UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

Marcela Soares Domingues

Uso de Business Intelligence como Ferramenta de Apoio em Previsões de Apostas de Jogos de Futebol

Recife

2021

Marcela Soares Domingues

Uso de Business Intelligence como Ferramenta de Apoio em Previsões de Apostas de Jogos de Futebol

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.

Aprovada em: 16 de Dezembro de 2021.

BANCA EXAMINADORA

Nome do Orientador: Gabriel Alves de Albuquerque Junior
Departamento de Estatística e Informática
Universidade Federal Rural de Pernambuco

Roberta Macêdo Marques Gouveia
Departamento de Estatística e Informática
Universidade Federal Rural de Pernambuco

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)

D671u Domingues, Marcela
Uso de Business Intelligence como Ferramenta de Apoio em Previsões de Apostas de Jogos de Futebol / Marcela Domingues. - 2021.
21 f. : il.

Orientador: Gabriel Alves de Albuquerque Junior.
Inclui referências.

Trabalho de Conclusão de Curso (Graduação) - Universidade Federal Rural de Pernambuco, , Recife, 2021.

1. Apostas esportivas. 2. Business Intelligence. 3. Análise de Dados. 4. Investimento em apostas. 5. Apostas de jogos de futebol. I. Junior, Gabriel Alves de Albuquerque, orient. II. Título

CDD

Uso de Business Intelligence como Ferramenta de Apoio em Previsões de Apostas de Jogos de Futebol

Marcela Soares Domingues [1]

[1] Departamento de Estatística e Informática – Universidade Federal Rural de Pernambuco
Rua Dom Manuel de Medeiros, s/n - CEP: 52171-90 - Recife - PE - Brasil

marcela.domingues@gmail.com

Abstract. *The number of Brazilians that currently make investments on online sports betting, mainly related to football teams, are huge. This practice is allowed in Brazil since 2018, according to the law 3.756/18. With this permission granted, added to the Brazilian passion for football games, this is an investment area that has been growing fast since then, reaching high levels of financial movements, making people to call on this practice as guaranty of extra or main source of incomes. To make an investment in an specific bet, the ideal is that you have some idea of the risks you are taking. Betting involves losses and sometimes the investments are worthless. To know that, is necessary to take into account some variables as historical data from football teams and players. The use of Business Intelligence, that is already used for some teams to study the behaviour of others, is also of great value when it comes to analyze which bets are worth to make an invesment. Through BI concepts, this paper presents the development of a dashboard containing graphs and reports with the intention to help on the bet analysis. Besides that, it also presents a predict model to find out the amount of goals in an specific match and the calculation of winning probabilities of each team in a given match.*
Keywords: *Sports Betting. Soccer Betting. Betting investment. Business Intelligence. Data Analysis.*

Resumo. São muitos os brasileiros que atualmente recorrem a casas de apostas esportivas online, principalmente de futebol, como forma de investimento. Essa prática vem sendo permitida no Brasil desde 2018, de acordo com a lei 13.756/18. Com a permissão liberada, somada a paixão do brasileiro pelo futebol, essa é uma área de investimento que vem crescendo muito desde então, atingindo altíssimos níveis de movimentação financeira, fazendo com que muitas pessoas recorram a essa prática como garantia de fonte de renda extra ou fixa. Para investir em determinada aposta, o ideal é que se tenha uma noção dos riscos que você vai assumir. Apostar envolve chances de perda, e por isso, nem sempre o investimento valerá a pena. Deve-se levar em conta algumas variáveis, como por exemplo dados históricos de times e jogadores de futebol. O uso de *Business Intelligence* (BI), que já é bastante utilizado por times de futebol para estudar comportamentos de outras equipes, também agrega bastante valor na hora de analisar as apostas nas quais se quer investir. Utilizando conceitos de BI, este trabalho apresenta

o desenvolvimento de um *Dashboard* (painel de controle) contendo gráficos e relatórios, a fim de facilitar o processo de análise das apostas, além de um modelo de predição para prever a quantidade de gols em um jogo e o cálculo de probabilidades de vitória de cada time em uma determinada partida.
Palavras-chave: Apostas esportivas. Apostas de jogos de futebol. Investimento em apostas. Business Intelligence. Análise de Dados.

1. Introdução

Não há dúvidas de que o esporte sempre foi uma grande paixão da humanidade. À medida que os esportes foram se tornando mais populares, e com o surgimento dos jogos olímpicos, a prática de apostas esportivas foi evoluindo. A ideia de poder ganhar dinheiro enquanto acompanha seu esporte favorito, fez com que essa prática se tornasse bastante popular. Bem diferente de um casino e de apostas mais conhecidas, como as da Mega Sena, que se baseiam principalmente na sorte, as apostas esportivas são atreladas hoje a um mercado especulativo [Appelbaum, 2019].

Nos últimos anos, casas de apostas *online* vêm investindo bastante no mercado brasileiro de futebol. Atualmente, é comum ver esses sites como patrocinadores da grande maioria das equipes de futebol da série A do campeonato brasileiro, por exemplo. A prática de apostas relacionadas a competições de futebol vem se tornando uma opção de investimento bastante popular e um mercado que vem movimentando bilhões de reais. A movimentação financeira nesta área mais do que dobrou de tamanho no Brasil entre os anos de 2018 e 2020, saindo de R$ 2 bilhões de reais para R$ 7 bilhões [Feitosa, 2020]. Conforme a Lei 13.756/18 [Temer *et al.*, 2018], que permite essa atividade no Brasil desde 2018, ela funciona como uma loteria em que o apostador tenta prever o resultado de eventos reais esportivos, como placar, número de cartões, quem fará o primeiro gol, além de várias outras informações.

Se o que for previsto pelo apostador condisser com a realidade, então o investimento feito retorna a ele, além de receber também uma quantia a mais. O valor dessa quantia é baseado no tipo da aposta e nas probabilidades conhecidas antes de o evento acontecer. As categorias de apostas existentes e o cálculo dessas probabilidades são explicados de maneira detalhada neste trabalho, com o intuito de achar indicadores que aumentem a probabilidade de acerto nas apostas. Além disso, alguns conceitos de *Business Intelligence* (BI) [Turban *et al.*, 2009] são abordados com a finalidade de fornecer uma ferramenta onde o usuário consegue fazer suas próprias análises, filtrando informações de acordo com sua necessidade. Também são apresentados, um modelo de previsão para prever a quantidade de gols em uma partida de futebol e o cálculo de probabilidades a fim de saber as chances de vitória de cada time em uma determinada partida.

A partir do momento que se conhece o funcionamento das apostas, o ideal é que se busque uma forma de aumentar as chances de acerto dos resultados, para que se tenha o mínimo de risco possível no investimento. Para que esse objetivo seja atingido, é recomendável que uma análise prévia seja feita, levando-se em consideração dados históricos de times, jogadores, campeonatos, etc. Diante desse cenário, o uso de BI como uma ferramenta de suporte a tomada de decisões é

extremamente relevante. Algumas outras técnicas como cálculo de probabilidades [Ross *et al.*, 2010] e *Machine Learning* [Hartshorn, 2016] também são abordadas a fim de que se consiga obter previsões seguras, que façam com que o apostador tenha chances de obter um bom retorno financeiro.

1.1. Justificativa e Motivação

Uma investigação detalhada de um grande conjunto de dados organizados, com informações históricas sobre performance de jogadores, times e campeonatos, ajuda bastante a diminuir os riscos na hora de elaborar a previsão de uma aposta, além de otimizar o tempo de todo o processo de análise prévia.

Normalmente, essas análises são feitas cruzando informações de diversos sites relacionados a campeonatos de futebol e alimentando algum tipo de planilha de controle, normalmente fornecidas por cursos de *trade* (comércio) esportivo, onde o apostador consegue criar fórmulas que calculam as probabilidades necessárias. Essas planilhas geralmente possuem uma quantidade tão grande de dados que acaba dificultando e atrasando o processo de análise, levando-se em consideração a baixa usabilidade desse tipo de ferramenta.

