José Matheus Souza De Oliveira

O uso de Auto ML nas apostas esportivas

Recife

2022

José Matheus Souza De Oliveira

O uso de Auto ML nas apostas esportivas

Monografia apresentada ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação

Universidade Federal Rural de Pernambuco – UFRPE
Departamento de Estatística e Informática
Curso de Bacharelado em Sistemas de Informação

Orientador: Cleviton

Recife
2022

Dados Internacionais de Catalogação na Publicação
Universidade Federal Rural de Pernambuco
Sistema Integrado de Bibliotecas
Gerada automaticamente, mediante os dados fornecidos pelo(a) autor(a)

S729u Souza de Oliveira, José Matheus
O uso de Auto ML nas apostas esportivas / José Matheus Souza de Oliveira. - 2022.
21 f. : il.

Orientador: Cleviton Vinicius Fonseca Monteiro.
Inclui referências.

Trabalho de Conclusão de Curso (Graduação) - Universidade Federal Rural de Pernambuco, Bacharelado em Sistemas da Informação, Recife, 2022.

1. Apostas Esportivas. 2. Aprendizado de Máquina. 3. Auto ML. 4. Investimento Esportivo. I. Monteiro, Cleviton Vinicius Fonseca, orient. II. Título

CDD 004

José Matheus Souza de Oliveira

O Uso de AutoML nas apostas esportivas

Artigo apresentado ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Aprovado em: 02 de junho de 2022.

BANCA EXAMINADORA

Cleviton Monteiro (Orientador)
Departamento de Estatística e Informática
Universidade Federal Rural de Pernambuco

Péricles Miranda
Departamento de Estatística e Informática
Universidade Federal Rural de Pernambuco

O uso de Auto ML nas apostas esportivas

José Matheus Souza De Oliveira [1], **Cleviton Vinicius Fonseca Monteiro** [1]

[1]Departamento de Estatıstica e Informática – Universidade Federal Rural de Pernambuco (UFRPE)
Rua Dom Manuel de Medeiros, s/n, - CEP: 52171-900 – Recife – PE – Brasil

{matheus007junior@gmail.com, cleviton@gmail.com}

Abstract. *The sports betting market has been growing effectively, and of the 40 teams that compete in the A and B series of the Brazilian Championship, 35 are sponsored by bookmakers. Observing the number of gamblers that are profitable, it was found that there is a great need for solutions that help these people to become profitable. In order to achieve this goal in the long term, the first step of this project is to evaluate the use of Auto Machine Learning (AutoML) solutions in two betting scenarios: both teams score, and the match has more or less than 2 goals. For this work, data from the 2017 Serie A Brazilian Football Championship were used. The championship has 38 rounds in total and 10 matches in each round, totaling 380 matches. To calculate the profit scenarios, two forms of evaluation were established: without restriction and with probability restriction. Comparing the results obtained in these two scenarios, scenario 2 in all results presented a profit, unlike what was achieved in scenario 1, in addition, we have the assertiveness rate that was much more positive in scenario 2, reaching a value of 73.68% while the maximum rate obtained in scenario 1 was 63.88%.*

Keywords: AutoML. Sports Betting. Bet365. Machine Learning. Pycaret.

Resumo. O mercado de apostas esportivas vem crescendo de forma efetiva, e dos 40 times que disputam as séries A e B do Campeonato Brasileiro, 35 são patrocinados por casas de apostas. Observando o número de apostadores que são lucrativos, foi constatado que existe uma grande necessidade por soluções que auxiliem essas pessoas a se tornarem lucrativas. Visando conseguir atingir esse objetivo no longo prazo, a primeira etapa desse projeto é avaliar o uso soluções de Auto Machine Learning(AutoML), em dois cenários de apostas: ambas equipes marcarem e a partida ter mais ou menos que 2 gols. Para esse trabalho, foram utilizados os dados do Campeonato Brasileiro de Futebol da Série A 2017. O campeonato possui 38 rodadas no total e 10 partidas em cada rodada, totalizando 380 partidas. Para calcular os cenários de lucro, foram estabelecidos duas formas de avaliação: sem restrição e com restrição de probabilidade. Comparando os resultados obtidos nesses dois cenários, o cenário 2 em todos os resultados apresentou lucro, diferentemente do que se alcançou no cenário 1, além disso, temos a taxa de assertividade que foi bem mais positiva no cenário 2, chegando a atingir um valor de 73,68% enquanto a taxa máxima obtida no cenário 1 foi de 63,88%.

Palavras-chave: Apostas Esportivas. Aprendizado de Máquina. Investimento Esportivo.

1. Introdução

Segundo o relatório da Consultoria Grand View Research (2021), o mercado de apostas esportivas vem crescendo de forma efetiva. Com valor aproximado 70 Bilhões de dólares, e previsão de crescimento de 10% a.a, esse mercado pode atingir a marca de 140 Bilhões de dólares. Para se ter uma dimensão desse crescimento, de acordo com Sabino & Gabriel (2022) dos 40 times que disputam as séries A e B do Campeonato Brasileiro, 35 são patrocinados por casas de apostas.

As grandes casas de aposta apresentam uma grande variedade de esportes, entre eles vários que são pequenos em popularidade como golfe e tênis, até os mais populares como basquete e o futebol (CARNEIRO, 2022). Em pesquisa realizada pela Globo (2021), mais de 50% das pessoas que responderam o estudo afirmaram ter começado as apostas esportivas durante a pandemia. Na mesma pesquisa, 52% das pessoas afirmaram que utilizam as apostas para aumentar a renda e 12% declararam que as apostas são a sua principal fonte de renda.

Todavia, um grande problema desse mercado, é que apenas 5% dos apostadores são lucrativos, ou seja, a grande maioria das pessoas que apostam perdem mais do que ganham, e isso pode ser ocasionado por diversos motivos, como a ausência de uma análise cautelosa dos dados, falta de informação ou até mesmo má gestão dos investimentos (IG, 2021).

Observando o número de apostadores que não são lucrativos, foi constatado que existe uma grande necessidade por soluções que auxiliem essas pessoas a se tornarem lucrativas, e o grande objetivo desse trabalho é desenvolver ferramentas que possam trazer mais clareza e assertividade na escolha das melhores oportunidades de apostas disponíveis.

Visando conseguir atingir esse objetivo no longo prazo, a primeira etapa desse projeto, e objetivo desse trabalho, é avaliar o uso de soluções de Auto Machine Learning(AutoML), para avaliar se com o uso dessas ferramentas é possível obter lucro no mercado de apostas.

Para avaliar os resultados do AutoML, os resultados obtidos foram comparados com os resultados do trabalho de Almeida (2019), sendo utilizado a mesma base de dados do Campeonato Brasileiro da série A de 2017 para a criação do modelo e execução dos testes.

