UNIVERSIDADE FEDERAL DO CEARÁ

INSTITUTO UNIVERSIDADE VIRTUAL

CURSO DE GRADUAÇÃO EM SISTEMAS E MÍDIAS DIGITAIS

ANTÔNIO GUILHERME DO NASCIMENTO PEREIRA

EXPLORANDO A EFICIÊNCIA DA APRENDIZAGEM DE MÁQUINA NA PREVISÃO DE RESULTADOS DO BRASILEIRÃO DA SÉRIE A: UMA ANÁLISE DE DADOS E ALGORITMOS

FORTALEZA

2025

ANTÔNIO GUILHERME DO NASCIMENTO PEREIRA

EXPLORANDO A EFICIÊNCIA DA APRENDIZAGEM DE MÁQUINA NA PREVISÃO DE RESULTADOS DO BRASILEIRÃO DA SÉRIE A: UMA ANÁLISE DE DADOS E ALGORITMOS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Sistemas e Mídias Digitais do Instituto Universidade Virtual da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Sistemas e Mídias Digitais.

Orientador: Prof. Dr. José Gilvan Rodrigues Maia

FORTALEZA

2025

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

P489e Pereira, Antônio Guilherme do Nascimento.
Explorando a eficiência da aprendizagem de máquina na previsão de resultados do brasileirão da série A : uma análise de dados e algoritmos / Antônio Guilherme do Nascimento Pereira. – 2025.
46 f.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Instituto UFC Virtual, Curso de Sistemas e Mídias Digitais, Fortaleza, 2025.
Orientação: Prof. Dr. José Gilvan Rodrigues Maia.

1. Aprendizagem de Máquina. 2. Futebol . 3. Previsão de Resultados. 4. Algoritmos. 5. Validação Cruzada. I. Título.

CDD 302.23

ANTÔNIO GUILHERME DO NASCIMENTO PEREIRA

EXPLORANDO A EFICIÊNCIA DA APRENDIZAGEM DE MÁQUINA NA PREVISÃO DE RESULTADOS DO BRASILEIRÃO DA SÉRIE A: UMA ANÁLISE DE DADOS E ALGORITMOS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Sistemas e Mídias Digitais do Instituto Universidade Virtual da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Sistemas e Mídias Digitais.

Aprovada em:

BANCA EXAMINADORA

Prof. Dr. José Gilvan Rodrigues Maia (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. José Wellington Franco da Silva
Universidade Federal do Ceará (UFC)

Prof. Dr. Ernesto Trajano de Lima
Universidade Federal do Ceará (UFC)

AGRADECIMENTOS

Primeiramente, agradeço a Deus, fonte de toda sabedoria, pois acredito que todas as coisas vêm Dele. Nós, como seres humanos, apenas tentamos compreender, com a inteligência que Ele nos deu, as maravilhas complexas e simples que Ele criou em nosso mundo.

Agradeço profundamente à minha família, em especial à minha mãe, Maria Jacinta Pereira do Nascimento, e ao meu pai, Francisco Antônio Pereira, que se dedicaram incansavelmente, dia e noite, para me proporcionar a oportunidade de estudar e obter uma educação digna. Sem o apoio e amor deles, nada disso seria possível.

Aos meus colegas de turma, que formaram uma verdadeira equipe ao longo dos diversos trabalhos realizados juntos, meu sincero agradecimento pela colaboração e companheirismo.

Sou grato também a todos os professores do curso de Sistemas e Mídias Digitais, que me proporcionaram valiosos ensinamentos, especialmente nas áreas de design, programação e artes multimídias, as quais foram fundamentais para o desenvolvimento das minhas habilidades profissionais e pessoais.

Um agradecimento especial ao professor Gilvan, que, com sua imensa paciência, orientação e experiência na área de aprendizado de máquina e inteligência artificial, contribuiu imensamente para a realização deste trabalho de TCC. Aulas com o professor Gilvan eram sempre estimulantes e enriquecedoras, tornando o aprendizado uma experiência prazerosa. Sua dedicação e conhecimento não apenas nos transmitiram o conteúdo necessário, mas também nos motivaram a buscar sempre mais.

“Se cheguei até aqui foi porque me apoiei no ombro de gigantes”

(Isaac Newton)

RESUMO

Com o avanço da revolução tecnológica, a aprendizagem de máquina tem desempenhado um papel cada vez mais relevante em diversas áreas do conhecimento, trazendo benefícios significativos para setores como saúde, educação, detecção de fraudes e análise de grandes volumes de dados. No contexto esportivo, especialmente no futebol, essa tecnologia vem sendo aplicada para compreender padrões complexos e tentar prever resultados com maior precisão. Diante desse cenário, este trabalho investiga a eficiência dos algoritmos de aprendizagem de máquina na previsão dos resultados do Campeonato Brasileiro da Série A, isto é, se a partidade será empate, derrota ou vitória. Busca-se compreender como esses modelos computacionais conseguem se ajustar e aprender a partir de dados históricos das partidas, analisando estatísticas para gerar previsões sobre confrontos futuros. Além disso, a pesquisa propõe uma comparação entre a capacidade preditiva dos algoritmos e a análise humana, avaliando se as técnicas de *Machine Learning* podem superar métodos tradicionais de previsão e a própria intuição dos especialistas. Para isso, foi realizada uma revisão de trabalhos anteriores e uma análise detalhada dos erros cometidos em abordagens preditivas anteriores, com o objetivo de aprimorar as estratégias utilizadas. O estudo contempla todas as etapas do processo, desde a obtenção e tratamento dos dados até a implementação dos modelos de aprendizado de máquina e a avaliação de seu desempenho. Diferentes algoritmos foram testados e comparados para identificar quais apresentam maior taxa de acerto e quais características influenciam diretamente a eficácia das previsões. Por meio dessa abordagem, espera-se contribuir para um melhor entendimento da aplicabilidade da aprendizagem de máquina no cenário esportivo, fornecendo *insights* que possam ser úteis tanto para estudiosos da área quanto para profissionais que utilizam previsões estatísticas, como apostadores e analistas esportivos.

Palavras-chave: Aprendizado de Máquina. Nível de eficiência. Futebol. Tecnologia.

ABSTRACT

With the advancement of the technological revolution, machine learning has become increasingly relevant in various fields of knowledge, bringing significant benefits to sectors such as healthcare, education, fraud detection, and big data analysis. In sports, particularly football, this technology has been applied to understand complex patterns and predict results more accurately. In this scenario, this work investigates the efficiency of machine learning algorithms in predicting the Brazilian Serie A Championship results. The goal is to understand how these computational models can adjust and learn from historical match data, analyzing statistics to generate predictions for future matchups. Furthermore, the research proposes a comparison between the predictive capabilities of algorithms and human analysis, evaluating whether machine learning techniques can surpass traditional prediction methods and the intuition of experts. To achieve this, a review of previous works was carried out along with a detailed analysis of the errors made in other predictive approaches, aiming to improve the strategies used. The study encompasses all stages of the proposed predictive process, from data collection and processing to implementing Machine Learning models and performance evaluation. Different algorithms will be tested and compared to identify which ones exhibit the highest accuracy rate and which characteristics directly influence the effectiveness of predictions. Our approach contributes to a better understanding of the applicability of machine learning in the sports field, providing insights that may be useful both for scholars in the area and for professionals who use statistical predictions, such as bettors and sports analysts.

Keywords

Machine Learning. Efficiency Level. Football. Technology.

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

AUC *Area Under the Curve*

CBF Confederação Brasileira de Futebol

MLP *Multilayer Perceptron*

RNA Rede Neural Artificial

ROC *Receiver Operating Characteristic*

SVM Máquina de Vetores de Suporte

SUMÁRIO

1 INTRODUÇÃO

O futebol não é apenas um esporte; ele representa uma poderosa ferramenta de inclusão social e identidade cultural. No Brasil, sua prática transcende as barreiras sociais, unindo pessoas de diferentes classes e origens em torno de uma paixão comum. Esse esporte é marcado por diversos campeonatos em diferentes países, no qual são expressadas a beleza, a alegria e a técnica dos jogadores em todos os lugares.

Esse esporte movimenta uma enorme quantidade de dinheiro durante todo um ano por meio das premiações das partidas e publicidades. A Confederação Brasileira de Futebol (CBF) para entender qual é o impacto do futebol na economia do Brasil identificou que esse esporte movimenta o total de quase 52,9 bilhões de reais na economia do mesmo, o que equivale a 0,72% da economia (CONTEúDO, 2024). Atualmente, parte dessa movimentação se deve também às plataformas de apostas em jogos esportivos, que pagam aos seus apostadores determinadas quantias por acertar o resultado de uma partida de futebol, isto é, se o resultado do jogo vai ser empate, derrota ou vitória. No entanto, são muitos fatores que podem determinar o resultado de uma partida. Assim, diante disso, fica virtualmente impossível analisar muitas partidas e ter um relatório com a maior probabilidade de determinado evento ocorrer nesse jogo.

1.1 Motivação

Assim, ao inserir uma área da tecnologia ainda um pouco explorada neste contexto de jogos, como aprendizagem de máquina, pode-se ajudar na previsão dessas partidas, no qual basicamente a aprendizagem de máquina é um ramo da Inteligência Artificial (IA) que se concentra no desenvolvimento de algoritmos e modelos capazes de aprender padrões a partir de dados (GALA, 2024). Esses algoritmos estão sendo empregados em grande escala em outros setores, como no ramo de e-commerces, tentando prever os próximos passos do usuário, ou até mesmo em *streams* de vídeo como a Netflix para indicar filmes que provavelmente as pessoas gostariam de assistir com base em suas informações que foram coletadas ao longo da plataforma.

Existem diversos algoritmos para Aprendizagem de Máquina que podem auxiliar na previsão, como uma árvore de decisão, que é uma estrutura hierárquica composta por elementos chamados nós. Cada nó interno, que é um ponto de decisão na árvore, representa um teste em um atributo, que é uma característica ou variável dos dados. Cada ramo simboliza o resultado desse teste, e cada nó folha, que é um ponto final na árvore, representa uma classe, que é a

categoria prevista, ou um valor numérico associado à previsão. A ideia é tentar prever baseado em estatísticas de jogos que já aconteceram antes do resultado do jogo.

Alguns autores têm trabalhado com esta tecnologia e têm obtido resultados interessantes, como o trabalho de (SCHMIDT, 2017), que, em resumo, reuniu dados objetivos de partidas de futebol e treinou modelos de aprendizado de máquina para aprender estes padrões, a fim de prever partidas futuras. Porém, cada campeonato tem sua especificidade, sobretudo o campeonato brasileiro, que em todas estas pesquisas não foi tão bem explorado.

Assim, o presente trabalho centra-se em explorar a eficiência destes algoritmos no contexto do Futebol, coletando dados de sites que oferecem informações das partidas que já aconteceram entre todos os times do brasileirão da série A, através do site *flashscore*, como chutes no gol, números de escanteios, bolas na trave, número de cartões amarelos, número de cartões vermelhos, a fim de verificar e medir a eficiência desses algoritmos nesse contexto das partidas de futebol.

O trabalho então centra-se nessa seguinte pergunta: Qual é a eficiência dos algoritmos de aprendizagem de máquina quando aplicados em estatísticas das partidas de jogos de futebol da série A? Ou seja, qual dos algoritmos de aprendizagem de máquina possui uma maior acurácia nos acertos dos resultados dos jogos do brasileirão, isto é, se a partida será vitória, derrota ou empate. Sendo acurácia uma métrica utilizada para avaliar o desempenho de um modelo de aprendizado de máquina. Ela indica a proporção de previsões corretas em relação ao total de previsões feitas.

1.2 Objetivos

O objetivo geral da pesquisa proposta é identificar a eficiência de diversos algoritmos de aprendizado de máquina, quando submetidos a uma grande quantidade de dados e estatísticas de jogos anteriores do campeonato brasileiro de Futebol da série A. Assim, serão abordados os seguintes objetivos específicos:

- Obter dados de sites esportivos sobre o histórico de partidas entre times de futebol da série A para a montagem da base de aprendizado;
- Identificar e selecionar algoritmos de aprendizado de máquina para a finalidade supracitada; e
- Comparar os algoritmos de aprendizado de máquina, verificando qual possui maior eficiência na previsão em relação à acurácia, verificando se a partida será empate, derrota ou

vitória.

1.3 Propósito e Ressalvas

Os autores deste trabalho ressaltam que o estudo desenvolvido tem finalidade exclusivamente acadêmica e não deve ser utilizado, sob hipótese alguma, como ferramenta oficial para previsão, realização de apostas ou tomada de decisões financeiras. Não há garantias de precisão ou confiabilidade nas previsões apresentadas, visto que resultados esportivos são influenciados por variáveis imprevisíveis e que os métodos ora avaliados também possuem limitações, ou mesmo potenciais falhas de implementação. Dessa forma, o autor se isenta de qualquer responsabilidade sobre o uso indevido das informações aqui expostas.

