Projeção de resultados no futebol brasileiro com o uso de machine learning

Rodrigo Alves[1], Wesley da Silva[2], José Pedro[3],

[1]Análise e desenvolvimento de sistemas - Faculdade Nova Roma

[2]Av. Adjar da Silva Casé. 800 - Indianópolis - Caruaru - PE, 5524-740 - Brasil

Abstract. *"Football is widely popular in Brazil and around the world, and along with the sport, the football market is a multi-billion-dollar industry both on and off the field. The aim of this study is to project match outcomes, a challenging task due to the sport's unpredictability. This study proposes using tree-based models for multi-class classification to predict whether the home team or the visiting team will win, or if the match will end in a draw. Utilizing data from the Brazilian Championship, or "Brasileirão Série A," including information about teams, rounds, and events occurring during matches, experiments were conducted to improve prediction performance. This study aims to contribute to the understanding of football dynamics through the application of Machine Learning techniques. "*

Resumo. *O Futebol é amplamente popular no Brasil e em todo o mundo e junto com o esporte o mercado do futebol é uma indústria bilionária tanto dentro de campo quanto fora dele. O objetivo deste estudo é mostrar uma projeção de resultados que se torna uma tarefa desafiadora, devido a imprevisibilidade do esporte. Nesse estudo foi proposto fazer uma projeção de resultados usando modelos baseados em árvores para a multiclassificação. Para que a previsão aponte se o time mandante ou visitante sairão com a vitória, ou o confronto termine em um empate. Utilizando dados do Campeonato Brasileiro de futebol, ou o "Brasileirão Série A". Incluindo informações sobre equipes, rodadas e eventos ocorridos durante os jogos, foi conduzido experimentos para melhorar o desempenho da previsão.*
Palavras-chaves*: Machine Learning. Futebol. Projeção. Predição. Multiclass. TreeBased.*

1. Introdução

O futebol é um dos esportes mais populares do mundo e também do Brasil com 3.5 bilhões de fãs espalhados por todo mundo [19]. A grandeza do futebol é evidenciada em eventos como a copa do mundo, de acordo com a FIFA [8], a Copa do Mundo de 2022, sediada no Qatar, teve um público de 3,4 milhões de torcedores, enquanto a audiência digital para a decisão do torneio foi estimada em mais de 1,5 bilhão de espectadores [16].

Essa grande audiência existente no futebol torna-o um mercado bilionário que engloba patrocínios, vendas de produtos relacionados aos times, direitos de transmissão e até mesmo o mundo das apostas esportivas. Segundo o levantamento da Sports Value [15], os clubes brasileiros juntos somam uma arrecadação superior a 33,2 bilhões de reais, havendo até mesmo clubes que faturam mais de 1 bilhão isolados no topo.

No futebol, a imprevisibilidade é recorrente [9], pois jogadores podem sofrer lesões repentinas devido ao desgaste físico, incidentes dentro ou fora de campo que podem ocasionar lesões sérias aos atletas, e isso pode causar desfalques aos times em momentos importantes para o rumo de uma partida, ou até mesmo da campanha de um time em um campeonato. Esses eventos causam reviravoltas nos jogos. Além disso, há casos em que árbitros falham em suas decisões e times considerados inferiores em nível tático causam outras reviravoltas emocionantes, transformando o curso de um jogo, onde até mesmo profissionais não conseguiram prever tais eventos [2]. Este artigo aborda a importância da análise de desempenho nos jogos esportivos, incluindo diversas abordagens analíticas, desde observações qualitativas tradicionais até as técnicas quantitativas modernas, enfatizando a necessidade de uma abordagem para que se compreenda melhor os fatores que influenciam o rendimento esportivo.

Com o auxílio da Machine Learning, as equipes têm a capacidade de estabelecer metas esportivas e objetivos financeiros para a temporada de jogos atual ou futuras [3]. A imprevisibilidade no contexto do futebol torna necessário identificar as previsões mais realistas, a fim de obter uma expectativa realista sobre o desempenho do clube. A abordagem visa principalmente a parte financeira da equipe, melhorando estrategicamente as decisões do clube para maximizar seu potencial competitivo [6]. Há uma necessidade de modernizar a gestão no futebol brasileiro, explorando práticas de gestão profissional e estratégias de governança, o que pode transformar os clubes de futebol em entidades mais eficientes e sustentáveis.

Nesse contexto apresentado, utilizar Machine Learning é uma alternativa para a predição de resultados de partidas de futebol. Utilizando uma base de dados do Transfermarkt, um website dedicado à modalidade de futebol e o mercado do esporte, é possível obter estatísticas sobre partidas disputadas entre times, rodada, público, faltas, chutes e outros tipos de lances que ocorrem dentro das regras do futebol [4].