Além disso, na avaliação foram trabalhados 2 modalidades de apostas:

- a primeira modalidade de aposta foi de ambas as equipes marcarem gol na partida;
- e a segunda modalidade foi a partida ter mais ou menos do que 2 gols.

1.1. Objetivos

1.1.1. Objetivo geral

Avaliar se o uso de AutoML pode auxiliar no crescimento de lucro de investidores no mercado de apostas esportivas.

1.1.2 Objetivos específicos

- Coletar dados das equipes e partidas do campeonato Brasileiro de Futebol Série A 2017;
- Replicar o método do estudo de Almeida (2019);

- Analisar se o uso do AutoMl é mais assertivo do que a utilização de um algoritmo em específico; e
- Analisar a lucratividade com o auxílio do AutoMl;

1.2. Organização do trabalho

Este trabalho foi organizado da seguinte forma: a Seção 1 apresenta uma introdução sobre o tema do estudo, sua justificativa e os objetivos do trabalho. Em seguida, a seção 2 discorre o referencial teórico que serviu de base para esse artigo. A seção 3 expõe os outros trabalhos relacionados com o tema que serviram para ajudar na formulação desse experimento. Já na Seção 4, detalha-se toda a metodologia utilizada no experimento. Na seção 5, são apresentados os resultados e discute-se esses dados relacionando à metodologia aplicada. Por fim, na sexta seção é apresentada a conclusão desse estudo e são listados os possíveis trabalhos futuros.

2. Referencial teórico

2.1. Apostas esportivas

Um dos meios possíveis de realizar apostas no mercado esportivo é através das grandes casas de apostas, como a Bet365. Dentro dessas casas de apostas, encontra-se uma gama de opções para se apostar nos mais variados tipos de eventos esportivos. Dentro de um só jogo de futebol, é possível realizar diversas modalidades de apostas, como a quantidade de gols, quantidade de escanteios, quantidade de cartões, quantidade de impedimentos e até quantidade de laterais.

Para cada aposta, é definido um valor a ser pago pelo acerto, esse valor é o que se conhece por ODD, que é calculado com base na probabilidade do evento acontecer. Por exemplo, para o jogo entre o Grêmio Novorizontino x Sport Recife, a Bet365, conforme a Figura 1, estava pagando 2.50 em maio de 2022 caso aconteça de ambas as equipes marcarem.

Figura 1. Valor da ODD para o jogo do Grêmio Novorizontino e Sport Recife na plataforma Bet365.

Fonte: Bet365 (2022).

Logo, se o apostador decide investir R$100,00 na aposta e acontece de ambas as equipes marcarem gols, o apostador vai receber R$250,00. Desse valor retornado, R$150,00 equivale ao lucro obtido na operação.

Em suma, existem dois tipos de apostas: a simples, exemplificada na Figura 1 acima, e que literalmente é a forma mais simples de se apostar, onde apenas um tipo de evento está envolvido na aposta; e o outro tipo é a aposta múltipla, que pode ser visualizada na figura 2 abaixo, onde o apostador pode combinar várias apostas dentro de uma só, e todas as ODDS são multiplicadas para dar à cotação final. Entretanto, na aposta múltipla, é necessário que todas as apostas combinadas sejam assertivas, senão basta um resultado dar errado que é perdido todo valor apostado.

Figura 2. Simulação de uma aposta do tipo múltipla na plataforma Bet365 (2022).
Fonte: Bet365 (2022).

2.2. Aprendizado de Máquina

Aprendizado de Máquina é uma área da Inteligência Artificial que tem como objetivo o desenvolvimento de técnicas computacionais sobre o aprendizado, bem como à construção de sistemas capazes de aprender com os dados e adquirir conhecimento de forma automática, tomando decisões baseadas em experiências acumuladas em problemas anteriores (MONARD & BARANAUSKAS, 2003).

Segundo Géron (2019), os sistemas de Aprendizado de Máquina podem ser classificados de acordo com a quantidade e o tipo de supervisão que recebem durante o treinamento. Existem quatro categorias principais de aprendizado: o supervisionado, o não supervisionado, o semi-supervisionado e o por reforço.

O aprendizado supervisionado é quando tentamos prever uma variável dependente a partir de uma lista de variáveis independentes. A classificação é uma tarefa típica do aprendizado supervisionado, os cenários de ambas equipes marcarem gols em um jogo é um bom exemplo disso, ele é treinado com muitos exemplos de partidas que possuem a classe (sim ou não) e deve aprender a classificar novas partidas como sim (ambas marcam) ou não (ambas não marcam) (GÉRON, 2019).

Já no aprendizado não supervisionado, os dados de treinamento não são rotulados. O sistema tenta aprender sem um professor. Trabalhando de forma independente, a máquina começa a analisar os dados e identifica padrões (GÉRON, 2019).

Por sua vez, no aprendizado semi-supervisionado, são algoritmos que podem lidar com dados de treinamento parcialmente rotulados, uma grande quantidade de dados não rotulados e um pouco de dados rotulados. Um exemplo, é a hospedagem de imagens com no Google Fotos, o aplicativo reconhece automaticamente que à pessoa (A) aparece nas fotos 1, 5 e 20 enquanto outra pessoa (B) aparece nas fotos 2, 5 e 19. Essa parte é a não supervisionada(agrupamento), o próximo passo é o usuário dizer quem são essas pessoas, com apenas um rótulo por pessoas ele é capaz de nomear todos (GÉRON, 2019).

Por fim, o aprendizado reforçado é baseado na experiência, a máquina deve lidar com o que errou antes de procurar a abordagem correta. Um exemplo é a recomendação de vídeos no youtube, após assistir um vídeo, a plataforma irá indicar outros títulos semelhantes, porém caso o usuário clique e não assista até o final, a máquina entende que

a recomendação não foi boa e irá tentar indicar outros vídeos na próxima vez(GÉRON, 2019).

2.3. Auto ML

Segundo Silva (2021), grandes empresas como Google, Facebook e Uber estão desenvolvendo ferramentas para automação de projetos de aprendizado de máquina. O AutoML surgiu com o objetivo de facilitar para que pessoas com pouco ou nenhum conhecimento em aprendizado de máquina possam criar modelos de alta qualidade.

O pipeline do AutoML consiste em vários processos: preparação dos dados, engenharia de recursos, geração de modelos e avaliação de modelos. O AutoML também pode auxiliar na busca dos melhores hiperparâmetros para cada modelo de aprendizado. Abaixo na Figura 3 está um fluxograma que detalha o pipeline do AutoML.