Além disso, este trabalho não incentiva ou recomenda a participação em jogos de azar ou apostas esportivas. O uso excessivo desses serviços pode levar a problemas financeiros e psicológicos, e recomenda-se que os interessados busquem práticas responsáveis. Qualquer problema técnico relacionado ao uso do produto também não é de responsabilidade do autor, sendo este trabalho amparado por um protótipo desenvolvido para fins de pesquisa.

1.4 Metodologia

Esta pesquisa foi direcionada à estudantes de aprendizado de máquina. O estudo teve como objetivo realizar uma análise detalhada e fundamentada sobre a aplicação de técnicas de aprendizado de máquina na previsão de resultados do **Campeonato Brasileiro da Série A**. Para isso, serão utilizados dados históricos das partidas, incluindo métricas como **chutes a gol**, **número de escanteios**, **faltas cometidas**, **cartões vermelhos**, **cartões amarelos**, entre outros fatores relevantes. Após isso, foi feito um um dataset com os dados que foram capturados e foram testados alguns algoritmos de aprendizagem de máquina nesses, a fim de obter o algoritmo que possuísse a maior capacidade para os acertos dos jogos.

1.5 Condução do Estudo

O estudo foi conduzido em etapas, conforme descrito a seguir: Após a construção de uma base teórica sólida, foi realizada a coleta de dados históricos de partidas do Brasileirão desde 2013. Para isso, será utilizado um código em Python utilizando técnicas de web scraping para capturar estatísticas relevantes. Depois, utilizando a biblioteca sklearn, que fornece diversas

implementações de algoritmos de aprendizado de máquina, será desenvolvido um código em Python para aplicar esses algoritmos aos dados coletados. A eficiência dos diferentes algoritmos será testada na previsão dos resultados do ano de 2023. O objetivo é identificar qual algoritmo apresenta melhor desempenho na previsão de vitórias, derrotas ou empates.

1.6 Características da Pesquisa

A pesquisa é de natureza prática, pois visa a criação de um novo projeto chamado "Score Rank", focado na previsão de resultados do Brasileirão, que é um campeonato ainda pouco explorado nas pesquisas bibliográficas. A abordagem utilizada é quantitativa, uma vez que envolve a análise de um grande volume de dados históricos das partidas do Brasileirão. Esses dados serão submetidos a algoritmos de aprendizado de máquina, com o objetivo de prever os resultados dos jogos, isto é, se é vitória, empate ou derrota.

Inicialmente, a pesquisa assumiu um caráter descritivo, focado na coleta de dados e na comparação da eficiência dos algoritmos de aprendizado de máquina. Em seguida, foi conduzida uma pesquisa experimental, que envolveu a manipulação de variáveis em um ambiente controlado, visando testar hipóteses e estabelecer relações de causa e efeito. O foco esteve na análise da eficiência dos algoritmos e seus desempenhos.

1.7 Avaliação Experimental

Para a avaliação da pesquisa, foi testado se depois de submetido os dados aos algoritmos de aprendizado de máquina, se o programa irá ter um bom desempenho ao tentar acertar os resultados da base que foi dividida para teste, analisando assim a sua eficiência e depois iremos comparar com os resultados de outros artigos, para assim verificar se tal trabalho chegou a níveis de acurácias superiores a de outros artigos, como o (SCHMIDT, 2017), (SANTOS, 2017), (IAGO, 2023) e (CAPOBIANCO, 2019).

2 REFERENCIAL TEÓRICO

Esta seção aborda conceitos fundamentais para a realização da pesquisa, como definição de termos técnicos e algumas pesquisas que já foram publicadas na área de previsão de jogos de futebol, utilizando aprendizado de máquina.

2.1 Apostas Esportivas

Apostas esportivas são práticas de jogo em que os apostadores tentam prever o resultado de eventos esportivos, colocando apostas em resultados que consideram mais prováveis. Esse fenômeno tem se expandido globalmente, abrangendo uma variedade de esportes como futebol, basquete, MMA, entre outros. As apostas podem ser realizadas em plataformas online, e as *odds* (ou cotações) são estabelecidas pelas casas de apostas, refletindo a probabilidade de um determinado resultado ocorrer (ALCANTARA, 2024). A popularidade das apostas esportivas tem levantado discussões sobre sua regulamentação, impactos econômicos e sociais, e questões éticas relacionadas ao vício em jogos.

Em alguns artigos que serão discutidos logo abaixo, como o de (IAGO, 2023), foi utilizado as probabilidades de determinado time ganhar, perder ou empatar das casas de apostas, que são as chamadas odds. Não se sabe ao certo como essas probabilidades foram calculadas, supõe-se que elas também se utilizam de aprendizado de máquina para tentar prever os resultados dos jogos com estatísticas passadas. Assim, se por exemplo, as porcentagens de acertos da aprendizagem de máquina forem muito relevantes, os apostadores poderão obter certa vantagem em comparação com as pessoas que apenas apostam sem levar em conta nenhum método de análise esportiva.

2.2 Aprendizado de Máquina

Aprendizado de máquina, conforme definido por (SAMUEL, 1959), é um campo de estudo dentro da inteligência artificial que se concentra no desenvolvimento de algoritmos de computação capazes de melhorar seu desempenho por meio de dados de treinamento, sem a necessidade de serem explicitamente programados para realizar tarefas específicas (SAMUEL, 1959). Em seu artigo, Samuel demonstra essa ideia utilizando o jogo de damas como exemplo, e mostra como o computador aprende a jogar e melhorar sua performance analisando e aprendendo com partidas anteriores.

Principais definições e conceitos utilizados em aprendizado de máquina:

- **Treinamento:** O processo pelo qual um algoritmo de aprendizado de máquina é alimentado com dados e ajusta seus parâmetros para melhorar seu desempenho em uma tarefa específica.
- **Dados de Treinamento:** Conjunto de dados utilizados para treinar o modelo. Esses dados contêm entradas e saídas conhecidas, permitindo que o modelo aprenda a mapear entradas para saídas.
- **Modelo:** O resultado do processo de treinamento. Um modelo de aprendizado de máquina é um conjunto de regras e parâmetros ajustados com base nos dados de treinamento para realizar previsões ou classificações em novos dados.
- **Algoritmos de Aprendizado:** Métodos matemáticos e estatísticos utilizados para treinar modelos de aprendizado de máquina. Alguns exemplos incluem regressão linear, árvores de decisão, máquinas de vetores de suporte (SVM), redes neurais artificiais e algoritmos de agrupamento (clustering).
- **Validação:** Processo de avaliar o desempenho do modelo em um conjunto de dados diferentes dos dados de treinamento para garantir que o modelo se generalize bem para novos dados.

Outros autores mais atuais (RUSSELL; NORVIG, 2021) afirmam que Machine Learning é o estudo de algoritmos que melhoram automaticamente através da experiência. Esses algoritmos constroem um modelo com base em dados, em vez de seguir instruções programadas explicitamente. Hoje em dia ela é empregada em diversas tarefas automatizadas como na identificação de fraudes, no qual é amplamente utilizado para identificar atividades fraudulentas em transações financeiras. É empregada na navegação e roteamento, no qual essas empresas de transporte utilizam para otimizar rotas e diminuir o tempo de espera ou então em processamento de linguagem natural, fazendo o computador reconhecer determinados padrões, sem a necessidade de programar passo a passo para isso, mas o algoritmo com os dados passados, aprende e é capaz de tomar decisões por si só.

2.3 Métricas de Desempenho

Considerando o contexto de Aprendizado de Máquina, a avaliação do desempenho de um modelo é crucial para determinar sua eficácia. Considerando os problemas de classificação, algumas métricas básicas são amplamente utilizadas: acurácia, precisão e *recall*.

A **acurácia** mede a proporção de previsões corretas em relação ao total de previsões. Ela é calculada pela fórmula:

$$\text{Acurácia} = \frac{\text{Verdadeiros Positivos (VP) + Verdadeiros Negativos (VN)}}{\text{Total de Amostras}}$$

Embora a acurácia seja uma métrica útil, ela pode ser enganosa em conjuntos de dados desbalanceados, onde uma classe é muito mais frequente do que a outra. A **precisão**, por sua vez, mede a proporção de previsões positivas corretas em relação ao total de previsões positivas. Ela é calculada pela fórmula:

$$\text{Precisão} = \frac{\text{Verdadeiros Positivos (VP)}}{\text{Verdadeiros Positivos (VP) + Falsos Positivos (FP)}}$$

A precisão é útil quando o custo de falsos positivos é alto. Por exemplo, em um sistema de detecção de spam, a precisão alta garante que poucos e-mails legítimos sejam marcados como spam. Assim, o **recall** (ou sensibilidade) mede a proporção de amostras positivas corretamente identificadas em relação ao total de amostras positivas reais. Essa métrica é calculada pela fórmula:

$$\text{Recall} = \frac{\text{Verdadeiros Positivos (VP)}}{\text{Verdadeiros Positivos (VP) + Falsos Negativos (FN)}}$$

O *recall* é útil quando o custo de falsos negativos é alto. Por exemplo, em um sistema de diagnóstico médico, o *recall* alto garante que poucas doenças sejam não detectadas.

Por fim, é importante destacar que a escolha da métrica de desempenho depende do problema específico e dos custos associados a diferentes tipos de erros. Neste trabalho, a acurácia é a métrica mais importante por não se tratar de um diagnóstico nem de um problema de filtragem, mas de classificação.

2.4 Naïve Bayes

Para explicar o algoritmo de *Naïve Bayes* conforme Zhang, é essencial compreender que o Naive Bayes é um método de classificação baseado no teorema de Bayes, que assume independência condicional entre os recursos. Segundo (ZHANG, 2004) , ele destaca que o algoritmo Naïve Bayes é uma técnica simples e eficaz para problemas de classificação, sendo

amplamente utilizado tanto na academia quanto no mercado de *machine learning*. Ele se baseia na suposição ingênua de que os recursos são independentes entre si e igualmente importantes para o resultado final. Isso significa que o algoritmo considera cada recurso como contribuindo de forma independente para a probabilidade de uma determinada classe, o que simplifica o processo de cálculo das probabilidades condicionais.

Além disso, Zhang ressalta que o Naïve Bayes é um algoritmo estatístico que utiliza análises de frequências e relações entre os recursos e as classes para fazer previsões inteligentes. Ele é capaz de aprender com os dados de treinamento e aplicar esse conhecimento para classificar novos dados de forma eficiente. Por fim, Zhang destaca que o Naive Bayes é uma excelente opção para iniciantes em aprendizado de máquina, oferecendo uma base sólida para compreender conceitos mais avançados posteriormente.

Figura 1 – Algoritmo de Nave Bayes

$$\mathbb{P}(Y = c|\mathbf{x}) = \frac{f(\mathbf{x}|Y = c)\mathbb{P}(Y = c)}{\sum_{s \in \mathcal{C}} f(\mathbf{x}|Y = s)\mathbb{P}(Y = s)}$$

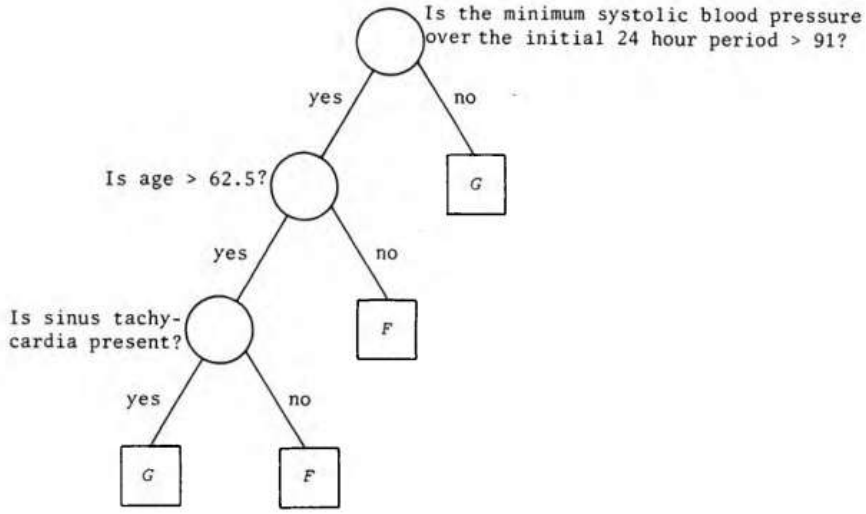
Fonte: (IZBICKI; SANTOS, 2020)

A fórmula do **Naïve Bayes** é baseada no **Teorema de Bayes**, que descreve como atualizar a probabilidade de uma hipótese com base em novas evidências.