O objetivo do nosso trabalho é criar uma predição de resultados no campeonato de futebol brasileiro (Brasileirão Série "A"), criando uma simulação de jogos onde determinará se naquela respectiva rodada se o time da casa ou o time visitante irá sair com a vitória, ou se a partida terminará em empate.

2. Fundamentação

2.1. Abordagem

Previsão de desempenho esportivo no futebol era uma tarefa que apenas alguns profissionais na área como analistas, matemáticos e treinadores faziam criando anotações e estatísticas, agora com o passar dos anos e a evolução técnologica, vem sendo implementado técnicas de *machine learning* para criar predições de resultados em jogos ou apontar quem será o campeão por exemplo. E cada vez mais vem obtendo resultados mais precisos, e contribuindo com essa proposta, realizamos uma predição de resultados dentro do campeonato de futebol brasileiro (Brasileirão Série A).

2.2. Métricas e técnicas

Grid SearchCV [1] é uma técnica usada para otimizar os hiperparâmetros de modelos de machine learning. Que realiza busca exaustiva em uma grade de parâmetros especifica-

dos, também inclui uma funcionalidade de *Cross-validation* enquanto busca os melhores hiperparâmetros que torna uma busca mais robusta e confiável.

Random forest [14] é um algoritmo de machine learning bastante utilizado, pertencente à categoria dos métodos de ensemble. Ele trabalha criando múltiplas árvores de decisão a partir de subconjuntos aleatórios do conjunto de dados de treinamento, introduzindo variabilidade e robustez ao modelo. O random forest utiliza todos os conjuntos possíveis para que cada árvore seja construída usando diferentes amostras e conjuntos de atributos, reduzindo assim a tendência de overfitting e melhorando a precisão.

XGBoost [6] é um algoritmo e aprendizado de máquina que melhora a precisão de modelos de previsão. Ele faz isso combinando várias árvores de decisão de forma eficiente e rápida para obter melhores resultados, usando técnicas avançadas como regularização e processamento paralelo, sendo muito utilizado para classificação e regressão.

One-hot-encoding [11] é uma técnica que transforma um conjunto de dados em grupos de bits, onde as combinações legais de valores têm apenas um bit alto (1) e todos os outros baixos (0). É utilizada para preservar a informação dos valores categóricos em modelos de machine learning.

Matriz de confusão [10] é uma ferramenta usada para avaliar o desempenho de modelos de classificação em machine learning. Ela é uma tabela que compara as previsões do modelo com os valores reais, organizando essas informações em quatro categorias: Verdadeiros Positivos (TP), Verdadeiros Negativos (TN), Falsos Positivos (FP) e Falsos Negativos (FN).

Precision e *recall* [5] são métricas usadas para avaliar o desempenho de modelos de classificação, especialmente focadas em desequilíbrios de classes. Precision mede a proporção de verdadeiros positivos (TP) sobre os exemplos que o modelo classifica como positivos. Recall mede a proporção de verdadeiros positivos sobre todos os exemplos que são realmente positivos.

RandomUnderSampler [22] é uma técnica de balanceamento de dados usada em problemas de Machine Learning, especialmente em classificação, quando há um desequilíbrio significativo entre as classes que pode prejudicar o desempenho dos algoritmos. Se a classe majoritária tende a dominar o modelo, isso resulta em previsões tendenciosas.

F1-score [21] é uma métrica de desempenho utilizada em casos de classificação para avaliar a precisão de um modelo. É a média harmônica entre *precision* e *recall*, que fornece um único valor que ajusta o balanceamento em ambos os aspectos. Medindo a proporção de verdadeiros positivos em relação ao total de postivos preditos, enquando sensibilidade mede a proporção de verdaderios positivos em relação ao total de positivos reais.

LightGBM [12] é um algoritmo de aprendizado de máquina baseado em árvores de decisão de gradiente (GBDT). Projetado para ser altamente eficiente e escalável, o LightGBM é capaz de lidar com grandes volumes de dados e oferece vantagens significativas em termos de velocidade e desempenho em comparação com outros métodos de boosting tradicionais.

2.3. Trabalhos relacionados

A ciência esportiva não é uma novidade. Em 2003, foi lançado o livro [13], baseado em fatos reais, onde um gerente chamado Michael Lewis, que atuava em um dos times com menor orçamento da Liga Americana de Beisebol (MLB), conseguiu com sua abordagem levar a equipe para os playoffs em 2002 e 2003. Ele obteve resultados surpreendentes utilizando a técnica de [18], que é a análise empírica do beisebol, especialmente as estatísticas do esporte, consistindo em medições tradicionais, como lançamentos, rebatidas e arremessos.

Alguns clubes ingleses já adotam a inteligência artificial, como o Chelsea e o Burnley, que jogam na English Premier League (EPL), e utilizam o aplicativo (ai.Scout). Esse aplicativo utiliza um método quantitativo [20], como o número de trocas de passes, posse de bola, finalizações e dribles. Com o auxílio da tecnologia, conseguem montar modelos estatísticos para identificar padrões. O aplicativo funciona como um olheiro, onde jovens enviam seus vídeos e uma IA atribui uma pontuação pelo desempenho.