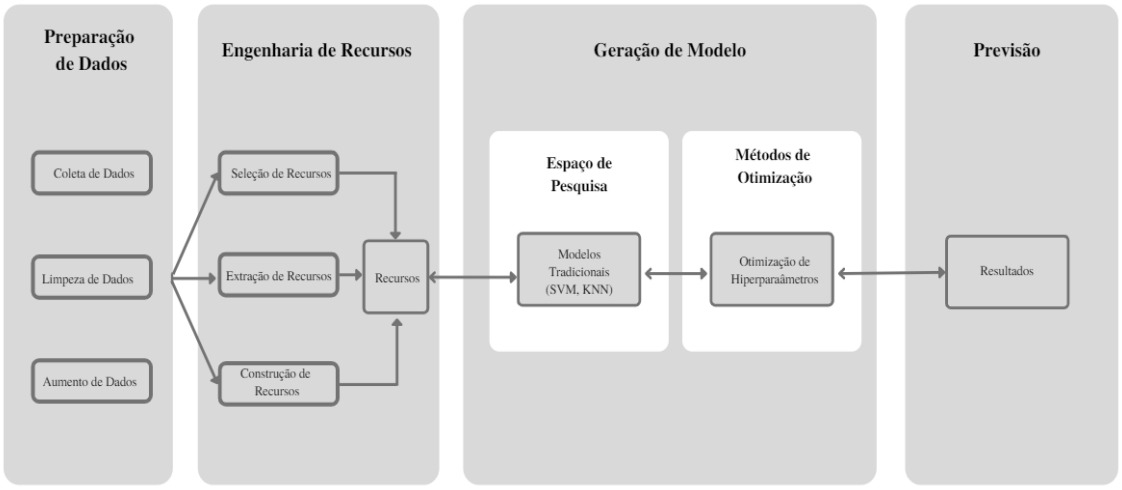

Figura 3. Fluxograma do pipeline do AutoML.

Fonte: Autor.

Segundo Herbst(2022), algumas das vantagens de utilizar AutoML, é a necessidade de apenas conhecimento superficial de Aprendizado de Máquina para se utilizar as ferramentas, automatizando passos, e diminuindo o tempo para se ter resultados iniciais. É possível utilizar em diferentes tipos de dados e problemas, com a possibilidade de incluir técnicas famosas como por exemplo(Redes Neurais, Redes neurais convolucionais). Existem algumas ferramentas open source disponíveis de Auto ML, alguns exemplos são: TPOT, Pycaret, GLM e XGBoost.

3. Trabalhos Relacionados

No estudo de Almeida (2019), foi realizado um trabalho utilizando o método Naive Bayes com três tipos de distribuições: Bernoulli, Gaussian e Multinomial. O modelo de avaliação foi dividido em quatro etapas: Treino Ajuste, Teste Validação, Treino e Teste, onde cada previsão depende dos resultados e dos dados da etapa anterior. Por exemplo, as rodadas 1, 2 e 3 são utilizadas para prever os resultados da rodada 4, e assim segue para todo o conjunto de dados.

Ainda sobre o estudo, foi utilizado uma base de dados do Campeonato Brasileiro de Futebol Série A de 2017, onde consta 380 jogos. Foram previstos 3 modalidades de apostas, a primeira é o fato de Ambas equipes marcarem gol, a segunda é se a partida vai ter mais ou menos de 2,5 gols e a terceira é se o resultado da partida é vitória do mandante,

empate ou vitória do visitante.

Como o método Naive Bayes usa em sua classificação a probabilidade do exemplo acontecer, foram criados cenários de restrição de probabilidade, para analisar quais os cenários são mais lucrativos sem restrição ou com restrição. Nos resultados desse trabalho(Tabela 1) os cenários 1 e 2 se apresentaram lucrativos sem e com restrição de probabilidade, já no cenário 3 não foi possível obter lucro mesmo com diferentes níveis de restrição, somente prejuízo.

Tabela 1. Resultados dos 3 cenários estudados no método Naive Bayes.

	Cenário 1		**Cenário 2**		**Cenário 3**	
Probabilidade	Lucro/Prejuízo		Lucro/Prejuízo		Lucro/Prejuízo	
	Médio	Total	Médio	Total	Médio	Total
Sem Restrição	5,11%	5,11%	5,51%	5,51%	-13,94%	-13,94%
Maior ou Igual a 60%	10,06%	4,04$	-1,41%	-2,76%	-15,40%	-16,50%
Maior ou Igual a 70%	2,02%	3,34%	8,01%	2,95%	-14,57%	-14,23%
Maior ou Igual a 80%	-8,54%	-1,59%	20,17%	7,86%	-11,26%	-4,66%
Maior ou Igual a 90%	19,30%	13,09%	-2,55%	12,24%	-23,13%	-15,23%
Maior ou Igual a 95%	-4,10%	-0,41%	-9,85%	5,96%	-22,72%	-20,58%
Maior ou Igual a 99%	36,05%	39,13%%	7,40%	13,29%	-11,32%	-13,29%

Fonte: Autor.

Em um outro estudo, SEHNEM *et al.* (2021), o autor utilizou o método Naive Bayes para a previsão das apostas, utilizando uma base de dados do Campeonato Brasileiro de Futebol do ano de 2010 até 2019. Para a execução desse trabalho, foi utilizado o software Orange que possui interfaces de comunicação que se conectam, fazem leitura do arquivo de treinamento e geram a previsão a partir de um arquivo de teste. O objetivo era prever o resultado da partida, vitória casa, empate ou vitória do visitante e o percentual de acerto foi de 53% resultando no final em lucro.

Em Duarte (2015), o campeonato analisado foi o de Portugal, a base utilizada possui dados desde o campeonato de 2009 até 2014, e o objetivo era a previsão do resultado da partida, vitória casa, empate ou vitória visitante. O autor realizou testes utilizando oito tipos de técnicas: C5.0, K-NN, Jrip, Random Forest, SVM com kernel linear, SVM com kernel gaussiano, Naive Bayes e Redes Neurais. Como resultado final, executando a previsão para um conjunto de dados de teste, a taxa de acerto foi de 45%.

4. Materiais e Métodos

4.1. Conjunto de dados

Para esse trabalho, foram utilizados os dados do Campeonato Brasileiro de Futebol da Série A 2017, obtidos com os autores do estudo de Almeida (2019). O campeonato possui 38 rodadas no total e 10 partidas em cada rodada, totalizando 380 partidas. Para cada jogo

disputado, existem 13 atributos na base de dados que vão ser utilizados para a previsão dos resultados conforme o quadro 1.

Quadro 1. Listra de atributos na base de dados do estudo.