2.4.0.1 ***Entendendo cada termo***

- P(HE) → *Probabilidade posterior*: A probabilidade da hipótese HHH ser verdadeira dado que a evidência EEE ocorreu. É o que queremos calcular.
- P(EH) → *Verossimilhança (Likelihood)*: A probabilidade de observar a evidência EEE caso a hipótese HHH seja verdadeira.
- P(H) → *Probabilidade prévia (Prior)*: A probabilidade inicial da hipótese HHH antes de considerar a evidência.
- P(E) → *Probabilidade marginal (Marginal likelihood)*: A probabilidade total da evidência EEE, considerando todas as hipóteses possíveis.

Figura 2 – Uma árvore de classificação utiliza critérios simples em cada nó interno para recortar recursivamente o espaço de características até identificar categorias em suas folhas. O exemplo reproduzido classifica o risco de um paciente: (F) para risco não alto e (G) para risco alto.



Fonte: (BREIMAN *et al.*, 2017)

2.5 Árvores de Decisão

A árvore de decisão é um modelo estatístico que utiliza um treinamento supervisionado, no qual o treinamento supervisionado é uma abordagem de aprendizado de máquina na qual o modelo é treinado utilizando um conjunto de dados rotulado. Isso significa que cada exemplo no conjunto de dados possui uma entrada (features) e uma saída esperada (rótulo ou label). Sendo usada para classificação e previsão de dados, no qual um problema complexo é decomposto em sub-problemas mais simples (GAMA, 2004; BREIMAN *et al.*, 2017). Ela é uma das técnicas mais eficientes aplicadas em vários campos científicos. A árvore de decisão utiliza uma estratégia de dividir-para-conquistar, na qual cada nó de decisão contém um teste para algum atributo, cada ramo descendente corresponde a um possível valor deste atributo, e cada folha está associada a uma classe. O algoritmo CART é um exemplo de algoritmo de classificação que utiliza a árvore de decisão, em qual as regras para dividir cada nó em dois nós filhos são definidas, o conjunto de regras para realizar a divisão de um nó em dois nós filhos é encontrado, e cada nó terminal é associado a uma classe ou a um valor preditivo no caso de regressão.

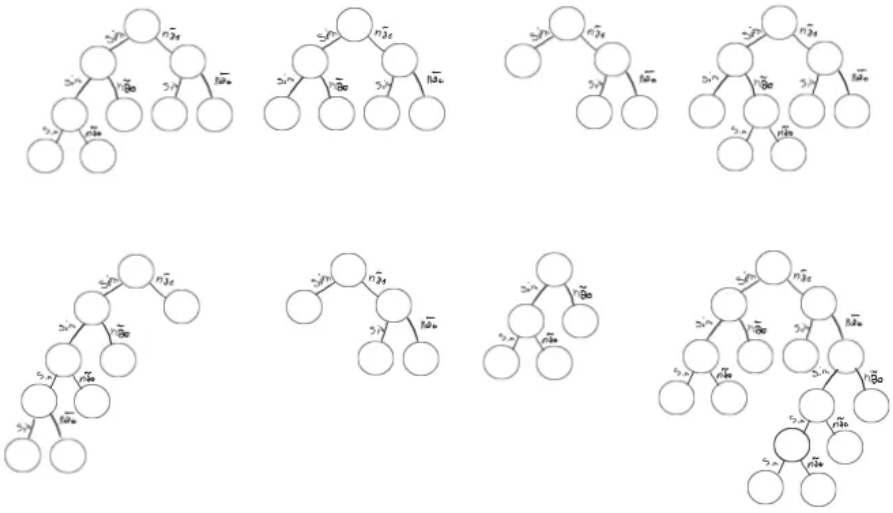
Além disso, as árvores de decisão são amplamente utilizadas em diferentes áreas, como saúde, finanças, astronomia, agricultura, e gestão de recursos, devido à sua flexibilidade, robustez, facilidade de compreensão e velocidade de processamento. Elas permitem, com base em

um conjunto de atributos extremamente diversificados, classificar populações, eventos, produtos, e posteriormente auxiliar na tomada de decisões. Em conclusão, as árvores de decisão são uma ferramenta poderosa de classificação que permite representar regras do tipo "se-então"que são facilmente compreendidas e aplicadas em diferentes áreas, como saúde, finanças, astronomia, agricultura, e gestão de recursos.

2.6 Florestas Randômicas

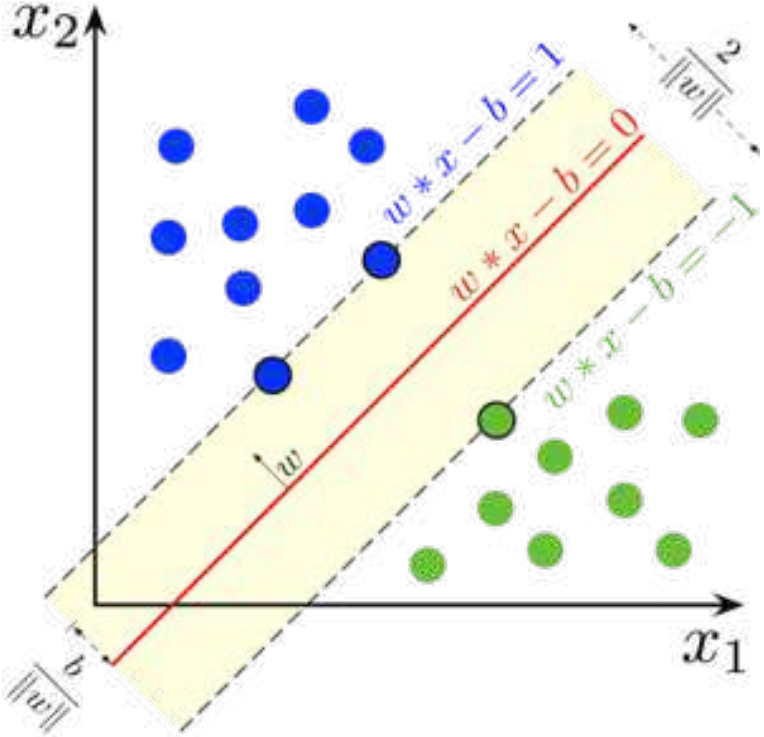
O algoritmo de Florestas Aleatórias, também conhecido como *Random Forests*, é um método de *Machine Learning* que combina múltiplas árvores de decisão para obter um resultado final mais robusto e preciso. Uma das ideias fundamentais por trás desse algoritmo é a utilização de uma técnica chamada Ensemble Learning, que consiste em combinar diferentes modelos para melhorar a eficácia da previsão (AFONSO, 2020).

Figura 3 – Para lidar com casos mais complexos, utiliza-se florestas já que estas reduzem o overfitting (o modelo funciona muito bem no treinamento mas apresenta resultados ruins no conjunto de testes) e conseguem lidar com um grande conjunto de dados. Cada árvore é diferente das outras e vota em alguma resposta. A resposta que tiver mais votos é considerada como a correta.



Fonte: (FIEDLER, 2020)

Figura 4 – Uma máquina de vetores de suporte determina um hiperplano de separação maximal definido em termos de amostras coletadas a partir do conjunto de treinamento, ou seja, os vetores de suporte. Esse critério pode ser relaxado quando as duas classes não são linearmente separáveis (i.e., não existe um hiperplano que classifique corretamente as duas classes para todas as amostras).

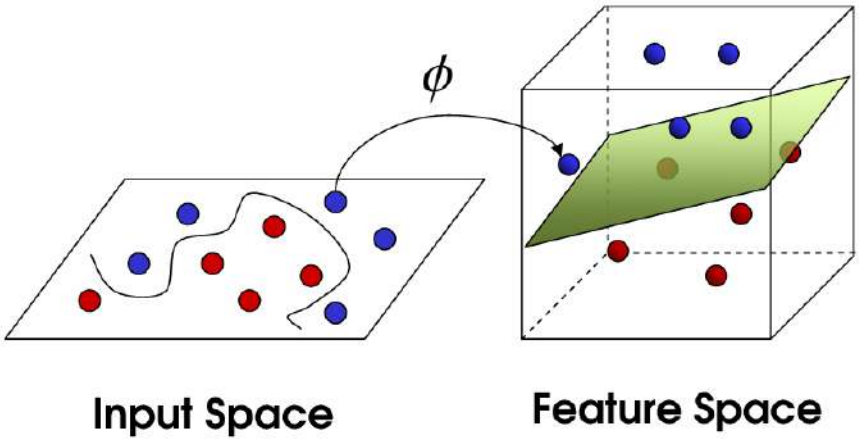


Fonte: (BREIMAN *et al.*, 2017)

2.7 Máquinas de Vetores de Suporte

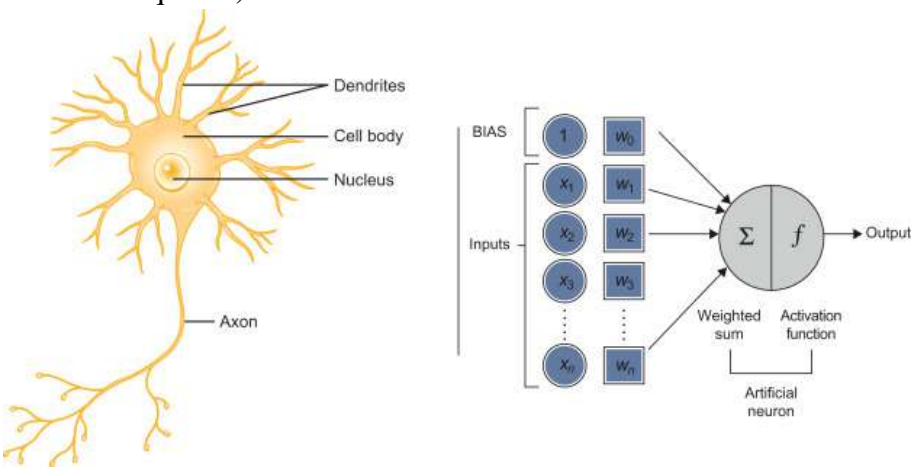
Uma Máquina de Vetores de Suporte, ou SVM, representa uma técnica de aprendizado de máquina amplamente utilizada para classificação e regressão. Introduzidas formalmente por Vladimir Vapnik e Alexey Chervonenkis na década de 1990, as SVMs são projetadas para encontrar um hiperplano que melhor separa diferentes classes em um espaço multidimensional. As SVMs funcionam ao mapear dados de entrada em um espaço de alta dimensão onde um hiperplano pode ser utilizado para diferenciar as classes. O objetivo é maximizar a margem entre as classes, ou seja, a distância entre o hiperplano e os pontos de dados mais próximos de cada classe, conhecidos como vetores de suporte. Essa abordagem é especialmente eficaz em situações onde há uma clara separação entre as classes.

Figura 5 – O potencial teórico do SVM está no uso de um classificador não linear. Para tanto, as amostras são mapeadas do espaço original, de dimensão finita e conhecida, para um espaço de dimensão potencialmente infinita (Espaço de Hilbert). Isso é feito usando dois vetores como entrada de uma função *kernel*, que substitui o produto escalar dos vetores correspondentes no Espaço de Hilbert. Esse artifício é conhecido na literatura como "*kernel trick*".



Fonte: (BREIMAN *et al.*, 2017)

Figura 6 – A computação em cada neurônio *perceptron* de uma Rede Neural Artificial é inspirada pelo neurônio biológico (à esquerda).



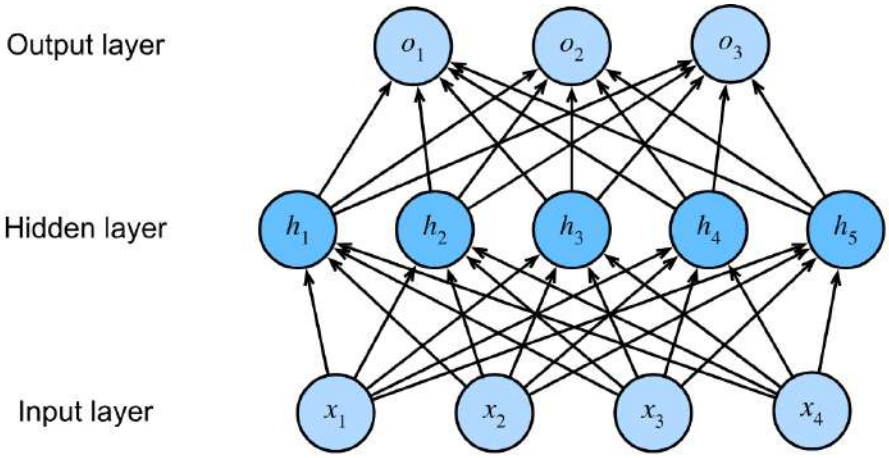
Fonte: (CHANG, 2020)

2.8 Redes Neurais Artificiais

O modelo computacional de RNA é fortemente inspirado no sistema nervoso biológico. Uma RNA é constituída por neurônios artificiais (ROSENBLATT, 1958) organizados em camadas e que são interconectados de acordo com uma topologia, produzindo ativações disparadas em cadeia que mimetizam o conceito biológico de sinapse (ZHANG *et al.*, 2023).