Em [17], é utilizado um modelo estatístico para prever os resultados das partidas da English Premier League (EPL), uma das ligas mais competitivas atualmente. Este modelo proposto que utiliza a estatística Bayesiana [7], permite utilizar conceitos probabilísticos capazes de gerar previsões contínuas com os dados para resultados de partidas de futebol, utilizando um conjunto de variáveis para predição.

3. Metodologia

Utilizando a metodologia CRISP-DM (Cross-Industry Standard Process for Data Mining) [Wirth and Hipp 2000] (Figure 1) que é uma metodologia estruturada para orientacão do processo de dados, amplamente utilizada em diversas industrias. Ela é composta por 6 fases, onde cada uma desempenhando um papel crucial no desenvolvimento de projetos data mining e machine learning.

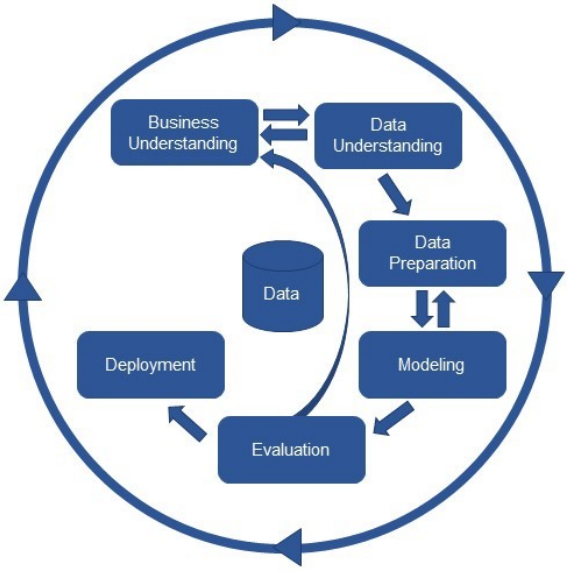

Figure 1. metodologia

Através das fases do CRISP-DM, foi possível compreender melhor o negócio e os seus dados, para enfim se ter uma noção melhor da base e de seu potencial para ser

trabalhada, primeiro é feito o entendimento do negócio que é uma etapa crucial, para que o entendimento dos dados esteja alinhado com o objetivo do negócio, o que é uma etapa fundamental no processo de análise e desenvolvimento para o modelo de *machine learning*. Essa etapa envolve a exploração, análise e interpretação dos dados dos dados, onde depois ocorrerá tomadas de decisões importantes sobre o pré-processamento e modelagem.

3.1. Base de dados

Em nosso projeto, a base de dados foi disponibilizada pelo Transfermarkt e contém dados da série A do campeonato brasileiro, abrangendo o período de 2003 a 2023, especificamente. Ela possui 35 colunas e 8079 linhas no total, estes dados abordam especificamente confrontos do campeonato brasileiro.

3.2. Pre Processamento

Após uma raspagem inicial nos dados, foi necessário remover os dados de 2003 a 2005 devido a uma mudança no regulamento e no calendário do Brasileirão. Anteriormente, eram 22 times e 42 jogos por edição, mas com o novo regulamento em 2006, passou a ser 20 times e 38 jogos. Regulamento este, que segue em vigência até os dias de hoje [4].

No tratamento da base foi identificado colunas númericas com ausência de 6% dos dados, nesses casos específicos foi realizada um preenchimento de dados nesses espaços faltantes com a média de dados registradas na base, a partir de um algoritmo de preenchimento automático da biblioteca pandas do python usando 94% dos dados preenchidos. Na próxima etapa, as colunas com menos de 50% dos dados preenchidos foram removidas, a alta proporção de dados faltantes resultaria em um desaproveitamento da capacidade preditiva do modelo de machine learning.

Propomos uma projeção a partir de novas colunas que foram geradas através do processamento de dados já existentes, para assim um formular dados que explicassem a dinâmica dos times ao longo do tempo. Especificamente foi calculado a média de gols e o saldo de gols acumulado dos times ao longo das rodadas.

Para cada confronto, foi calculado o saldo de gols dos times mandantes e visitantes. O Saldo de gols é definido como a diferença entre os gols marcados pelo time e os gols sofridos.

Com isso foi gerado a coluna de Saldo de gols do time mandante que é a diferença entre os gols marcados pelo time mandante e os gols que ele sofreu do time visitante e a de saldo de gols do time visitante que funciona da mesma maneira, porém com a lógica da diferença entre os gols marcados pelo time visitante e os gols que sofreu do time mandante. A partir disso, foi feita uma função para capturar o saldo acumulado ao longo da temporada, gerando uma coluna de saldo de gols acumulados para cada time. A cada rodada do campeonato o saldo é atualizado, somando-se o saldo da partida atual ao saldo acumulado que estava presente até a rodada anterior. No início de cada ano, o saldo acumulado é resetado para zero, refletindo o início de uma nova temporada.

