**, ainda sendo de dimensão $n \times n$, com a entrada W_{ij} sendo o peso da aresta a qual conecta o vértice i ao vértice j.

Outra definição presente nas redes complexas é o **direcionamento** das arestas. Em redes não direcionadas, não importa se a conexão é realizada do vértice i para o vértice j ou vice-versa, sendo que a matriz adjacência será sempre simétrica nesse caso ($A_{ij} = A_{ji}$ ou $W_{ij} = W_{ji}$). Já redes direcionadas, a aresta a ser formada entre os nós depende do vértice do início e do final, ou seja, pode existir uma conexão saindo do vértice i até o j mas não existindo uma ligação de j até i, fazendo com que a matriz adjacência não seja simétrica.

2.2 Representação Gráfica

A visualização de uma rede complexa, a fim de facilitar o entendimento do ambiente como um todo, é realizada por meio de grafos da seguinte maneira:

- Vértices: representados por pontos ou círculos, geralmente contendo alguma identificação próxima ao objeto. A disposição dos pontos no plano pode seguir alguma disposição real, como por exemplo a localização geográfica, ou pode ser feita de forma aleatória.

- Arestas: representadas por um seguimento de rede conectando dos nós no caso sem direção, ou por meio de uma seta quando a rede for direcionada. Pode apresentar cores e tamanhos distintos dependendo do peso entre os vértices, contendo ou não a identificação do seu valor.

Por exemplo, as matrizes adjacência 1 e 2 representando um grafo sem pesos:

Matriz 1:

$$\begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Matriz 2:

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Pelo fato da Matriz 1 ser não direcionada, temos uma matriz simétrica formada apenas por zeros e uns, pelo fato de não estarmos considerando a presença de peso nas arestas. Já a Matriz 2, pode-se notar a não simetria em suas entradas, indicando uma rede direcionada. A Figura 2.1 representada o grafos destas duas matrizes.

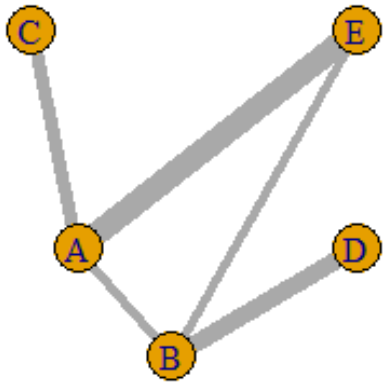
(a) Rede matriz adjacência 1 (b) Rede matriz adjacência 2

Figura 2.1: Representação gráfica das redes correspondentes às matrizes adjacência 1 e 2.

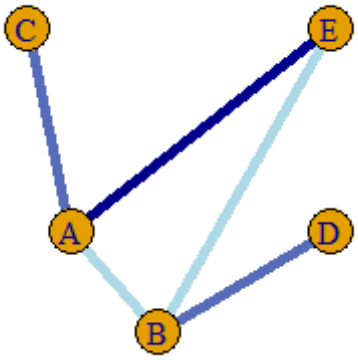
Ainda, caso as redes possuíssem pesos na arestas, a representação gráfica das redes com essa ponderação poderiam ser realizadas das seguintes formas: colocando o valor do peso sobre a aresta, mudando a espessura das arestas de acordo com os respectivos pesos, ou mudando as cores das arestas de acordo com algum padrão relacionado com o valor do peso. Por exemplo, seja uma matriz adjacência 3 da forma:

$$\begin{bmatrix} 0 & 1 & 3 & 0 & 8 \\ 1 & 0 & 0 & 5 & 1 \\ 3 & 0 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 & 0 \\ 8 & 1 & 0 & 0 & 0 \end{bmatrix},$$

logo, algumas possíveis representações para a rede em questão podem ser realizadas como descrito na Figura 2.2, onde podemos representar o peso mudando o tamanho das arestas ou aumentando gradativamente as cores das mesmas, respectivamente.



(a) Rede matriz adjacência 3 por espessura



(b) Rede matriz adjacência 3 por cor

Figura 2.2: Representação gráfica das redes correspondentes à matriz adjacência 3.

2.3 Redes Bipartidas

Dentre os diversos tipos de rede, há as chamadas redes bipartidas. Este tipo de rede provém, na teoria dos grafos, dos grafos bipartidos em que, por definição, seu conjunto de vértices pode ser dividido em outros dois subconjuntos de nós disjuntos, ou seja, um vértice pode pertencer a apenas um dos subconjuntos. Além disso, o conjunto de arestas não podem conectar nós de um mesmo subconjunto. Dessa forma, uma rede bipartida G que antes possuía apenas os conjuntos V e E de vértices e arestas, respectivamente, terá agora dois subconjuntos disjuntos U e V de nós, portanto $G = (V, U, E)$. Além de um novo subconjunto, a matriz adjacência de uma rede bipartida ainda será quadrada, mas de ordem $(m + n) \times (m + n)$, em que m é o número de vértices presentes no subconjunto V, e n é o número de vértices presentes no subconjunto U.

Para a identificação de uma rede bipartida, é possível dividir os vértices do grafo em questão em dois subconjuntos e, caso não exista a presença de nenhuma aresta dentro dos subgrupos, o grafo será bipartido, essa característica é denominada **bipartição**. Por exemplo, a Figura 2.3 representa um grafo arbitrário com oito nós sendo que é possível dividir a rede em dois conjuntos disjuntos $U = \{A, B, C\}$ e $V = \{D, E, F, G, H\}$, nos quais não existe a presença de arestas conectando os vértices de um mesmo grupo. Mesmo com essa divisão realizada nos nós da rede bipartida, podemos apresentar ainda redes com ponderamento em seus vértices, ou seja, redes com peso, e também a presença ou não de direcionamento.

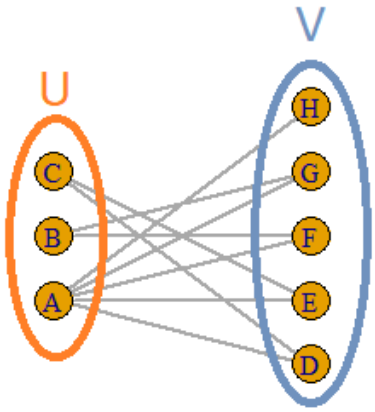


Figura 2.3: Grafo bipartido com seus subconjuntos separados.

Ainda, uma característica que deve estar presente em grafos bipartidos é apenas a presença de **ciclos pares**. O ciclo de um grafo se refere ao caminho realizado para sair de um determinado vértice e retornar ao mesmo sem repetir algum nó no caminho. Nas redes bipartidas, devido ao fato de não ser possível vértices de um mesmo grupo se conectarem, para que um ciclo seja realizado, obrigatoriamente devemos sair do subconjunto o qual pertence o primeiro nó, partir para o outro subgrupo e voltar ao original, realizando este passo-a-passo sem repetir os vértices até chegar no nó inicial. Logo, é impossível realizar esse caminho em um número ímpar de passos.

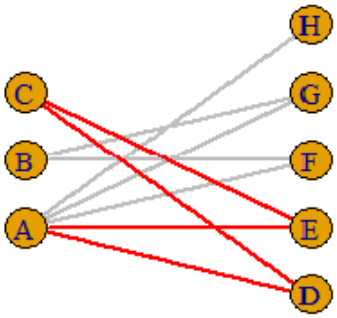


Figura 2.4: Representação de um ciclo par.

A Figura 2.4 mostra um exemplo de ciclo em um grafo bipartido. Neste caso, o ciclo iniciou-se no vértice C e seguiu o caminho C-E-A-D-C, sem repetir nenhum vértice no passo-a-passo. Pode-se notar que o ciclo em questão possui comprimento par de 4 arestas por se tratar de uma rede bipartida.

Capítulo 3

Banco de Dados

3.1 Construção da Rede

O banco de dados utilizado para o respectivo estudo foi extraído do pacote **nflfastR** (SebastianCarl, 2024) presente no repositório CRAN do software R (R Core Team, 2019). Neste pacote, inúmeros dados referentes a temporadas da NFL desde 1999 podem ser encontrados, bem como estatísticas a respeito dos jogos de um time, seus respectivos jogadores, o calendário de partidas, entre outras funções. Para este estudo, os dados "playbyplay" serão utilizados, dados estes que contemplam estatísticas jogada por jogada em todos os jogos de uma temporada da NFL.

A dinâmica de um jogo de futebol americano consiste em o time que possui a posse de bola atravessar o campo a fim de chegar à parte final do mesmo, chamada de *endzone*, anotando o chamado *touchdown*. Para isso, cada time possui quatro oportunidades para percorrer ao menos dez jardas, iniciando a contagem da linha de *scrimmage* até a linha de primeira descida, e, caso este feito não seja realizado, a bola é devolvida ao adversário, e caso a meta seja alcançada, o time possui novas quatro oportunidades para alcançar dez jardas. Para percorrer todo campo, é possível dividir as estratégias em dois tipos de jogadas: o jogo corrido, em que certo jogador tenta correr com a bola para ganhar a maior quantidade de jardas possíveis, e o jogo passado, onde um jogador, denominado *quarterback*, lança a bola para os demais a fim de obter as jardas necessárias (GE, 2017).

Devido ao fato da dinâmica das partidas envolver posses de bolas definidas, ou seja, em cada jogada um time detém o controle da bola, as análises são facilitadas neste caso, com o tipo de jogadas bem definidas e, com isso, os dados "playbyplay" possuem grande usabilidade para a análise de dados na NFL. Esta base de dados contém, em cada uma

das linhas, a jogada realizada por uma equipe em determinada posse de bola, contando com a presença de 372 variáveis para cada jogada, indicando por exemplo, a *id* da jogada, o time que a realizou, o tipo de jogada (passe ou corrida), o resultado da jogada, aplicação de faltas, quantas jardas de ganho, se o time anotou ou não pontos, dentre outras demais variáveis. Em média, por temporada, existem mais de 45.000 jogadas realizadas.

Para o uso da base de dados nas aplicações deste estudo, primeiramente foi realizado um tratamento na base de dados "playbyplay" devido à sua vasta quantidade de informações. Neste estudo, apenas as jogadas de passe serão consideradas para a construção da rede, além de que apenas partidas de temporada regular serão incluídas, sem a presença de jogos de pós temporada, a fim de padronizar a quantidade de partidas por temporada em cada uma das equipes.

A construção da rede bipartida será anual, ou seja, para cada equipe teremos uma rede por temporada envolvendo todas as jogadas de passe do time, a qual será baseada da seguinte maneira: o subconjunto U à esquerda irá conter somente os jogadores que realizaram, ou tentaram realizar passes (serão considerados apenas os *quarterbacks* neste caso), enquanto que o subconjunto V à direita terá os jogadores que receberam (ou foram alvo) dos passes provenientes dos jogadores do subconjunto U.

Logo, a rede será direcionada e com peso, em que teremos uma aresta conectando dois vértices se tivermos pelo menos um passe entre os dois jogadores na temporada. O peso da aresta será determinada por seis variáveis presentes no banco de dados, sendo elas:

- **Passes Completos:** aqui, o peso da aresta será representado pela quantidade de passes entre os dois jogadores durante a temporada;

- **Jardas Positivas:** refere-se ao ganho de jardas das jogadas provenientes de passes entre os dois jogadores representados pelos vértices;

- **Comprimento do passe:** Refere-se a quantidade de jardas que o bola percorreu no ar até chegar ao recebedor;

- **Jardas após a recepção:** Indica quantas jardas foram percorridas pelo recebedor após este receber um passe do passador;

- **Número de Touchdowns:** Quantos dos passes realizados entre os dois jogadores resultaram em um *touchdown*;

- **Taxa de sucesso nos passes (%):** Razão entre o número de passes completos com o número de tentativa de passes entre dois jogadores.

Para exemplificar, suponha que temos dois jogadores i e j, em que as jogadas de passe entre eles estão representadas na Tabela 3.1.

Tabela 3.1: Jogadas de passe entre dois jogadores em uma temporada

Passador	Recebedor	Completo	Jardas	Comprimento	Jds. Pós Recepção	TD
i	j	1	23	3	20	0
i	j	1	2	5	0	0
i	j	1	6	5	5	1

Note que nem sempre Comprimento + Jds. pós recepção são o número total de jardas, pois o passe pode ter sido realizado bem atrás da linha de *scrimmage*.

Neste exemplo, o peso da aresta nas seis ocasiões são: Passes Completos (3), Jardas Positivas (31), Comprimento do Passe (13), Jardas Pós Recepção (25), Número de *touchdowns* (1), Taxa de sucesso nos passes (100).

Dessa forma, para cada time, em cada temporada, teremos seis diferentes redes a serem analisadas. Para cada equipe, em cada temporada, um novo conjunto de dados foi construído contemplando todos os jogadores que realizaram passes (subconjunto U), bem como seus respectivos jogadores que receberem ou foram alvos desses passes (subconjunto V), com a presença de todas as seis variáveis citadas, ou seja, todos os tipos de peso da rede.

Logo, esta nova base de dados irá conter, em cada uma das linhas, uma aresta referente à rede que está sendo analisada e, com esses dados, é possível construir a matriz adjacência da rede bipartida. A fim de exemplificar a construção da rede, temos a equipe *Minnesota Vikings* no ano de 2023, a qual possuiu quatro diferentes jogadores passando a bola, e catorze diferentes jogadores a receber passes. Dessa forma, a matriz adjacência terá dimensão 18×18, em que cada entrada será o peso da aresta referente a um vértice de U com um vértice de V. Se o número de passes fosse escolhido como a variável nesta rede, a Figura 3.1 representa a rede desta equipe na temporada de 2023, com a quantidade de passes sendo o peso de cada uma das arestas que conecta os dois diferentes subconjuntos da rede bipartida.