Dependendo da quantidade de camadas de uma RNA, o modelo se enquadra como modelo clássico de aprendizado de máquina ou modelo de Aprendizado Profundo (GOMES; ZHU, 2021; ZHANG *et al.*, 2023).

Figura 7 – Em uma RNA do tipo MLP, cada camada interconectada de uma rede processa a informação, de modo que a saída de cada um de seus neurônio é usada como entrada pelos neurônios da camada seguinte.



Fonte: (ZHANG *et al.*, 2023)

Uma RNA recebe um conjunto grande de informações completas sobre um determinado escopo que se deseja prever ou gerar informação. É a partir desses dados, ou seja, os vetores que alimentam a camada de entrada, que o modelo realiza processamentos e produz saídas.

O objetivo principal de uma rede neural é aprender a relação entre os dados por meio de uma série de operações matemáticas e estatísticas. Contudo, para alcançar essas operações, foram necessários anos de pesquisa e testes que até hoje permeiam em serem validados. No contexto de um aprendizado de máquina inspirado no cérebro humano, é possível dar ênfase às suas semelhanças e diferenças após a interpretação das seguintes definições: O aprendizado humano é um processo pelo qual os indivíduos adquirem conhecimento, habilidades, atitudes ou comportamentos. É uma mudança relativamente permanente no comportamento, resultado da experiência ou da prática; o processo de aprendizado de máquina envolve a exposição do sistema a grandes volumes de dados, nos quais se identificam padrões, realizam-se inferências e aprimora-se seu desempenho conforme é exposta a mais informações. Ela é usada em problemas de visão computacional e também em problemas de linguagem natural (GOODFELLOW *et al.*, 2016; ZHANG *et al.*, 2023).

3 TRABALHOS RELACIONADOS

Foi realizada uma pesquisa no Google Acadêmico usando como chave a *string* de busca "*previsão de resultados do campeonato brasileiro com machine learning*", pode-se encontrar uma variedade de publicações que me forneceram muitas dicas a respeito de como realizar o presente trabalho. Foram cerca de 20 artigos, sendo que os critérios para selecionar estes artigos foram principalmente pelo ano de publicação. É razoável imaginar que quanto mais recente for o trabalho, maior a probabilidade de que tenha recorrido a tecnologias melhores ou mais modernas.

Os trabalhos foram organizados conforme o nível de interesse e a profundidade das informações fornecidas. Muitos artigos não apresentavam uma explicação clara sobre a construção do dataset, principalmente por envolverem dados de empresas. Essa falta de detalhamento dificultou a reprodução dos experimentos por outras pessoas. Assim, a organização dos trabalhos teve como objetivo otimizar o processo de análise e facilitar a absorção das lições aprendidas.

3.1 Schmidt (2017)

Assim, o primeiro artigo analisado foi publicado por Schmidt (2017). O autor realizou uma análise abrangente de diversos artigos na área, destacando como a aprendizagem de máquina tem sido aplicada para prever resultados de campeonatos de futebol. Essa análise incluiu a descrição das bases de dados utilizadas e os algoritmos empregados nesses estudos.

Para sua pesquisa, Schmidt utilizou uma base de dados extensa, composta por 5.570 partidas de 39 equipes. Os dados incluem uma variedade de estatísticas de jogos anteriores, tais como:

- Número de gols das equipes;
- Número de chutes a gol das equipes;
- Número de faltas das equipes;
- Número de escanteios das equipes; e
- Número de cartões amarelos e vermelhos das equipes.

Além das estatísticas de jogos, Schmidt também integrou dados sobre as habilidades individuais dos jogadores. Esses dados foram obtidos por meio de técnicas de *web scraping* da plataforma FIFA. As habilidades consideradas foram:

- Habilidade geral;
- Habilidade de goleiros;
- Habilidade de defensores;
- Habilidade de meio-campo;
- Habilidade de atacantes;
- Habilidade de passe; e
- Habilidade de corrida.

A média dessas habilidades foi calculada para cada jogador, permitindo uma análise mais detalhada do impacto das capacidades individuais no desempenho da equipe. Além disso, foram coletados dados de *odds* de apostas de sete empresas diferentes, oferecendo uma visão abrangente sobre as expectativas de resultados dos jogos.

Schmidt utilizou diversos algoritmos de aprendizagem de máquina para analisar os dados e prever os resultados dos jogos. Os principais algoritmos empregados foram:

- Redes Neurais Artificiais (RNAs), utilizadas para relacionar os atributos que representam as estatísticas das equipes com os resultados das partidas; e
- *Random Forest*, que utiliza múltiplas árvores de decisão para a classificação dos resultados.

A análise realizada pelo autor sugere o potencial da aprendizagem de máquina na previsão de resultados de jogos de futebol, destacando a importância de utilizar uma variedade de dados estatísticos e habilidades dos jogadores.

O uso de algoritmos como RNA e *Random Forest* mostrou-se particularmente eficaz na modelagem de resultados, oferecendo *insights* valiosos para pesquisadores e profissionais da área.

O referencial teórico provido pela pesquisa de Schmidt (2017) serviu como uma base sólida para futuras investigações e aplicações de técnicas de aprendizagem de máquina no esporte. Ao final do artigo, o autor discorre sobre os resultados para os dois algoritmos utilizados, nos quais foi possível chegar a um aproveitamento de 58,77% considerando três categorias de resultados possíveis: vitória, empate e derrota.

Tal resultado pode não parecer muito expressivo à primeira vista, mas é importante destacar que esse desempenho supera a probabilidade randômica. No entanto, essa acurácia se mostrou relativamente baixa quando comparada a outros artigos. Note-se que esses trabalhos utilizaram atributos subjetivos, tais como o momento da equipe, que são atributos mais complicados de se conseguir.

3.2 Iago (2023)

Em um artigo mais recente, Iago (2023), que investigou a eficácia de um modelo de predição para resultados de jogos de futebol, utilizando dados históricos de mais de 100.000 partidas de diversas ligas. O autor coletou informações detalhadas sobre esses jogos, incluindo gols, chutes no alvo, chutes fora do alvo, ataques perigosos, ataques e posse de bola, bem como as cotações (*odds*) oferecidas por sites de apostas.

Foram empregados diversos algoritmos de aprendizado de máquina para a construção e validação do modelo de predição, destacando-se Regressão Logística, K-Neighbors, *Decision Tree*, Gaussian NB, XGBoost, *Random Forest* e Redes Neurais Artificiais. Além disso, técnicas de otimização e validação, como *Grid Search* e validação cruzada (*cross validation*), foram utilizadas para aprimorar a precisão dos modelos.

O objetivo do artigo era identificar possíveis falhas nas probabilidades oferecidas pelas casas de apostas. No entanto, uma das conclusões do autor é que as probabilidades são continuamente ajustadas ao longo da partida e conforme as apostas são realizadas: essa dinâmica dificulta a identificação de inconsistências. Apesar disso, o estudo encontrou algumas situações em que as cotações pareciam estar desajustadas.

O autor reconhece que, para aprimorar a acurácia do modelo preditivo, que é definida como a proporção de previsões corretas feitas pelo modelo em relação ao total de previsões (acurácia), seria necessário incorporar mais variáveis contextuais, como a motivação dos jogadores e desfalques na equipe. Essas variáveis adicionais poderiam fornecer uma visão mais completa e precisa sobre os fatores que influenciam os resultados dos jogos e, consequentemente, as cotações das apostas. Este estudo forneceu uma base importante para a análise crítica das probabilidades oferecidas pelas casas de apostas e sugere direções para pesquisas futuras que possam incluir variáveis mais complexas e contextuais.

3.3 Santos (2017)

Outro estudo se propõe o uso de redes neurais artificiais como uma abordagem eficaz para prever resultados esportivos, com um foco específico nos eventos ocorridos entre 2014 e 2017 (SANTOS, 2017). Essa pesquisa investigou a viabilidade e a precisão desse método estatístico aplicado ao cenário esportivo, considerando as flutuações de elenco e outros fatores variáveis ao longo do tempo.

Com relação aos resultados obtidos, o autor reporta uma taxa de acerto de 50% nas previsões. Isso, segundo o autor, foi acompanhado de um retorno de +2,02% sobre o capital apostado, o que por si é uma métrica questionável. Esse desempenho positivo, apesar das mudanças significativas nas equipes ao longo das temporadas, sugere que o uso de dados históricos como entradas para redes neurais pode fornecer resultados satisfatórios na previsão de eventos esportivos. Este estudo contribui para o avanço do campo de análise esportiva, destacando o potencial das técnicas de aprendizado de máquina na tomada de decisões relacionadas ao investimento e à estratégia esportiva.

3.4 Rein e Memmert (2016) e Capobianco *et al.* (2019) e

(CAPOBIANCO *et al.*, 2019) analisaram trabalhos anteriores e verificaram que geralmente foram analisados dados disponíveis no final da partida nos trabalhos anteriores, tais como o número de gols ou o número de cartões vermelhos. Esse problema também foi discutido por Rein e seus colaboradores (REIN; MEMMERT, 2016). Esses autores afirmaram que um dos principais problemas na análise esportiva é a falta de dados disponíveis e relevantes, o que está se tornando um obstáculo para a modelagem da tomada de decisões táticas em esportes de equipe.

Os autores também fizeram referência a estudos anteriores sobre a predição de resultados de futebol que utilizaram diversas técnicas de aprendizado de máquina. Por exemplo, um estudo utilizou a engenharia de características, que é o processo de usar conhecimento de domínio para selecionar e transformar variáveis brutas em características que melhor representem o problema que está sendo resolvido por um modelo de aprendizado de máquina, resultando em uma melhoria no desempenho do modelo, e análise exploratória de dados para identificar os fatores mais importantes na predição de resultados da *Premier League* inglesa.

A abordagem proposta no trabalho do autor diferiu de estudos anteriores ao considerar novos tipos de dados, como a disposição espacial dos jogadores no campo, e ao focar em uma classificação binária (vitória ou não vitória). A metodologia proposta é dividida em duas fases principais: a construção do modelo e a predição de resultados em duas etapas. Na fase de construção do modelo, dados brutos são adquiridos a partir de relatórios de partidas e, após um processo de limpeza e pré-processamento, formam um conjunto de características usado para treinar modelos de classificação.

O conjunto de dados utilizado foi construído a partir de relatórios de partidas da Série A italiana da temporada 2017-2018, contendo 98 atributos e 378 instâncias. Cada instância

representa uma partida específica da liga, com diversos tipos de atributos como:

- Informação sobre a posse de bola em cada tempo;
- Disposição espacial dos jogadores em campo;
- Gols marcados e sofridos; e
- Outras informações relevantes sobre as equipes.

Um total de 20 características foram selecionadas através do método de análise de componentes principais com busca heurística *best-first*. Entre essas características estão a distância percorrida pelos jogadores, a velocidade média, a posse de bola em diferentes períodos da partida e a recuperação de bolas no meio-campo.

Para avaliar a eficácia do vetor de características, foram escolhidos seis algoritmos de classificação: J48, SMO, RepTree, Random Tree, RandomForest e MultilayerPerceptron. A avaliação foi realizada utilizando *k-fold cross-validation*, com $k = 20$, permitindo uma estimativa robusta dos modelos construídos.

Os resultados experimentais indicaram que o algoritmo RandomForest apresentou o melhor desempenho, com uma precisão média de 0,857 e *recall* de 0,750 na predição de partidas vencidas e uma precisão média de 0,862 na predição do número de gols. Outras métricas, como taxa de falsos positivos (FP rate), *F-Measure* e área sob a curva *Receiver Operating Characteristic* (ROC), i.e., *Area Under the Curve* (AUC), também foram consideradas para avaliar a performance dos algoritmos.