Além do saldo de gols, foi calculado a média de gols por partida para cada time. Esta métrica é obtida dividindo-se o saldo acumulado pelo número de partidas jogadas até aquela rodada. Aassim, foi obttida a média de gols por partida para cada time, tanto

para os times mandantes quando para os visitante. Esta média também é recalculada por rodada.

No final do pré-processamento, foi obtido o aproveitamento de 5917 linhas de dados e 27 colunas usando o período de anos entre 2006 a 2023.

Dicionário de dados após pre processamento

Atributo	tipo	Descrição
ano-campeonato	Int	ano em que foi realizado o campeonato
rodada	Int	rodada do campeonato
estadio	String	estadio do jogo
arbitro	String	arbitro do jogo
publico	Double	publico no estadio
time-mandante	String	time que joga como o mandante
time-visitante	String	time que joga como visitante
tecnico-mandante	String	tecnico do time da casa
tecnico-visitante	String	tecnico do time visitante
colocacao-mandante	Int	colocacao na tabela da equipe mandante
colocacao-visitante	Int	colocacao na tabela da equipe visitante
valor-equipe-titular-mandante	Float	valor em reais da equipe mandante
valor-equipe-titular-visitante	Float	valor em reais da equipe visitante
idade-media-titular-mandante	Double	idade media dos jogadores do time mandante
idade-media-titular-visitante	Double	idade media dos jogadores do time visitante
gols-mandante	Int	gols do time mandante
gols-visitante	Int	gols do time visitante
gols-1-tempo-mandante	Int	gols no primeiro tempo do time mandante
gols-1-tempo-visitante	Int	gols no primeiro tempo do time visitante
saldo-gols-mandante	Double	saldo de gols do time mandante
saldo-gols-visitante	Double	saldo de gols do time visitante
saldo-acumulado-mandante	Double	saldo acumulado ao longo das rodadas do time mandante
saldo-acumulado-visitante	Double	saldo acumulado ao longo das rodadas do time mandante
partidas-jogadas-mandante	Int	partidas jogadas
partidas-jogadas-visitante	Int	partidas jogadas
media-gols-mandante	Double	media de gols time mandante
media-gols-visitante	Double	media de gols time visitante

3.3. Construção

Três modelos foram testados, todos eles baseados em árvore para este problema de multiclasse,para definir vitória do visitante, vitória do mandante ou empate. Que foram complementadas com técnicas como o *RandomUnderSampler* para o balanceamento da base e o *Grid SearchCV* para a otimização dos hiperparâmetros. *O Grid SearchCV* pode ser encontrado na biblioteca do Scikit-learn, já o *RandomUnderSampler* pode ser encontrando na biblioteca do imblearn. O modelo foi treinado a partir da realização de cada rodada, utilizando como conjunto de treino as partidas ocorridas durante todos os anos do campeonato para preparar o modelo para predição das partidas do ano posterior. As

features selecionadas para o modelo foram: Colocação do mandante e visitante, Saldo de gols acumulado do mandante e do visitante, as partidas jogadas pelo visitante e pelo mandante, média de gols do time mandante e visitante e por fim o saldo acumulado do time mandante e do visitante até a determinada rodada, essas *features* foram selecionadas pensando em colunas relacionadas ao desempenho, mas que evitassem enviesamento dos dados, evitando assim uma coluna de gols do mandante e do visitante, por exemplo. Onde ele já teria os resultados e possivelmente replicaria eles com enviesamento nos próximos confrontos entre os times. As colunas alvo foram feitas através de uma lógica onde se gols mandante fossem maior que gols visitante na partida, a vitória iria ser do mandante, o inverso dessa situação também foi aplicada e por fim, se os gols mandante e visitante fossem iguais, geraria a coluna de empate, ficando assim: empate, vitória mandante e vitória visitante. Foram calculadas métricas de precisão, sensibilidade e F1-score e acurácia do modelo a partir das previsões feitas no conjunto de teste. Essa abordagem permite uma análise do desempenho dos times ao longo do campeonato para realizar a predição do resultado das partidas.

4. Resultados

Nesta seção, é apresentado o resultado obtido através de técnicas de *machine learning*, mostrando quais melhores desempenharam e as configurações utilizadas para a execução desses modelos, bem como a otimização desses modelos e suas métricas de avaliação, a partir desta avaliação foi feita a seleção de qual modelo foi o escolhido e o por quê de ele se encaixar melhor com o objetivo do projeto.

4.1. Grid Search

O Grid Search foi utilizado para a otimização de hiperparâmetros neste projeto, as configurações utilizadas estão sendo demonstradas na tabela abaixo. Com estas configurações, três modelos de *machine learning* foram avaliados após a otimização dos hiperparâmetros: *RandomForest*, *LightGBM*, e *XGBoost*.