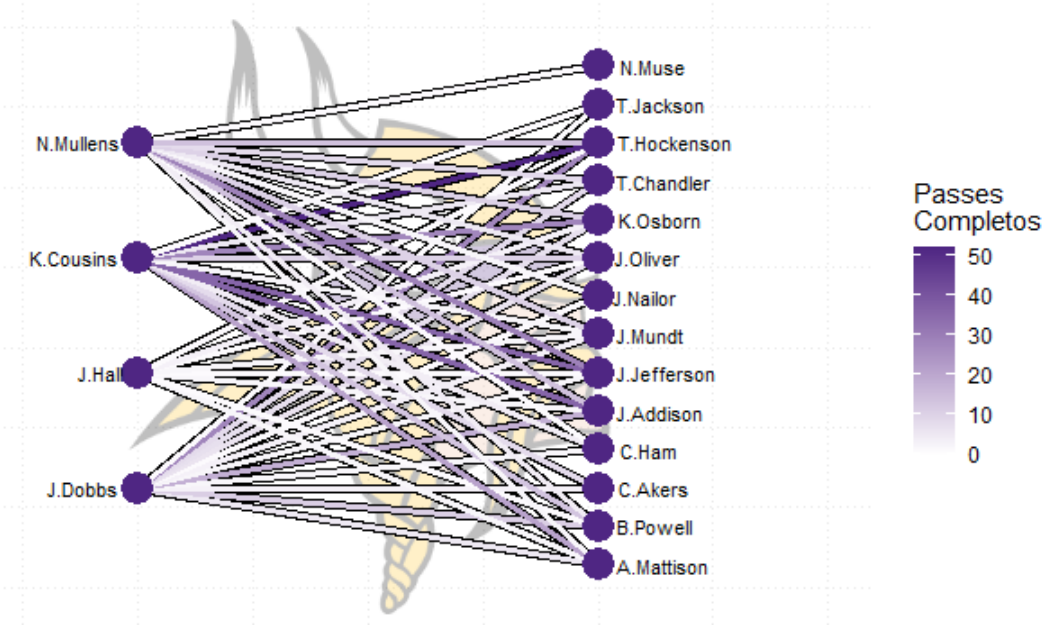


Figura 3.1: Rede passes completos - Minnesota Vikings - 2023

Pode-se destacar, que diferentes tipos de peso produzem diferentes tipos de rede, como por exemplo o que pode ser visto na Figura 3.2 quando utiliza-se a variável passes para *touchdown* para a construção da rede.

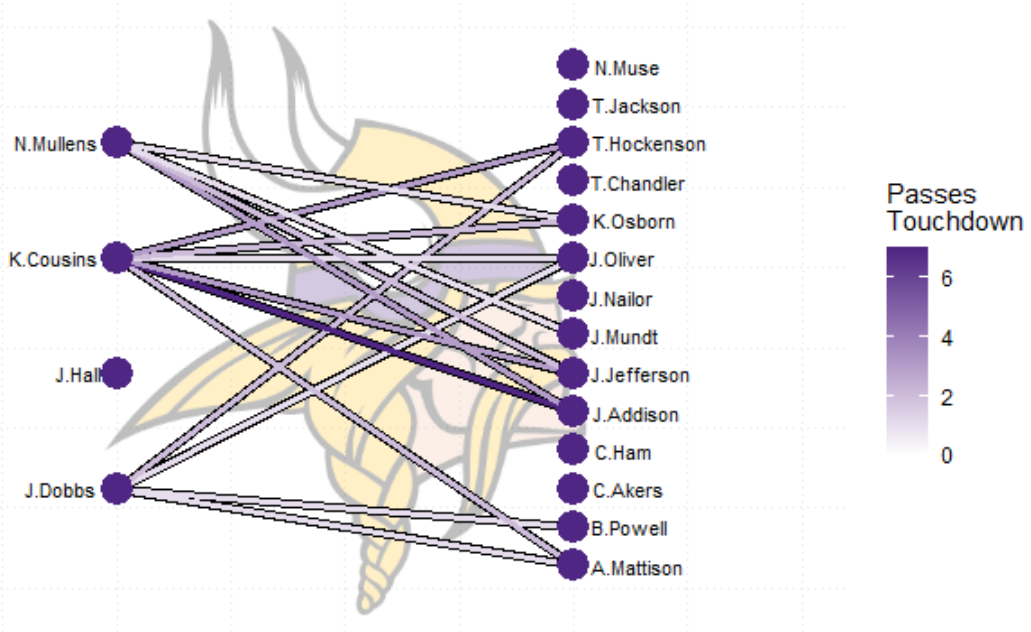


Figura 3.2: Rede passes para *touchdown* - Minnesota Vikings - 2023

Com as redes construídas para os seis diferentes tipos de peso, as quatro medidas de centralidade serão aplicadas (Seção 4), formando assim o banco de dados final, aquele que será utilizado nos modelos de classificação. Neste conjunto, cada observação será

uma equipe em uma temporada, contendo valores das medidas de centralidade nas redes e ainda o valor da resposta (campanha positiva ou não positiva). Ao todo, temos 32 equipes em 15 temporadas de 2009 até 2023, sendo assim, 480 observações presentes no banco de dados.

Capítulo 4

Medidas de Centralidade

Uma das formas de compreender como a estrutura de vértices e arestas se comporta é fazendo o uso de medidas de centralidade globais para as redes complexas. Enquanto que medidas de centralidade locais são utilizadas para verificar a importância de um determinado nó para o funcionamento da rede, as medidas de centralidade globais quantificam como a rede como um todo está estruturada. Para cada medida global, uma determinada característica da rede será levada em conta, sendo essa enfatizada na medição a ser realizada e, com isso, o uso de várias medidas diferentes garantem um maior entendimento de vários aspectos a serem destacados no sistema representado por uma rede complexa.

Dentre as medidas utilizadas neste trabalho, foram selecionadas aquelas que se adequam ao tamanho da rede a ser estudada e que focam no fato da rede em questão ser com peso e bipartida. Porém, algumas medidas para o caso sem ponderação e que podem ser aplicadas em qualquer tipo de rede foram também selecionadas para uma maior quantidade de covariáveis a serem incorporadas nos modelos de classificação posteriormente.

Muitas das medidas descritas a seguir possuem um grande uso na parte da ecologia, mais precisamente no campo de cadeias alimentares, as quais podem ser representadas como redes bipartidas em que cada nó representa uma espécie e as arestas são relações de presa-predador. Dessa forma, muitas das medidas que possuíam conceitos relacionados com o campo da ecologia foram traduzidas para o campo da teoria dos grafos para facilitar o entendimento. A implementação dessas no software R se deu a partir do pacote **bipartite** (Dormann, 2020).

4.1 Força

Umas das medidas mais simples de ser calculada em uma rede complexa é a força (ou grau no caso sem peso). O objetivo principal desta métrica é quantificar as conexões que estão sendo realizadas na rede, ou seja, no caso com peso, somar todas essas interações entre os vértices presentes.

Devido ao fato de que uma rede que possui uma quantidade maior de nós tende a possuir uma soma dos seus pesos maior do que uma rede com um número menor de vértices, é comum calcular a força média para que seja possível a comparação entre redes com diferentes tamanhos. Em redes bipartidas, a forma de calcular a força segue a mesma, a soma da matriz adjacência, porém sem a necessidade de somar as entradas em vértices pertencentes ao mesmo subgrupo, já que este valor será nulo. Logo, seja P o subconjunto da rede referente aos passadores e R o subconjunto referente aos recebedores, então a força pode ser calculada da seguinte maneira:

$$S = \frac{1}{n} \sum_{p=1}^{I} \sum_{r=1}^{J} W_{pr} \,. \tag{4.1}$$

em que I é o número de nós da rede pertencentes ao subconjunto P e em que J é o número de nós da rede pertencentes ao subconjunto R, com W_{pr} correspondendo ao peso da aresta que sai do vértice p até o vértice r segundo à certo tipo de peso, como dito no Capítulo 3.

4.2 Densidade de Ligação

Uma importante medida para associar com os pesos de uma rede bipartida é a quantidade média do peso das arestas para cada vértice em cada um dos subgrupos do grafo. Levando em conta os pesos, Bersier *et al.* (2002) implementou uma média ponderada para calcular essa característica do subgrupo P denominada Generalidade:

$$G = \sum_{r=1}^{J} \frac{W_{.r}}{S} 2^{H_r}, \tag{4.2}$$

em que W_r corresponde à soma da coluna r da matriz adjacência W, e S é o peso total da rede (força). O valor H_r é denominado Entropia de Shannon, o qual é proveniente da teoria da informação, medindo a incerteza de um sistema segundo a uma distribuição de

probabilidade. Este valor de entropia é dado por:

$$H_r = -\sum_{p=1}^{I} \frac{W_{pr}}{W_{.r}} \log_2 \left(\frac{W_{pr}}{W_{.r}} \right), \tag{4.3}$$

É possível notar que os valores $\frac{W_{pr}}{W_{.r}}$ representam uma distribuição de probabilidade, indicando a probabilidade de um vértice r possuir uma aresta com o vértice p, ou seja, na rede deste estudo, é a probabilidade de que, dado que um jogador r recebe um passe, este passe seja de um jogador p. Dessa forma, a entropia de Shannon mede o quanto de incerteza teremos ao selecionar aleatoriamente um passe recebido pelo jogador r e predizer de qual jogador de P o passe saiu. Quando as arestas de r possuem o mesmo peso para todos os vértice de P, isto é, a distribuição de probabilidade é uniforme assumindo o valor $\frac{1}{I}$ para todos os valores de p, o valor do índice atinge seu máximo, ou seja, temos a presença de um sistema totalmente incerto:

$$H_r = -\sum_{p=1}^{I} \frac{1}{I} \log_2 \frac{1}{I} = -\log_2 \frac{1}{I} = \log_2 I \tag{4.4}$$

Dessa forma, e a fim de recuperar a medida original dos eventos, o recíproco 2^{H_r} do índice foi utilizado na generalidade em (4.3), pois para o valor máximo de entropia:

$$H_r = \log_2 I \rightarrow 2^{H_r} = I. \tag{4.5}$$

Analogamente, é possível definir uma medida que calcula esta mesma média para o subgrupo P trocando os índices da expressão referentes a cada subconjunto p e r e o número de vértices em cada um deles, com isso, a medida obtida será a vulnerabilidade.

$$V = \sum_{p=1}^{I} \frac{W_{p.}}{S} 2^{H_p}, \tag{4.6}$$

A fim de calcular um índice que generalize as medidas acima para toda a rede, é possível calcular a Densidade de ligação do grafo bipartido, a qual é uma média da generalidade e vulnerabilidade:

$$LD = 0.5(G + V). \tag{4.7}$$

4.3 Assimetria de força de interação

Em redes bipartidas, também é de interesse quantificar quão balanceado o sistema está com relação ao peso de suas arestas. Por exemplo, duas redes podem possuir uma força semelhante, mas uma delas pode ter as interações balanceadas, ou seja, os pesos são bem distribuídos entre os pares de vértices, com a informação se propagando de uma maneira mais eficiente, enquanto a outra pode ser pouco balanceada, com pesos se concentrando em poucas arestas.

Primeiramente, Bascompte *et al.* (2006) define uma medida de dependência entre os vértices de diferentes subgrupos da rede bipartida. Basicamente, a dependência de um nó de P para um nó de R equivale à proporção do peso das arestas da ligação entre esses dois vértices com o peso total atrelado ao vértice de P, dessa forma:

$$d_{pr} = \frac{W_{pr}}{W_{p.}}. \tag{4.8}$$

Vale ressaltar, que para calcular a dependência invertendo os subgrupos, ou seja, calcular esta medida para os nós de R com relação a P, basta inverter os índices p e r relacionados a cada um dos subconjuntos na equação.

Agora, para calcular o quanto a rede está desbalanceada, a assimetria da força de interação entre os vértices pode ser calculada. Esta métrica é calculada para cada par de vértices separadamente, atuando como uma medida de dissimilaridade entre as duas dependências entre vértices de subconjuntos diferentes:

$$AS_{pr} = \frac{d_{pr} - d_{rp}}{d_{pr} + d_{rp}}. \tag{4.9}$$

Quanto mais distante as dependências entre essas medidas, mais assimétrica será a rede com relação a estas dependências. Em termos do problema, esta dependência mede o quão semelhante é a proporção de passes que saem de p e vão para r com relação a todos os passes feitos por p, com a proporção de passes recebidos por r que são provenientes de p com relação a todos os passes recebidos por r. Para que a medida seja estendida para toda a rede e não para apenas pares de vértices separadamente, a média de todas os pares possíveis deve ser tomada.

4.4 Equidade de interação

Outra medida que pode ser utilizada para caracterizar uma rede bipartida é a equidade de interação. Esta medida descreve o padrão das arestas em uma rede bipartida entre os subgrupos de interesse, indicando que quanto mais próximo de 1 a métrica está, as arestas com seus respectivos pesos possuem a mesma abundância (proporção) em todos os pares de vértices da rede.

Para o cálculo deste índice, Mark Hill (Hill, 1973) combinou diversas medidas que quantificam a diversidade em uma comunidade (a diversidade de arestas em uma rede, neste caso) em uma única expressão que foi denominada de números de Hill. A expressão depende de uma constante q a qual pode assumir os valores $\{0, 1, 2\}$ da seguinte maneira:

$$ {}^{q}D = \left(\sum_{p}^{I} \sum_{r}^{J} \left(\frac{W_{pr}}{S} \right)^{q} \right)^{\frac{1}{1-q}}, \tag{4.10} $$

em que quando $q = 0$ temos a força total da rede ($N_0 = S$), para q próximo a 1 há o antilogaritmo da medida de incerteza de Shannon:

$$ N_1 = \exp \left(\sum_{p}^{I} \sum_{r}^{J} \frac{W_{pr}}{S} \log \left(\frac{W_{pr}}{S} \right) \right), \tag{4.11} $$

e quando $q = 2$, tem-se o recíproco do índice de Simpson:

$$ N_2 = \left(\sum_{p}^{I} \sum_{r}^{J} \left(\frac{W_{pr}}{S} \right)^{2} \right)^{-1}, \tag{4.12} $$

índice este que representa a probabilidade de que duas arestas selecionadas aleatoriamente sejam a mesma aresta, levando em conta o peso entre elas. Para o caso da rede de estudo, é a probabilidade com que 2 passes realizados sejam provenientes do mesmo par passador-recebedor.

Com os números de Hill, diversos índices foram construídos para o cálculo da equidade de iteração. Inicialmente, Hill (1973) propôs a razão de Hill, a qual levava em conta apenas os índices N_1 e N_0, em que $E_{10} = \frac{N_1}{N_0}$, sendo modificado posteriormente por Hurlbert (1971), onde este subtraiu valor mínimo de cada índice na fórmula de modo que:

$$ F_{10} = \frac{N_1 - 1}{N_0 - 1}. \tag{4.13} $$

Como N_0 é a quantidade de interações na rede, seu valor mínimo é 1, enquanto que o menor valor para a entropia de Shannom é 0, resultando em $N_1 = \exp 0 = 1$. Rotenberry (1978) posteriormente incorporou o recíproco do Índice de Simpson no cálculo da equidade, de modo que $E_{21} = \frac{N_2}{N_1}$, o qual foi modificado por Alatalo (1981) subtraindo novamente pelo valor mínimo de cada um dos índices:

$$F_{21} = \frac{N_2 - 1}{N_1 - 1}, \tag{4.14}$$

em que o N_2 é mínimo quando possuímos apenas uma aresta na rede, resultando em probabilidade 1, para a seleção aleatória de dois passes pertencentes ao mesmo par de vértices. O cálculo da equidade de interação utilizado nos cálculos deste trabalho será o proposto por Alatalo presente em (4.14).