A análise dos dados revelou que as técnicas de aprendizado de máquina são capazes de prever com sucesso os resultados das partidas de futebol, com a RandomForest se destacando por seu uso de múltiplas árvores de decisão para melhorar a precisão da classificação. Além disso, a metodologia proposta permite prever o resultado da partida e o número de gols antes do final do jogo, oferecendo aos treinadores a possibilidade de ajustar suas táticas em tempo real.

3.5 Conclusões Preliminares

Os trabalhos descritos anteriormente serviram de base para a construção deste projeto e, através deles, foi possível perceber que somente dados das partidas podem não ser suficientes para alcançar uma previsão eficiente. No entanto, este trabalho adotará uma abordagem prática, utilizando atributos objetivos que podem ser obtidos com maior facilidade na internet, como número de chutes ao gol, número de escanteios e número de faltas.

O objetivo é verificar se a utilização desses dados objetivos é suficiente para alcançar

bons resultados na previsão dos resultados do Brasileirão. Dessa forma, este estudo se concentrará na análise e combinação desses atributos, avaliando a eficiência dos algoritmos de aprendizado de máquina na previsão dos resultados das partidas.

4 DESENVOLVIMENTO DO PROJETO

Nesta seção, será descrito o desenvolvimento do projeto desde a captura dos dados para a construção dos datasets e também a apresentação dos seus respectivos resultados.

4.1 Seleção dos Dados e Processamento

Para a seleção dos dados utilizados neste projeto, foi realizada uma pesquisa criteriosa em plataformas que disponibilizam informações sobre partidas de futebol. Como parte desse processo, foi desenvolvido um algoritmo em Python para extrair os dados dessas fontes, com foco em partidas do Campeonato Brasileiro da Série A. A principal API utilizada para essa coleta foi a API-Football, acessível por meio do serviço RapidAPI.

Após a extração dos dados, foi estruturado um banco de dados PostgreSQL denominado `score_rank` para armazenar as informações coletadas. Esse banco de dados foi modelado com as seguintes tabelas:

- `teams`: responsável por armazenar os dados dos times participantes.
- `matches`: contém informações sobre as partidas realizadas, incluindo os times que competiram e estatísticas relevantes de cada confronto.
- `statistics`: registra as estatísticas detalhadas dos jogos.

Posteriormente, após o treinamento dos modelos de aprendizado de máquina e a geração das previsões, foram adicionadas duas novas tabelas:

- `results_2023`: armazena os resultados reais das partidas ocorridas no ano de 2023.
- `predictions`: contém os dados relacionados às previsões feitas pelo modelo, indicando se a previsão foi bem-sucedida ou não para cada partida.

Na figura 1, é apresentada a estrutura do banco de dados em formato de esquema.

Para uma melhor compreensão, irei explicar melhor os dados que são armazenados na tabela `statistics`, pois ela contém os dados que serão colocados no algoritmo de aprendizagem de máquina, a fim de que ela possa predizer o resultado dos jogos.

A tabela `statistics` contém informações detalhadas sobre o desempenho das equipes em partidas de futebol, armazenando métricas quantitativas relacionadas às finalizações, faltas e outros aspectos do jogo. A seguir, são descritos os atributos presentes na tabela:

- `id`: Identificador único da estatística registrada para uma determinada partida.
- `team_id`: Identificador da equipe à qual as estatísticas registradas pertencem.

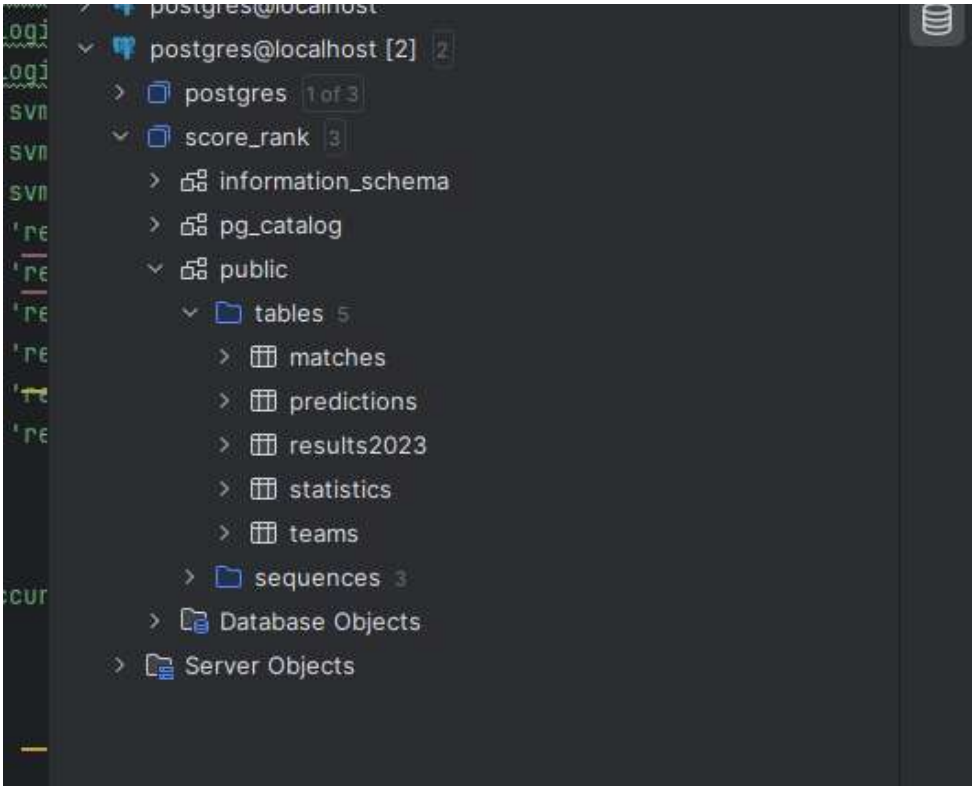


Figura 8 – Arquitetura do banco de dados

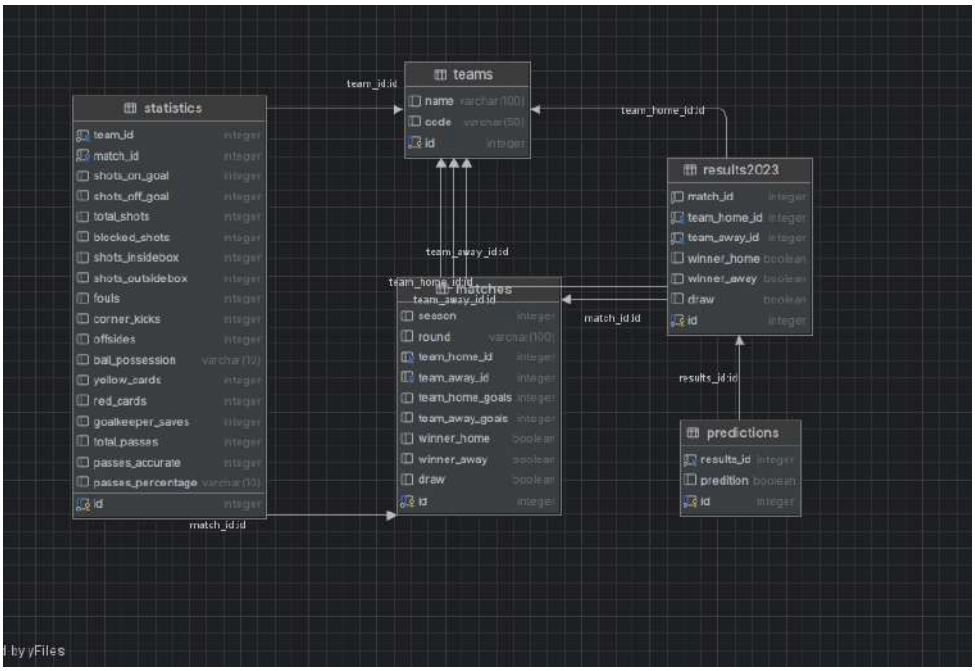


Figura 9 – Diagrama do banco de dados

- `match_id`: Identificador da partida na qual os dados estatísticos foram coletados.
- `shots_on_goal`: Número de finalizações da equipe que foram direcionadas ao gol e exigiram uma defesa do goleiro adversário ou resultaram em gol.
- `shots_off_goal`: Número de finalizações que foram para fora do gol, sem a necessidade de intervenção do goleiro.

- `total_shots`: Quantidade total de finalizações realizadas pela equipe na partida, englobando chutes no gol, para fora e bloqueados.
- `blocked_shots`: Número de finalizações que foram interceptadas por jogadores adversários antes de alcançar o goleiro.
- `shots_insidebox`: Número de finalizações realizadas dentro da grande área adversária, refletindo a proximidade da equipe ao gol no momento do chute.
- `shots_outsidebox`: Número de finalizações feitas fora da grande área, indicando tentativas de gol a partir de longa distância.
- `fouls`: Quantidade total de faltas cometidas pela equipe durante a partida.

Essas estatísticas são fundamentais para a análise do desempenho das equipes, permitindo a aplicação de técnicas de aprendizado de máquina para prever resultados de partidas e identificar padrões estratégicos no futebol.

Após a realização da obtenção de dados, foi iniciada a construção do projeto em Python, com a instalação da biblioteca `sklearn`, que é uma biblioteca conveniente por concentrar os principais algoritmos utilizados na aprendizagem de máquina.

O primeiro passo foi pegar os dados que estão armazenados no banco de dados e transformá-los em um formato que fosse possível que a máquina entendesse. Para isso, para todos os times, foi gerado um arquivo com extensão `csv`, onde cada coluna estava ao lado da respectiva coluna dos dados do time adversário, como por exemplo, esse caso abaixo.

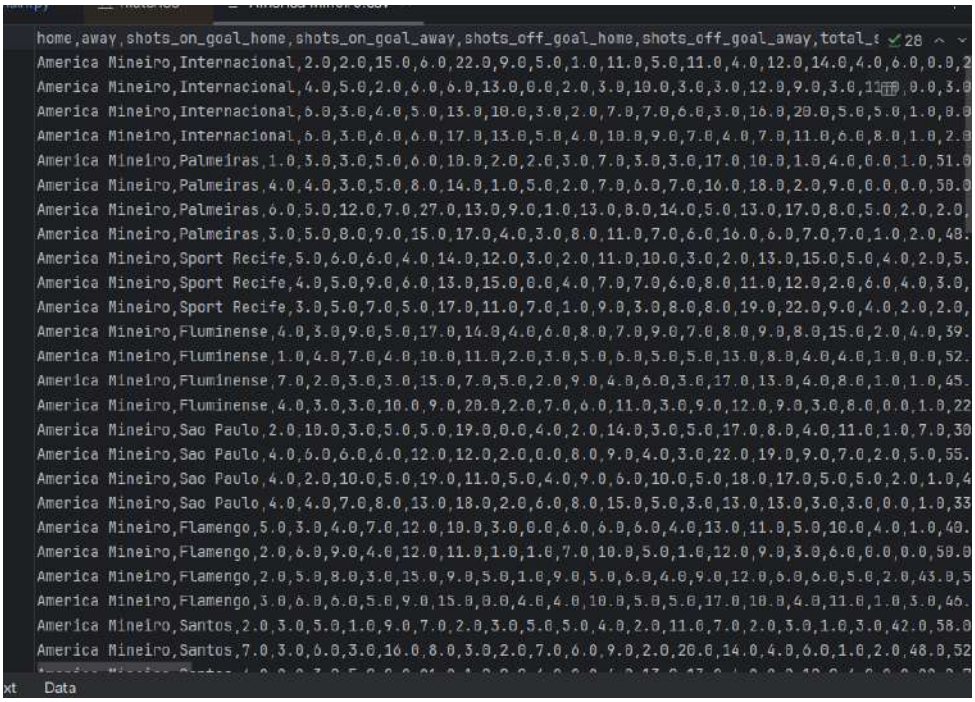


Figura 10 – CSV contendo os dados de todas as partidas realizadas por um time

Assim se por exemplo temos `shots_on_goal_home` que são chutes no gol do time da casa e do outro lado temos `shots_on_goal_away` que são chutes no gol do time adversário, fazendo com que assim cada uma das propriedades dos jogos estejam lado a lado.

Após esse passo, é preciso processar os dados, pois na aprendizagem de máquina é preciso que os dados estejam dentro dos padrões, um dos problemas é o fato de possuir variáveis nominais dentro da base de teste (são variáveis que não são números, como por exemplo o nome do time que está jogando, exemplo Fortaleza e Flamengo, assim é preciso converter o mesmo para um número conveniente) ou haver variáveis zeradas na base de treinamento. Isso pode ser um problema quando se trata desses algoritmos, que são implementações de fórmulas matemáticas, se houver um zero, devido às multiplicações que ocorrem podem gerar resultados tendenciosos, então foi feito um processamento, no qual se determinado atributo vier zerado, é colocado no lugar dele a média daquele atributo, levando em consideração todas as partidas disputadas e também nessa média é levado em consideração o resultado do jogo, por exemplo, se naquela partida, aquele atributo que veio zerado foi uma vitória, então a substituição desse valor zerado será pela média daquele atributo, no qual o resultado do jogo foi vitória.