ATRIBUTO	DESCRIÇÃO
QUALIDADE_ELENCO_CASA	Atributo numérico de 1 a 3, onde serve de parâmetro para a qualidade do elenco do time da casa. Essa pontuação foi definida pelos especialistas do site Globo Esporte (GLOBO, 2017).
QUALIDADE_ELENCO_VISITANTE	Atributo numérico de 1 a 3, onde serve de parâmetro para a qualidade do elenco do time visitante. Essa pontuação foi definida pelos especialistas do site Globo Esporte (GLOBO, 2017).
COLOCACAO_CASA	Colocação do time mandante no campeonato.
COLOCACAO_VISITANTE	Colocação do time visitante no campeonato.
RANKING_ANUAL_CBF_CASA	Colocação do time mandante no ranking anual da Confederação Brasileira de Futebol, disponibilizado no site da organização.
RANKING_ANUAL_CBF_VISITANTE	Colocação do time visitante no ranking anual da Confederação Brasileira de Futebol, disponibilizado no site da organização.
COMPETICOES_SIMULTANEAS_CASA	Número de competições em que a equipe da casa está participando em paralelo ao campeonato brasileiro.
COMPETICOES_SIMULTANEAS_VISITANTE	Número de competições em que a equipe visitante está participando em paralelo ao campeonato brasileiro.
MEDIA_GOLS_FEITOS_CASA	Média de gols feitos pela equipe mandante da partida.
MEDIA_GOLS_FEITOS_VISITANTE	Média de gols feitos pela equipe visitante da partida.
MEDIA_GOLS_SOFRIDOS_CASA	Média de gols sofridos pela equipe mandante.
MEDIA_GOLS_SOFRIDOS_VISITANTE	Média de gols sofridos pela equipe visitante.
CLASSICO	Atributo que identifica se a partida é um clássico ou não. Exemplo: Sport x Santa Cruz

Fonte: Autor.

4.2. Modalidades de apostas utilizadas

Duas modalidades de apostas foram utilizadas no modelo de avaliação, a primeira é o de

ambas equipes marcarem e a segundo é onde o jogo possui mais de 2 gols. Essas duas modalidades foram escolhidas pois também estão presentes no trabalho de Almeida(2019).

A modalidade 1 é conhecida no mercado de apostas brasileiro como: Ambas Marcam (AM) e internacionalmente como Both Team To Score: (BTTS), as classes que definem são:

- Sim: Quando ambas as equipes marcam, por exemplo os placares de 1x1,2x1 e 3x2.
- Não: Quando ambas ou uma equipe não marca, por exemplo os placares de 0x0, 1x0 e 3x0.

A modalidade 2 é conhecida no mercado de apostas brasileiro como: Acima de 2,5 Gols e Abaixo de 2,5 gols e internacionalmente como: Over 2.5 goals e Under 2.5 goals, as classes que definem são:

- Acima: Quando na partida a soma dos gols é maior que 2, por exemplo: 2x1, 3x0, 2x2.
- Abaixo: Quando na partida a soma dos gols é menor ou igual a 2, por exemplo: 0x0,1x0,2x0.

4.3. Metodologia

Como ferramenta de Auto ML, foi escolhido o Pycaret. Por alguns motivos, como: A linguagem utilizada ser python, ter bastante material disponível e pela facilidade de conseguir executar testes e criar modelos de forma rápida. O módulo escolhido foi o de classificação, que segundo PYCARET (2022) é um módulo de aprendizado de máquina supervisionado usado para classificar elementos em grupos, cujo objetivo é prever rótulos de uma classe categórica que são discretos e não ordenados.

Baseando-se em PYCARET (2022), o módulo de classificação pode ser usado para problemas binários ou multiclasse, ele fornece recursos de pré-processamento e possui mais de 18 algoritmos prontos para uso.

Abaixo está a lista de todos os algoritmos do módulo de classificação:

- Logistic Regression
- K Neighbors Classifier
- Naive Bayes
- Decision Tree Classifier
- SVM - Linear Kernel
- SVM - Radial Kernel
- Gaussian Process Classifier
- MLP Classifier
- Ridge Classifier
- Quadratic Discriminant Analysis
- Ada Boost Classifier
- Gradient Boosting Classifier
- Linear Discriminant Analysis
- Extra Trees Classifier
- Extreme Gradient Boosting
- Light Gradient Boosting Machine
- CatBoost Classifier

4.3.1. Preparação dos dados

Na etapa de preparação de dados, foi necessário remover as colunas de cotações das apostas, pois não iam ser utilizadas para a previsão, porém estavam presente na base, essas colunas somente irão ser usadas posteriormente no cálculo de lucro/prejuízo. Também nessa etapa foi realizado a transformação de alguns dados utilizando o label encoder, transformando colunas categóricas em númericas, como por exemplo: Sim ou Não(Ambas Marcam), Sim ou não (Over 2,5), ambas foram convertidas em 0(Não) e 1(Sim).

4.3.2. Criação do modelo

O modelo utilizado para avaliação se divide em 2 etapas: Treino com as rodadas anteriores e teste com a próxima rodada. Para prever por exemplo os resultados da rodada 5, são utilizados no treino todos os dados das rodadas 1 até 4, o quadro 2 detalha esse modelo. Para cada nova previsão(Nova rodada), o Pycaret é usado e define qual algoritmo, e quais parâmetros dele, teve melhor acurácia no treino realizado com todas as rodadas anteriores e ele é utilizado para a previsão da nova rodada.

Quadro 2. Listra de atributos na base de dados do estudo.

Treino	Teste
Rodadas 1-2	Rodada 3
Rodadas 1-3	Rodada 4
Rodadas 1-4	Rodada 5
Rodadas 1-5	Rodada 6
Rodadas 1-6	Rodada 7
Rodadas 1-7	Rodada 8
Rodadas 1-8	Rodada 9
Rodadas 1-9	Rodada 10

Fonte: Autor.

Dentro da etapa 1 (treino), foi necessário definir os dados que vão ser processados, onde constam no Quadro 1. Já na Figura 4 mostra a utilização do Pycaret, onde foi definido o conjunto de dados a ser utilizado e o target que é a classe que o modelo irá prever o resultado.

```
reg = setup(data = rodada1a2, target = 'CLASSE_AMBAS'
```

	Description	Value
0	session_id	4894
1	Target	CLASSE_AMBAS
2	Target Type	Binary
3	Label Encoded	None
4	Original Data	(20, 17)
5	Missing Values	False
6	Numeric Features	13
7	Categorical Features	3
8	Ordinal Features	False
9	High Cardinality Features	False
10	High Cardinality Method	None

Figura 4. Saída do método setup.
Fonte:Autor.