A grande maioria das plataformas de dados esportivos fornecem as informações desejadas pelo apostador, mas nem sempre ele consegue encontrar todas as informações relevantes juntas em um só lugar e de maneira organizada. Além disso, um grande número de informações nestas plataformas só estão disponíveis para quem possui assinatura, e os preços dos planos geralmente são bem altos e disponíveis apenas em moedas como o euro ou a libra, tornando-se uma opção inviável para muitos.

Tendo em vista os problemas citados acima, a criação de um *dashboard* que reúna todos os dados necessários para se realizar uma predição, e que possua uma boa variedade de gráficos e relatórios que consigam transformar esses dados em informações relevantes de maneira interativa e visualmente agradável, tornam o processo de análise bem mais intuitivo e preciso. Além disso, a utilização das técnicas de *Machine Learning* e probabilidades garantem bons resultados quando se trata de realizar previsões, como podemos ver em alguns trabalhos relacionados citados na seção 2 deste trabalho.

1.2. Objetivos

Um dos principais objetivos deste trabalho é o desenvolvimento de um *dashboard*, onde é possível realizar a análise de diversos dados históricos relacionados a jogadores, equipes e campeonatos de futebol, facilitando a definição de previsões por apostadores. Além disso, o desenvolvimento de um modelo para prever a quantidade de gols feitos em um jogo (parâmetro de grande importância em apostas) e o cálculo da probabilidade de vitória de um time em uma determinada partida, também fazem parte dos objetivos principais do trabalho. Alguns pontos são levados em consideração no processo: (I) estudo sobre o domínio do problema; (II) análise sobre a ferramenta de desenvolvimento do *dashboard; (III)* definição de indicadores que ajudem a prever resultados; (IV) análise do algoritmo de regressão a ser utilizado; (V) elaboração de gráficos e relatórios.

A partir do *dashboard* criado é possível visualizar de maneira organizada, através de gráficos e relatórios, todos os dados necessários para ajudar o usuário na tomada de decisão ao fazer uma aposta. Além disso, o desenvolvimento de um modelo que seja capaz de realizar a previsão da quantidade de gols esperados em uma determinada partida, levando-se em consideração alguns indicadores importantes como a média de gols dos times como mandantes e visitantes.

1.3. Organização do Trabalho

O conteúdo deste artigo está dividido em 6 seções. Na primeira, tem-se uma breve introdução sobre o objeto de estudo, assim como a justificativa e objetivos que levaram a realização deste trabalho. Na seção 2, uma análise de trabalhos correlatos, que buscam mostrar estudos anteriores a respeito de previsões assertivas em apostas esportivas, e também, sobre o uso de *Business Intelligence* nesse mercado. Na terceira, pode-se encontrar todo o referencial teórico a respeito do trabalho A seção 4 contém a metodologia utilizada na realização do trabalho. Na seção 5 tem-se a descrição dos resultados obtidos, e logo após, a seção 6, onde é descrita a conclusão do estudo e trabalhos futuros.

2. Trabalhos Relacionados

Nesta seção são abordados alguns trabalhos relacionados a *Business intelligence* no contexto de apostas esportivas, onde pode-se encontrar pontos relevantes relacionados ao tema deste artigo.

No artigo [da Costa *et al.*, 2017]*,* é apresentado um estudo sobre análise de dados esportivos, apresentando diversos conceitos relevantes e descrevendo todo o processo de seleção, coleta, tratamento, armazenamento e consumo dos dados. De acordo com o trabalho, todo esse ciclo é conhecido como Processo de Descoberta de Conhecimentos em Bases de Dados (em inglês, *Knowledge Discovery in Databases* - KDD). O KDD pode ser definido como um processo não trivial, de extração de informações, previamente desconhecidas e potencialmente úteis a partir de um conjunto de dados [Gama, 2010]. Assim, muitas pesquisas na área de computação aplicadas em análise de dados esportivos segue implicitamente esse processo. O trabalho apresenta uma introdução ao tema, tentando definir qual é o papel da preditibilidade e aleatoriedade nos resultados esportivos. Além disso, também destaca a relação entre a análise de dados esportivos e os mercados de aposta. Por fim, chegam a conclusão que estamos entrando em uma nova era, onde os dados exercem uma grande influência. É dito que o esforço humano ainda é essencial em análises esportivas, mas em breve estaremos lidando com uma quantidade tão grande de dados que será inviável que este trabalho seja feito por pessoas. Ainda segundo as considerações finais do artigo, as máquinas serão mais capacitadas para avaliar os dados e sugerir mudanças em estratégias de jogo, por exemplo, fazendo com que engenheiros e cientistas de dados assumam um papel de destaque nos bastidores.

Já o objeto de estudo do trabalho [Goddard e Asimakopoulos, 2004] se baseia em um modelo matemático que calcula o valor de *odds (*probabilidades dentro de um determinado evento) [Hilgler, 2013] em apostas de jogos de futebol. O modelo é usado para testar a eficiência das *odds* fixas no mercado de apostas. Assim que uma

aposta é feita, as *odds* são fixas, mas o *bookmaker* (responsável por estabelecer o valor das *odds*) pode ajustar esse valor à medida que a hora da partida se aproxima, de modo a equalizar o volume de apostas feitas em qualquer um dos times. Nesse estudo, chega-se a conclusão que se o modelo de previsão gera probabilidades que não estão refletidas nas probabilidades citadas pelo *bookmaker,* então elas falham em satisfazer critérios do padrão de eficiência.

Algumas publicações focam o estudo das características de jogadores e times de futebol para realizar previsões, como é o caso do trabalho [Stübinger *et al.*, 2019], onde eles utilizam um grande volume de dados relacionados a jogadores de diferentes campeonatos europeus para entender de maneira mais direta de que forma esses dados influenciam nos resultados dos jogos. Para atingir o objetivo do trabalho, são utilizadas algumas técnicas de *Machine Learning*, como Árvores de decisão, *Boosting (BOO)* [Schapire e Freund, 2014], Máquina de vetor de suporte (SVM) [de Almeida e da Cunha, 2015]e Regressão linear [Diniz e Thiele, 2021]. Utilizando essa abordagem o resultado final do estudo concluiu que retornos econômicos e estatisticamente significativos de 1,58% foram alcançados.

No artigo [Mattera, 2021], o autor explica que pretende fornecer uma estrutura simples para obter previsões precisas para resultados binários em partidas de futebol. Em estatística, uma abordagem comum para modelagem e previsão de séries temporais binárias pode ser encontrada na Média Móvel Autorregressiva Binária, um caso especial da classe de modelos lineares generalizados especialmente projetados para dados de séries temporais, chamada média móvel autorregressiva generalizada [Nielsen, 2021]. Para mostrar a utilidade do método estatístico proposto, dois experimentos empíricos para a *Premier League* inglesa e para a Série A italiana são fornecidos para prever cartões vermelhos, *Under/Over* e eventos *Goal/No Goal*. No final, são enfatizadas as vantagens de uma estratégia baseada em evitar apostas no caso de eventos muito incertos. Com isso, a ausência de recomendação, devido ao raciocínio probabilístico, permite aumentar a precisão das apostas e reduzir os custos, sendo visto como um ponto forte da abordagem proposta.

No trabalho [Beal *et al.*, 2021], é apresentado um conjunto de dados de *benchmark* [Tableau, 2021] focado em aplicativos e resultados de um conjunto de modelos básicos de Processamento de Linguagem Natural e Aprendizado de Máquina para previsão de resultados de partidas em jogos de futebol. Com isso, é fornecida uma linha de base para a acurácia da previsão que pode ser alcançada explorando dados de partidas e artigos contextuais de jornalistas esportivos. O conjunto de dados concentra-se em um período de tempo representativo ao longo de 6 temporadas da *Premier League*, e inclui prévias de jogos do jornal *The Guardian*. Os modelos apresentados neste artigo alcançaram uma precisão de 63,18%, mostrando um aumento de 6,9% com relação aos métodos estatísticos tradicionais.

Através da análise dos trabalhos citados, foram definidos alguns recursos a serem utilizados neste artigo, como o processo KDD na análise de dados , técnicas de *Machine Learning* para previsões de resultados relacionados a apostas esportivas e cálculos de probabilidades.

3. Referencial Teórico

Nesta seção são apresentados os principais conceitos relacionados ao contexto deste trabalho, que ajudam no entendimento do desenvolvimento do estudo.

