UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E COMPUTAÇÃO CIENTÍFICA
DEPARTAMENTO DE MATEMÁTICA APLICADA

Carlos Danilo Tomé

Predição de resultados esportivos utilizando Aprendizado de Máquina

Campinas
28/08/2020

Carlos Danilo Tomé

Predição de resultados esportivos utilizando Aprendizado de Máquina

Monografia apresentada ao Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas como parte dos requisitos para obtenção de créditos na disciplina Projeto Supervisionado, sob a orientação do(a) Prof. Dr. Laércio Luis Vendite.

Resumo

Neste projeto foi utilizado técnicas de aprendizado de máquina para predizer resultados esportivos em que a classe alvo da predição é o resultado da partida para o time mandante assumindo então, três tipos de saídas para o modelo vitória, empate ou derrota para o time mandante. Os algoritmos utilizados são da classe multiclass classification, serão testados quatro modelos diferentes nativos da biblioteca scikit-learn distribuída na linguagem Python.

Abstract

In this project, machine learning techniques were used to predict sports results in which the target class of the prediction is the result of the match for the home team, assuming then three types of outlets for the winning, draw or defeat model for the home team. The algorithms used are from the multiclass classification class, four different native models from the scikit-learn library distributed in the 'Python' language will be tested.

Conteúdo

1 Introdução

Os esportes coletivos são conhecidos pelo alto grau de imprevisibilidade até mesmo entre o campo dos especialistas existem posições totalmente divergentes quando o assunto é predizer, isto é palpitar, sobre o resultado de uma partida desses esportes. A FIFA realizou em 2017 uma pesquisa com todas as 211 associações que possuem vinculo com a instituição, e neste censo foi constatados que, se contabilizados praticantes, atletas e técnicos, cerca de 265 milhões de pessoas são praticantes regulares deste esporte, um dos mais populares do planeta. FIFA [2017]

Esse grande interesse no esporte é ainda acentuado devido as rivalidades regionais e internacionais que são construídas através da história e que adicionam à crônica do esporte mais emoção, todo esse contexto faz com que o futebol seja campo para grandes movimentações financeiras, seja por meio de contratos de patrocinadores, direitos de imagem, transferência de jogadores e também pela apostas esportivas. O interesse constante neste esporte que faz parte importante de diversos países ao redor do globo aumenta a relevância, seja para analise jornalistica ou apostas, da capacidade de se prever resultados esportivos.

Sendo assim a intenção deste projeto é medir através da métrica de precisão qual modelo tem melhor performance para problemas de classificação múltipla, em que a classe alvo do nosso problema é descobrir o resultado da partida para o clube mandante.

Neste projeto serão utilizados algoritmos de aprendizado de máquina implementados na linguagem python, com o suporte das bibliotecas scikit-leark e pandas, respectivamente Pandas [2020] e Scikit-Learn [2020], ambas bibliotecas open-source amplamente utilizadas para este tipo de projeto.

2 Desenvolvimento

No espectro de estudo de algoritmo de machine learning existem diversos problemas que podem ser resolvidos através de aprendizado supervisionado, ou seja, é fornecido ao modelo dados históricos e a classe alvo (resposta do modelo) para estes eventos, no caso do estudo deste projeto a classe alvo é o resultado da partida construído a partir do placar da partida (evento estudado). Essa construção é feita da seguinte forma se a quantidade de gols do time mandante for maior que a quantidade de gols do time visitante então a classe alvo recebe vitória, para o caso em que essas features são iguais a classe alvo recebe empate e, por fim, caso não se encaixe nas duas hipóteses anteriores é recebido o valor derrota para a classe alvo.

Para os modelos identificados que se encaixam na resolução deste problema a distribuição do scikit-learn adota algumas premissas que maximizam a capacidade dos modelos em generalizar as situações para as classes alvo sem que haja um problemas comuns de problemas de aprendizado de maquina como, por exemplo, overfitting caso em que o modelo se adapta muito bem aos dados de treino mas tem dificuldades em acertar dados que não estão na base de dados de treino, estas premissas adotam algumas padronizações nos dados de entrada em que as principais são: os dados estejam padronizados entre 0 e 1 (para que não haja peso maior em atributos que possuem maior módulo) e que exista um balanceamento das classes alvo, ou seja, para que exista uma distribuição parecida para a quantidade de respostas da classe alvo dos dados de entrada do modelo (os dados de entrada possuam uma quantidade parecida de eventos para vitória, empate e derrota).