4.5 Outras medidas

Há ainda, a presença de outras medidas menos complexas que as anteriores presentes na literatura que captam algumas estruturas importantes dos grafos bipartidos. Pode-se tomar, por exemplo, a quantidade de vértices presentes em toda a rede, ou até mesmo em cada um dos seus subgrupos para obter indicativos a respeito da dimensão do sistema retratado na rede bipartida. Além disso, com essas informações é possível calcular a assimetria de uma rede desconsiderando os pesos, diferentemente da medida da Seção 4.3. Para este trabalho, mais duas medidas, tamanho do subconjunto a esquerda e tamanho do subconjunto a direita serão utilizadas.

Em Dormann *et al.* (2009), uma grande quantidade de índices são apresentados, tanto a nível de toda a rede, quanto aos níveis de vértices e arestas, envolvendo redes com e sem peso. A escolha das medidas citadas acima neste trabalho foi resultados de duas análises. Primeiramente, muitas das medidas são sensíveis quando tratamos de redes pequenas, como é o caso das redes bipartidas deste estudo e, com isso, muitas das medidas apresentam problemas em seus cálculos para essas dimensionalidades. Logo, mesmo as medidas escolhidas ainda sendo sensíveis ao tamanho do grafo, seus cálculo não apresentará problemas para os casos tratados neste trabalho.

Ainda, Dormann *et al.* (2009) realizou um estudo para analisar o quão correlacionados estão os índices apresentados. Pelo fato de que os mesmos serão utilizados como

covariáveis em modelos de classificação posteriormente, a multicolinearidade pode ser um problema caso tenhamos métricas muito correlacionadas. Dessa maneira, a escolha das medidas busca reduzir este problema selecionando métricas pouco correlacionadas.

4.6 Aplicação na Rede

Dessa maneira, com todas as seis redes para cada time em cada temporada construídas, é possível construir o conjunto de dados o qual será utilizado nos modelos de classificação descritos no Capítulo 5. Em cada uma das seis redes serão calculados os valores das quatro medidas globais presentes, resultando em 24 covariáveis em cada uma dos observações. Além disso, a amostra levará em conta os times desde a temporada de 2009 até 2023, contando com 480 observações. Por exemplo, para a rede construída anteriormente, as covariáveis extraídas estão representadas na Tabela 4.1.

Tabela 4.1: Covariáveis Minnesota Vikings - 2023

	Força	**Assimetria**	**Equidade**	**Densidade**
Passes	424	0.161	0.708	5.478
Jardas	4729	0.216	0.693	4.884
Jardas Aéreas	3003	0.210	0.721	4.091
Sucesso (%)	3369.544	0.220	0.969	7.304
Passes p/ TD	30	0.145	0.787	3.162
Jardas Após	1926	0.224	0.755	5.678

Além de possuirmos os valores 4 e 14 para o tamanho do subconjunto a direita e ao subconjunto a esquerda.

Ainda, para cada observação tem-se o valor da variável resposta binária que representa se o time possuiu campanha positiva (venceu mais jogos do que perdeu) ou negativa:

$$Y_k = \begin{cases} 1\,, & \text{se a equipe } k \text{ possuiu campanha positiva na respectiva temporada} \\ 0\,, & \text{caso contrário} \end{cases}, \tag{4.15}$$

Dessa maneira, a base de dados final já está devidamente tratada, e os métodos de classificação e de seleção de variáveis podem ser aplicadas nos dados a fim de realizar uma seleção de variáveis e verificar se as redes bipartidas nos dados da NFL bem como

suas medidas globais são úteis para explicar os resultados de uma temporadas de um time na liga. Vale ressaltar ainda, que como o objetivo deste trabalho é a seleção de variáveis que expliquem bem a variável resposta, não será necessária a divisão do banco de dados em conjuntos de treino e teste a fim de verificar seu poder preditivo. Logo, caso a construção de um classificador fosse de interesse, o procedimento de *data splitting* ou validação cruzada poderiam ser aplicados.

Capítulo 5

Modelos de Classificação

Dentre os modelos de aprendizado supervisionado, há a presença dos modelos de classificação. Tais modelos, segundo Izbicki e dos Santos (2020), são importantes para a realização de predições quando a variável resposta a ser estudada não é quantitativa como nos modelos de regressão clássicos, mas sim qualitativa. Assim, os modelos de classificação se empenham em construir um classificador dado por uma função de regressão $f(x)$ com boas propriedades preditivas. As várias aplicações deste tipo de modelo incluem situações como verificar se um indivíduo é bom ou mal pagador de acordo com algumas de suas características, detectar se um paciente possui ou não determinada doença a partir de informações a respeito dele, identificar imagens de acordo com seus padrões, entre outros exemplos.

5.1 Regressão Logística

5.1.1 Definição Geral

Para modelos de regressão linear clássicos, necessitamos que a variável resposta seja quantitativa para que o modelo seja ajustado aos dados. No entanto, para demais tipos de variáveis resposta, Paula (2004) mostra que modelos lineares generalizados podem ser utilizados para que seja possível relacionar a resposta com valores de covariáveis e estimar os parâmetros da regressão. Para dados categóricos, quando a resposta assume apenas dois possíveis valores, uma alternativa para um ajuste de modelo de regressão é a Regressão Logística.

Nos modelos de regressão, a relação entre o vetor de covariáveis $\mathbf{X}$ e a variável depen-

dente ou resposta Y se dá pela modelagem da média desta condicionada ao valores das covariáveis. Para modelos clássicos de regressão, a relação se deve ao vetor de parâmetros $\boldsymbol{\beta}$ da seguinte maneira:

$$E[Y|x] = \mathbf{X}\boldsymbol{\beta} = \beta_0 + \sum_{i=1}^{d} \beta_i x_i, \tag{5.1}$$

em que d representa o número de covariáveis do modelo.

Quando tratamos de regressão logística, a variável resposta Y possui distribuição *Bernoulli*:

$$p(y) = p^y(1-p)^{1-y}, \tag{5.2}$$

isto é, a variável Y possui espaço amostral $\{0, 1\}$, em que p é a probabilidade de sucesso desta variável. Para esta distribuição, a esperança da variável aleatória, a qual será modelada pelo modelo de regressão, é a própria probabilidade de sucesso, ou seja, $E[Y] = P(Y = 1) = p$.

No entanto, esta probabilidade não poderia ser estimada utilizando os modelos clássicos de regressão, pois a parte sistemática do modelo $\eta = \mathbf{X}\boldsymbol{\beta}$ pode assumir valores em $\mathbb{R}$, garantindo que o a probabilidade de sucesso também assuma valores em $\mathbb{R}$, o que não seria possível devido ao fato de estarmos tratando de uma probabilidade, a qual apenas assume valores no intervalo $\{0, 1\}$, além é claro, da suposição de normalidade não ser satisfeita. Dessa forma, o uso de um modelo linear generalizado, a regressão logística, deve ser aplicada neste caso.

Nos modelos MLG, uma função de ligação é utilizada para relacionar a esperança da variável resposta com a parte sistemática do modelo, de modo que: $E(Y|x, \beta) = g^{-1}(\eta)$, em que g é a função de ligação. Quando a resposta possui distribuição *Bernoulli*, existem algumas funções de ligação que podem ser utilizadas, dentre elas o complemento log-log, o probito e o logito. O logito é também denominado de função de ligação canônica, o qual possui algumas propriedades para facilitar a estimação dos parâmetros. Sua função é do tipo:

$$g(P(Y = 1|x, \beta)) = \log\left(\frac{P(Y = 1|x, \beta)}{1 - P(Y = 1|x, \beta)}\right) = \eta = \beta_0 + \sum_{i=1}^{d} \beta_i x_i. \tag{5.3}$$

Pela Figura 5.1, pode-se notar que para os valores do componente sistemático variando no conjunto $\mathbb{R}$, os valores da probabilidade de sucesso irão convergir para 0 ou para 1, satisfazendo o intervalo dos possíveis valores de uma probabilidade.

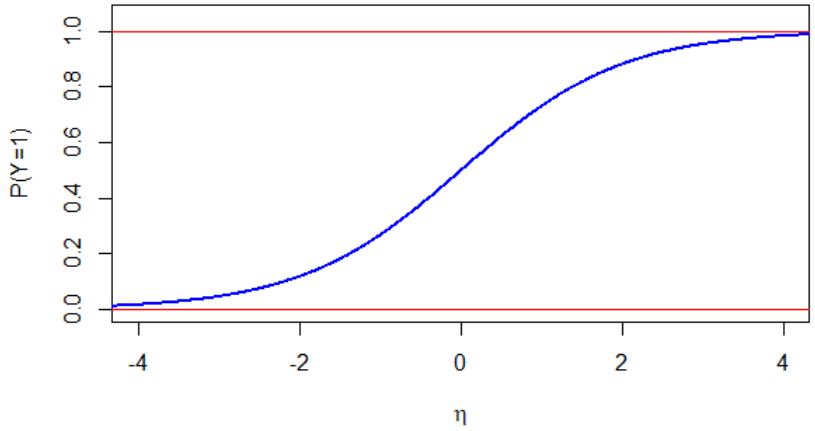


Figura 5.1: Possíveis valores de $P(Y = 1|x, \beta)$ em função de η.

Uma outra importância do logito para a regressão logística é a facilidade da interpretação dos parâmetros. Pelo fato da resposta assumir apenas valores 0 ou 1, temos que $P(Y = 0|x, \beta) = 1 - P(Y = 1|x, \beta)$, ou seja, a probabilidade de sucesso é o complementar do fracasso. Dessa forma, podemos reescrever (5.3) da seguinte maneira:

$$\log\left(\frac{P(Y=1|x,\beta)}{P(Y=0|x,\beta)}\right) = \eta = \beta_0 + \sum_{i=1}^{d} \beta_i x_i, \tag{5.4}$$

em que a razão das probabilidade de sucesso e fracasso $\frac{P(Y=1|x,\beta)}{P(Y=0|x,\beta)}$ é denominada Odds. Logo, para cada indivíduo k, temos sua Odds dada por $\exp \eta_k$, em que este valor representa o quanto a probabilidade de sucesso desta observação é maior do que a probabilidade de fracasso. Ainda, para a comparação de dois grupos, variando o valor de uma covariável e fixando as demais, é possível utilizar da razão entre duas odds e verificar o quanto a odds de um grupo é maior que a do outro grupo.

Utilizando da função logito, a estimação dos parâmetros possui sua forma simplificada. Para os modelos MLG, o método de mínimos quadrados deve ser substituído pelo método dos mínimos quadrados ponderados, o qual é proveniente do método Escore de Fisher em um processo iterativo, isto se dá pois a relação entre os parâmetros e a variável resposta não é mais linear, ou seja, agora depende de uma função de ligação e, com isso, métodos para a maximização da verossimilhança devem ser utilizados.

Seja uma amostra independente identicamente distribuída (i.i.d) de $(\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_n, Y_n)$

em que $\mathbf{x}_k = (1, x_{k1}, \ldots, x_{kd})$ é o vetor das covariáveis do k-ésimo indivíduo, temos que a estimação por máxima verossimilhança não possui fórmula analítica fechada, exigindo, portanto, o uso de métodos numéricos iterativos para o cálculo dos parâmetros. Inicialmente, o método numérico pode ser pensado para este caso da seguinte maneira:

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + ((-U'_\beta)^{-1})^{(m)} U_\beta^{(m)}, \tag{5.5}$$

em que $U_\beta^{(m)}$ = é o vetor escore na m-ésima iteração e $((-U'_\beta)^{-1})^{(m)}$ é a matriz $d \times d$ das primeiras derivadas das funções escore para cada par de variávies. Uma possível simplificação no processo é utilizar o método de Escore de Fisher, substituindo a matriz das derivadas pelo oposto de sua esperança $E[-((-U'_\beta)^{-1})] = \mathbf{K}_{\beta\beta}^{-1}$, de modo que o processo iterativo agora é dado por:

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + (\mathbf{K}_{\beta\beta}^{(m)})^{-1} U_\beta^{(m)}. \tag{5.6}$$

Por meio de manipulações algébricas provenientes do uso de uma função de ligação canônica, a expressão final para o cálculo iterativo dos parâmetros $\boldsymbol{\beta}$ se dá por meio do método dos mínimos quadrados ponderados da forma:

$$\boldsymbol{\beta}^{(m+1)} = (\mathbf{X}^T \mathbf{V}^{(m)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{(m)} \mathbf{z}^{(m)} \tag{5.7}$$

em que $\mathbf{X}$ é a matriz contendo todas as covariáveis dos indivíduos da amostra, com as observações sendo indicadas nas linhas e as variáveis de cada β nas colunas. Ainda, $\mathbf{V}$ é a matriz de variâncias $diag\{p_1(1-p_1), \ldots, p_n(1-p_n)\}$ e $\mathbf{z} = (z_1, \ldots, z_n)^t$ é a variável dependente modificada que, para o k-ésimo indivíduo possui forma $z_k = \eta_k + y_k - p_k / p_k(1-p_k)$. Com isso, o processo iterativo se dá inicialmente colocando o valor da média da distribuição p_k, de cada um dos k indivíduos, com sendo o valor da variável resposta y_k, e com isso, já se torna possível o cálculo de $\eta_k^0 = g(y_k)$ e da matriz $\mathbf{V}^0$. Dessa forma, após a primeira iteração, calcula-se $\boldsymbol{\beta}^1$ em (5.7), e atualiza os valores de $\eta_k^1 = \sum_{i=1}^d \beta_0 + \beta_i^1 x_{ik}$ de $p_k^1 = g^{-1}(\eta_k^1)$, consequentemente atualizando $\mathbf{z}^1$ e $\mathbf{V}^1$, repetindo o processo iterativo até a convergência.

Para a convergência do método, é comum, ao invés de realizar um número fixo de iterações, estabelecer um critério de parada, o qual compara o valor da estimativa atual do vetor de parâmetros com a anterior e, caso não ocorra uma mudança significativa

no valor dos coeficientes estimados, o processo iterativo é finalizado. Um dos possíveis critérios é comparar a proporção da mudança absoluta que ocorreu em cada um dos parâmetros e verificar se esta excede um erro escolhido da seguinte maneira:

$$\sum_{i=1}^{d} \left(\frac{\beta_i^m - \beta_i^{m-1}}{\beta_i^{m-1}} \right)^2 \leq \epsilon, \tag{5.8}$$

em que ϵ é o valor do erro escolhido.

Após a estimação dos coeficientes, é possível verificar quais das variáveis presentes no modelo são de fato significativas para explicar a variável resposta, procedimento este que será discutido posteriormente na Seção 5.1.2. Após isso, para verificar se o modelo ajustado os dados está adequado para as devidas análises, é possível realizar uma análise de diagnóstico para identificar possíveis inconsistências.