3.1. Apostas Esportivas

No ramo das apostas esportivas, existem duas modalidades comuns: o *Punting* e o *Trading*. O *Punting* é o estilo clássico de aposta, e mais conhecido, que é o que tenta prever qual equipe sairá vitoriosa em uma partida, por exemplo. Além disso, é possível prever se a partida terminará empatada, o número de cartões que a partida terá, a quantidade de gols e diversas outras opções. Nessa modalidade, o apostador disputa com a casa de apostas. Caso sua previsão não seja correta, o investimento feito é perdido, e em caso de acerto, a casa de apostas deverá devolver ao apostador a quantia investida, além de pagar um valor a mais, dependendo da probabilidade de o evento ocorrer [Appelbaum, 2019].

Já o *Trading*, é uma modalidade um pouco mais complexa de apostas. Se assemelha bastante a uma Bolsa de Valores, onde o apostador trabalha com variações nas cotações dos eventos esportivos. Neste tipo de aposta, o apostador não necessariamente faz uma análise prévia das equipes, fazendo e encerrando as apostas de acordo com o que está acontecendo na partida, em tempo real.

Dentro do contexto deste estudo, sempre que apostas forem mencionadas, elas serão da modalidade *Punting*, visto que nesta opção o apostador precisa realizar uma análise prévia das equipes e jogadores, podendo se beneficiar com o objeto de estudo deste trabalho.

Quando se trata de apostas esportivas, uma das palavras mais citadas é a *odd*. Uma *odd* nada mais é do que uma cotação relacionada a um time em uma aposta de futebol, por exemplo. Elas determinam a probabilidade de um evento específico acontecer, ou seja, quanto maior for a *odd*, menor será a chance daquele evento se tornar real. O valor calculado dessas probabilidades varia de acordo com cada casa de aposta, mas normalmente se aproximam bastante um do outro, estando sempre dentro de uma margem de erro.

O retorno sobre o valor investido em determinada aposta está diretamente ligado às *odds.* Cada aposta tem um valor de *odd* associado a ela. Por exemplo, se um time conhecido por ser muito vitorioso enfrenta um time considerado fraco, a aposta no time de melhor desempenho é uma opção muito óbvia, e quem investir nela estará se arriscando muito pouco, e com isso, obtendo um baixo retorno financeiro. Neste exemplo, o time considerado mais vitorioso possui um valor de *odd* menor do que o time mais fraco. A partir daí, podemos perceber que quanto menor o valor da *odd*, maior será a probabilidade do evento acontecer.

3.2. Business Intelligence

O *Business Intelligence* combina análise de negócio, *data mining*, visualização de dados, ferramenta de dados e infraestrutura, e as melhores práticas para ajudar organizações ou pequenos negócios no momento da tomada de decisões. Na prática, é possível perceber o uso de BI quando existe uma visualização organizada dos

dados, e esses dados ajudam a direcionar decisões, descartar opções ineficientes e a se adaptar rapidamente ao mercado [Turban *et al.*, 2009].

O termo B*usiness Intelligence* abrange também os métodos e processos de coleta e armazenamento de dados, além da análise de dados de negócios ou atividades que precisem de melhora na performance, como foi dito anteriormente. Ao longo dos anos, essa prática foi melhorada e passou a incluir cada vez mais processos e atividades que ajudam a melhorar o desempenho. Alguns dos principais exemplos são:

- **Métricas de desempenho**: comparação de desempenhos atuais com dados históricos, normalmente usando painéis personalizados;
- **Criação de relatórios**: compartilhamento da análise de dados com *stakeholders*;
- **Consulta**: consultas específicas aos dados, onde o *Business Intelligence* puxa as respostas dos conjuntos de dados.
- **Visualização de dados**: transformar a análise de dados em representações visuais, como tabelas, gráficos e relatórios, para que os dados sejam consumidos com facilidade.

3.3. KDD

O objetivo principal do processo de *KDD* (*Knowledge Discovery in Databases*) é extrair conhecimento de grandes bases de dados. A descoberta do conhecimento envolve uma sequência de fases, como a coleta de dados, processamento, tratamento, mineração e interpretação do resultado final da extração do conhecimento. O KDD se trata de um processo iterativo, visto que pode ser repetido quantas vezes for necessário, em busca de melhores resultados [Gama, 2010]. Para se utilizar deste processo é necessário que se tenha objetivos bem definidos e que se saiba quais resultados se deseja atingir. Na Figura 1 tem-se uma demonstração das etapas do processo KDD.

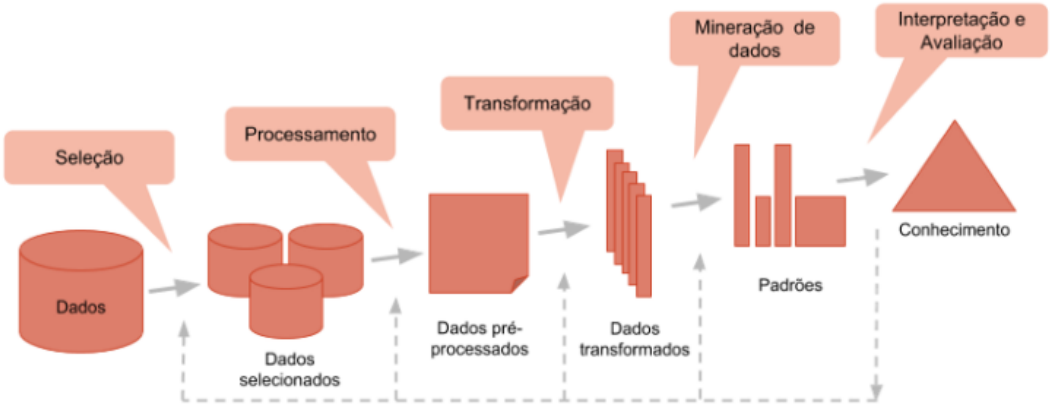

Figura 1. Etapas do processo KDD. Fonte: [Moura, 2019]

A etapa de seleção consiste em selecionar um conjunto de dados que fará

parte do processo de análise. Esses dados podem vir de bases variadas como planilhas e sistemas e possuir dados com diferentes formatos.

Na etapa de processamento é feita a verificação da qualidade dos dados armazenados. Toda a base passa por um processo de limpeza, correção ou remoção de dados inconsistentes, verificação de dados ausentes ou incompletos e de *outliers* (anomalias).

Na transformação são aplicadas algumas técnicas como normalização, agregação, criação de novos atributos, redução e sintetização dos dados. Nessa etapa os dados ficam disponíveis agrupados em um mesmo local para que se possa aplicar os modelos de análise.

É na etapa de mineração que se constrói modelos ou se aplica técnicas de mineração de dados. Os objetivos dessas técnicas são verificar uma hipótese e descobrir novos padrões. Além disso, a descoberta pode ser dividida em: preditiva e descritiva. Esses modelos geralmente são aplicados e refeitos inúmeras vezes dependendo do objetivo do projeto.

Finalmente, na etapa de interpretação, é feita a validação do modelo através da avaliação do desempenho, utilizando dados de teste, ou seja, que não foram aplicados na fase de treino.

3.4. Aprendizado de Máquina

O aprendizado de máquina é um método de análise de dados, dentro da área de inteligência artificial, que se baseia na ideia de que as máquinas podem aprender determinadas informações, através de dados históricos, identificando padrões e tomando decisões com o máximo de automatização [Carvalho *et al.*, 2021]. Existem 3 tipos de aprendizado de máquina. São eles: supervisionado, não supervisionado e aprendizado por reforço.

O aprendizado supervisionado se baseia na regressão básica ou classificação. Na classificação o algoritmo é treinado para classificar os dados de entrada em variáveis discretas. Os dados de entrada de treinamento são passados ao algoritmo com um rótulo de “classe”. Por exemplo, os dados de treino podem consistir em características específicas de um determinado jogador. No momento em que se é apresentado um novo conjunto de características, o algoritmo saberá informar se esse jogador atua como um atacante ou não. Já na regressão, a máquina recebe um conjunto de dados para treino, onde ela obtém conhecimento através desses dados, como por exemplo, informações relacionadas a partidas de futebol que já ocorreram. Através do que foi absorvido de conhecimento pela máquina, ela será capaz de realizar previsões, conseguindo identificar qual é a média de gols de um time, por exemplo [Mueller e Massaron, 2019].