Parâmetros e valores usados no Grid Search

Modelo	**Parâmetros**	**Valores**
RandomForest	n_estimators	[100, 200]
	max_depth	[10, 20, None]
	min_samples_split	[2, 5]
	min_samples_leaf	[1, 2]
	bootstrap	[True, False]
	criterion	[gini, entropy]
LightGBM	max_depth	[3, 6, 9]
	learning_rate	[0.1, 0.01, 0.001]
	n_estimators	[100, 200, 300]
Xgboost	max_depth	[3, 6, 9]
	learning_rate	[0.1, 0.01, 0.001]
	n_estimators	[100, 200, 300]
	objective	[multi:softmax]

Table 1. Fonte: Autor

4.2. Execução dos modelos

Foi feita a execução de 3 modelos de *machine learning*, todos os modelos baseados em árvores de decisão: *Random Forest, Light GBM e XgBoost*. Todos os modelos foram executados usando o *Random Under Sampler* para que fosse realizada o balanceamento das classes que foram alvo da previsão neste estudo. Os modelos foram executados com e sem a otimização de hiperparâmetros para comparar também o impacto que a otimização feita com o *Grid SearchCV* causaria as métricas de avaliação das classes e modelos, todas as execuções foram feitas em cima da versão final da base de dados.

4.3. Avaliação dos modelos

Para avaliar os modelos de *machine learning* utilizados neste estudo (*Random Forest, LightGBM e XGBoost*), foram analisadas como eles se saiam nas métricas de desempenho pré estabelecidas nas seções anteriores do estudo: precisão, sensibilidade, F1-score e acurácia.

4.3.1. Modelo XGBoost

Aqui foi avaliado a precisão, sensibilidade e F1-Score das classes com e sem a otimização no modelo XGBoost, nas tabelas abaixo é possível ver estes resultados com mais clareza.
Resultados sem Otimização:

Classe	Precisão	Sensibilidade	F1-Score
Empate	0.31	0.36	0.33
Vitória mandante	0.69	0.54	0.61
Vitória visitante	0.42	0.53	0.47

Table 2. Fonte: Autor

Resultados com Otimização:

Classe	Precisão	Sensibilidade	F1-Score
Empate	0.35	0.37	0.36
Vitória mandante	0.72	0.58	0.64
Vitória visitante	0.45	0.60	0.52

Table 3. Métricas de Precisão, Sensibilidade e F1-Score para cada classe do modelo XGBoost sem e com otimização.

Table 4. Fonte: Autor

4.3.2. Modelo LightGBM

Aqui foi avaliado a precisão, sensibilidade e F1-Score das classes com e sem a otimização no modelo LightGBM, nas tabelas abaixo é possível ver estes resultados com mais clareza.

Resultados sem Otimização:

Classe	Precisão	Sensibilidade	F1-Score
Empate	0.32	0.36	0.34
Vitória mandante	0.70	0.57	0.63
Vitória visitante	0.44	0.56	0.49

Table 5. Fonte: Autor

Resultados com Otimização:

Classe	Precisão	Sensibilidade	F1-Score
Empate	0.35	0.38	0.37
Vitória mandante	0.70	0.57	0.63
Vitória visitante	0.46	0.60	0.52

Table 6. Métricas de Precisão, Sensibilidade e F1-Score para cada classe do modelo LightGBM sem e com otimização.

Table 7. Fonte: Autor

4.3.3. Modelo Random Forest

Aqui foi avaliado a precisão, sensibilidade e F1-Score das classes com e sem a otimização no modelo Random Forest, nas tabelas abaixo é possível ver estes resultados com mais clareza.

Resultados sem Otimização:

Classe	Precisão	Sensibilidade	F1-Score
Empate	0.30	0.34	0.32
Vitória mandante	0.70	0.56	0.62
Vitória visitante	0.44	0.56	0.50

Table 8. Fonte: Autor

Resultados com Otimização:

Classe	Precisão	Sensibilidade	F1-Score
Empate	0.33	0.36	0.35
Vitória mandante	0.71	0.59	0.64
Vitória visitante	0.47	0.59	0.52

Table 9. Métricas de Precisão, Sensibilidade e F1-Score para cada classe do modelo Random Forest sem e com otimização

Table 10. Fonte: Autor

4.3.4. Acurácia dos Modelos

Foi avaliado também a acurácia dos modelos, por ser uma métrica com uma interpretabilidade mais acessível. Como é uma métrica geral, ela não apresenta o valor de acurácia por classes, então nas tabelas abaixo, fica evidenciado nas colunas os resultados que o modelo atingiu pela avaliação dessa métrica.