Após a definição do conjunto de dados e da classe que o modelo irá prever, é executado a comparação dos modelos. Nessa etapa é realizado o treino e avaliação do desempenho de todos os modelos disponíveis na biblioteca do Pycaret usando validação cruzada. O resultado dessa etapa pode ser observada na Figura 5, onde consta uma grade de pontuações dos algoritmos que tiveram melhor desempenho, ordernadas pela coluna de acurácia.

```
melhores_modelos1a6 = compare_models(errors="raise")
```

	Model	Accuracy	AUC	Recall	Prec.	F1
lightgbm	Light Gradient Boosting Machine	0.675	0.3000	0.0000	0.0000	0.0000
dummy	Dummy Classifier	0.675	0.3000	0.0000	0.0000	0.0000
svm	SVM - Linear Kernel	0.605	0.0000	0.1500	0.1500	0.1333
ridge	Ridge Classifier	0.605	0.0000	0.2500	0.4000	0.2967
lr	Logistic Regression	0.580	0.3500	0.2000	0.3000	0.2300
dt	Decision Tree Classifier	0.575	0.2167	0.1000	0.2000	0.1300
gbc	Gradient Boosting Classifier	0.575	0.4000	0.1000	0.2500	0.1400
rf	Random Forest Classifier	0.555	0.4500	0.0000	0.0000	0.0000
et	Extra Trees Classifier	0.555	0.3333	0.0667	0.2000	0.1000
knn	K Neighbors Classifier	0.550	0.2750	0.0000	0.0000	0.0000
qda	Quadratic Discriminant Analysis	0.490	0.2500	0.1667	0.1833	0.1400
lda	Linear Discriminant Analysis	0.490	0.2833	0.2167	0.2500	0.2067
ada	Ada Boost Classifier	0.460	0.3333	0.0833	0.2000	0.1167
nb	Naive Bayes	0.435	0.2500	0.1000	0.1500	0.1200

Figura 5. Resultado do uso da função Compare_Models.
Fonte:Autor.

4.3.3. Previsão

Nesta etapa é onde acontece a previsão dos resultados, são passados como parâmetros o modelo treinado e a rodada teste e como resultado são obtidas as colunas Label e Score, a primeira é o resultado previsto e a segunda é a probabilidade da classe prevista acontecer. Na Figura 6 é possível observar quais foram as métricas obtidas do algoritmo utilizado na previsão da nova rodada, sendo a acurácia a principal métrica de avaliação com base no trabalho de Almeida(2019), onde na figura 6 consta 0.8, equivalente a 80% de acerto na previsão da rodada, das partidas 10 esse algoritmo acertou 8. A previsão obtida nessa rodada pode ser observada na Tabela 2.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
	Light Gradient Boosting Machine	0.8	0.7857	0.3333	1.0	0.5	0.4118	0.5092

Figura 6. Métricas obtidas da avaliação do teste da rodada 7.
Fonte:Autor.

Casa	Visitante	Classe_Over	Label	Score
1	2	0	0	0.7430
16	17	0	0	0.6137
19	5	1	0	0.6285
0	3	1	0	0.6137
10	14	0	0	0.8338
15	13	0	0	0.6230
7	9	0	0	0.6413
6	18	1	1	0.5240
8	4	0	0	0.6267
11	12	0	0	0.6230

Tabela 2. Resultado obtido do teste na rodada 7.
Fonte:Autor.

Para calcular os cenários de lucro, foram estabelecidas duas formas de avaliação. A primeira é sem restrições, todos os palpites que o modelo prever irão ser contabilizados no cálculo de lucro ou prejuízo. A segunda forma de avaliação é através do uso de restrições de probabilidade, onde são agrupadas as previsões de acordo com percentual de probabilidade dado pelo Pycaret, cujos valores definidos foram 60%, 70%, 80%, 90%, 95% e 99%.

Para calcular se o modelo criado resultou em lucro ou prejuízo, foi estipulado que cada aposta teria um valor fixo de R$ 10,00. Cada aposta mal sucedida resulta em uma perca de R$10,00 e cada aposta com sucesso é calculado da seguinte forma: Valor da ODD multiplicado pelo valor fixo de cada aposta, subtraído pelo valor investido da aposta. Na Tabela 3 é possível visualizar um exemplo do resultado do cálculo.

Tabela 3. Exemplo de cálculo de Lucro/Prejuízo.

Resultado	Valor Investido	Valor da ODD	Retorno	Lucro/Prejuízo
Acertou	R$10,00	1,9	R$ 19,00	R$ 9,00
Acertou	R$10,00	2,1	R$ 21,00	R$ 11,00
Errou	R$10,00	1,7	R$ 0	- R$ 10,00
Errou	R$10,00	2	R$ 0	- R$ 10,00
Acertou	R$10,00	1,8	R$ 18,00	R$ 8,00
Errou	R$10,00	1,9	R$ 0	- R$ 10,00
Acertou	R$10,00	2	R$ 20,00	R$ 10,00
Acertou	R$10,00	1,7	R$ 17,00	R$ 7,00
Errou	R$10,00	1,8	R$0	- R$10,00
Acertou	R$10,00	2,3	R$23,00	R$13,00
Total	R$ 100,00	-	R$ 118,00	R$ 18,00

Fonte:Autor.

Observando o exemplo da Tabela 3, o lucro total foi de R$18,00 equivalente a 18% de lucro sobre o total investido. Já nos cenários de restrição de probabilidade, o cálculo é um pouco diferente, pois ao invés de usar todos os palpites no cálculo, só é utilizado os que estão dentro do grupo da porcentagem estabelecida. Na Tabela 4, observa-se a exemplificação do cálculo.

Tabela 4. Exemplo de cálculo de Lucro/Prejuízo com restrição de probabilidade para 90%.

Resultado	Valor Investido	Valor da ODD	Retorno	Lucro/Prejuízo
Acertou	R$10,00	1,9	R$ 19,00	R$ 9,00
Baixa Probabilidade	-	-	-	-
Errou	R$10,00	1,7	R$ 0	- R$ 10,00
Baixa Probabilidade	-	-	-	-
Acertou	-	-	-	-
Baixa Probabilidade	-	-	-	-
Acertou	R$10,00	2	R$ 20,00	R$ 10,00
Acertou	R$10,00	1,7	R$ 17,00	R$ 7,00
Baixa Probabilidade	-	-	-	-
Baixa Probabilidade	-	-	-	-
Total	R$ 40,00		R$ 56,00	R$ 16,00

Fonte:Autor

Na coluna Resultado da Tabela 4, os valores que estão como Baixa Probabilidade, são os palpites cujo a probabilidade está abaixo de 90%. Observa-se que no exemplo acima, o lucro total foi de R$16,00 equivalente a 40% de lucro sobre o total investido. Além disso, pode-se notar que nesse cenário existiu uma menor necessidade de investimento, pois foram apostados um número menor de jogos, porém o retorno em termos de porcentagem foi superior ao sem restrição.

5. Resultados e Discussão

Como citado anteriormente no tópico 5.1 para cada jogo disputado, existem 13 atributos na base de dados que foram utilizados para a previsão dos resultados. Esses resultados previstos foram divididos em duas modalidades de apostas: Ambas Marcam e Over 2,5 gols. O modelo utilizado para avaliação foi divido em 2 etapas: Treino com as rodadas anteriores e teste com a próxima rodada, para cada nova previsão(Nova rodada),

o Pycaret foi usado e definiu qual algoritmo teve melhor resultado e esse é utilizado para a previsão.