Já no aprendizado não supervisionado, a máquina começa a analisar os dados de maneira independente, identificando padrões e aprendendo a separar o que é um atacante de um zagueiro, por exemplo. Neste caso, como a máquina aprende por si só conceitos que nunca viu antes, o processo acaba sendo mais lento.

No aprendizado por reforço o ensinamento é feito com base na experiência, onde a máquina deve realizar treinos, aprender através dos erros e alcançar o mais

próximo possível da abordagem correta. Um exemplo desse tipo de aprendizado é a recomendação de *streamings* de música baseada no que você costuma ouvir. Após escutar um determinado estilo de música, a plataforma assume que este é um estilo do seu gosto musical e passa a fazer recomendações de artistas similares.

Neste trabalho é abordado o aprendizado de máquina do tipo supervisionado, utilizando o modelo de algoritmos de regressão *Random Forest,* que ajuda nos cálculos de previsões relacionadas às apostas.

3.5. Random Forest

O algoritmo do tipo *Random Forest* consiste em um grande número de árvores de decisões individuais que operam como um conjunto. Uma das grandes vantagens desse algoritmo é o fato dele poder ser utilizado tanto para classificação quanto para regressão. Na classificação, cada árvore retorna uma classe de previsão e a classe mais comum se torna o modelo de predição. Já na regressão, é tirada a média de todos os valores previstos por cada árvore chegando ao valor de predição final. Um grande número de árvores (modelos) relativamente não correlacionados operando como um grupo, em geral tem um desempenho melhor do que qualquer um dos modelos individuais [Hartshorn, 2016]. A Figura 2 demonstra o processo do *Random Forest*:

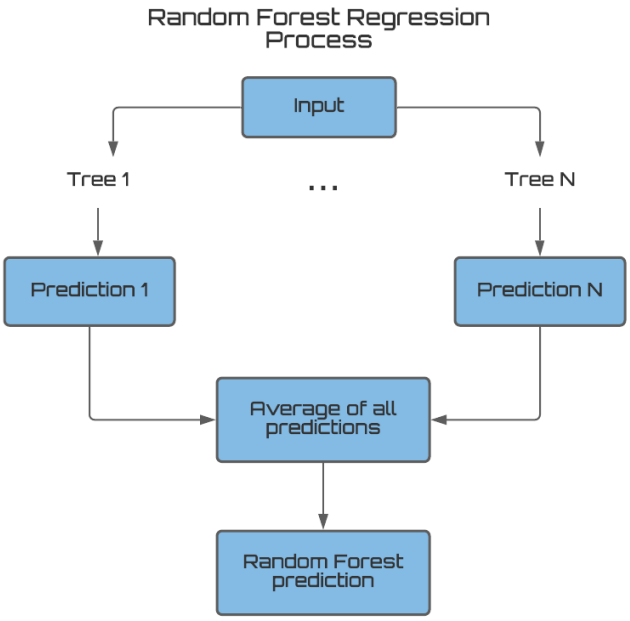

Figura 2. Demonstração do processo do algoritmo *Random Forest Regression.* Fonte: autora.

Ao criar as árvores, o R*andom Forest* adiciona aleatoriedade ao modelo, buscando sempre a melhor característica em algum subconjunto aleatório de características. Este processo, que cria uma grande variedade, é o que faz com que modelos melhores sejam gerados. Neste trabalho a abordagem utilizada para a predição da quantidade de gols esperados em uma determinada partida será a de regressão.

4. Materiais e Métodos

Através dos conceitos apresentados anteriormente, foi possível dar início à implementação deste trabalho. Durante o desenvolvimento, foram utilizadas técnicas referentes ao processo *KDD,* seguindo etapas como seleção, processamento, transformação, mineração e interpretação dos dados. O desenvolvimento foi dividido da seguinte forma:

- Definição da base de dados a ser utilizada;
- Definição de ferramenta de criação do *dashboard;*
- Formatação dos dados;
- Processamento/Transformação dos dados;
- Implementação de algoritmo de aprendizado de máquina;
- *Upload* dos dados na ferramenta escolhida;
- Criação de variáveis calculadas a partir de indicadores selecionados;
- Visualização/Interação dos dados.

4.1. Base de Dados

A base de dados utilizada neste trabalho, possui dados referentes aos times, jogadores, goleiros, juízes e partidas do principal campeonato de futebol da Inglaterra, a *Premier League*. As informações captadas se resumem apenas ao campeonato, ainda em andamento, do ano de 2021, desde a primeira rodada até a rodada mais recente no momento da implementação. A prioridade foi utilizar uma quantidade de dados suficiente (dados relacionados às doze primeiras rodadas do campeonato) para que se conseguisse demonstrar, de maneira funcional, a viabilidade e relevância do presente trabalho.

Na Figura 3 pode-se ver a modelagem das tabelas de dados utilizadas na criação do *dashboard*. Esses dados foram retirados de uma plataforma que possui uma grande base relacionada a dados esportivos, o *InStat* [Ivanskiy, 2007], onde foi possível realizar o *download* através de arquivos no formato *.xlsx.*

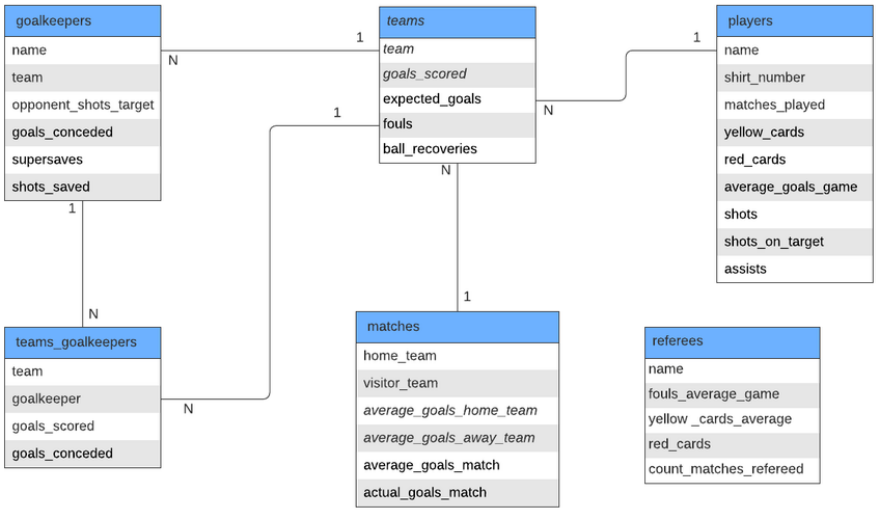

Figura 3. Modelagem das tabelas presentes no dashboard. Fonte: autora.

4.2. Metodologia

A ferramenta escolhida para a criação do d*ashboard*, batizado de *FootBoard*, foi o *Google Data Studio* [Google, 2016]*,* devido aos diversos benefícios oferecidos relacionados à manipulação e visualização de dados. O processo de formatação se deu primeiramente com a conversão dos arquivos *.xlsx* para o formato aceito no *Data Studio*, *.csv* (separado por vírgula), além de alteração da codificação dos caracteres para o tipo UTF-8 e a remoção de caracteres especiais.

Com os conjuntos de dados já disponíveis no *dashboard,* foi possível criar gráficos e relatórios com informações relevantes através da interface do *Data Studio*, que oferece diversas opções de manipulação dos dados, assim como vários componentes visuais, como gráficos e tabelas.

Para o cálculo das *odds*, foram utilizados valores da categoria de cada time e da influência que cada um exerce no mando de campo (desempenho do time ao jogar em casa ou fora). Essa atribuição de valores é bem subjetiva, podendo variar dependendo do usuário, ou seja, cada apostador pode atribuir valores diferentes a essas variáveis. As opções de categoria disponíveis são de A+/A/A- a D+/D/D-, onde A representa um time muito bom, e D um time muito ruim. Os valores correspondentes às probabilidades de vitória, empate ou derrota de cada time foram retirados de uma planilha, utilizada por um usuário real, com dados fixos para cada combinação de categorias, como mostrado na Figura 4.