Figura 11 – Método de tratamento de dados

Depois foi feita a divisão de elementos previsores e classes. Ou seja, foi separado o que são atributos e o que são resultados.

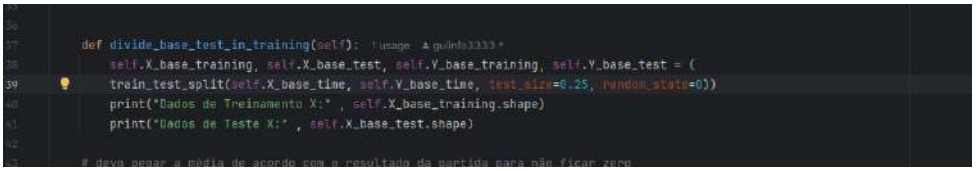


Figura 12 – Separação entre elementos previsores e classes

Outro passo é executar o que chamamos de *Label Encoding*, que é quando transformamos variáveis nominais em números, como por exemplo o nome dos times, pois como se trata

de algoritmos matemáticos, é preciso que só existam números.

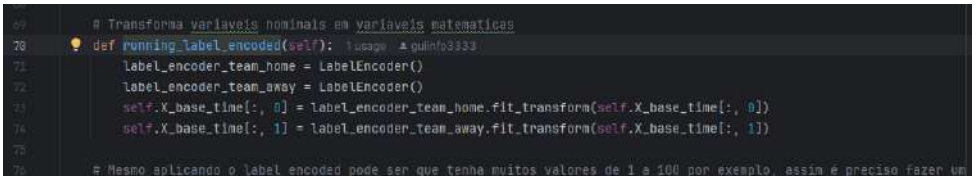


Figura 13 – Execução da técnica Label Encoding

Em seguida, é executada outra técnica na aprendizagem de máquina que chamamos de *One Hot Encoding*, que é uma técnica aplicada ao *Label Encoding*, pois na base de dados podem ter valores que são muito altos e outros muito baixos. Então, para evitar que o algoritmo seja tendencioso, é preciso equilibrar os dados de forma que eles não estejam muito distantes, essa é a técnica do *One Hot Encoding*.

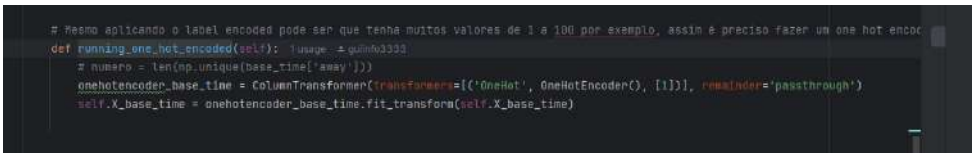


Figura 14 – Execução da técnica One Hot Encoding

E por fim, o último passo do tratamento de dados é dividir nossos dados em dados de treino e dados de teste, a fim de que a máquina possa aprender os resultados baseados nos diferentes algoritmos de aprendizagem de máquina. No exemplo abaixo estamos definindo que 25% dos dados serão para teste e 75% dos dados serão para treinamento.

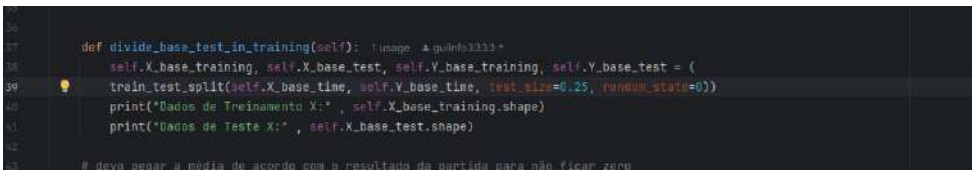


Figura 15 – Divisão entre dados de teste e dados de treino

Depois de ter feito essas validações, foi iniciada a etapa do treinamento em si, que consiste em aplicar essa base de dados em um algoritmos de aprendizagem de máquina, conforme na figura abaixo. Ao todo, dos mais de 3.039 registros, foram divididos em 2448 registros para treinamento e 816 registros para teste. Desses registros são 750 empates, 778 derrotas e 1511 vitórias. Como podemos observar, o número de vitórias é muito maior que o número de derrotas e empate, o que pode gerar um problema de overfiting, o qual é quando o algoritmo aprende muito determinados padrões, porém para outros casos que são diferentes dele, erra em demasia. Enfim,

para tentar amenizar esse problema foram utilizado alguns parâmetros no próprio algoritmo de aprendizagem de máquina, no qual é rebalanceado os dados de forma que um peso é atribuido a cada uma das classes, fazendo com que assim essa disparidade na quantidade dos registros deixe de ser um problema e a máquina possa aprender, diminuindo os riscos de overfiting, como na figura abaixo, no qual colocamos o classweight = 'balanced' no algoritmo de árvores de descisão.

A figura Figura 23 mostra a implementação dos algoritmos de aprendizagem de máquina,utilizando a base de treinamento que foi montada a partir do nosso trabalho. Nas figuras abaixo, é mostrado os hiperparâmetros que foram obtidos através de uma técnica chamada Grid Search, que é uma técnica utilizada para encontrar os melhores hiperparâmetros de um modelo de aprendizado de máquina. Ele funciona testando exaustivamente todas as combinações possíveis de valores para esses hiperparâmetros dentro de um determinado conjunto (grid).

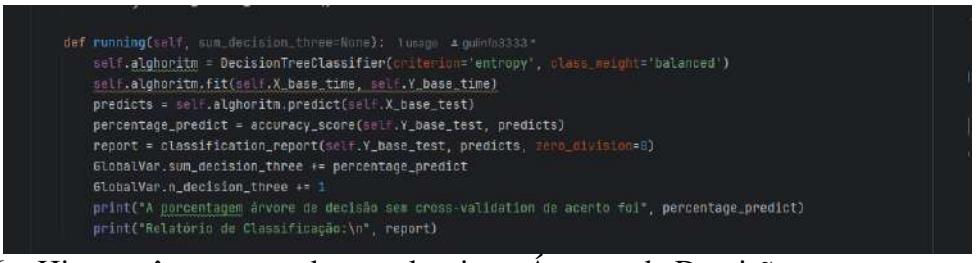


Figura 16 – Hiperparâmetros usados no algoritmo Árvores de Descisão

Figura 17 – Hiperparâmetros usados no algoritmo KNN

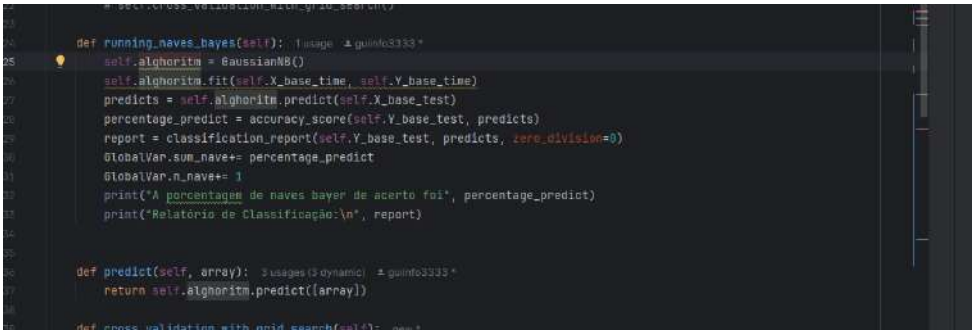


Figura 18 – Hiperparâmetros usados no algoritmo Nave Bayes

```
def running(self):  1 usage  ± guilhfo3333 *
    #Posso configurar a quantidade de árvores que vai ser utilizada
    try:
        self.forest= RandomForestClassifier(n_estimators=5, criterion='entropy', random_state=0, class_weight='balanced')
        self.forest.fit(self.X_base_time, self.Y_base_time)
        predicts = self.forest.predict(self.X_base_test)
        percentage_predict = accuracy_score(self.Y_base_test, predicts)
        report = classification_report(self.Y_base_test, predicts, zero_division=0)
        GlobalVar.sum_random_forest+= percentage_predict
        GlobalVar.n_random_forest+= 1
        print("A porcentagem florestas randômicas de acerto foi", percentage_predict)
        print("Relatório de Classificação:\n", report)
    except ValueError as e:
        print(f" {e}.")
```

Figura 19 – Hiperparâmetros usados no algoritmo Random Forest

```
#hidden_layer_sizes é calculado a partir da forma (quantidade de atributos + __1) / 2, nesse caso (65 + 1) / 2
def running(self):  1 usage  ± guilhfo3333 *
    self.alghoritm = MLPClassifier(verbose=True, max_iter = 1000, tol=0.000010,
                                   hidden_layer_sizes = (33,33))

    self.alghoritm.fit(self.X_base_time, self.Y_base_time)
    predicts = self.alghoritm.predict(self.X_base_test)
    percentage_predict = accuracy_score(self.Y_base_test, predicts)
    report = classification_report(self.Y_base_test, predicts, zero_division=0)
    GlobalVar.sum_neu+= percentage_predict
    GlobalVar.n_neu+= 1
    print("A porcentagem rede neural de acerto foi", percentage_predict)
    print("Relatório de Classificação:\n", report)

def generate_matrix_confusion(self):  1 usage  ± guilhfo3333 *
```

Figura 20 – Hiperparâmetros usados no algoritmo Redes Neurais

```
def running(self):  1 usage  ± guilhfo3333 *
    self.alghoritm = LogisticRegression(random_state=1, class_weight='balanced')
    self.alghoritm.fit(self.X_base_time, self.Y_base_time)
    predicts = self.alghoritm.predict(self.X_base_test)
    percentage_predict = accuracy_score(self.Y_base_test, predicts)
    GlobalVar.sum_logistic_regression+= percentage_predict
    GlobalVar.n_logistic_regression+= 1
    report = classification_report(self.Y_base_test, predicts, zero_division=0)
    print("A porcentagem da regressão logística de acerto foi", percentage_predict)
    print("Relatório de Classificação:\n", report)

def predict(self, array):  3 usages (3 dynamic)  ± guilhfo3333
```

Figura 21 – Hiperparâmetros usados no algoritmo Regressão Logística

4.2 Resultados

Nesta seção, apresentamos os resultados obtidos a partir da aplicação de diferentes algoritmos de aprendizado de máquina na previsão de resultados das partidas do Brasileirão Série A. O objetivo foi avaliar o desempenho de cada modelo com base na taxa de acertos obtida nos testes realizados.

Primeiramente, será mostrado a matriz de confusão para cada um dos algoritmos que foram utilizados, no qual é uma ferramenta essencial para avaliar o desempenho de modelos de classificação. Ela fornece uma visão detalhada das previsões corretas e incorretas, facilitando a análise dos erros do modelo, no qual a classe 1 significa empate, classe 0 significa derrota e a classe 2 significa vitória. Para os dados de teste, foram utilizados 816 registros.

A seguir serão mostrados os resultados obtidos em porcentagem de acertos, depois de submetidos à aprendizagem de máquinas pelos seguintes algoritmos abaixo, os quais foram escolhidos por conta do desempenho relatado pelos trabalhos relacionados:

Figura 22 – Hiperparâmetros usados no algoritmo SVM

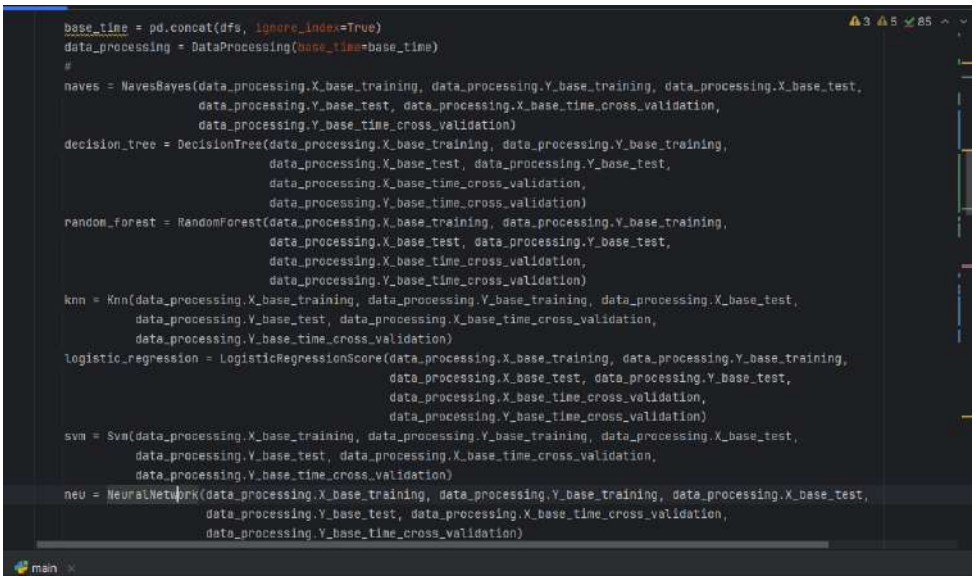


Figura 23 – Rodando o treinamento da base de dados

4.3 Análise Comparativa dos Algoritmos

Algoritmo	**Taxa de Acerto (%)**
Naive Bayes	52%
Árvore de Decisão	95%
Florestas Randômicas	83%
KNN	51%
Regressão Logística	68%
SVM	69%
Rede Neural	66%

Tabela 1 – Desempenho dos algoritmos analisados.