5.1. Cenário sem restrição de probabilidade

Na modalidade de aposta 1, que reflete o Ambas Marcam, com dois possíveis resultados: sim ou não, obteve-se através da avaliação experimental utilizando o Pycaret, uma assertividade de 51,94% nos resultados dos jogos, entretanto, o prejuízo adquirido nesse cenário foi de 4,96% em relação ao valor total investido.

Já na modalidade de aposta 2, que reflete as classes acima e abaixo de 2,5 gols na partida, com dois possíveis resultados: acima ou abaixo, foi obtida uma taxa de assertividade de aproximadamente 57%. Essa assertividade apresentou um lucro de 4,65% em relação ao valor total investido nas apostas.

5.2. Cenários com restrições de probabilidade

5.2.1. Cenário 1 – Ambas Marcam (AM)

Para os cenários com restrições de probabilidade, temos inicialmente uma restrição de 60% onde foi alcançado um prejuízo de 2,14%. Neste, 167 jogos fizeram parte do grupo analisado, obtendo-se uma assertividade aproximada de 53,89%. Para restrição de 70%, obteve-se um lucro de 6,19% em relação ao valor investido. Este grupo teve 64 jogos investidos com uma taxa de 57,84% de acerto, ou seja 37 dos 64 jogos. Ao subir a restrição para taxa de 80%, o percentual de lucro aumentou para 6,42% e nesse grupo foi investido apostas em 57 partidas com uma taxa de acerto de 57,89%.

Elevando um pouco mais a restrição, dessa vez para 90%, obteve-se também lucro. Entretanto, o percentual ficou abaixo dos casos com restrições de 70 e 80%, a taxa de lucratividade ficou próxima a 6% e com um percentual de 57,7% de assertividade para um grupo de 45 partidas apostadas.

Por sua vez, com restrição de 95%, a margem de lucro alcançou 5,6% para um grupo de 40 partidas e assertividade de 57,5%. Por fim, na restrição de 99%, o experimento obteve o seu melhor desempenho com um percentual lucrativo de 17,33%em relação ao valor total investido. Neste último caso, o grupo possuiu 36 jogos com uma taxa de assertividade de 63,88%. Para uma melhor visualização dos resultados alcançados nesse cenário é possível visualizar de forma resumida na tabela abaixo (Tabela 5) o tamanho do grupo analisado, seu respectivo Lucro/Prejuízo e a taxa de assertividade para cada restrição aplicada.

Tabela 5. Dados obtidos na análise do cenário 1.

Cenário 1			
Restrição	Grupo de partidas	Lucro/Prejuízo	Assertividade
Sem restrição	360	-4,96%	52,22%
Maior ou igual a 60%	167	-2,14 %	53,89%
Maior ou igual a 70%	64	6,19 %	57,84%
Maior ou igual a 80%	57	6,42 %	57,89%
Maior ou igual a 90%	45	6,0 %	57,70%
Maior ou igual a 95%	40	5,6 %	57,50%
Maior ou igual a 99%	36	17,33 %	63,88%

Fonte: Autor.

5.2.2. Cenário 2 – Acima/Abaixo 2,5 Gols

Neste cenário, ao aplicar restrição de 60% de probabilidade, foi possível obter um lucro de 8,39% para um total de 292 jogos investidos com uma taxa de acerto de 59,44%. Já na restrição de 70% o lucro aumentou em relação a restrição de 60%, sendo possível

chegar em 12,43% de lucro investindo em 44 partidas com uma assertividade de 63,63%.

Na restrição de 80% continuou o crescimento do lucro, chegando a 16,53% com uma taxa de acerto de 66,66%, para um total de 36 partidas investidas. Continuando o crescimento do percentual de restrição para 90%, o percentual lucrativo desse grupo alcançou a taxa de 25,79% de lucro em 19 jogos com uma taxa de acerto de 73,68%.

Para uma melhor visualização dos resultados alcançados nesse cenário é possível visualizar de forma resumida na tabela abaixo (Tabela 6) o tamanho do grupo analisado, seu respectivo Lucro/Prejuízo e a taxa de assertividade para cada restrição aplicada.

Tabela 6. Dados obtidos na análise do cenário 2.

Cenário 2			
Restrição	Grupo de partidas	Lucro/Prejuízo	Assertividade
Sem restrição	360	4,65%	59,44%
Maior ou igual a 60%	292	8,39 %	62,32%
Maior ou igual a 70%	44	12,43 %	63,63%
Maior ou igual a 80%	36	16,53 %	66,66%
Maior ou igual a 90%	19	25,79 %	73,68%
Maior ou igual a 95%	18	23,61 %	72,22%
Maior ou igual a 99%	17	21,35 %	70,58%

Fonte: Autor.

Continuando a análise dos resultados desse cenário, na restrição de 95% o lucro continuou alto, porém um pouco abaixo da restrição de 90%, chegando a 23,61% de lucro aplicados ao investimento em apostas para um grupo de 18 partidas atingindo o valor de 72,22% de assertividade. Por fim, a restrição de 99% também foi um pouco menor que a restrição de 90% e também da restrição de 95%, chegando ao valor de 21,35% de margem de lucro, em 17 partidas investidas com uma assertividade de 70,58%.

Comparando os resultados obtidos nos dois cenários, o cenário 2 em todos os resultados apresentou lucro, diferentemente do que se alcançou no cenário 1, conforme a Figura 8 evidência.