Nesta análise, inicialmente é ideal verificar se existe a presença de pontos aberrantes no modelo, também chamados de pontos de alavanca, pontos esses que possuem valor muito discrepante dos demais e podem influenciar demais na predição de novas observações pelo modelo, causando inconsistências no mesmo. Para a sua identificação, pode-se construir a matriz chamada $\hat{\mathbf{H}}$ chapéu:

$$\hat{\mathbf{H}} = \hat{\mathbf{V}}^{1/2}\mathbf{X}(\mathbf{X}^T\hat{\mathbf{V}}\mathbf{X})^{-1}X^T\hat{\mathbf{V}}^{1/2} \tag{5.9}$$

onde os elementos de sua diagonal indicam a influência em futuras predições. Pode-se, então, construir um gráfico de índices para esses valores e verificar a existência de pontos acima do limite calculado.

Além disso, há a detecção de pontos influentes, os quais influenciam na estimação dos coeficientes do modelo. Para isso, a distância de Cook é um bom diagnóstico para identificar se uma observação está acima de um certo limite, podendo estar causando inconsistências nas estimações.

Por fim, o cálculo de resíduos e análises descritivas desses podem ser úteis na detecção de instabilidades nas suposições do modelo ajustados, dentre elas, a independência e a homoscedasticidade. Dentre os resíduos, há os resíduos studentizado de Pearson (t_{Si}) e Deviance (t_{Di}), os quais combinados a gráficos com o índice das observações, podem fornecer indicativos de dependência caso não haja um padrão aleatório nos pontos. Além disso, gráficos do valor ajustado com os resíduos podem fornecer indícios de uma má

escolha da função de ligação ou até mesmo heterogeneidade dos dados. Por fim, o gráfico de envelope simulado garante problemas no modelo, caso muitos pontos estejam foram do envelope construído.

5.1.2 Seleção de Variáveis

Em muitos casos, uma base de dados pode possuir uma grande quantidade de covariáveis e ser incorporadas no modelo, sendo que nem todas elas são significativas para explicar a variável resposta, o que atrapalha o desempenho preditivo do modelo. Além disso, (Izbicki e dos Santos, 2020) também apresenta a justificativa de que em dados com alta dimensionalidade, ou seja, o número de covariáveis é maior do que o número de observações, o método dos mínimos quadrados não pode ser utilizado para estimar os coeficientes, pois $\mathbf{X}^T\mathbf{X}$ não é mais invertível.

Para este estudo, o caso de alta dimensionalidade não será um problema devido ao grande tamanho amostral que será utilizado. Dessa maneira, mesmo com um grande número de covariáveis, a estimação dos parâmetros do modelo será possível. Porém, é de interesse selecionar o modelo o qual contenha apenas as covariáveis que de fato são significativas para explicar a resposta.

Uma alternativa, é selecionar o melhor modelo por algum critério, por exemplo o AIC e BIC. Neste cenário, o modelo apenas com o intercepto é ajustado, partindo-se para o seguinte com apenas uma covariável, realizando todas as 2^d possíveis combinações e seleciona o modelo segundo aos critérios anteriores. No entanto, para grandes quantidades de variáveis, o método acima se torna inviável devido à enorme quantidade de possíveis modelos.

Dessa forma, modelos automáticos de seleção de variáveis podem ser aplicados para altas dimensões de banco de dados. Um dos métodos é a seleção forward, em que inicia-se com a comparação de modelos com uma covaríavel em cada selecionando o melhor segundo algum critério (AIC, BIC ou p-valor) e, após isso, ir adicionando uma segunda variável repetindo os critérios para a escolha do melhor modelo até que a inserção de novas variáveis não melhorem o ajuste. Ainda, existe o método backward com ideia semelhante ao anterior, apenas que neste caso o processo se inicia com o modelo com todas as covariáveis, retirando uma e escolhendo o melhor modelo, o qual apresentou menos piora segundo os critérios definidos, até que a remoção das variáveis não impactam no ajuste do modelo (Hocking, 1976).

Além disso, uma junção entre os dois métodos acima, o stepwise, é preferível por englobar uma maior quantidade de modelos a serem testados. Neste caso, o método altera entre o método forward e backward, verificando repetidamente a inclusão e exclusão de variáveis no modelo, até que não exista nenhum ganho ao remover ou excluir alguma covariável presente. Vale ressaltar que esses procedimentos não são adequados em dados com alta dimensionalidade.

Uma outra alternativa, são os **métodos de encolhimento** (shrinkage methods), os quais aplicam uma penalização nos coeficientes a fim de reduzirem o seu respectivo valor para zero e, consequentemente, remover a variável correspondente ao parâmetro do modelo. Essas técnicas são mais adequadas em conjunto de dados com alta dimensionalidade e em casos que as variância das covariáveis é alta. A seguir, os três principais métodos de encolhimento serão apresentados.

Para estes procedimentos, também denominados métodos de penalização, há uma restrição na estimação dos coeficientes a fim de "forçar" alguns dos parâmetros estimados para o valor zero e com isso, removê-los do modelo. No caso do Lasso, proposto por Tibshirani (1996), a restrição se dá na soma dos valores absolutos dos coeficientes, de modo que $\sum_{i=1}^{d} |\beta_i| \leq b$. Ao impor esta penalização, muitos dos coeficientes terão seus valores estimados próximos a zero, para que a soma de seus valores não ultrapassem o valor de b imposto, as variáveis relacionadas a estes parâmetros serão retiradas do modelo.

Em alguns critérios como AIC e BIC, o número de parâmetros é penalizado, sem levar em conta o valor da estimação. No Lasso, ao ser imposta a restrição sobre a norma L_1, o método capta a ideia de quem uma pequena mudança nos valores estimados dos coeficientes não possui um grande impacto na complexidade do modelo. Logo, o Lasso penaliza os parâmetros que possuem valores muito próximo a zero, diferentemente dos critérios AIC e BIC, onde apenas a igualdade a zero é levada em consideração.

Nos modelos clássicos de regressão, nos quais os β's são estimados pelo método dos mínimos quadrados, encontrando o valor dos parâmetros que minimiza a soma do quadrado dos erros. Incorporando a restrição L_1 do Lasso, a estimação dos novos coeficientes β's fica da forma:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left(\sum_{k=1}^{n} \left(y_k - b_0 - \sum_{j=1}^{d} b_j x_{k,j} \right)^2 + \lambda \sum_{j=1}^{d} |b_j| \right), \tag{5.10}$$

em que λ é um hiper parâmetro o qual irá controlar a intensidade da penalização a ser

aplicada. Pode-se notar que quanto maior o valor deste parâmetro, maior será a restrição e, com isso, mais variáveis serão retiradas do modelo e, por outro lado, para valores pequenos de λ, uma maior ênfase será dada ao método dos mínimos quadrados e mais variáveis ficarão no modelo final.

Dessa maneira, um valor intermediário para o λ deve ser escolhido para que o melhor modelo seja selecionado. Uma das maneiras de realizar esta escolha é utilizando um método de validação cruzada, o k-fold, e escolher o modelo que possuir um menor erro de predição da seguinte forma:

1. Separa-se a base de dados original em k subconjuntos mutualmente exclusivos, de preferência com o mesmo tamanho ou semelhante.

2. Retira-se um dos k subconjuntos e estima-se o modelo (com Lasso) para cada um dos valores de λ em um intervalo escolhido.

3. Para cada modelo, calcula-se o erro médio de predição (EMP) no subconjunto que foi deixado de fora, ou seja, que não foi utilizado na construção do modelo. O EMP é dado por EMP$= \frac{\sum_{i=1}^{K}(y_i - \hat{y}_i)^2}{K}$, em que K é o número de observações referentes ao subconjunto deixado de fora.

4. Para cada valor de λ, calcula-se a média do erro médio de predição em todos os k subconjuntos que ficaram de fora. Por fim, escolhe-se o valor de λ com o menor erro calculado.

Ainda, outros métodos de validação cruzada podem ser utilizados para encontrar o melhor valor do hiper parâmetro. Por exemplo, a variação do k-fold, leave-one-out, em que apenas uma observação é retirada por vez para realizar a validação cruzada, é uma alternativa.

Pode-se destacar ainda, que no Lasso, as covariáveis precisam ser padronizadas antes de realizar os procedimentos descritos acima, pois dependendo da magnitude de algumas delas, podemos ter valores naturalmente menores para os parâmetros mesmo antes de se aplicar a restrição. Além disso, a estimativa dos coeficientes por Lasso é viesada devido à penalização e, caso for de interesse, é possível retirar o viés estimando o modelo sem a restrição após a retirada das covariáveis pelo Lasso.

Já no método Ridge (Hoerl e Kennard, 1970), a seleção das variáveis é realizada de uma maneira semelhante do Lasso, com apenas uma modificação na penalização a ser

realizada. Nesta seleção, a penalização é aplicada na norma L_2: $\sum_{i=1}^{d} \beta_i^2 \leq b$. Com isso, a estimação dos parâmetros nos modelos clássicos é dada por:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left(\sum_{k=1}^{n} \left(y_k - b_0 - \sum_{j=1}^{d} b_j x_{k,j} \right)^2 + \lambda \sum_{j=1}^{d} b_j^2 \right), \tag{5.11}$$

em que a estimação do hiper parâmetro λ é feita da mesma maneira que o Lasso, por validação cruzada.

Mesmo com ideias semelhantes, o Lasso e o Ridge podem levar a seleções de variáveis bem distintas entre si, dado que o Ridge realiza uma penalização mais suave, deixando no modelo uma quantidade de covariáveis maior que o Lasso, o qual aplica uma penalização mais restritiva.

Outro método de penalização, o ElasticNet pode ser considerado uma junção entre o Lasso e o Ridge, em que os dois tipos de penalização, em L_1 e L_2 são considerados na estimação dos parâmetros β's. Dessa forma, os coeficientes estimados seguem:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left(\sum_{k=1}^{n} \left(y_k - b_0 - \sum_{j=1}^{d} b_j x_{k,j} \right)^2 + \lambda_1 \sum_{j=1}^{d} |b_j| + \lambda_2 \sum_{j=1}^{d} b_j^2 \right) \tag{5.12}$$

Neste método, temos dois hiper parâmetros a serem selecionados. No entanto, é possível seguir um procedimento semelhante ao k-fold, realizando duas validações cruzadas para encontrar o melhor valor de λ_1 e λ_2.

5.2 Árvores de Classificação

5.2.1 Definição Geral

Dentre os métodos não paramétricos para a realização de predições e seleção de variáveis, as árvores de regressão se destacam por seu grande poder interpretativo. Seja $f(\mathbf{x})$ a função de regressão de uma determinada variável resposta y atrelada às suas d covariáveis, então, as árvores irão predizer a função de regressão com base na estrutura representada na Figura 5.2. Neste esquema, os retângulo em azul são denominados nós da árvore, enquanto as casas em roxo são as folhas e, por exemplo, se a covariável x_1 de um indivíduo k for menor do que um valor c, ou seja, se a condição do nó for satisfeita, segue-se para o caminho da esquerda, e caso esta não o for, o caminho da direita deve ser

acessado.

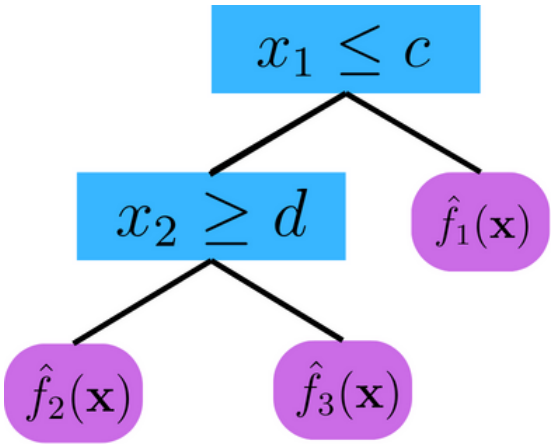


Figura 5.2: Esquema de uma árvore de regressão.

O processo é repetido até que uma folha da árvore seja localizada, a qual irá conter o valor estimado da função de regressão para o indivíduo k com o determinado valor das covariáveis que levaram àquela folha em questão. Com isso, a ideia principal das árvores é particionar o espaço $\mathbb{R}^d$ formado pelas d covariáveis presentes no modelo de modo que haja j subregiões $R_1, \ldots, R_j$, em que cada uma delas contenha uma quantidade de indivíduos em que suas folhas sejam as mesmas na árvore construída.

Por exemplo, se tomarmos a árvore exemplificada anteriormente com as covariáveis x_1 e x_2 no intervalo $\{0, 1\}$ com valores de $c = d = 0.4$, teremos uma repartição no espaço das variáveis preditoras conforme mostrado na Figura 5.3.

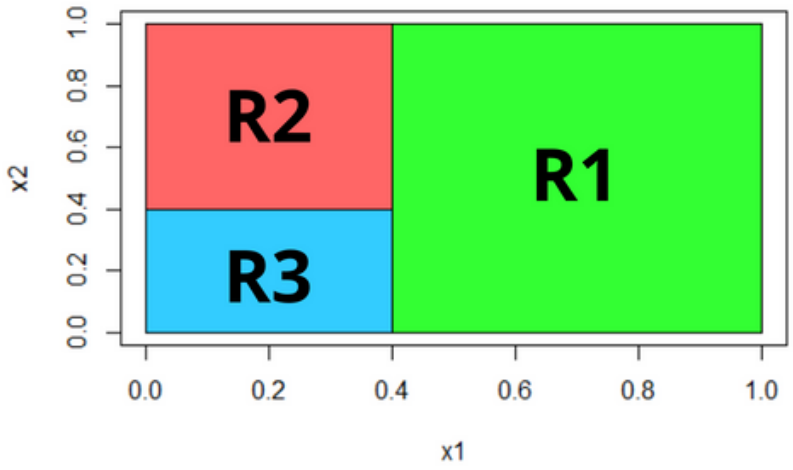


Figura 5.3: Regiões no espaço das covariáveis da árvore referente à Figura 5.2.

Com as três regiões formadas, a predição para os valores em cada uma delas será dada

pela média entre todas as respostas pertencentes a cada uma das regiões, ou seja, para certa região R_l, a função de regressão estimada é dada por:

$$f(\mathbf{x}) = \frac{1}{|i : x_i \in R_l|} \sum_{i:x_i \in R_l} y_i. \tag{5.13}$$

No entanto, o caso anterior é utilizado quando a natureza da variável resposta é quantitativa, assim, para casos em que a resposta é de natureza binária, as árvores de classificação são utilizadas, em que a estimação se dá pela moda das respostas em cada uma das regiões formadas pelas árvores, de modo que a equação 5.13 se modifica para:

$$f(\mathbf{x}) = \text{moda}\{y_i : x_i \in R_k\}. \tag{5.14}$$

Uma característica importante de uma árvore de regressão é a sua pureza $P(T)$, a qual irá indicar o quão heterogênea estão as regiões $R_1, \ldots, R_j$ construídas, ou seja, o quão distantes estão os valores das respostas em indivíduos de uma mesma região. Para que o ajuste de uma árvore gere bons resultados preditivos, é necessário que a pureza seja mínima, de modo que as observação sejam homogêneas em cada partição. No entanto, encontrar uma árvore específica T que possua pureza mínima é computacionalmente custoso, devido à grande quantidade de divisões que podem ser realizadas no espaço $\mathbb{R}_d$.