Resultados sem Otimização:

Modelo	Acurácia
XGBoost	0.49
LightGBM	0.50
Random Forest	0.50

Table 11. Fonte: Autor

Resultados com Otimização:

Modelo	Acurácia
XGBoost	0.53
LightGBM	0.52
Random Forest	0.52

Table 12. Métrica de Acurácia para cada modelo sem e com otimização

Table 13. Fonte: Autor

4.4. Discussão

Os resultados apresentam que o modelo apresentou uma melhoria em quase que todas as suas métricas, no entanto alguns modelos apresentaram uma pequena melhora no desempenho em métricas específicas enquanto outros se mantiveram estáveis em relação aos seus resultados antes da otimização ser aplicada. XGBoost apresentou a maior acurácia (0.53) e teve o melhor desempenho na classe "Vitória mandante" em termos de precisão (0.72) e F1-Score (0.64). LightGBM e Random Forest mostraram resultados muito próximos, mas o Random Forest teve ligeiramente melhor desempenho em termos de sensibilidade e F1-Score para a classe "Vitória mandante".

4.4.1. Escolha do Modelo

Com base nas análises, decidimos aprofundar no modelo XGBoost após suas otimizações, antes das otimizações era o modelo que apresentava a pior performance geral por ter a vantagem em métricas como acurácia por exemplo mas principalmente por ter desempenhado acima em métricas de avaliação como: precisão e F1-Score nas classes de Vitória mandante e e vitória visitante, tendo apenas o LightGBM com resultados mais parelhos comparados ao XGBoost, porém numa avaliação geral, o XGBoost apresentou um balanceamento melhor e uma distância menor entre as classes, apesar da pouca diferença entre os modelos, por desempenho geral o XGBoost foi escolhido pela leve melhora em desempenho.

4.4.2. Modelo XGBoost

A matriz de confusão é gerada utilizando dados de validações diferentes do modelo de treinamento, nele e possível observar os resultados reais e de previsões de desempenho do modelo *XGBoost*.