Equipe 1	Equipe 2	% equipe 1	% empate	% equipe 2
A	A	35%	30%	35%
A	B	52%	25%	23%
A	C	68%	19%	13%
A	D	80%	13%	7%
A	E	89%	7%	4%
A	A-	40%	29%	31%
A	B-	56%	24%	20%
A	C-	71%	18%	11%
A	D-	82%	12%	6%
A	E-	91%	6%	3%
A	A+	31%	29%	40%
A	B+	48%	26%	26%
A	C+	65%	20%	15%
A	D+	78%	14%	8%
A	E+	87%	8%	5%
A+	A	40%	29%	31%
A+	B	56%	24%	20%
A+	C	71%	18%	11%
A+	D	82%	12%	6%

Figura 4. Demonstração da tabela com valores de categoria relacionados às probabilidades de vitória de cada time. Fonte: [da Aposta, 2021].

Já a influência no mando de campo é representada pelos valores de 0 a 3, onde o usuário escolhe um valor mais alto para o time que julgar mais influente naquela partida. Na Figura 5 tem-se uma demonstração da fórmula utilizada no *Data Studio* para o cálculo das *odds*. Para o desenvolvimento do cálculo, foram levados em consideração os valores representados na tabela da Figura 4, referentes a cada combinação de categorias, além dos valores de influência no mando de campo.

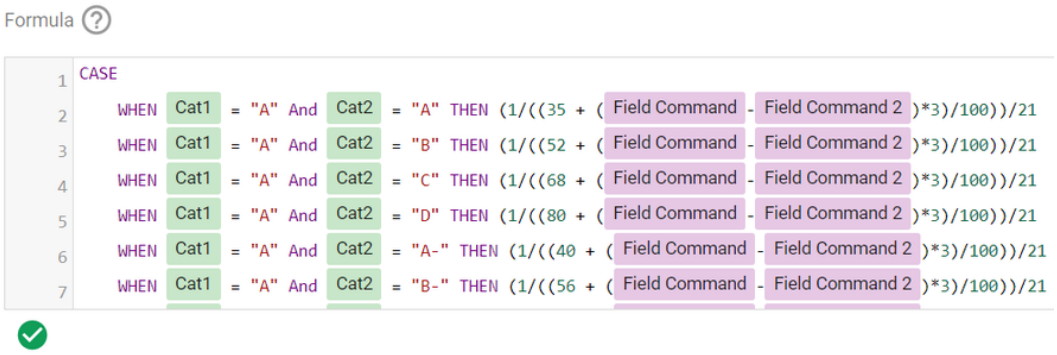

Figura 5. Demonstração da fórmula para cálculo das *odds*. Fonte: autora.

Além da utilização de arquivos no formato *.xlsx*, utilizados como principal base de dados, também foi utilizada a licença gratuita do banco de dados *Google BigQuery* [Google, 2010], como demonstrado nas Figuras 6 e 7, levando-se em consideração a facilidade de conexão com o *Data Studio* e as APIs (*Application Programming Interface*) existentes para manipulação dos dados utilizando scripts na linguagem de programação *Python*. Um dos motivos para utilização do banco de dados, além dos arquivos no formato *.xlsx,* foi a necessidade de recuperação dos dados na base para realizar manipulações e o desenvolvimento do algoritmo de aprendizado de máquina *Random Forest*, visando à criação de um modelo de regressão para realizar o cálculo de predições que tentam definir a quantidade de gols em uma determinada partida, que é um parâmetro de grande importância nas apostas de futebol.

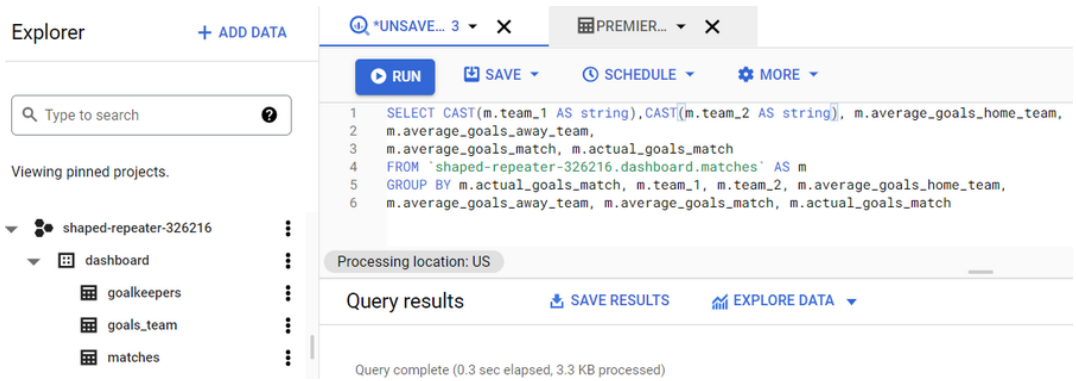

Figura 6. Interface do BigQuery. Fonte: autora.

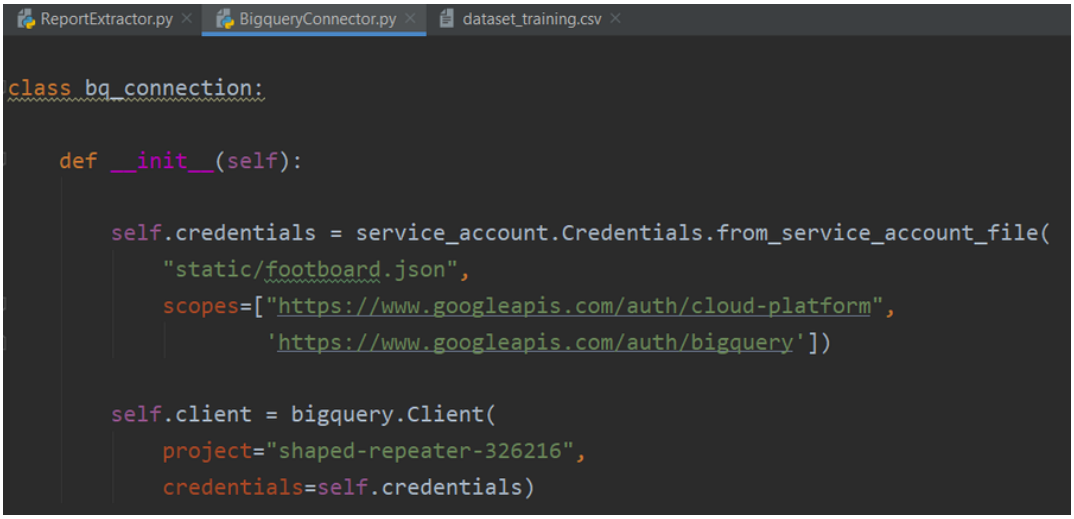

Figura 7. Classe de conexão com o *BigQuery*. Fonte: autora.

Após o *upload* dos arquivos *.csv* formatados, como foi descrito na subseção 4.2, além dos dados inseridos no *BigQuery* através dos scripts em *Python*, foi possível ter acesso aos *datasets* na interface do *Data Studio* e realizar a manipulação dos dados, envolvendo a criação de relatórios e gráficos capazes de transmitir facilmente as informações necessárias para análises, além de permitir que o usuário forneça dados de entrada que auxiliam no cálculo de variáveis dinamicamente.

Para o desenvolvimento do modelo de previsão de gols feitos em uma partida, foram utilizados alguns parâmetros como dados de treino:

- Time mandante;
- Time visitante;
- Média móvel de gols do time mandante quando está jogando em casa, considerando apenas os últimos cinco jogos do time;
- Média móvel de gols do time visitante quando está jogando fora, considerando apenas os últimos cinco jogos do time;
- Média móvel de gols previstos para a partida;
- Quantidade real de gols da partida entre os times informados.

Com os dados informados, o modelo tenta prever qual será a média de gols em uma partida específica. Visto que o algoritmo não aceita parâmetros no formato de texto, foi preciso utilizar uma função para convertê-los em valores binários. Esse processo de conversão dos valores é conhecido como *one-hot encoding,* e foi realizado através da função *get_dummies* da biblioteca de manipulação de arquivos *Pandas*. Na figura 8 temos um exemplo de saída dos dados após a realização do processo de *one-hot encoding*.