Dessa forma, a construção de uma árvore de regressão é dividida em dois procedimentos. Primeiramente, constrói-se uma árvore extensa com muitos nós e após isso, realiza-se a poda de cada um deles de acordo com certo critério de pureza. Para a árvore de regressão, o critério de pureza é o erro quadrático médio, ou seja, a variância dentro de cada região, já para a classificação, com uma resposta com C categorias, o Índice de Gini é utilizado:

$$P(T) = \sum_{R} \sum_{c \in C} \hat{p}_{R,c}(1 - \hat{p}_{R,c}), \tag{5.15}$$

em que este $\hat{p}_{R,1}$ é a proporção de respostas de valor c na região R, sendo que $P(T)$ é mínimo quando todas as proporções são 0 ou 1.

Com a pureza definida, para o primeiro nó da árvore, uma variável x_i e uma partição t_1 são escolhidas de modo a minimizarem o valor de $P(T)$. Após isso, a árvore possuirá duas divisões R_1 e R_2, onde uma delas será escolhida, bem como uma covariável x_i e outra partição que minimize $P(T)$. Realiza-se o procedimento recursivamente até certa escolha de parada, por exemplo até que haja menos de uma quantidade de observações em cada

região. Em seguida, os nós serão um por um podados da árvore, medindo em cada poda o erro preditivo no conjunto de validação, em que por fim, será escolhida a árvore com menor valor do seu erro em sua poda. Este procedimento evita o super-ajuste, em que a variância do estimador de $f(\mathbf{x})$ é alta, realizando um balanceamento com o viés da função, resultando em melhores predições.

5.2.2 Seleção de Variáveis

Uma das vantagens do uso das árvores de classificação é a seleção automática de variáveis, pois o procedimento já escolhe quais covariáveis inserir nas condições presentes nas árvores a fim de minimizar a função $P(T)$, ou seja, tornar a árvore mais homogênea. Além disso, o processo já leva em conta a interação entre as variáveis, já que para acessar alguma folha é possível realizar o caminho passando por várias condições de diferentes covariáveis.

Um procedimento proveniente da construção de árvores, o **bagging**, visa diminuir a variância das predições resultantes deste procedimento e, além disso, é possível quantificar a importância das variáveis presentes para a explicação e predição da resposta. Primeiramente, quando há dois estimadores $\hat{f}_1(\mathbf{x})$ e $\hat{f}_2(\mathbf{x})$ para a função de regressão $f(\mathbf{x})$, sendo estes não viesados, não correlacionados e com variâncias iguais, então pode-se realizar uma combinação entre as predições anteriores:

$$\hat{f}(\mathbf{x}) = \frac{\hat{f}_1(\mathbf{x}) + \hat{f}_2(\mathbf{x})}{2}, \tag{5.16}$$

sendo que o risco desta nova predição é menor que o risco tanto do primeiro, quanto do segundo estimador.

Baseado na equação 5.16, o bagging cria uma quantidade B de reamostras bootstrap da amostra original e para cada uma delas, tem-se um estimador para o valor de uma predição baseado em uma árvore completa, ou seja, sem que a poda tivesse sido realizada. Com isso, a nova função de regressão estimada tem forma:

$$\hat{f}(\mathbf{x}) = \frac{1}{B}\sum_{i=1}^{B} \hat{f}_i(\mathbf{x}). \tag{5.17}$$

Através desta ideia, é possível atribuir uma medida da importância das covariáveis baseado em quanto essa trouxe de redução para a soma de quadrado dos resíduos (SQR) em cada

divisão da árvore baseada na variável. Se uma variável x_i dividiu a árvore t em t_1 e t_2, a importância de x_i é dada por:

$$I(x_i) = \text{SQR}_t - \text{SQR}_{t_1} - \text{SQR}_{t_2}. \tag{5.18}$$

Assim, repete-se este cálculo para todas as vezes que a covariável foi utilizada para realizar a divisão na árvore levando em conta todas as B árvores construídas nas amostras bootstraps, calculando a média de todas as reduções.

Capítulo 6

Aplicação

Para a realização da aplicação, cujo objetivo é verificar se as medidas de centralidade são significativas nos modelos de classificação, explicando bem a resposta de campanha positiva, uma análise descritiva foi realizada, a fim de obter-se indícios sobre possíveis relações das variáveis com a resposta, além de investigar multicolinearidade e um possível tratamento para este fator. Após isso, a aplicação de uma regressão logística via lasso será realizada nos diferentes cenários propostos na análise descritiva, verificando, em cada caso, se selecionamos ou não medidas de centralidade que expliquem bem a variável resposta. Após isso, a seleção de variáveis via Árvore de Classificação será realizada, bem como um ranking de importância de cada uma delas via Floresta Aleatória. Por fim, uma conclusão será apresentada comparando os resultados obtidos em cada um dos método propostos.

É importante destacar que, como o objetivo de estudo é inferencial, ou seja, não desejamos construir um bom classificador e compará-lo com outros, mas sim apenas verificarmos quais covariáveis são significativas para explicar a resposta, não iremos dividir o conjunto de dados em treino e teste, mas escolhendo os valores dos hiper parâmetros via validação cruzada nos métodos. Pela questão inferencial, uma análise de diagnóstico do método paramétrico (regressão logística) será feita para garantir que o modelo está adequado.

6.1 Análise Descritiva

Antes da aplicação dos modelos de classificação propostos, uma análise descritiva dos dados pode ser útil na identificação de alguns indicativos de relações entre as covariáveis entre si e até mesmo com a variável resposta, além da identificação de algumas observações discrepantes, os *outliers*. Primeiramente, vale ressaltar que não há a presença de nenhum

valor faltante em nenhuma observação do conjunto de dados, todas as covariáveis e respostas estão completas.

Então, é ideal verificar como está o balanceamento dos dados segundo à variável resposta, isto é, se a frequência de times com campanha positiva é semelhante com times de campanha negativa.

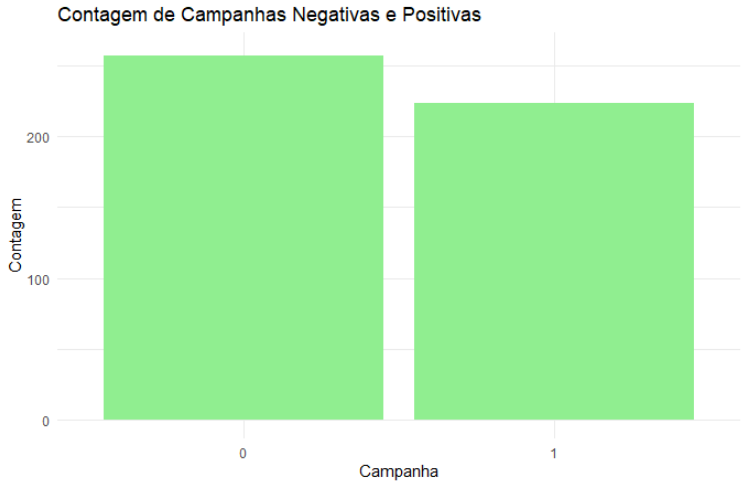


Figura 6.1: Frequência de valores (1 = Campanha Positiva, 0 = Caso Contrário).

Pela Figura 6.1 é possível perceber que as respostas 0 (campanha negativa) e 1 (campanha positiva) possuem valores semelhantes, com 257 e 223 valores respectivamente. Com isso, temos um balanceamento de 0.535% de respostas assumindo valores 0, e 0.464% assumindo valores 1 no conjunto de dados.

Ainda, com relação à variável resposta, iremos supor sua independência mesmo que o número de vitórias de um time dependa do número de derrotas de outro, tendo indícios de uma relação, mesmo que não muito forte, de independência, a qual será checada no diagnóstico da regressão logística.

A fim de relacionar as covariáveis de estudo com os valores da variável resposta individualmente para obter algum indicativo do efeito da covariável, é possível realizar a construção de boxplots para cada valor da resposta e verificar se a distribuição das caixas muda de resposta para resposta, ou seja, para indivíduos com valor 0 e 1. Por exemplo, a Figura 6.2 indica que a força em geral apresenta valores maiores em todos os tipos de peso exceto o sucesso nos passes, além de ser possível identificar alguns *outliers* em alguns gráficos. Ainda, a diferença maior entre as respostas se dá quando tratamos da força no peso *touchdown* neste caso, mostrando indicativos de uma tendência dos times com campanha postiva possuírem uma rede com este peso com uma força maior.

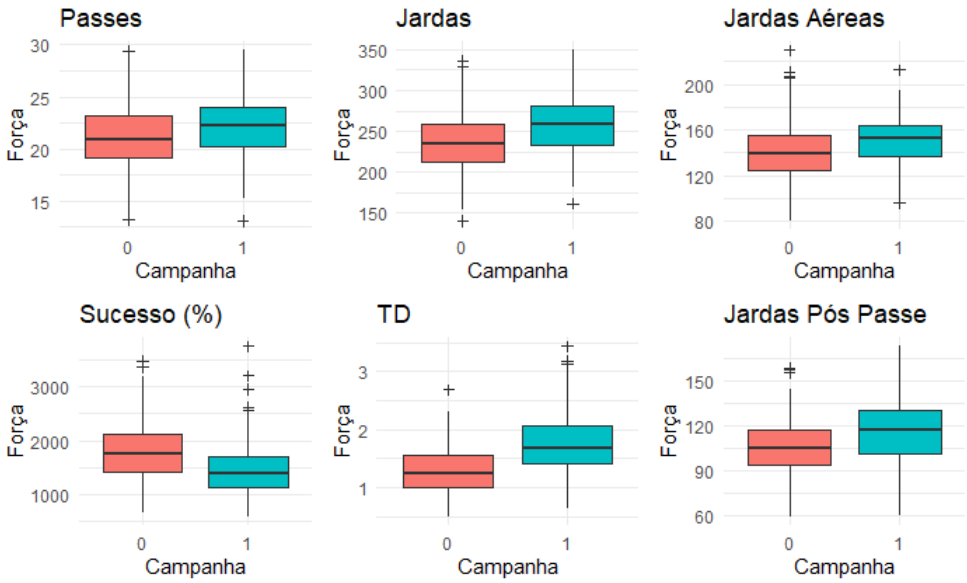


Figura 6.2: Boxplot para a medida de centralidade Força (1 = Campanha Positiva, 0 = Caso Contrário).

Já com relação à assimetria na força de ligação, pela Figura 6.3 é possível notar que as caixas seguem um padrão semelhante para todos os tipos de peso, mostrando indícios de que talvez estas covariáveis possuam um efeito semelhante na resposta. Porém, novamente, os valores desta medida apresentam valores maiores de média, mediana e quartis para o caso em que time possuiu campanha positiva na temporada.

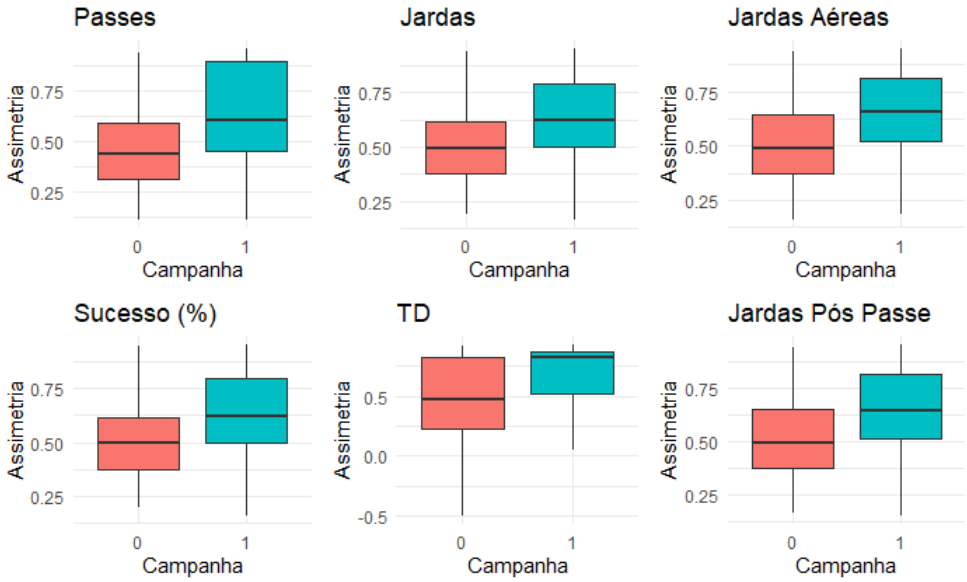


Figura 6.3: Boxplot para a medida de centralidade Assimetria (1 = Campanha Positiva, 0 = Caso Contrário).

No entanto, para o caso da medida de centralidade densidade mostrada na Figura

6.4, neste caso os menores valores das covariáveis se encontram nos times de campanha positiva. Ainda, estes valores são muitos próximos em pesos como sucesso e *touchdown*, mostrando indícios de que tais covariáveis não influenciem a variável resposta. Por fim, para o caso da medida de equidade de iteração, temos valores muito próximos de medidas resumo como média, mediana, e quartis, apenas com passes e jardas apresentando uma ligeira diferença entre esses valores, conforme é possível notar na Figura 6.5.

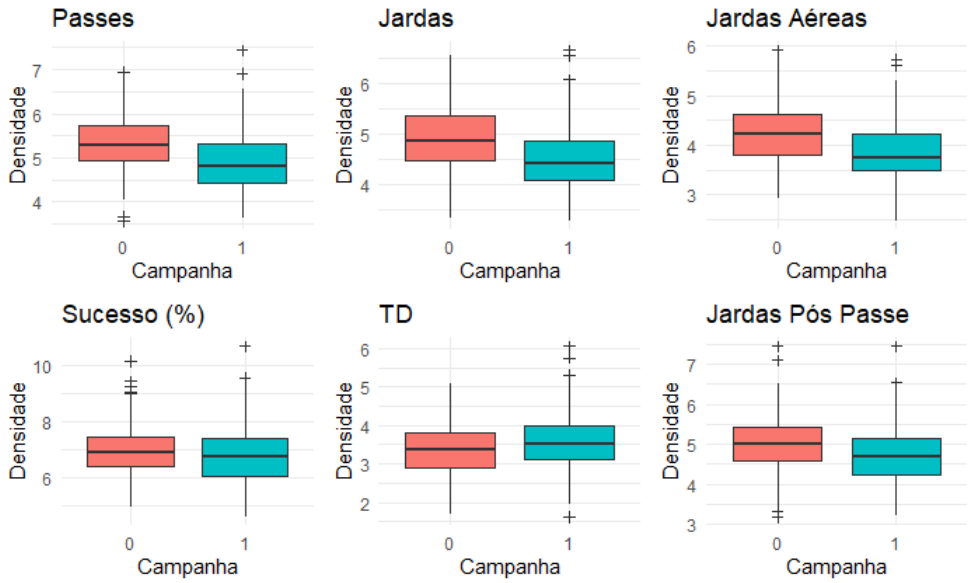


Figura 6.4: Boxplot para a medida de centralidade Densidade (1 = Campanha Positiva, 0 = Caso Contrário).