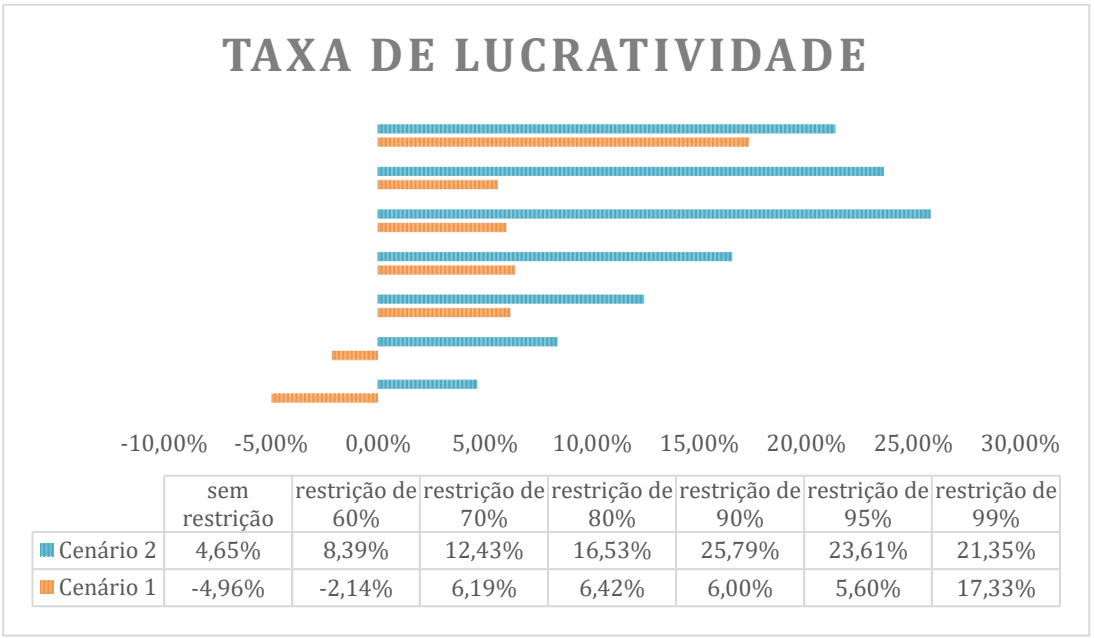

Figura 8. Gráfico representativo da taxa de lucratividade dos cenários avaliados.
Fonte:Autor.

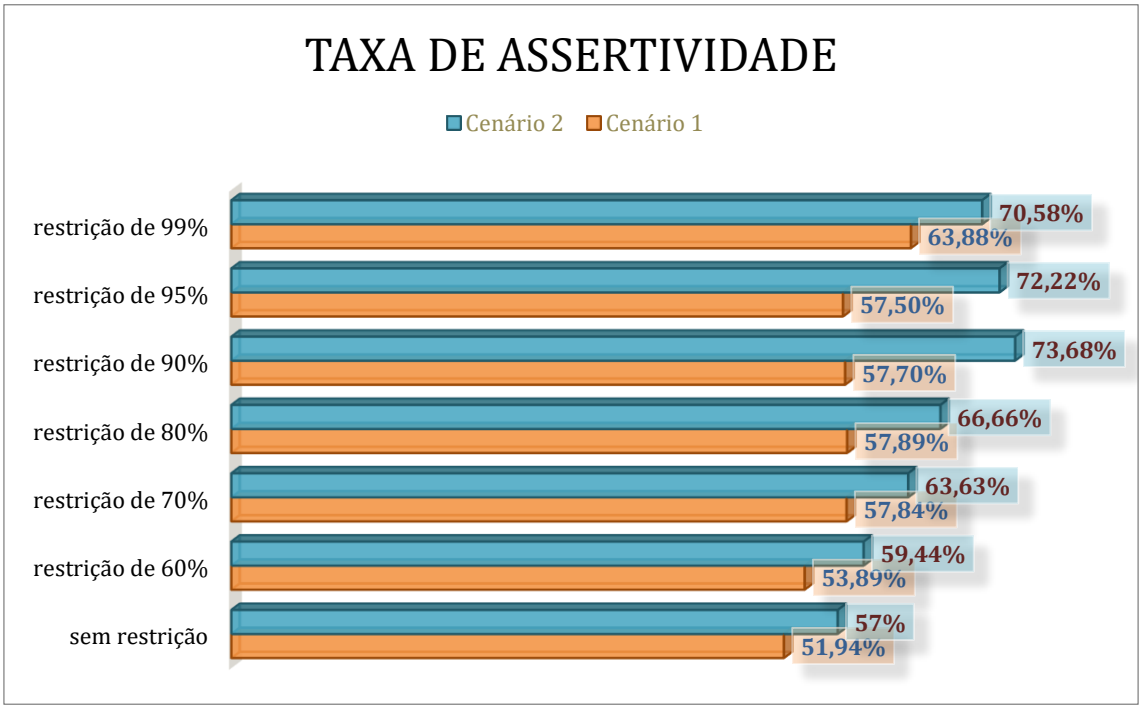

Figura 9. Gráfico representativo da taxa de assertividade dos cenários avaliados.
Fonte:Autor.

Em comparativo também, temos a taxa de assertividade que foi bem mais positiva no cenário 2, chegando a atingir um valor de 73,68% enquanto a taxa máxima obtida no cenário 1 foi de 63,88% conforme é possível visualizar através do gráfico apresentado na figura acima (Figura 9).

5.3. Comparação dos trabalhos

Como parte do objetivo é validar o desempenho do AutoML, foram comparados os resultados obtidos com o experimento apresentado com o trabalho de Almeida (2019). Ambos os trabalhos usaram a mesma base de dados do Campeonato Brasileiro de Futebol da Série A de 2017 e foi seguida a mesma estratégia, onde para prever uma nova rodada, o treinamento é realizado com os dados de todas as rodadas anteriores.

Tabela 7. Comparativo das taxas de lucro do trabalho modelo com o experimento.

	Cenário 1		**Cenário 2**	
Probabilidades	Almeida (2019)	Experimento	Almeida (2019)	Experimento
Sem restrição	**5,11%**	-4,96%	**5,51%**	4,65%
>= 60%	**4,04%**	-2,14%	-2,76%	**8,39%**
>= 70%	3,34%	**6,19%**	2,95%	**12,43%**
>= 80%	-1,59%	**6,42%**	7,86%	**16,53%**
>= 90%	**13,09%**	6%	12,24%	**25,79%**
>= 95%	-0,41%	**5,60%**	5,96%	**23,61%**
>= 99%	**39,13%**	17,33%	13,29%	**21,35%**

Fonte:Autor.

Observando a Tabela 7, é possível notar que o desempenho do experimento no cenário 2 foi bastante superior ao apresentado no trabalho de Almeida (2019), onde em 7 casos de lucro, o experimento obteve resultados melhores em 6 deles. Destaca-se que

esses resultados foram mais que o dobro dos mesmos cenários do trabalho modelo.

Já no cenário 1, o trabalho de Almeida (2019) teve melhor desempenho em 4 dos 7 possíveis casos. Porém, isso não altera o fato que o experimento saiu lucrativo em 5 casos e em apenas 2 apresentou prejuízo.

O resultado encontrado foi o esperado, pois com a utilização do AutoML, foi possível testar e utilizar vários algoritmos diferentes na obtenção dos melhores resultados, logo o esperado era obter desempenho superior ao trabalho realizado pelo Almeida (2019). Os resultados obtidos desse trabalho contribuem para a área de dados, pois fornece evidências que o uso de AutoML pode ser vantajoso pela facilidade, agilidade e resultado obtido no final.

6. Conclusões e Trabalhos Futuros

O objetivo principal deste trabalho, é avaliar se o com a utilização de AutoML nas apostas esportivas é possível conseguir um desempenho lucrativo e superior ao obtido a partir do uso individual de um modelo de aprendizagem de máquina, como fez o trabalho de Almeida (2019).