A análise dos resultados evidencia que o algoritmo de Árvore de Decisão apresentou a melhor performance, alcançando uma taxa de acerto de 95%. Esse resultado pode ser atribuído à sua capacidade de lidar bem com dados categóricos e à facilidade de interpretação das regras geradas pelo modelo. Isso pode ser visto na Tabela 1, que exibe as médias de acertos obtidas por cada um dos algoritmos analisados.

Em seguida, Florestas Randômicas obteve uma precisão de 83%, destacando-se

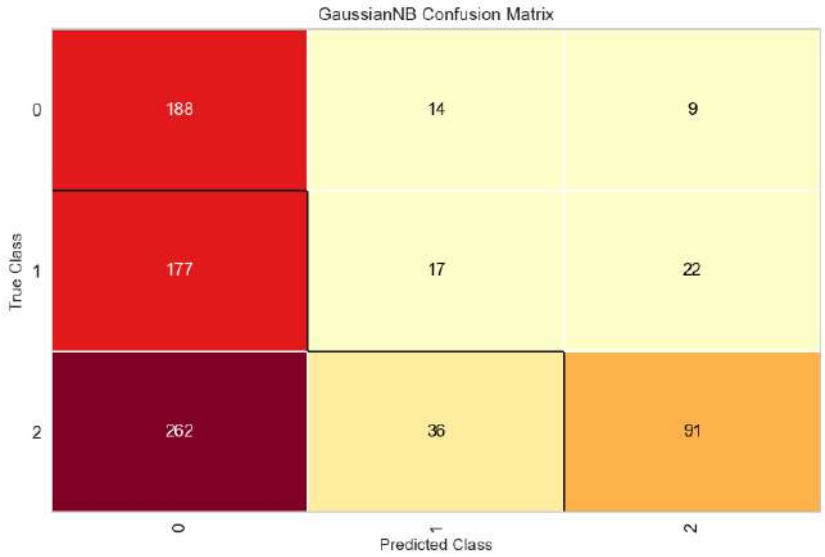


Figura 24 – Matriz de Confusão Naves Bayes.

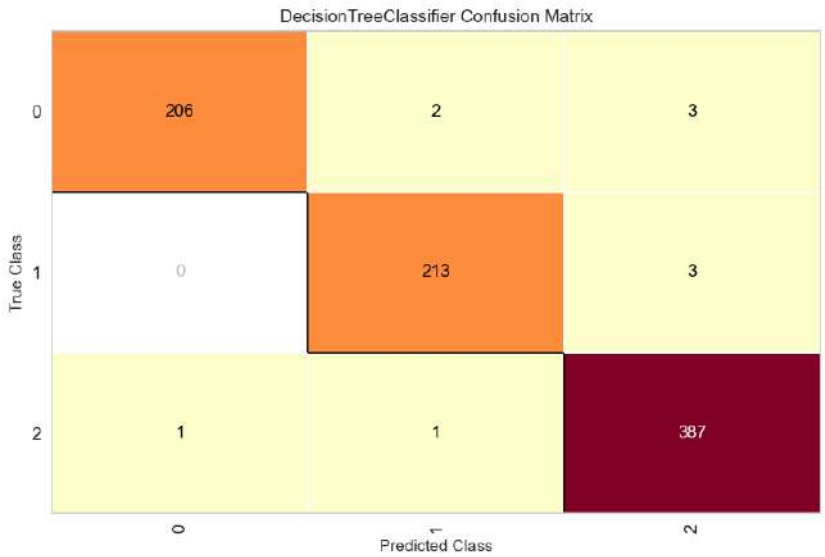


Figura 25 – Matriz de Confusão Árvores de Descisão.

como um dos métodos mais robustos devido à sua abordagem baseada em múltiplas árvores de decisão.

Os algoritmos SVM (69%) e Regressão Logística (68%) apresentaram desempenhos intermediários, enquanto as Redes Neurais (66%) ficaram ligeiramente abaixo. Já o Naive Bayes (52%) e o KNN (51%) tiveram os piores desempenhos, sugerindo que esses métodos podem não ser os mais adequados para a previsão dos resultados do campeonato.

Apesar de resultados bastante satisfatórios, corre-se o risco de durante a aprendizagem de máquina gerar overfiting, então uma forma de testar se a máquina realmente conseguiu aprender e não apenas decorar os registros, foi utilizada uma técnica chamada validação cruzada,

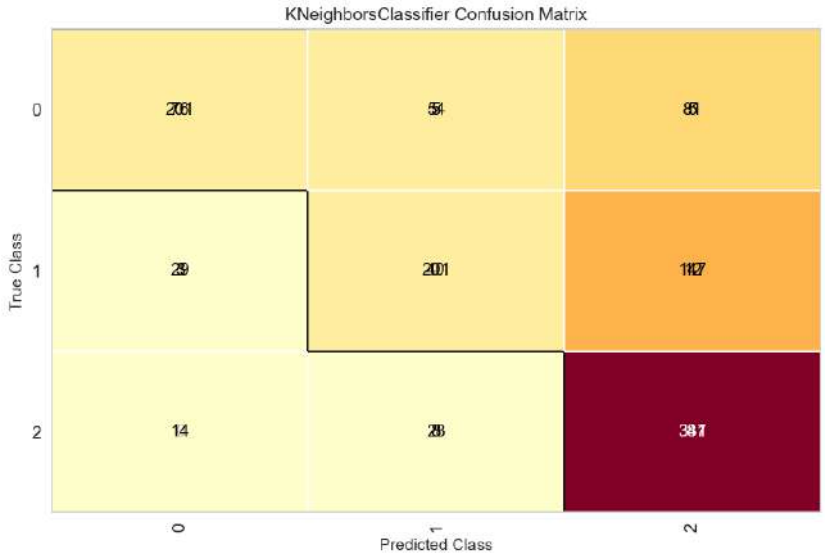


Figura 26 – Matriz de Confusão KNN.

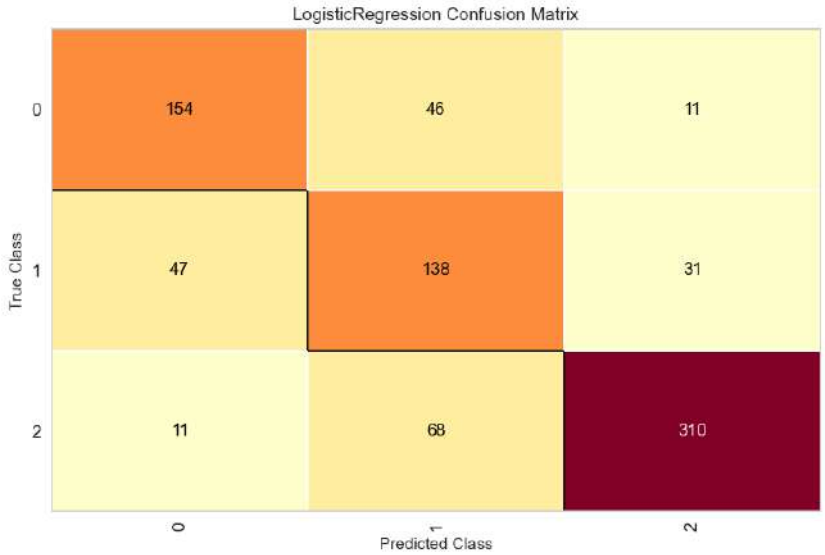


Figura 27 – Matriz de Confusão Regressão Logística.

que consiste em não dividir inicialmente a base em dados de teste e dados de treinamento, de uma forma que todos os dados sejam testados como dados de teste e também de treinamento, garantindo assim que o algoritmo não se adapte muito bem somente a uma base de treinamento. Assim teremos uma ideia, se realmente está acontecendo overfiting, uma vez que as médias de cada uma desses conjuntos estiverem bem dispersas.

Foi executada a validação cruzada somente para os 2 algoritmos com maior quantidade de acertos que foram Florestas Randômicas e Árvores descisão devido ao alto custo computacional.

Os resultados indicam que o algoritmo **Random Forest** apresentou uma média ligei-

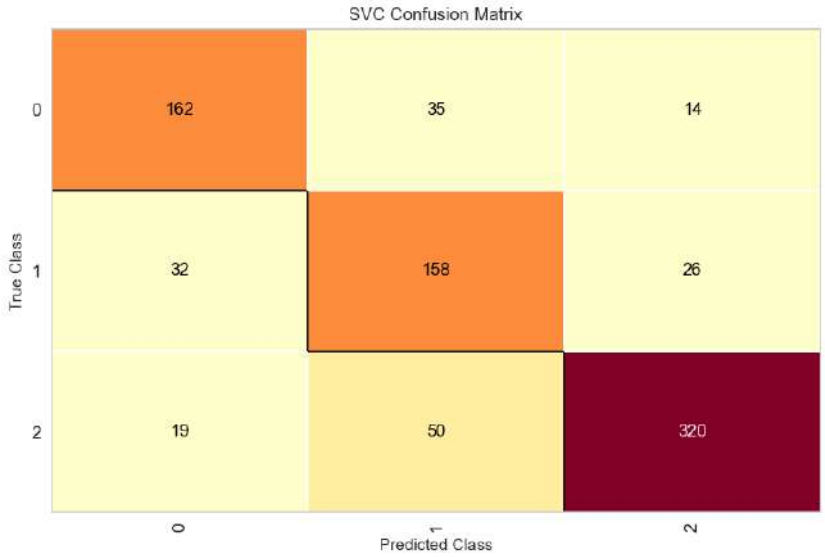


Figura 28 – Matriz de Confusão SVC.

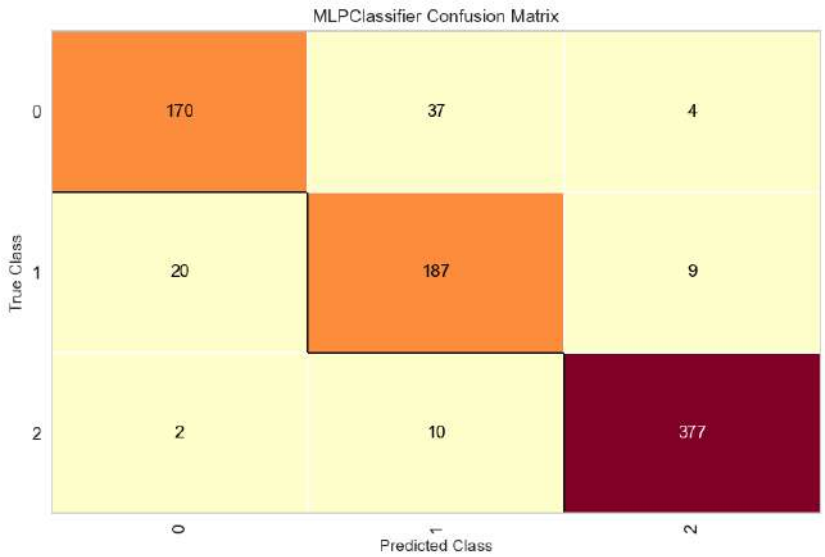


Figura 29 – Matriz de Confusão Redes Neurais.

ramente superior em relação à **Árvore de Decisão**, demonstrando maior precisão e estabilidade conforme podemos observar na tabela 2.

O desvio padrão reduzido em ambas as abordagens indica que os resultados são consistentes. Além disso, o Random Forest obteve tanto o maior valor máximo quanto o maior valor mínimo, reforçando sua confiabilidade.

Entretanto, apesar do alto desempenho, é necessário avaliar se esse resultado pode ser generalizado para outras temporadas e competições, ou se está sujeito a overfitting (superajuste) ao conjunto de dados específico utilizado neste trabalho.