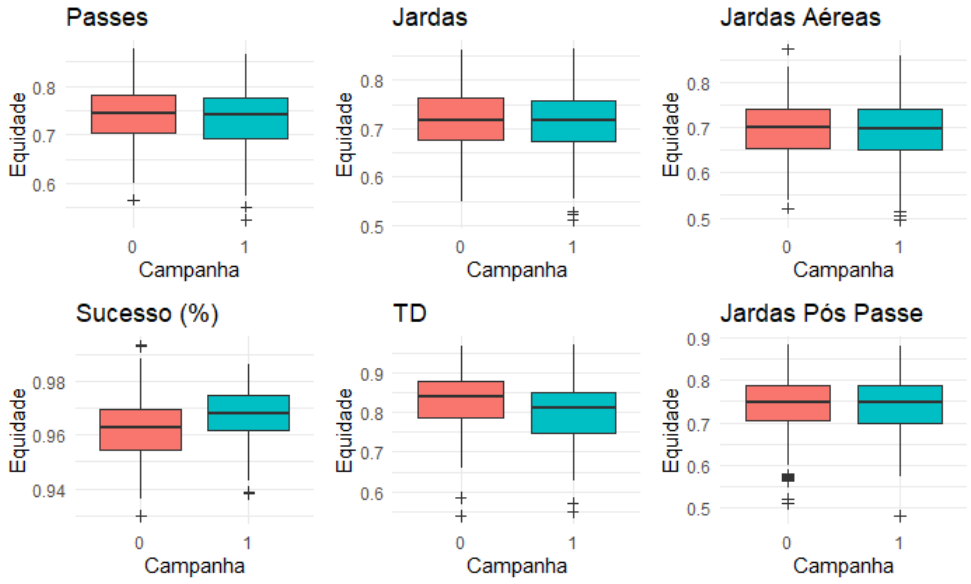


Figura 6.5: Boxplot para a medida de centralidade Equidade (1 = Campanha Positiva, 0 = Caso Contrário).

Ainda, o tamanho do subconjunto a esquerda (*n-esq*), ou seja, dos passadores e a direita (*n-dir*), dos recebedores, também serão utilizados como covariável neste caso. A fim de verificar a relação da variável com a resposta, temos tabelas de contingência destes valores, já que possuímos poucos valores distintos em cada subconjunto.

Tabela 6.1: Tabela de contingência - Passadores

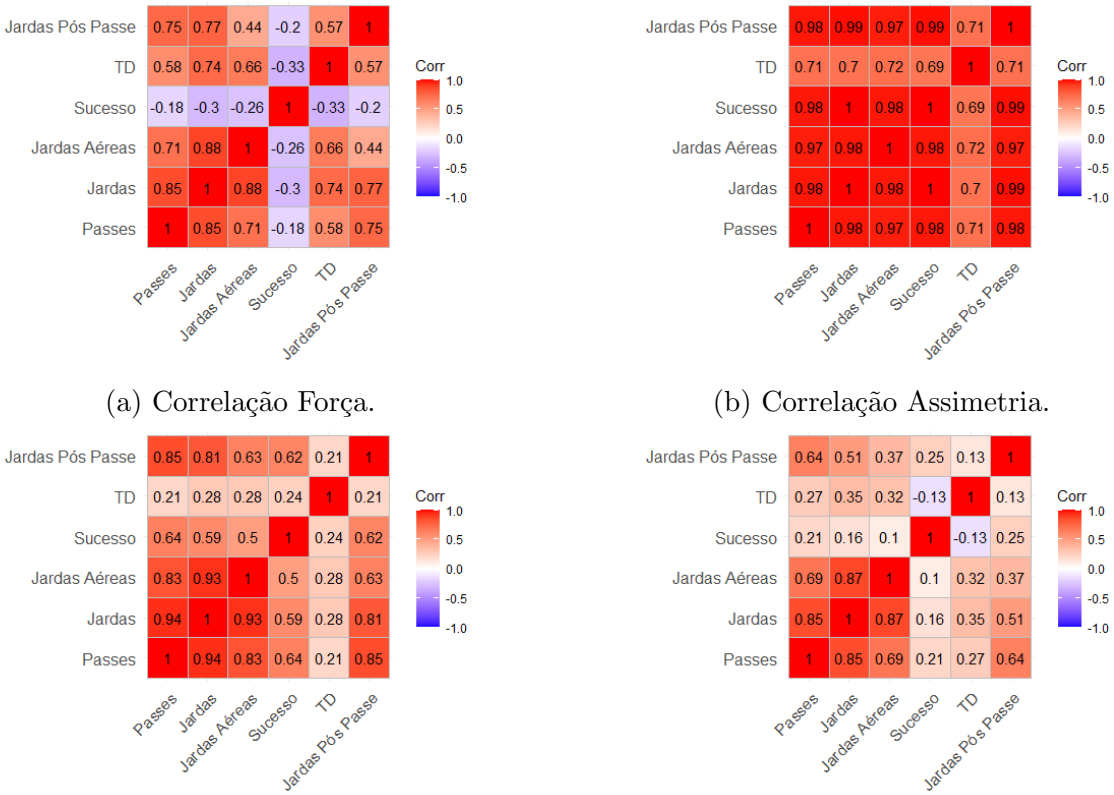
	Resposta	
Passadores	0	1
1	7.00%	15.69%
2	47.08%	65.91%
3	35.01%	13.90%
4	10.11%	4.03%
5	0.77%	0.44%

A Tabela 6.1 apresenta o total marginal à resposta em cada valor da covariável passadores. Pode-se perceber que a porcentagem de valores é bem semelhante com relação ao número de passadores em times com campanhas positivas e negativas. Já a tabela 6.2, a qual representa os recebedores, também possue valores muito semelhantes com relação à resposta nos valores da covariável.

Tabela 6.2: Tabela de contingência - Recebedores

	Resposta			**Resposta**	
Recebedores	0	1	**Recebedores**	0	1
10	0%	0.44%	18	8.94%	10.76%
11	0.77%	1.34%	19	4.28%	2.24%
12	4.28%	7.17%	20	3.11%	0.44%
13	8.17%	12.10%	21	1.94%	0.89%
14	14.39%	17.48%	22	1.16%	0.89%
15	20.62%	20.62%	23	0.38%	0.89%
16	16.73%	14.79%	25	0.38%	0%
17	14.78%	9.86%			

Além da relação com a variável resposta, é de interesse verificar como as covariáveis estão correlacionadas entre si.



(a) Correlação Força.

(b) Correlação Assimetria.

(c) Correlação Densidade.

(d) Correlação Equidade.

Figura 6.6: Matriz de correlação para as diferentes medidas de centralidade.

Para isso, as covariáveis foram divididas segundo o tipo de medida de centralidade e, em seguida, a matriz de correlação entre elas foi calculada, a qual está presente na Figura 6.6. Pode-se perceber que algumas medidas possuem variáveis muito correlacionadas, como é o caso da assimetria da força de ligação, onde todos os tipos de peso possuem uma grande correlação entre si. Fora isso, as demais medidas apresentam correlação moderada, apenas em casos onde tempos jardas, passes e jardas aéreas que possuímos uma correlação maior, algo que pode impactar na qualidade dos resultados posteriormente. Logo, a abordagem utilizada nas aplicações será a seguinte: em cada método, iremos ajustar o modelo considerando todas as covariáveis presentes no banco de dados; posteriormente, ao verificar a seleção dessas variáveis em cada caso, algumas serão removidas de acordo com seus valores de correlação e a significância ou não no ajuste anterior; por fim, as variáveis serão retiradas de modo que não haja nenhuma correlação maior que 0.8 dentre elas, e o modelo é ajustado novamente. Após isso, os resultados serão devidamente analisados.

No entanto, a correlação dentro dos pesos, ou seja, com relação às diferentes medidas não é forte, por exemplo, a Figura 6.7 ilustra a correlação de diferentes medidas dentro da rede com o peso passes completos. É possível perceber os baixos valores da correlação

entre as variáveis.

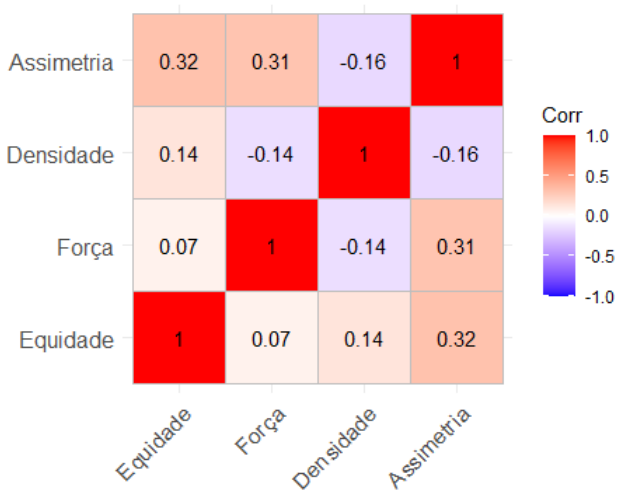


Figura 6.7: Correlação das medidas dentro da rede de passes completos

Dessa maneira, apenas a correlação entre os tipos de redes em cada medida deve ser estudada para este banco de dados.

6.2 Regressão Logística

Para o primeiro ajuste de modelos, a regressão logística via Lasso e ElasticNet foi utilizada para a seleção das variáveis de interesse. Por mais que estes métodos apresentem bons resultados mesmo com a presença de multicolinearidade nas variáveis, três diferentes cenários foram determinados para um melhor estudo das seleção das medidas de centralidade para explicar a resposta de campanha positiva das equipes em uma temporada.

No cenário 1, todas as 26 covariáveis serão levadas em conta no modelo Lasso e ElasticNet. Na Tabela 6.3, temos algumas covariáveis marcadas que foram selecionadas em cada um dos métodos. Foram selecionadas 16 de 26 variáveis para o Lasso, enquanto o ElasticNet selecionou uma a mais, podemos perceber que todas as variáveis selecionadas por este último método também foram escolhidas pelo Lasso. Mesmo com a presença de alta correlação entre algumas medidas, é possível notar já algumas evidências de início de que as covariáveis podem ser úteis para explicar a variável resposta.

Tabela 6.3: Variáveis selecionadas para o Cenário 1

Rede	Medida	Lasso	E.Net	Rede	Medida	Lasso	E.Net
Passes Completos	Força	■	■	Jardas Pós Recepção	Força	■	■
	Densidade	■	■		Densidade	■	■
	Assimetria		■		Assimetria		
	Equidade	■	■		Equidade	■	■
Jardas Totais	Força			Taxa de Sucesso (%)	Força		
	Densidade	■	■		Densidade		
	Assimetria				Assimetria		
	Equidade				Equidade	■	■
Jardas Aéreas	Força	■	■	Passes p/ TD	Força	■	■
	Densidade				Densidade	■	■
	Assimetria	■	■		Assimetria	■	■
	Equidade	■	■		Equidade	■	■
-	Passadores			-	Recebedores	■	■

Para o segundo cenário, algumas covariáveis serão retiradas pela alta correlação com alguma outra presente no modelo levando em conta a sua significância. Por exemplo, a medida de centralidade assimetria possui grande correlação em diferentes tipos de rede e, neste caso, iremos remover todas essas medidas, deixando apenas a assimetria para rede de passes para *touchdown* (pois a correlação é baixa) e a assimetria para rede de jardas áreas, pois foi a variável que apresentou uma maior significância no cenário 1. Dessa forma, este tipo de seleção será realizada, retirando apenas algumas variáveis muito correlacionadas que não apresentaram bons resultados no cenário 1 em detrimento de outras medidas correlacionadas a esta. Como mostrado na Tabela 6.4, ambos os métodos selecionaram as mesmas covariáveis, algo semelhante ao cenário 1. Ainda, mesmo com a remoção de algumas medidas correlacionadas, um grande número de variáveis ainda permaneceu no modelo, 14 de 20 totais, o que indica novamente uma evidência da importância das medidas de centralidades das redes para a explicação da resposta.

No cenário seguinte, o cenário 3, medidas serão removidas de modo que não haja nenhuma correlação maior que 0.8 entre as covariáveis. Isso será feito de modo que, além de observar as correlações, os ajustes realizados nos dois cenários anteriores serão levados em conta ao escolher qual das covariáveis correlacionadas serão mantidas. Por exemplo, se duas delas apresentarem correlação maior que 0.8, iremos verificar qual das

duas apresentou um melhor resultado nos cenários anteriores e, então, esta será mantida.

Tabela 6.4: Variáveis selecionadas para o Cenário 2 (NA para variáveis não utilizadas neste cenário)

Rede	**Medida**	**Lasso**	**E.Net**	**Rede**	**Medida**	**Lasso**	**E.Net**
Passes Completos	Força			Jardas Pós Recepção	Força		
	Densidade	**NA**	**NA**		Densidade		
	Assimetria	**NA**	**NA**		Assimetria	**NA**	**NA**
	Equidade				Equidade		
Jardas Totais	Força			Taxa de Sucesso (%)	Força		
	Densidade				Densidade		
	Assimetria	**NA**	**NA**		Assimetria	**NA**	**NA**
	Equidade				Equidade		
Jardas Aéreas	Força			Passes p/ TD	Força		
	Densidade	**NA**	**NA**		Densidade		
	Assimetria				Assimetria		
	Equidade				Equidade		
-	Passadores			-	Recebedores		

Levando em conta o cenário 3, a Tabela 6.5 traz as variáveis selecionadas:

Tabela 6.5: Variáveis selecionadas para o Cenário 3 (NA para variáveis não utilizadas neste cenário)

Rede	**Medida**	**Lasso**	**E.Net**	**Rede**	**Medida**	**Lasso**	**E.Net**
Passes Completos	Força	**NA**	**NA**	Jardas Pós Recepção	Força		
	Densidade	**NA**	**NA**		Densidade		
	Assimetria				Assimetria	**NA**	**NA**
	Equidade				Equidade		
Jardas Totais	Força			Taxa de Sucesso (%)	Força		
	Densidade	**NA**	**NA**		Densidade		
	Assimetria	**NA**	**NA**		Assimetria	**NA**	**NA**
	Equidade	**NA**	**NA**		Equidade		
Jardas Aéreas	Força	**NA**	**NA**	Passes p/ TD	Força		
	Densidade				Densidade		
	Assimetria	**NA**	**NA**		Assimetria		
	Equidade				Equidade		
-	Passadores			-	Recebedores		

Neste cenário, novamente várias variáveis foram selecionadas, mesmo sem a presença de alta correlação entre elas. Podemos ressaltar que a rede de passes para *touchdown* teve todas as medidas de centralidade selecionadas em todos os cenários.