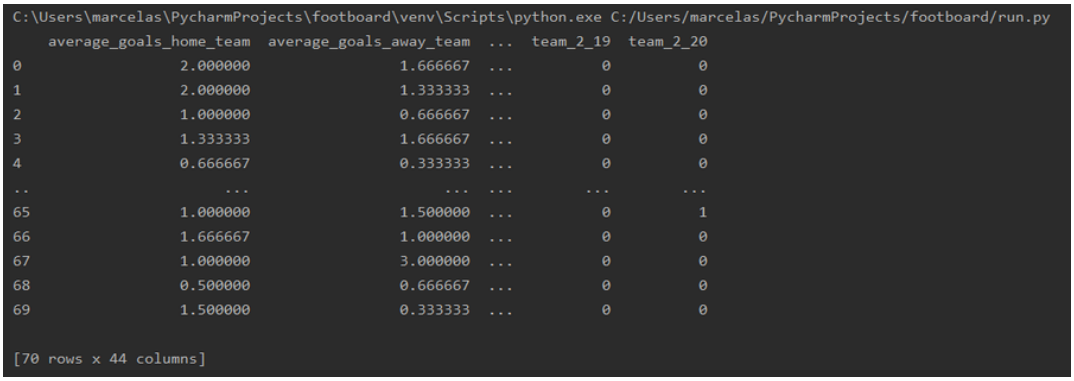

Figura 8. Demonstração do dataframe com os valores utilizados no algoritmo *Random Forest*. Fonte: autora.

Na Figura 8, é possível visualizar o *dataframe* com os parâmetros utilizados para a criação do modelo de regressão. Com a conversão dos valores no processo de *one-hot encoding*, para cada valor relacionado a um time, foram criadas colunas onde é atribuído o valor 1 caso o time da linha específica seja referente àquele determinado valor, caso contrário, o valor atribuído é 0.

5. Resultados

Esta seção está dividida em duas subseções referentes ao desenvolvimento do d*ashboard* e do algoritmo de previsão *Random Forest,* respectivamente.

5.1. Dashboard

A partir do processo de desenvolvimento descrito na seção anterior, foi possível implementar o *FootBoard*, que como explicado anteriormente, é um dos principais objetivos deste trabalho. Através do *dashboard* desenvolvido foi possível ter acesso aos diversos dados presentes na base de dados especificada na subseção 4.1.

A seguir, têm-se alguns exemplos das páginas criadas e de como funciona a dinâmica de cada uma delas no *FootBoard*. Ele está dividido em 5 seções (páginas):

- **Teams:** Mostra dados estatísticos de cada time participante do campeonato;
- **Players:** Mostra dados estatísticos de cada jogador do campeonato, tanto titulares quanto reservas.;
- **Goalkeepers:** Mostra dados estatísticos de cada goleiro do campeonato, tanto titulares quanto reservas;
- **Referees:** Mostra dados estatísticos de cada juiz participante do campeonato;
- **Matches:** Mostra dados estatísticos de todas as partidas da temporada de 2021 realizadas até o momento da criação do *dashboard*.

Em cada uma das seções, o usuário tem acesso às informações através de gráficos e relatórios, podendo filtrar as informações que mais lhe interessam e realizar comparações de maneira visualmente agradável.

5.1.1. Odds

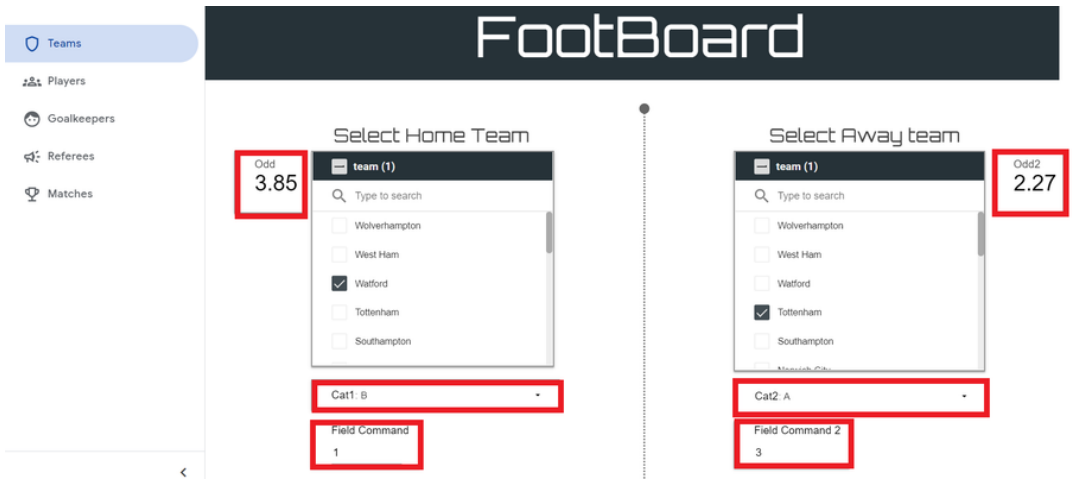

Figura 9. Demonstração da seção de times no *Footboard*. Fonte: autora.

Na Figura 9, pode-se visualizar uma demonstração da seção de times no *FootBoard*. As partes destacadas em vermelho são as variáveis referentes ao cálculo das *odds* de cada time. Para se calcular os valores dessas *odds*, deve-se selecionar uma categoria, de A+/A/A- a D+/D/D- para cada time, levando-se em consideração o desempenho deles, assim como informar um peso, de 1 a 3, para o mando de campo de cada um deles. Os valores de categoria e de mando de campo são utilizados como fatores de pesos diferentes ao se calcular uma *odd*, e se baseiam em tabelas de valores pré-definidos, onde valores de A a D se referem a uma escala de times muito bons a muito ruins, e os valores de 0 a 3 representam a influência do time no mando de campo, podendo ser nula, baixa, média ou alta. A partir deste cálculo podemos visualizar as *odds* referentes a cada time. Nota-se que o time com menor probabilidade de vencer o jogo, possui um valor de *odd* maior, que como explicado anteriormente, é o valor que será multiplicado pelo valor investido, caso o apostador vença a aposta. Deve-se salientar que os valores de categoria e mando de campo informados, não são atribuídos de fato ao time selecionado, servem apenas para facilitar a visualização e a análise do usuário.

5.1.2. Seções

Os times também podem ser filtrados e selecionados, fazendo com que apenas as informações referentes as opções selecionadas sejam mostradas nos gráficos e relatórios, como visto na Figura 10, que mostra a relação entre a quantidade de gols esperados por um determinado time, e a quantidade real de gols marcados por este mesmo time.

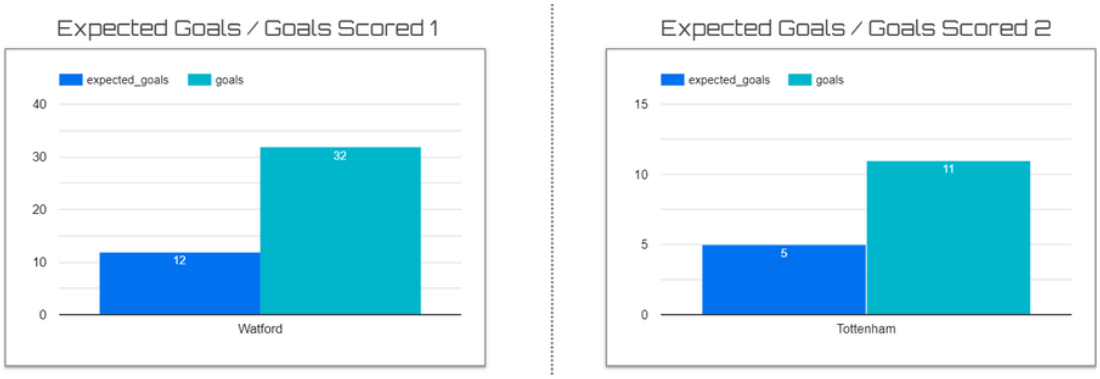

Figura 10. Demonstração dos gráficos que mostram a relação entre gols esperados/gols marcados de um determinado time. Fonte: autora.