Além disso, a inclusão de fatores subjetivos, como o momento do time, lesões

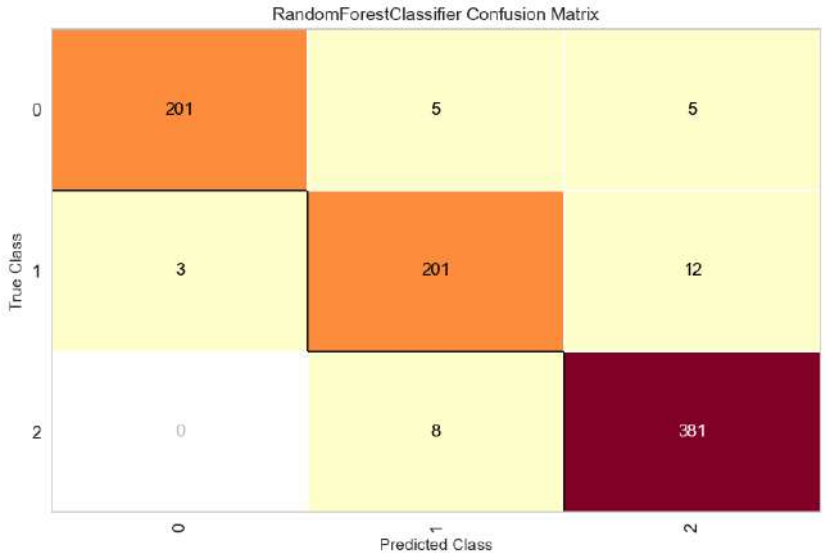


Figura 30 – Matriz de Confusão Florestas Randômicas.

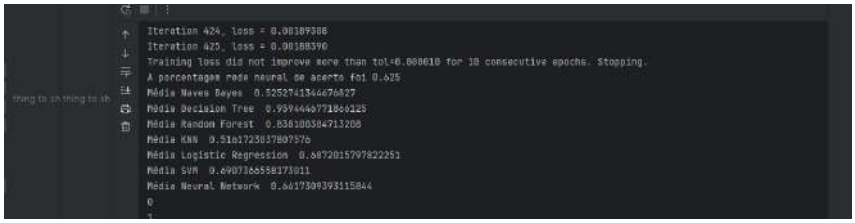


Figura 31 – Média dos algoritmos.

Tabela 2 – Resultados comparativos entre Árvore de Decisão e Random Forest aplicando a técnica de validação cruzada

Estatística	**Árvore**	**Random Forest**
Contagem (count)	30	30
Média (mean)	0,9854	0,9884
Desvio padrão (std)	0,0011	0,0012
Mínimo (min)	0,9835	0,9862
1º quartil (25%)	0,9844	0,9874
Mediana (50%)	0,9856	0,9882
3º quartil (75%)	0,9862	0,9893
Máximo (max)	0,9871	0,9908

e motivação dos jogadores, ainda não foi explorada em profundidade. Estudos como o de Iago (2023) indicam que tais variáveis podem ter um impacto significativo na previsão, e essa abordagem pode ser um caminho para futuros aprimoramentos do modelo.

Dessa forma, este estudo reforça a importância de continuar explorando diferentes fontes de dados e técnicas de aprendizado de máquina para aprimorar a previsão de resultados do Campeonato Brasileiro.

Assim para trabalhos futuros, é preciso aumentar a base de dados para os novos

Figura 32 – Aplicando validação cruzada nos algoritmos de Florestas Randômicas e Árvores de Descisão.

campeonatos brasileiros, de forma que a base sempre fique a mais atualizada possível e testar com novos registros para verificar se essa porcentagem realmente não está com overfitting e também testar com novos registros que não estejam nem na base de treinamento e também não estejam na base de teste.

5 CONSIDERAÇÕES FINAIS

Diante das análises realizadas neste trabalho, foi possível observar uma significativa melhoria na eficiência dos resultados em comparação com outros algoritmos. O modelo desenvolvido obteve uma taxa de acerto superior a 90%, um desempenho competitivo. Esse desempenho é especialmente promissor considerando que muitos estudos não detalham claramente a construção dos *datasets* utilizados. Porém é preciso considerar a possibilidade de estar acontecendo overfiting, mesmo que este trabalho tenha utilizado técnicas de validação cruzada para verificar se o modelo não está apenas decorando os registros,é preciso que seja adicionado novos casos de teste, que sejam diferentes do que estejam tanto na base de treinamento, quanto na base de teste, o que não foi feito neste trabalho.

Diferentemente de outros trabalhos, este estudo adotou uma abordagem transparente, documentando todas as etapas do processo, desde a concepção do *dataset* até a implementação dos algoritmos de aprendizado de máquina. O principal diferencial metodológico foi a consideração da singularidade de cada partida, utilizando exclusivamente os dados das partidas anteriores do mesmo time, sem levar em conta os jogos mais recentes contra outros adversários.

Entre os algoritmos testados, as árvores de decisão demonstraram o melhor desempenho, atingindo a maior acurácia. Esse resultado se destaca quando comparado aos resultados obtidos em estudos anteriores, evidenciando a eficiência da abordagem proposta neste trabalho.

Neste trabalho, foi possível atingir os objetivos específicos propostos, trazendo contribuições relevantes para a previsão de resultados do Brasileirão da Série A utilizando algoritmos de aprendizado de máquina.

O primeiro objetivo consistia na extração de dados de partidas de futebol, o que foi alcançado com sucesso por meio de técnicas de web scraping e do uso de APIs gratuitas, como a FlashScore. Ao todo, foram coletados dados de mais de 3.039 partidas do futebol brasileiro, abrangendo o período de 2013 a 2022.

O segundo objetivo envolvia a identificação de algoritmos de aprendizado de máquina adequados para o problema proposto. Após a revisão de diversos artigos científicos, foram selecionados os seguintes algoritmos: Naive Bayes, Árvores de Decisão, Florestas Randômicas, Máquinas de Vetores de Suporte e Redes Neurais. Essa escolha foi fundamentada na análise das características e desempenhos desses algoritmos em cenários similares.

Por fim, em relação ao terceiro objetivo, após a etapa de limpeza e preparação dos dados, foi realizada a aplicação prática dos algoritmos para cada time. A metodologia adotada

apresentou um diferencial importante ao considerar apenas os jogos históricos de cada equipe, analisando a linha temporal específica em que as duas equipes estavam inseridas. Essa abordagem evita o uso indiscriminado de todos os jogos anteriores, proporcionando uma análise mais precisa e contextualizada na previsão dos resultados.

Com essa estratégia, foi possível avaliar a eficiência dos algoritmos de aprendizado de máquina para cada time, permitindo que eles aprendessem de forma mais assertiva com os dados de partidas anteriores e gerando métricas mais realistas e consistentes para a previsão de resultados.

Assim para futuros trabalhos, é preciso que o dataset sempre esteja atualizado com novos dados das partidas que forem acontecendo, bem como também testar com novos registros que sejam diferentes dos que já estão armazenados tanto na base de treinamento e na base de teste, para verificar a hipótese de estar acontecendo overfiting.

REFERÊNCIAS

AFONSO, M. Monografia (Graduação em Engenharia), **Avaliação de modelos de machine learning para predição da temperatura crítica de supercondutores**. 2020. Acesso em: 23 jul. 2024. Disponível em: <https://sistemas.eel.usp.br/bibliotecas/monografias/2020/MEF20011.pdf>.

ALCANTARA, A. C. Um novo fenômeno esportivo: refletindo sobre apostas. **Revista Educação Pública**, v. 23, n. 11, março 2024. Disponível em: <https://educacaopublica.cecierj.edu.br/artigos/23/11/um-novo-fenomeno-esportivo-refletindo-sobre-apostas>.

BREIMAN, L.; FRIEDMAN, J.; OLSHEN, R. A.; STONE, C. J. **Classification and regression trees**. [S.l.]: Routledge, 2017.

CAPOBIANCO, G. Combining support vector machines with genetic algorithms to optimize portfolio selection. In: **International Conference on Agents and Artificial Intelligence (ICAART)**. SCITEPRESS, 2019. p. 212–219. Acesso em: 27 jul. 2024. Disponível em: <https://www.scitepress.org/Papers/2019/73075/73075.pdf>.

CAPOBIANCO, G.; GIACOMO, U. D.; MERCALDO, F.; NARDONE, V. Can machine learning predict soccer match results? In: **International Conference on Agents and Artificial Intelligence (ICAART)**. SCITEPRESS, 2019. p. 237–244. Acesso em: 24 abril 2024. Disponível em: <https://www.scitepress.org/PublishedPapers/2019/73075/73075.pdf>.

CHANG, A. C. Chapter 2 - history of artificial intelligence. In: CHANG, A. C. (Ed.). **Intelligence-Based Medicine**. Academic Press, 2020. p. 23–27. ISBN 978-0-12-823337-5. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780128233375000020>.

CONTEúDO, E. Cene Produtora de. **Futebol: um dos grandes aliados da economia**. 2024. Acesso em: 07 out. 2024. Disponível em: <https://www.campograndenews.com.br/artigos/futebol-um-dos-grandes-aliados-da-economia>.

FIEDLER, V. **Floresta Aleatória: Uma Rápida Explicação**. 2020. Acesso em: 24 fev. 2025. Disponível em: <https://medium.com/@viniFiedler/floresta-aleat%C3%B3ria-uma-r%C3%A1pida-explica%C3%A7%C3%A3o-a232d3de672e>.

GALA, A. S. **Aprendizado de Máquina: entenda o que é o Machine Learning**. 2024. Acesso em: 07 out. 2024. Disponível em: <https://www.handtalk.me/br/blog/aprendizado-de-maquina/>.

GAMA, J. Árvores de decisão. In: **Conferência Nacional em Interação Pessoa-Máquina**. [s.n.], 2004. Acesso em: 22 jul. 2024. Disponível em: <https://www.dcc.fc.up.pt/~ines/aulas/MIM/arvores_de_decisao.pdf>.

GOMES, B. L.; ZHU, K. H. **Entenda o que são redes neurais e como as máquinas aprendem com a inteligência humana**. 2021. Acesso em: 19 jul. 2024. Disponível em: <https://blog.itaipuparquetec.org.br/redes-neurais-pti/>.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A.; BENGIO, Y. **Deep learning**. [S.l.]: MIT press Cambridge, 2016. v. 1.

IAGO. **Um modelo de previsão de resultados de futebol utilizando Machine Learning**. 2023. Disponível em: <https://lume.ufrgs.br/bitstream/handle/10183/261788/001172666.pdf?sequence=1>. Acesso em: 15 fev. 2025. Disponível em: <https://lume.ufrgs.br/bitstream/handle/10183/261788/001172666.pdf?sequence=1>.

IZBICKI, R.; SANTOS, T. M. d. **Aprendizado de máquina: uma abordagem estatística**. 1. ed. [S.l.]: Nome da Editora, 2020. 272 p. ISBN 978-65-00-02410-4.

REIN, R.; MEMMERT, D. Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. **SpringerPlus**, Springer, v. 5, p. 1–13, 2016.

ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. **Psychological review**, American Psychological Association, v. 65, n. 6, p. 386, 1958.

RUSSELL, S. J.; NORVIG, P. **Artificial Intelligence: A Modern Approach**. 3. ed. [S.l.]: Prentice Hall, 2021.

SAMUEL, A. L. Some studies in machine learning using the game of checkers. **IBM Journal of Research and Development**, v. 3, n. 3, p. 210–229, 1959. Acesso em: 27 jul. 2024. Disponível em: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5392560>.

SANTOS, F. M. d. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação), **Redes neurais artificiais para previsão de resultados no Campeonato Brasileiro de Futebol**. 2017. Acesso em: 27 maio 2024. Disponível em: <https://felipemsantos.com/download/projects-footballneuralnetworks-felipe.pdf>.

SCHMIDT, H. L. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação), **Uso de técnicas de aprendizado de máquina no auxílio em previsão de resultados de partidas de futebol**. 2017. Acesso em: 25 maio 2024. Disponível em: <https://repositorio.unisc.br/jspui/bitstream/11624/2157/1/Henrique%20Luis%20Schmidt.pdf>.

ZHANG, A.; LIPTON, Z. C.; LI, M.; SMOLA, A. J. **Dive into deep learning**. [S.l.]: Cambridge University Press, 2023.

ZHANG, H. The optimality of naive bayes. In: **Proceedings of the 17th International FLAIRS Conference**. [s.n.], 2004. Acesso em: 24 jul. 2024. Disponível em: <https://www.cs.unb.ca/~hzhang/publications/FLAIRS04ZhangH.pdf>.