Para uma melhor interpretação deste cenário, uma regressão logística será aplicada sem a presença de regularizações com as variáveis selecionadas no cenário 3, de modo que a estimativa dos parâmetros seja desvieseda, possibilitando a interpretação das *odds*, além da realização de uma análise de diagnóstico para verificar se as suposições do modelo são satisfeitas. Como o objetivo deste estudo é puramente inferencial, é importante que o modelo de fato seja adequado aos dados.

Com relação às ODDS, o exponencial do coeficiente estimado a respeito de uma covariável (Razão de ODDS) indica o aumento (ou diminuição) na probabilidade de campanha positiva quando aumentamos em uma unidade o valor da variável, se as demais se mantiverem constantes. Para esse caso, iremos analisar as medidas de centralidade mais significativas para a regressão logística. Começando pela força *touchdown*, sua razão de ODDS possui valor de $\exp(2.39) = 11.01$ indicando que a cada aumento de 1 unidade na força de redes para passes para *touchdown* aumenta em 11.01 vezes a probabilidade da equipe ter campanha positiva no final da temporada. Por outro lado, o coeficiente relacionado com a medida densidade na rede de jardas aéreas possue valor de -0.89, o que implica em uma razão de ODDS de 0.41, indicando que a cada aumento unitário dessa medida, as chances de campanha positiva diminuem em 0.41.

Partindo para a análise de diagnóstico, esta etapa é de extrema importância para verificar se os resultados de seleção de variáveis estão adequados. Primeiramente, iremos verificar se há a presença de muitos valores *outliers* que podem impactar nos resultados obtidos.

A Figura 6.8 mostra, a esquerda, um gráfico de dispersão dos valores ajustados com a medida h (diagonal da matriz $\mathbf{H}$), a qual representa possíveis pontos de alavanca (pontos aberrantes) e, neste caso, apenas 2 pontos chamam a atenção, porém o valor da medida está bem próximo da nuvem de pontos, indicando a não necessidade de preocupação com esses pontos. No gráfico da direita, temos a distância de Cook com relação ao índice das observações a fim de que seja possível detectar pontos influentes, mas, neste caso, nenhum ponto apresenta um valor grande o bastante nesta medida para ser considerado com *outlier*.

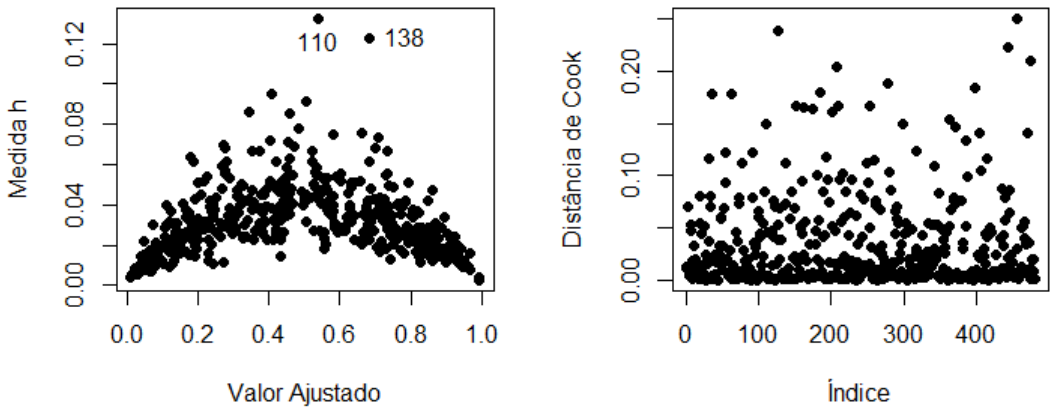


Figura 6.8: Pontos de Alavanca e Distância de Cook.

Ainda, existem duas suposições a serem estudadas: a homoscedasticidade dos dados e a independência da variável resposta. Para isso, podemos fazer uso de um gráfico de índice por resíduo Componente do Desvio, onde se os pontos estiverem dentro do intervalo, a suposição de homoscedasticidade pode ser considerada válida, além de que se não houver nenhum padrão nos dados, ou seja, a dispersão for aleatória, a suposição de independência também pode ser considerada válida. A Figura 6.9 indica que as suposições podem ser consideradas válidas.

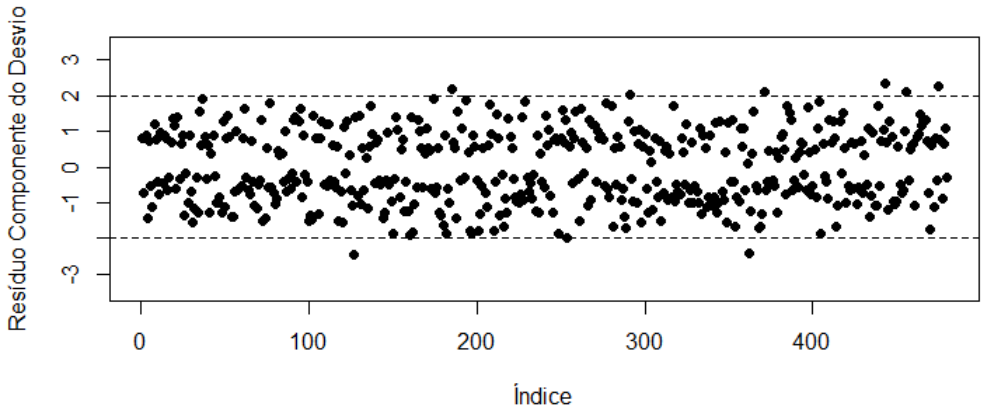


Figura 6.9: Gráfico de índice por resíduo Componente do Desvio.

Finalmente, para a análise do ajuste como um todo, podemos utilizar o gráfico de envelope simulado de Atkinson (1981), como mostrado no Figura 6.10. Como todos os pontos estão dentro do envelope, isto indica que o modelo está bem adequado aos dados.

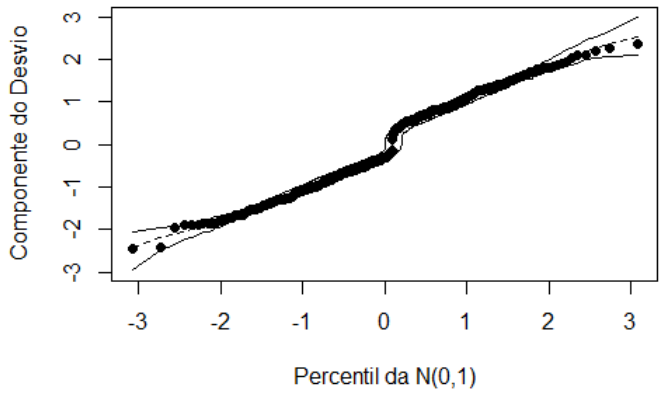


Figura 6.10: Gráfico de envelope simulado.

Dessa maneira, com base nas análises de diagnóstico realizadas, podemos concluir que de fato o modelo está bem adequado a este conjunto de dados com as variáveis referentes ao terceiro cenário, indicando que as variáveis selecionados neste modelo conseguem ser úteis para explicar a variável resposta.

6.3 Árvores de Classificação

Para as árvores de classificação, iremos realizar os mesmo três cenários realizados na regressão logística, ou seja, iniciando com todas as variáveis, depois removendo algumas correlacionadas que não possuem tanto efeito inicialmente, e por fim, removendo mais variáveis de modo que não haja nenhuma correlação maior que 0.8 entre as covariáveis.

No ajuste de cada uma das árvores para os três cenários, iremos realizar o ajuste em todo o conjunto e realizar a poda de acordo com o hiperparâmetro *cp* (complexity parameter), o qual será escolhido por validação cruzada e irá realizar a poda da árvores posteriormente. Este parâmetro só irá permitir a divisão da árvore se a redução no erro for maior de que seu valor. A construção da árvore será feita utilizando o pacote **rpart** do software R.

Dito isso, para o primeiro cenário (todas as variáveis), o esquema da árvore está ilustrado na Figura 6.11, onde temos as condições e em cada folha, a saída do modelo, bem como a quantidade de observações com valor 0 à esquerda, e com valor 1, à direita.

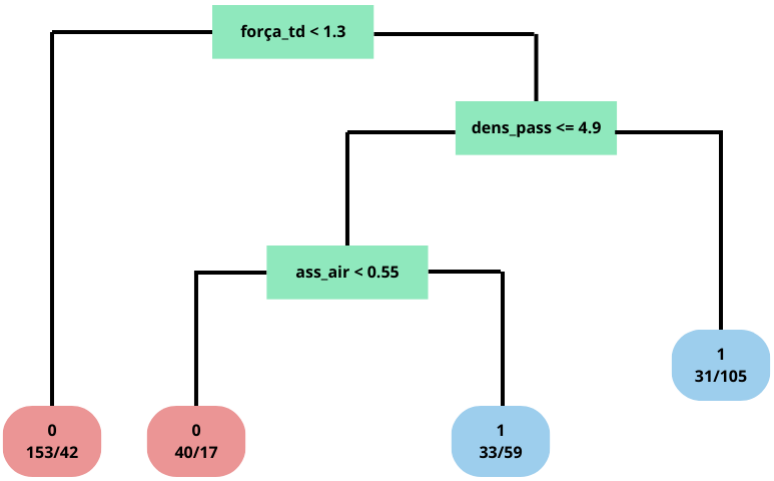


Figura 6.11: Árvore de Classificação para o cenário 1.

Neste caso, o método realizou a seleção de três variáveis: força *touchdowns*, densidade passes e assimetria jardas aéreas. Com relação à regressão logística, todas estas covariáveis também foram selecionadas no primeiro cenário, reforçando evidências de que estas explicam bem a variável resposta.

Para os próximos dois cenários, o esquema da árvore foi o mesmo para ambos os casos, o qual está ilustrado na Figura 6.12, em que apenas a variável força *touchdown* foi selecionada. Analisando a importância desta variável na regressão logística e na floresta (próxima seção), tem-se que a força da rede passes para *touchdown* é bem significativa para explicar a resposta.

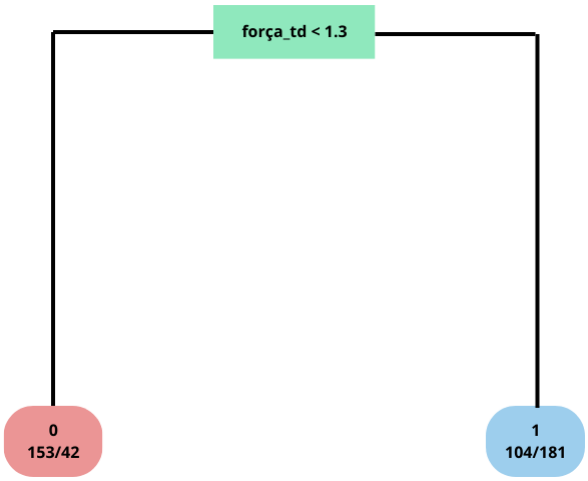


Figura 6.12: Árvore de Classificação para os cenários 2 e 3.

Dessa forma, a fim se as demais variáveis eram selecionadas na árvore de classificação,

um novo ajuste no cenário 2 foi realizado desconsiderando a variável força *touchdown*, de modo a ser possível identificar se outras covariáveis eram importantes nesse cenário.

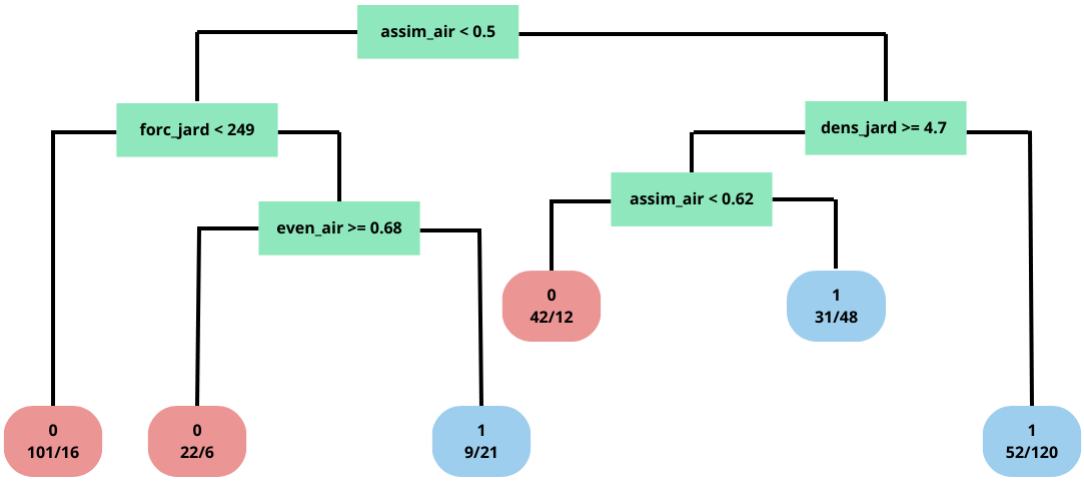


Figura 6.13: Árvore de Classificação para o cenário 3 (sem força td).

Pela Figura 6.13, é possível perceber que existem outras variáveis no cenário 2 selecionadas para explicar a resposta (caso contrário não teríamos nenhuma condições na árvore), são estas: assimetria jardas aéreas, força jardas totais, equidade járdas aéreas e densidade jardas totais. Novamente, assim como no cenário 1, assimetria da rede jardas aéreas foi selecionada.

Dessa maneira, para o ajuste das árvores de classificação levando em conta os três cenários, a medida de centralidade força na rede de passes para *touchdown* apresentou-se de fato significativa em todos os caso e, além desta, outras variáveis foram selecionadas para explicar a resposta de campanha positiva, com destaque para a assimetria na rede de jardas aéreas. Logo, temos evidências de que as medidas de centralidade são úteis para explicar a resposta quando utilizamos modelos de Árvores de Classificação.

6.4 Floresta Aleatória

Por mais que o método de Floresta Aleatória não seleciona variáveis, este é útil ao retornar uma medida de importância para cada uma das covariáveis, medindo o quanto estas foram importantes para diminuir a medida escolhida para o erro de predição. Neste caso, analisaremos o índice de Gini e a acurácia.

Por meio das florestas, um ajuste será realizado, novamente, para cada um dos cenários para que seja possível comparar se os resultados obtidos foram semelhantes, além de

quantificar a quão bem cada variável consegue explicar a resposta.

Para o cenário 1, pela Figura 6.14 temos que a medida de centralidade força para rede de passes para *touchdown* novamente apresenta evidências de ser significativa para explicar a variável resposta, além das outras duas medidas selecionadas pela árvore (densidade jardas aéreas e densidade passes) também apresentarem uma alta importância. Ainda, equidade sucesso e assimetria jardas aéreas apresentaram valores semelhantes com relação a sua importância.