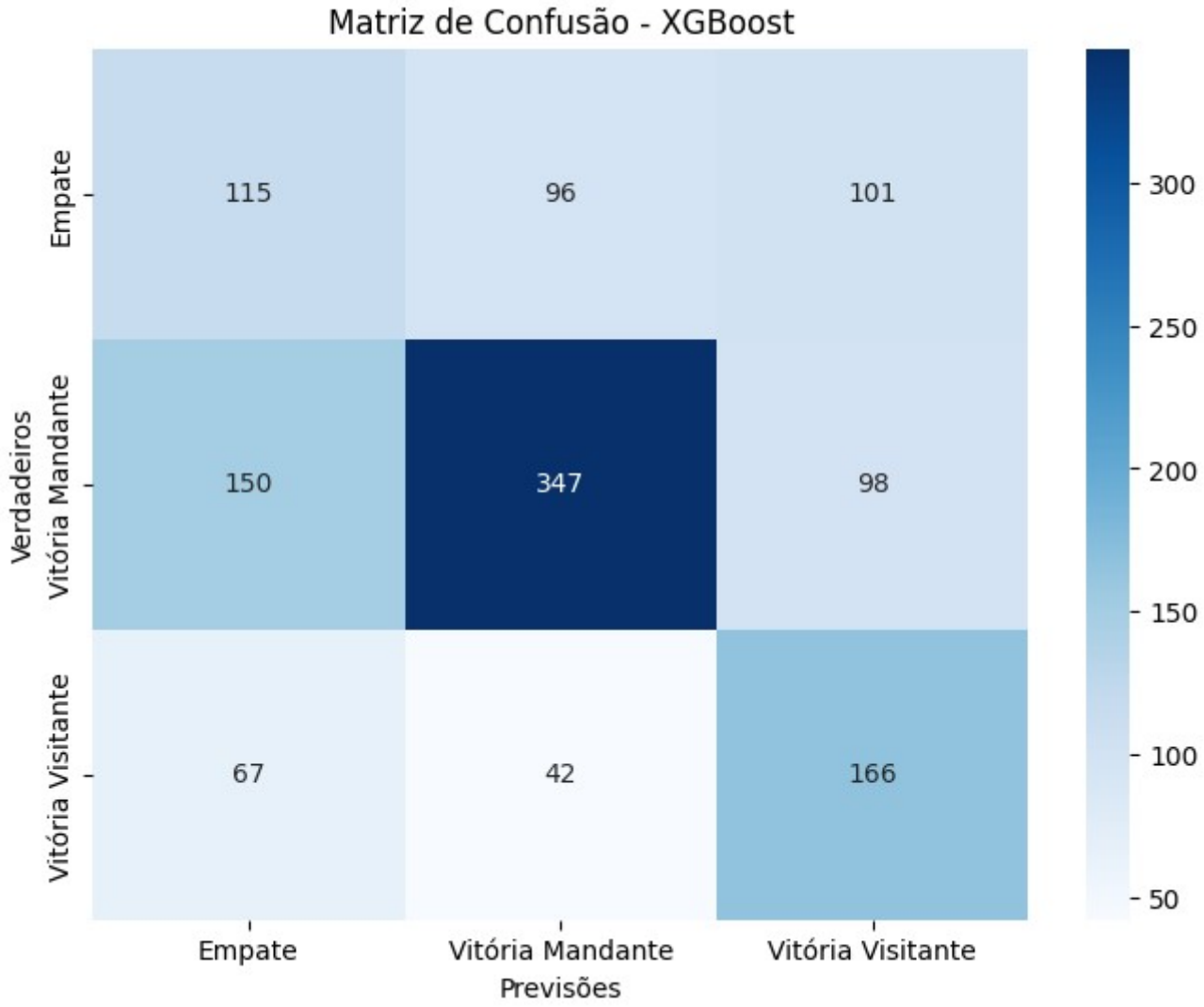

Figure 2. Matriz de confusão

Figure 3. Fonte: Autor

Na análise da matriz de confusão foi observado que o modelo teve um desempenho abaixo na previsão de empates, com 115 acertos em 312 jogos (36,86%). Na análise do modelo de previsão de vitórias mandantes teve um desempenho bem melhor, com 347 acertos em 595 jogos (58,32%). Já o modelo de previsão de vitórias visitantes teve um desempenho um pocuo maior que vitórias mandantes, com 166 acertos em 275 jogos (60,36%).

5. Conclusão

O resultado de partidas no futebol brasileiro frequentemente surpreendem, com certos times, sendo colocados favoritos em um confronto, entretanto o determinado time acaba amargando uma derrota para o um adversário teoricamente mais fraco naquela partida. Este estudo usufruiu modelos de *machine learning* para tentar prever esses resultados, utilizando dados que cobrem de 2006 a 2023 de confrontos do Campeonato Brasileiro

A análise revelou que a projeção de resultados com algoritmos de *machine learning* pode ser uma ferramenta útil para entender possíveis cenários e obter previsões mais realistas em partidas de futebol.

Neste projeto foram avaliados três modelos de *Machine Learning*: *XGBoost*, *LightGBM* e *Random Forest*. Todos são modelos baseados em árvore de decisão, os três apresentaram valores próximos em questão de desempenho nas métricas de avaliação escolhidas, porém o XGBoost obteve um ligeiro destaque com uma acurácia de 0.53% já os modelos LightGBM e Random Forest obtiveram os dois 0.52%.

Os modelos avaliados também apresentaram bons índices de sensibilidade em suas predições. O Random Forest obteve a maior resultado de sensibilidade com 0.59, já o *XGBoost* obteve 0.58 enquanto o *LightGBM* obteve 0.57.

Em virtude da escolha do melhor modelo para o projeto, o *XGBoost* obteve melhores resultados de precisão (Quantidade de acertos em relação ao total de tentativas) alcançando 0.72 em vitória mandante. Já no f1-score para vitória mandante, a métrica escolhida para o modelo por ser uma média harmônica entre precisão e sensibilidade, já que havia um desequilíbrio entre as classes. Foi onde o XGBoost melhor desempenhou, mesmo que com resultado ligeiramente melhor que os outros modelos.

Por mais que o F1-Score de 0.64, tenha sido na classe de vitória mandante, 19 rodadas dos jogos são jogados como mandante enquanto os outros 19 são jogados pelos visitantes. ao menos em metade das rodadas do campeonato, as previsões podem ser significativamente mais precisas do que um palpite aleatório. Pois a métrica não está tão próxima de 50% ou 0,50. Também foi um dos motivos que fez com que fosse escolhido a métrica F1-Score, pois além de ser uma métrica mais balanceada que apresentava um bom desempenho, era uma das métricas que mais se distanciava do aleatório.

5.1. Passos Futuros

Para aprimorar este estudo, futuras pesquisas podem focar em várias áreas. Primeiro, há a possibilidade de aperfeiçoar os modelos para não apenas prever os resultados das partidas, mas também simular toda a tabela do campeonato através das simulações de partidas assim que o calendário da CBF fosse divulgado. Isso permitiria com que projeções mais detalhadas fossem feitas pelos times para determinar metas para a temporada, podendo fazer um planejamento envolvendo resultados financeiros e esportivos para os clubes, pois assim metas já estariam estabelecidas desde o começo do ano, podendo assim orientar e estruturar o time para bater a meta estabelecida pelo modelo antes do campeonato começar.

Além disso, se vê essencial melhorar a qualidade dos dados em que são trabalhados, e que eles sejam atualizados. Embora a base de dados utilizada tenha sido extensa,

houve uma quantia significativa de dados faltantes que, se estivessem preenchidos, poderiam melhorar o desempenho dos modelos. A coleta de dados mais completos e detalhados, especialmente sobre fatores táticos e estratégicos dos jogos, pode proporcionar uma base mais robusta e detalhada para as predições.