A Figura 11 mostra a seção de jogadores no FootBoard. Nesta seção é possível filtrar e selecionar um determinado jogador e visualizar diversas informações referentes apenas a ele nos gráficos e relatórios. Além disso, ao selecionar um jogador, o filtro de times também é atualizado automaticamente, informando o time do qual o jogador faz parte, e vice-versa, ou seja, ao se selecionar um time, será possível visualizar apenas os jogadores referentes a este time.

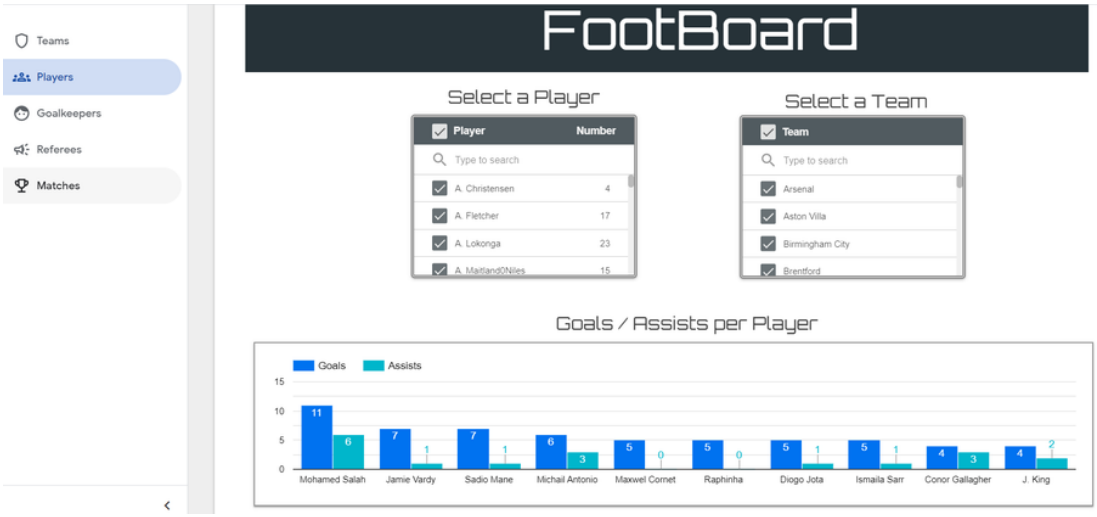

Figura 11. Demonstração da seção de jogadores. Fonte: autora.

Assim como na seção de jogadores, na página de goleiros, mostrada na Figura 12, conseguimos filtrar a opção desejada e visualizar o time do qual o goleiro selecionado faz parte. Da mesma forma, pode-se selecionar o time para obter a lista dos goleiros.

Figura 12. Demonstração da seção de goleiros. Fonte: autora.

Na seção de juízes podemos ter acesso à media da quantidade de faltas marcadas por um determinado juiz em uma partida, assim como a quantidade de cartões amarelos e vermelhos distribuídos por jogo, como mostra a Figura 13.

Figura 13. Demonstração da seção de juízes. Fonte: autora.

Na seção de partidas, demonstrada na Figura 14 é possível visualizar todos os jogos disputados, obtendo dados referentes a média de gols feitos de cada time, considerando os jogos como mandante e como visitante, assim como a média de gols esperados na partida e a quantidade real de gols da partida em questão. Também é possível filtrar o time mandante e o visitante e visualizar apenas os dados referentes a esta partida.

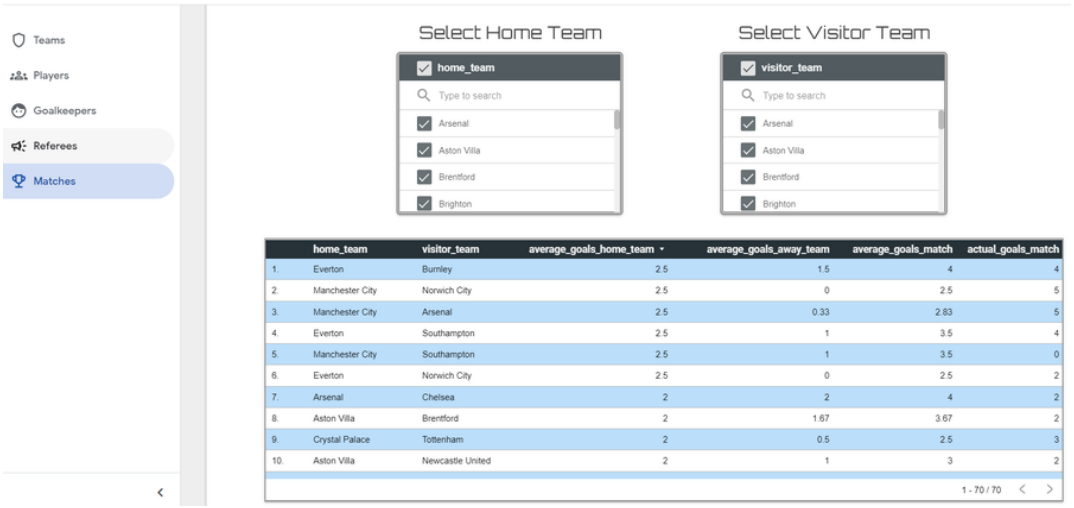

Figura 14. Demonstração da seção de partidas. Fonte: autora.

5.2. Previsão de Gols

Com relação ao modelo de regressão *Random Forest* gerado, o modelo tenta prever a quantidade de gols em uma partida através dos parâmetros de entrada informados. O modelo obteve um bom desempenho com os dados de treino, tendo um valor de erro de 0,48 e a acurácia em torno de 78%, como mostrado na Figura 15.

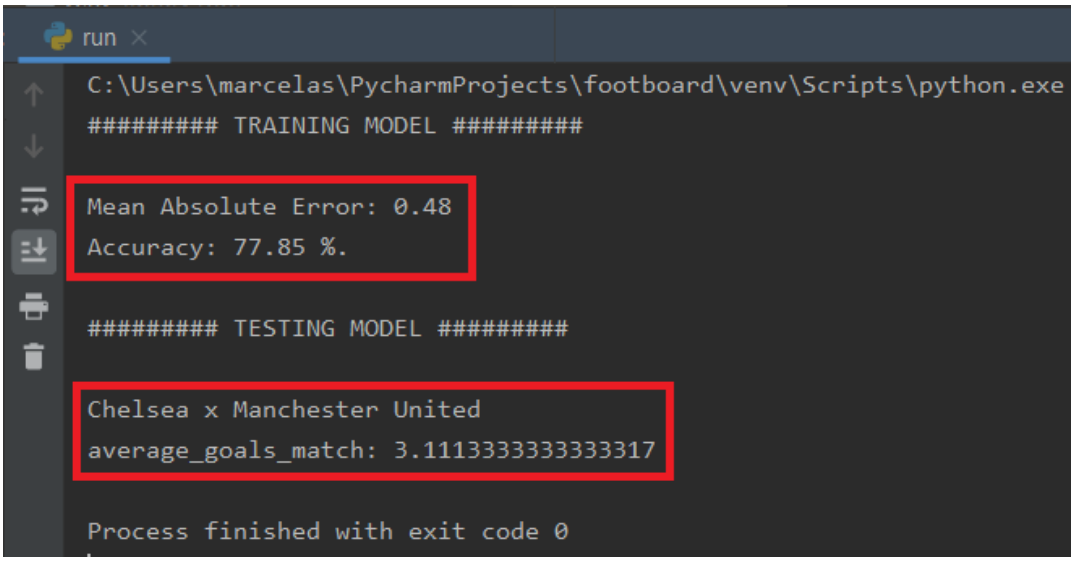

Figura 15. Demonstração de resultados de treino e teste do modelo *Random Forest*. Fonte: autora.

Os parâmetros escolhidos como dados de entrada informados pelo usuário para o modelo foram:

- Time mandante;
- Time visitante.