A metodologia aplicada nesse estudo para a base de dados do campeonato Brasileiro de Futebol Série A (2017) apresentou um cenário mais lucrativo que o trabalho de Almeida (2019), além disso, inferiu-se que aplicando as restrições de probabilidade foi possível alcançar um maior percentual de lucro.

O confronto entre os resultados desse experimento e o estudo de Almeida (2019) destacou que o resultado obtido no cenário 2 no estudo atual possui uma lucratividade até 300% maior que o modelo. Para o cenário 1, apesar do modelo apresentar melhor desempenho em 4 dos 7 jogos, o experimento com AutoML alcançou resultados satisfatórios.

O uso do AutoML foi bastante importante nos resultados deste trabalho, pois foi possível por meio do Pycaret utilizar vários algoritmos de forma automatizada ganhando tempo, pois sem o uso do AutoML levaria mais tempo para iterar, otimizar, testar e gerar modelos em vários algoritmos diferentes. Destaca-se que essa tecnologia possibilitou ao autor, que não é da área de dados, alcançar situações de lucro, cenário esse alcançado por apenas 5% do total dos apostadores esportivos.

O experimento utilizou a mesma estratégia do estudo de Almeida (2019), não implementando algumas técnicas que poderiam ter melhorado o desempenho. Segundo Barella (2019), algumas limitações podem prejudicar a acurácia de um algoritmo de classificação, entre eles o desbalanceamento da quantidade de exemplos nas classes de um conjunto de dados, que acaba gerando dificuldade para os modelos classificarem classes com poucos representantes(Classes minoritárias). Assim sendo, para trabalhos futuros, o autor pretende:

- realizar o balanceamento dos dados.
- aumentar o número de campeonatos, como consequência possuir mais dados para treinar o modelo;
- Incluir mais dados sobre os times, como por exemplo: Quantidade de chutes no gol, quantidade de escanteios, jogadores machucados, gastos com salários, dados de campeonatos anteriores;
- incluir dados individuais dos jogadores, como por exemplo: número de gols, quantidade de assistências, média de cartões tomados; e
- incluir mais modalidades de apostas, como quantidade de escanteios no jogo,

quantidade de cartões e quantidade de chutes no gol;

7. Referências

ALMEIDA, L.A.B. Previsão de resultados de jogos do Campeonato Brasileiro de Futebol utilizando aprendizagem de máquina. Universidade Federal Rural de Pernambuco (UFRPE) - Unidade Acadêmica de Garanhuns, 2019.

BARELLA, V.H. Técnicas para o problema de dados desbalanceados em classificação hierárquica. USP – São Carlos, p. 20, 2019. Disponível em :<https://www.teses.usp.br/teses/disponiveis/55/55134/tde-06012016-145045/publico/VictorHugoBarella_dissertacao_revisada.pdf>. Acesso em: 27/05/2022.

BET365. Esportes. Disponível em:<https://bet365.com/#/HO/>. Acesso em:11/05/2022.

CARNEIRO, J. Os esportes mais previsíveis em apostas esportivas. Redação Esportes, 2022. Disponível em:< https://www.jesocarneiro.com.br/esporte/os-esportes-mais-previsiveis-em-apostas-esportivas.html>. Acesso em: 10/05/2022

DUARTE, L. 1X2 – PREVISÃO DE RESULTADOS DE JOGOS DE FUTEBOL. Faculdade de Engenharia Universidade do Porto, 2015. Disponível:<https://repositorio-aberto.up.pt/bitstream/10216/79327/2/35444.pdf>. Acesso em: 14/05/2022

GÉRON, A. Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow. Alta Books, Rio de Janeiro, 2019.

GLOBO. Ranking dos elencos do Brasileirão 2017. Disponível em:<http://app.globoesporte.globo.com/futebol/brasileirao-serie-a/guia/avaliacao-de-elencos-brasileirao-2017/index.html#palmeiras>. Acesso em: 16/05/2022

GLOBO. O mercado de apostas esportivas. Esporte para sentir, 2021. Disponível em:<https://negociostvriosul.com.br/Imagens/MidiaKit/MidiaKit_295_pdf_20210729110045.pdf>. Acesso em: 16/05/2022

GRAND VIEW RESEARCH. Sports Betting Market Size, Share & Trends Analysis Report By Platform (Online, Offline), By Type (Fixed Odds Wagering, eSports Betting), By Sports Type (Football, Basketball), By Region, And Segment Forecasts, 2021.

IG. A realidade das apostas esportivas e das casas de apostas. Esporte IG, 2021. Disponível em:<https://esporte.ig.com.br/futebol/2021-10-21/a-realidade-das-apostas-esportivas-e-das-casas-de-apostas.html>. Acesso em: 18/05/2022

MONARD, M.C.; BARANAUSKAS, J.A. Conceitos sobre Aprendizado de Máquina, Cáp. 4, 2003. Disponível em:<dcm.ffclrp.usp.br/~augusto/publications/2003-sistemas-inteligentes-cap4.pdf>. Acesso em: 19/05/2022

PYCARET. Classification - PyCaret Official. Disponível em:<https://pycaret.readthedocs.io/en/latest/api/classification.html>. Acesso em: 19/05/2022

PYCARET. Quickstart - PyCaret Official. Disponível em:<https://pycaret.gitbook.io/docs/get-started/quickstart#classification>. Acesso em: 20/05/2022

SABINO, A.; GABRIEL, J. Com alto investimento no futebol, sites de apostas esperam regulamentação. Yahoo Esportes, 2022. Disponível em:<https://esportes.yahoo.com/noticias/com-alto-investimento-no-futebol-170200943.html>. Acesso em: 20/05/2022

SEHNEM, R.; FROZZA, R.; BAGATINI, D.D.S.; PERACONI, S. Análise de Variáveis em Partidas de Futebol: Previsão de Resultados com Naïve Bayes e Poisson. Universidade de Santra Cruz do Sul, 2021.

SILVA, L.L.F. Uso de Automated Machine Learning (Auto ML) em Sistemas de Recomendação. 2021. 11f. Trabalho de Conclusão de Curso (Artigo), Curso de Ciência da Computação, Centro de Engenharia Elétrica e Informática, Universidade Federal de Campina Grande - Paraíba - Brasil, 2021. Disponível em: <http://dspace.sti.ufcg.edu.br:8080/jspui/handle/riufcg/24996>. Acesso em: 19/05/2022

HERBST, Cynthia. O que é AutoML e quais são as suas vantagens? Eldorado, 2022. Disponível em:<https://www.eldorado.org.br/en/blog/o-que-e-automl-e-quais-sao-as-suas-vantagens/>. Acesso em: 07/06/2022