Uma abordagem que pode vir a ser interessante para o desempenho dos modelos, pode ser quebrar as decisões em sub-decisões binárias. Por exemplo, ao em vez de fazer a previsão diretamente para o resultado final da partida, o modelo poderia primeiro prever se vai haver uma vitória do time da casa ou não, e em seguida, se haverá uma vitória do time visitante ou não. Esse método de decomposição das previsões em decisões binárias pode ajudar o algoritmo na tomada de decisões e tirando um pouco a complexidade do código, facilitando talvez a tomada de decisão.

Por fim, a implementação de uma prototipação prática, com uma interface que permita a simulação dos resultados em tempo real, com dados atualizados, que possa fornecer informações úteis e validar ainda mais a eficácia dos modelos propostos. Essa aplicação prática pode transformar a teoria em uma ferramenta prática para clubes de futebol, ajudando-os a tomar decisões estratégicas baseadas em dados preditivos.

5.2. Considerações Finais

Este estudo demonstra como pode vir a ser importante o uso de *machine learn* no futebol, não só com a predição de resultados mas como a experiência dos profissionais que atuam na área pode se beneficiar enormemente dessa análise, a incorporação desta tecnologia pode gerar contribuição para uma compreensão mais lógica do jogo com o auxílio da predição baseada em dados, ajudando no melhor planejamento e na tomada de decisões por parte da gestão dos clubes e técnicos, principalmente.

References

[1] Diagnóstico médico eficiente de doenças cardíacas humanas usando técnicas de aprendizado de máquina com e sem gridsearchcv.

[2] A análise da performance nos jogos desportivos. revisão acerca da análise do jogo. *Revista Portuguesa de Ciências do Desporto*, 2001, 01 2001.

[3] Rahul Baboota and Harleen Kaur. Predictive analysis and modelling football results using machine learning approach for english premier league. *International Journal of Forecasting*, 35(2):741–755, 2019.

[4] Fabio Augusto Barbieri, Larissa Cerignoni Benites, and Samuel de Souza Neto. Os sistemas de jogo e as regras do futebol: considerações sobre suas modificações. *Motriz: Revista de Educação Física*, pages 427–435, 2009.

[5] Michael Buckland and Fredric Gey. The relationship between recall and precision. *Journal of the American society for information science*, 45(1):12–19, 1994.

[6] E. Carravetta. *Modernização da Gestão no Futebol Brasileiro*. AGE, 2006.

[7] Ricardo Sandes Ehlers. Introdução à inferência bayesiana. *URL: http://www. leg. ufpr. br/% 7Epaulojus/CE227/ce227. pdf*, 2003.

[8] FIFA. Fifa world cup qatar 2022 in numbers, 2022.

[9] Christian Gaum and Ralf Prohl. On the worlds of football and the core of the game. *German Journal of Exercise and Sport Research*, 48(2):201–210, 2018.

[10] Laura Ferreira Helou and Beatriz Caiuby Novaes. Utilização da matriz de confusão na indicação de aparelho de amplificação sonora individual. *Distúrbios da Comunicação*, 17(2), 2005.

[11] Adil Yousef Hussein, Paolo Falcarin, and Ahmed T Sadiq. Enhancement performance of random forest algorithm via one hot encoding for iot ids. *Periodicals of Engineering and Natural Sciences*, 9(3):579–591, 2021.

[12] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[13] Michael Lewis. *Moneyball: The art of winning an unfair game*. WW Norton & Company, 2004.

[14] Steven J Rigatti. Random forest. *Journal of Insurance Medicine*, 47(1):31–39, 2017.

[15] SportsValue. Valuation: Top 30 clubes do brasil 2023 - 4ª edição, 2023.

[16] Rolling Stone. Mais de 1 bilhão: Os incríveis números de audiência da copa do mundo 2022, 2022.

[17] Ben Ulmer, Matthew Fernandez, and Michael Peterson. Predicting soccer match results in the english premier league. *Doctoral dissertation, Doctoral dissertation, Ph. D. dissertation, Stanford*, 2013.

[18] César Soto Valero and Mabel González Castellanos. Sabermetría y nuevas tendencias en el análisis estadístico del juego de béisbol. *Retos: nuevas tendencias en educación física, deporte y recreación*, (28):122–127, 2015.

[19] Eleni Veroutsos. The most popular sports in the world. *WorldAtlas*, 2023.

[20] Jacques Wainer et al. Métodos de pesquisa quantitativa e qualitativa para a ciência da computação. *Atualização em informática*, 1(221-262):32–33, 2007.

[21] Reda Yacouby and Dustin Axman. Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. In *Proceedings of the first workshop on evaluation and comparison of NLP systems*, pages 79–91, 2020.

[22] Todd Zhou and Hong Jiao. Exploration of the stacking ensemble machine learning algorithm for cheating detection in large-scale assessment. *Educational and Psychological Measurement*, 83(4):831–854, 2023.