Além dos times informados, dois outros parâmetros são passados ao modelo: média de gols do time da casa e do time visitante, considerando os últimos cinco jogos de cada um. Através do nome de cada time informados, o algoritmo consegue encontrar os valores da quantidade média de gols tanto do time da casa, quanto do visitante nos últimos cinco jogos. Esses valores são então passados como parâmetros de entrada para a função de predição do modelo, junto com os nomes dos times. Com o modelo treinado, alguns testes foram realizados, fazendo comparações entre os valores previstos e os valores reais de cada partida como pode ser visto na Figura 16.

Home Team	Away Team	Actual Goals	Prediction	Error
Arsenal	Newcastle	2	2.86	-0.86
Liverpool	Southampton	4	3.04	0.96
Norwich	Woverhampton	0	2.6	-2.6
Crystal Palace	Aston Villa	3	2.65	0.35
Brighton	Leeds	0	1.32	-1.32
Brentford	Everton	1	2.49	-1.49
Manchester City	West Ham	3	2.9	0.1
Leicester	Watford	6	2.36	3.64
Chelsea	Manchester United	2	3.1	-1.1

Figura 16. Tabela comparativa entre quantidade real de gols em uma partida e os valores de predição pelo modelo gerado. Fonte: autora.

Como visto na tabela comparativa da Figura 16, a maioria dos valores previstos (78%) ficou dentro de uma boa margem de erro, ente 0 e 1.5 gols de diferença para o valor real da partida. Para se considerar um valor previsto com margem de erro entre 0 e 1.9 como resultado considerável, foi levado em consideração que alguns outros fatores como desfalques no ataque e na zaga de cada time, por exemplo, podem influenciar diretamente na quantidade de gols feitos em uma partida, tanto para mais quanto para menos. Alguns outros fatores como arbitragem e importância do jogo para cada time também devem ser considerados pelo apostador ao fazer sua análise. Ou seja, por mais que os resultados do modelo consigam chegar perto dos valores reais, muitas vezes, o que vai definir a escolha do apostador, é a junção dessas previsões com a conclusão de toda a análise feita por ele, considerando todos os fatores envolvidos na marcação de gols.

A análise feita considera as apostas onde o apostador tenta descobrir a quantidade exata de gols em uma partida. Porém, também existe outra modalidade, chamada linha de gols, onde é possível prever se a partida terá menos ou mais gols do que o valor especificado pela casa de apostas. Por exemplo, se a linha de gols for de 2.5 gols, o apostador pode decidir se deve apostar que naquela partida específica terá uma quantidade de gols abaixo ou acima de 2.5. Para esse tipo de aposta, deve-se levar em consideração o valor da linha de gol utilizado para comparação dos resultados.

6. Conclusão e Trabalhos Futuros

Este trabalho tem como principal objetivo o desenvolvimento de um *dashboard* com dados estatísticos relacionados a campeonatos de futebol, possibilitando ao usuário realizar análises de maneira personalizada, além da criação de um modelo que prevê a quantidade de gols em uma determinada partida, através de técnicas de aprendizado de máquina e o cálculo de probabilidades de vitória de um time em uma determinada partida.

Através do que foi relatado na seção de resultados, pôde-se concluir que o uso de *Business Intelligence* na área de apostas esportivas é de grande relevância, visto que após todo o processo de coleta, formatação e processamento dos dados, foi possível demonstrar uma interface de fácil entendimento e interação, facilitando o desenvolvimento de análises e substituindo métodos com menos usabilidade, como planilhas Excel, por exemplo, além de ser de grande ajuda na tomada de decisões. Apesar dos fatores positivos do *Data Studio*, notou-se durante a implementação do *dashboard*, uma necessidade de realizar fórmulas mais complexas e menos engessadas para o cálculo de variáveis dinâmicas, visto que a ferramenta oferece limitações no desenvolvimento de campos calculados. Para suprir tais limitações, o desenvolvimento de cálculos de algumas variáveis foram realizados através de scripts na linguagem *Python*.

O modelo gerado para predição da linha de gols em uma partida, utilizando o algoritmo *Random Forest*, também apresentou resultados satisfatórios, obtendo um bom desempenho, tendo sua acurácia em torno de 78%. Além disso, o cálculo das chances de vitória de um time em um jogo específico, as *odds*, se aproximaram bastante dos valores reais após a realização de comparações.

Com relação às melhorias e trabalhos futuros, alguns pontos a serem implementados de modo a complementar o uso do *Footboard* se destacam:

- Ao invés de utilizar planilhas como base principal, realizar toda a extração de dados e cálculo de variáveis necessárias através de scripts, podendo inserir os dados finais no *BigQuery* e acessá-los no Data Studio.
- Implementação de API ligada a casas de apostas esportivas, fazendo com seja possível realizar a comparação das *odds* calculadas pelo apostador e as *odds* calculadas pela casa de apostas.
- Implementação de código para realizar a previsão da variação dos valores das *odds* de cada time de uma aposta em tempo real.
- Automatização do processo de *upload* dos dados ao *Data Studio*, visto que acontece ao menos uma rodada de jogos por semana em um campeonato de futebol, o ideal é que o *dashboard* seja alimentado com novos dados semanalmente.

Referências

Appelbaum, J. (2019). *The Everything Guide to Sports Betting: From Pro Football to College Basketball, Systems and Strategies for Winning Money*. Everything.

Beal, R., Middleton, S. E., Norman, T. J., e Ramchurn, S. D. (2021). Combining Machine Learning and Human Experts to Predict Match Outcomes in Football: A Baseline Model.

Carvalho, A. C. P., Faceli, K., Lorena, A. C., Gama, J., e Almeida, T. A. (2021). *Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina*. LTC.

da Aposta, C. (2021). Tabela de Categorias de Desempenho dos Times.

da Costa, Í. B., Pires, C. E. S., e Marinho, L. B. (2017). Sports Analytics: Mudando o Jogo. *sol.sbc.org.br*.

de Almeida, M. B. e da Cunha, M. J. (2015). *Máquina de Vetores de Suporte (Support Vector Machines): Uma Introdução*. eBook Kindle.

Diniz, E. S. e Thiele, J. (2021). *Modelos De Regressão Em R*. Clube de Autores.

Feitosa, L. (2020). Mercado de Apostas em Alta no Brasil.

Gama, J. (2010). *Knowledge Discovery from Data Streams*. CRC Press.

Goddard, J. e Asimakopoulos, I. (2004). Forecasting Football Results and the Efficiency of Fixed-odds Betting.

Google (2010). Google BigQuery.

Google (2016). Google Data Studio.

Hartshorn, S. (2016). *Machine Learning With Random Forests And Decision Trees: A Visual Guide For Beginners*. Ebook Kindle.

Hilgler, M. (2013). *Texas Hold'em. Odds e Probabilidades. Estratégias de Limit, No-Limit e Torneios*. Raise.

Ivanskiy, A. (2007). InStat. Web Platform.

Mattera, R. (2021). Forecasting Binary Outcomes in Soccer.

Moura, K. (2019). Processo KDD.

Mueller, J. P. e Massaron, L. (2019). *Aprendizado de máquina para leigos*. Alta Books.

Nielsen, A. (2021). *Análise Prática de Séries Temporais: Predição com Estatística e Aprendizado de Máquina*. Alta Books.

Ross, S., Conti, A. R. D., e Pertence Júnior, A. (2010). *Probabilidade: Um Curso Moderno com Aplicações*. Bookman, 8ª edition.

Schapire, R. E. e Freund, Y. (2014). *Boosting: Foundations and Algorithms (Adaptive Computation and Machine Learning series)*. MIT Press.

Stübinger, J., Mangold, B., e Knoll, J. (2019). Machine Learning in Football Betting: Prediction of Match Results Based on Player Characteristics. *Applied Sciences*.

Tableau (2021). Benchmarking.

Temer, M., Guardia, E. R., Conalgo Junior, E. P., de Almeida Pedrozo, C. M. M., da Silva, L. C. F., Araújo Junior, J., e Jungmann, R. (2018). Lei 13.756/18.

Turban, E., Sharda, R., Aronson, J. E., e King, D. (2009). *Business Intelligence: Um Enfoque Gerencial para a Inteligência do Negócio*. Bookman.