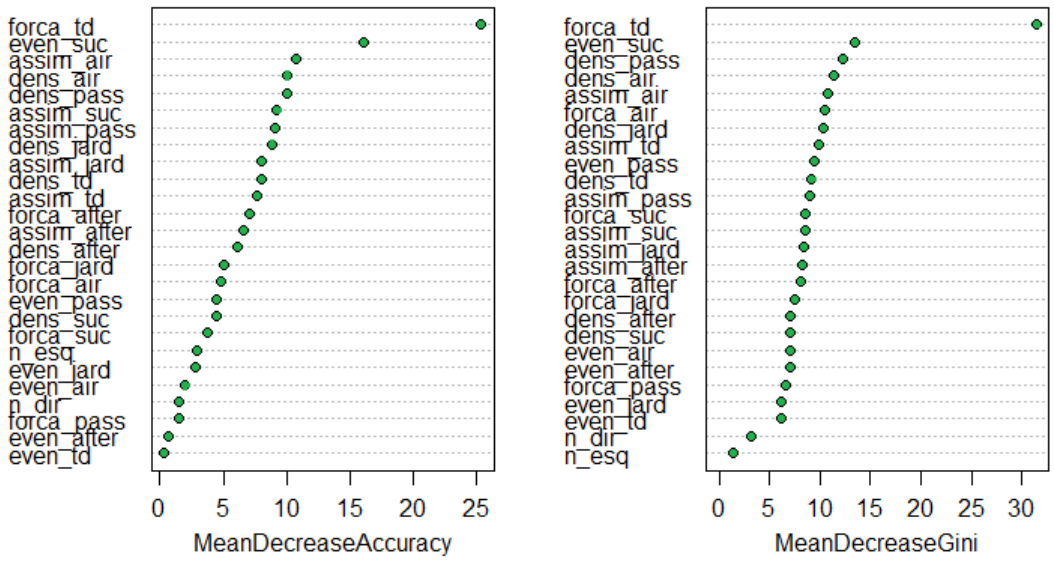


Figura 6.14: Medida de importância para o cenário 1.

Seguindo para o cenário 2, novamente força *touchdown* apresentou o maior valor das métricas, como mostrado na Figura 6.15, além de possuirmos outras medidas como equidade sucesso, assimetria jardas aéreas e densidade jardas totais com notável importância, mesmo não sendo selecionadas na árvore de classificação. Observando a Figura 6.16, temos que a força *touchdown* apresenta uma importância notavelmente maior que as demais, assim como visto no cenário 2, mesmo que algumas covariáveis apresentem bons valores de importância, como equidade sucesso, densidade jardas aéreas (no cenário 3), o que corrobora com o fato de que, no cenário 3, apenas a força foi levada em conta nas árvores de classificação, mas ao retirá-la do modelo, outras variáveis foram selecionadas para explicar a resposta de campanha positiva.

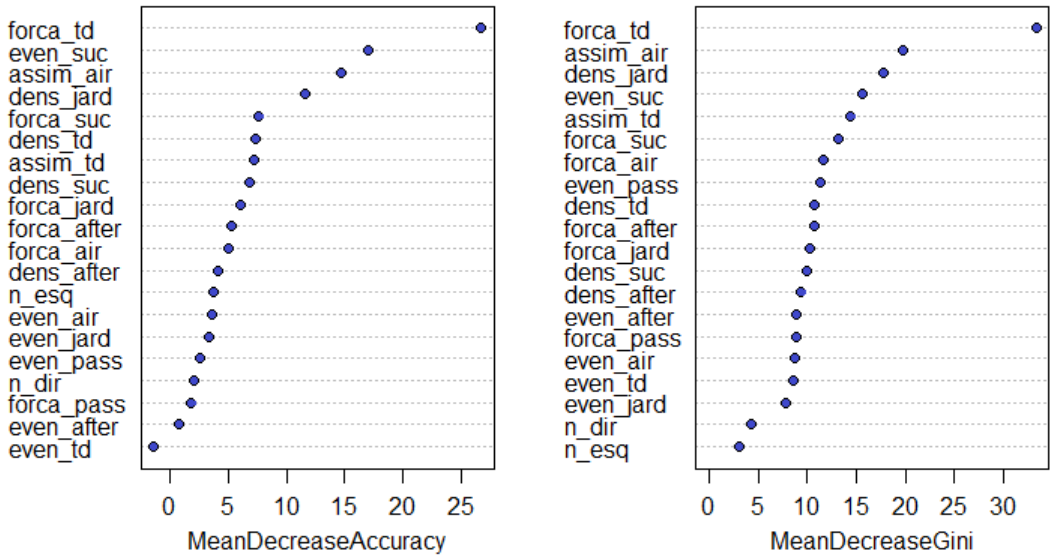


Figura 6.15: Medida de importância para o cenário 2.

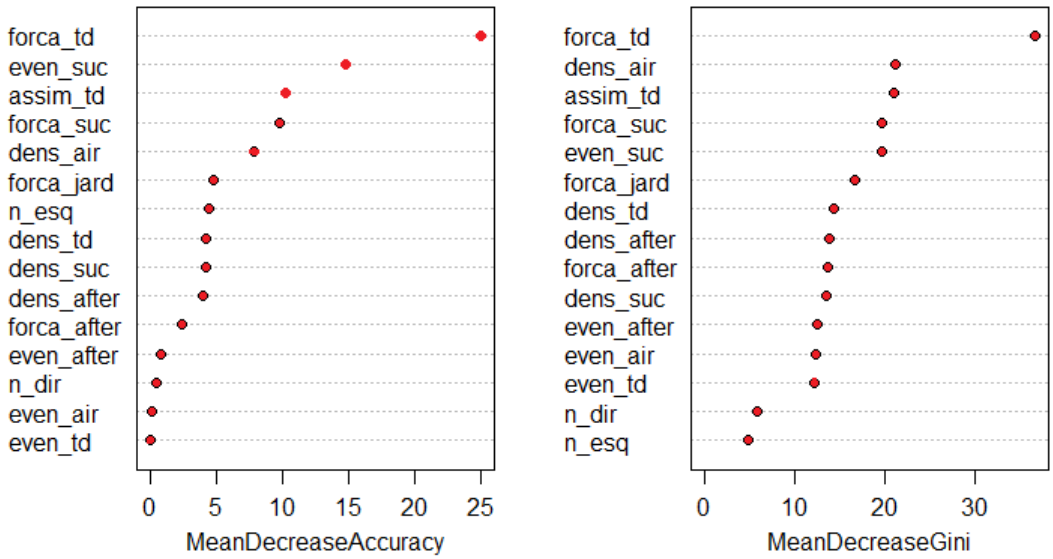


Figura 6.16: Medida de importância para o cenário 3.

Dessa forma, temos que os resultados obtidos analisando as medidas de importância da floresta aleatória foram úteis para colaborar com os resultados obtidos em outros métodos, visualizando a importância das medidas em cada cenário.

6.5 Conclusão

Após a aplicação realizada, já na análise descritiva é possível tomar conclusões a respeito de todo o processo de construção da rede com as medidas de centralidade globais, pois, dentro das medidas, observamos uma correlação muito grande entre os diferentes tipos de peso, de modo que algumas das redes representam a mesma informação em algumas das medidas. Por outro lado, a correlação entre as diferentes medidas nas redes é pequena, indicando que as medidas de centralidade capturam informações diferentes a respeito da estrutura da rede.

Com relação a significância das medidas, para explicar a resposta, a Tabela 6.6 mostra as variáveis selecionadas no último cenário, com uma escala de cores representando a importância das variáveis no método de Floresta Aleatória, quanto mais escuro, mais importante.

Tabela 6.6: Variáveis selecionadas no Cenário 3

Rede	**Medida**	**Lasso**	**E.Net**	**Árvores**	**Floresta**
Passes Completos	Assimetria				
	Equidade				
Jardas Totais	Força				
Jardas Aéreas	Densidade				
	Equidade				
Jardas Pós Recepção	Força				
	Densidade				
	Equidade				
Taxa de Sucesso	Força				
	Densidade				
	Equidade				
Passes p/ TD	Força				
	Densidade				
	Assimetria				
	Equidade				
-	Passadores				
-	Recebedores				

Pela Tabela 6.6, é possível perceber que, mesmo após a remoção de algumas medidas por conta da multicolinearidade, a seleção de variáveis no métodos de regularização da regressão logística, além da seleção da 1 variável na árvores (além das selecionadas no cenário 2, como explicado anteriormente), e a importância segundo floresta aleatórias (acurácia) corroboram para evidências de que as medidas de centralidade globais são úteis na explicação da resposta. Além disso, em pelo menos em um dos métodos, todas as medidas foram selecionadas uma vez, indicando que todas fornecem boas informações a respeito da rede complexa da equipe em uma temporada.

Capítulo 7

Conclusões Gerais

Com o final de todas as etapas realizadas, é possível, agora, verificar se os objetivos definidos anteriormente foram de fatos obtidos. Primeiramente, é de interesse verificar se as redes complexas são capazes de explicar a campanha de um time da NFL em uma temporada por meio de suas medidas de centralidade globais. A partir do ajuste dos modelos de classificação, de fato foi possível observar medidas que são significativas para explicar esta variável resposta e, por mais que algumas covariáveis selecionadas foram diferentes em alguns métodos, em todos pelo menos uma das medidas foram selecionadas como significativas.

Além disso, saber quais das medidas de centralidade que se mostraram significativas era um dos objetivos. Aqui, por mais que a correlação era muito alta entre alguns dos diferentes tipos de pesos, entre as medidas este valor se mostrou baixo, indicando que as diferentes medidas capturaram informações diferentes a respeito da rede e, nos modelos de classificação, todas as quatro medidas foram selecionadas em pelo menos um dos métodos ajustados, indicando que todas foram úteis para explicar a resposta de campanha positiva. Para o jogo em si, saber quais medidas foram significativas e em qual tipo de peso é útil para traçar possíveis planos de jogo para aumentar o desempenho da equipe. Por exemplo, quanto maior a medida densidade de jardas aéreas, menor a chance de uma campanha positiva, assim, a fim de aumentar o desempenho, diminuir esta medida na rede do jogo passado é importante para o sucesso de uma equipe. Ainda, entender quais as medidas mais significativas em detrimento de outras pode ser útil ao traçar quais das medidas de centralidade vale mais a pena de se trabalhar na rede em um primeiro momento.

Logo, a metodologia de utilizar redes complexas combinadas com modelos de classificação para explicar características presentes no jogo de futebol americano, mais pre-

cisamente as campanhas positivas, apresentou bons resultados neste caso. Assim, futuras análises envolvendo redes complexas neste esporte podem apresentar bons resultados também, a partir desta primeira análise.

Referências Bibliográficas

Alatalo, R. V. (1981). Problems in the measurement of evenness in ecology. *Oikos*, páginas 199–204.

Atkinson, A. C. (1981). Two graphical displays for outlying and influential observations in regression. *Biometrika*, **68**(1), 13–20.

Bascompte, J., Jordano, P. e Olesen, J. M. (2006). Asymmetric coevolutionary networks facilitate biodiversity maintenance. *Science*, **312**(5772), 431–433.

Bersier, L.-F., Banašek-Richter, C. e Cattin, M.-F. (2002). Quantitative descriptors of food-web matrices. *Ecology*, **83**(9), 2394–2407.

Da Rocha, L. E. (2009). Structural evolution of the brazilian airport network. *Journal of Statistical Mechanics: Theory and Experiment*, **2009**(04), P04020.

de Oliveira Salim, M. e Brandao, W. C. (2018). Predicting the success of nfl teams using complex network analysis. Em *ICEIS (1)*, páginas 135–142.

de Stefano, E., de Oliveira Farroco, L., Lima, G. B. A., Parrancho, A., Gavião, L. O. e Principe, V. A. (2020). Decision trees for the prediction of outcome of soccer games-historical data analysis. *Brazilian Journal of Development*, **6**(1), 4719–4732.

Dormann, C. F. (2020). Using bipartite to describe and plot two-mode networks in r. *R package version*, **4**, 1–28.

Dormann, C. F., Fründ, J., Blüthgen, N. e Gruber, B. (2009). Indices, graphs and null models: analyzing bipartite ecological networks.

Echegoyen Blanco, I., Martín Buldú, J., Busquets, J., Martínez, J., Herrera-Diestra, J. L., Galeano, J. e Luque, J. (2018). Using network science to analyse football passing networks: Dynamics, space, time, and the multilayer nature of the game.

Félix, L. G., Barbosa, C. M., Vieira, V. d. F. e Xavier, C. R. (2019). A social network analysis of football with complex networks. Em *Anais Estendidos do XXV Simpósio Brasileiro de Sistemas Multimídia e Web*, páginas 47–50. SBC.

GE, R. (2017). Guia da nfl: entenda como funciona o futebol americano. Acesso em: 09 05 2024.

Gifford, M. e Bayrak, T. (2023). A predictive analytics model for forecasting outcomes in the national football league games using decision tree and logistic regression. *Decision Analytics Journal*, **8**, 100296.

Hill, M. O. (1973). Diversity and evenness: a unifying notation and its consequences. *Ecology*, **54**(2), 427–432.

Hocking, R. R. (1976). A biometrics invited paper. the analysis and selection of variables in linear regression. *Biometrics*, páginas 1–49.

Hoerl, A. E. e Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**(1), 55–67.

Hurlbert, S. H. (1971). The nonconcept of species diversity: a critique and alternative parameters. *Ecology*, **52**(4), 577–586.

Izbicki, R. e dos Santos, T. M. (2020). *Aprendizado de máquina: uma abordagem estatística*. Rafael Izbicki.

Mata, A. S. d. (2020). Complex networks: a mini-review. *Brazilian Journal of Physics*, **50**, 658–672.

Newman, M. (2018). *Networks*. Oxford university press.

Paula, G. A. (2004). *Modelos de regressão: com apoio computacional*. IME-USP São Paulo.

Prasetio, D. *et al.* (2016). Predicting football match results with logistic regression. Em *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*, páginas 1–5. IEEE.

PRICE, D. (1965). Networks of scientific papers. *Science (New York, NY)*, **149**(3683), 510–515.

Price, D. d. S. (2011). Networks of scientific papers. Em *The Structure and Dynamics of Networks*, páginas 149–154. Princeton University Press.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rahman, M. H. A. A., Mustapha, A., Razali, N. e Fauzi, R. (2018). Bayesian approach to classification of football match outcome. *International Journal of Integrated Engineering*, **10**(6).

Rejmanek, M. e Starỳ, P. (1979). Connectance in real biotic communities and critical values for stability of model ecosystems. *Nature*, **280**(5720), 311–313.

Rotenberry, J. T. (1978). Components of avian diversity along a multifactorial gardient. *Ecology*, **59**(4), 693–699.

SebastianCarl (2024). *nflfastR*. R package version 4.6.1.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **58**(1), 267–288.

Xavier, F. J. (2024). Prediction of football match outcomes from passing network structure.

Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of anthropological research*, **33**(4), 452–473.