Dentro das boas práticas de um projeto de aprendizado de máquina existem etapas pré estabelecidas que devem ser observadas antes de efetivamente treinar os modelos, inicialmente uma analise exploratória da base de dados que o objetivo é entender melhor as especificidades da base de dados, dos atributos e suas características como formato entrada se elas são do tipo inteiro, racional ou string. Posteriormente é feito o pré-processamento dos dados onde o objetivo é padronizar os dados de entrada para que se encaixem nas premissas adotadas nos modelos, também é feito a criação de novos atributos para os eventos já que existem informações pertinentes ao problema em que não

estão presentes na base de dados retiradas da fonte de dados. Por fim, é feito o treinamento do modelo onde são testados diversos hiper-parâmetros e são selecionados através do algoritmo grid-search os melhores destes parâmetros que generalizam as respostas do modelo e possuam a maior pontuação de acerto para a base de dados de teste e por fim é feita a avaliação dos modelos através de destes para diversas bases de dados e situações afim de entender qual foi a melhor performance.

2.1 Base de Dados

Neste projeto foi utilizado a base de dados "European Soccer Database"Mathien [2016] disponível no repositório aberto da plataforma de competições de aprendizado de máquina Kaggle, neste repositório o autor compila dados de três fontes distintas.

- **EnetScore [Enetscores, 2016]** : Neste repositório são retirados via API os placares das partidas, formação inicial(jogadores que iniciaram a partida), formação final (jogadores que final a partida) e dados de contexto da partida como data, time mandante e etc.

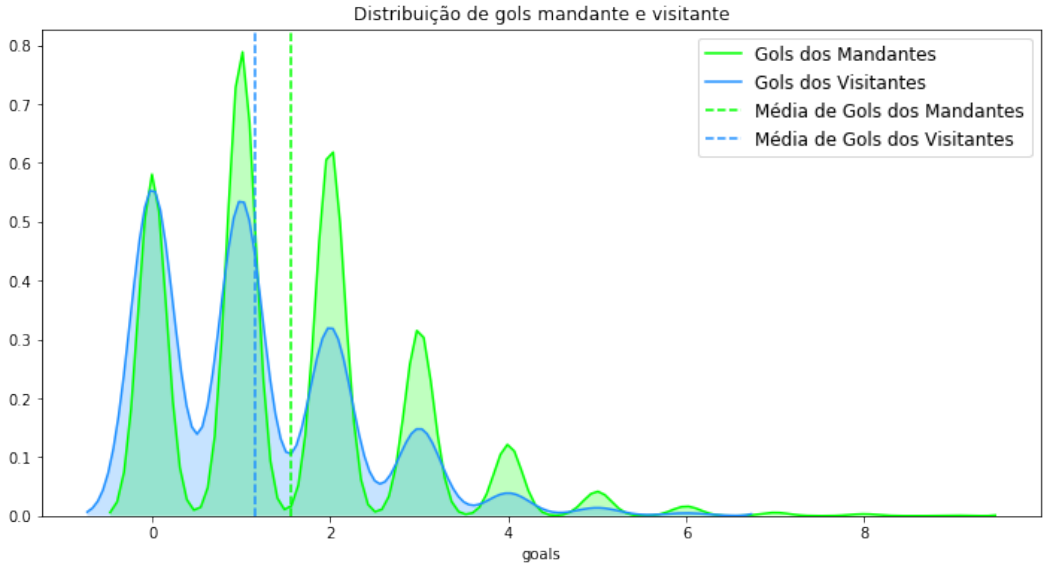
- **Football Data [Data, 2016]** : Neste repositório são retirados via API dados das probabilidades de aposta para 10 plataformas principais de apostas no Reino Unido, essas probabilidades estão distintas por resultado, ou seja, probabilidade de vitoria do mandante, empate e derrota para todas as 10 plataformas de apostas. Estes dados são extremamente relevantes para o modelo pois adicionam à base de dados a analise técnica de casas de apostas que definem essas probabilidades com base em modelos de predição somados à analise de especialistas no esporte. Como o objetivo é predizer eventos que contam com um alto grau de imprevisibilidade e também é razoável partir da premissa que existem variáveis importantes envolvidas numa partida que não o scout e dados históricos.

- **Fifa Data [Sofifa, 2016]** : Neste repositório o autor retira dados do jogo "FIFA", nestes dados estão incluídos overall dos jogadores e times. Estes dados não serão utilizados durante a modelagem.

Após a retirada de dados das fontes listadas o próximo passo realizado foi um estudo minucioso dos atributos que serão utilizados nesse modelo, aqui entenda-se os atributos como dados da partida onde essas características possuam algum tipo de relação com a classe alvo do problema, isto é, quais atributos dos eventos influenciam no placar

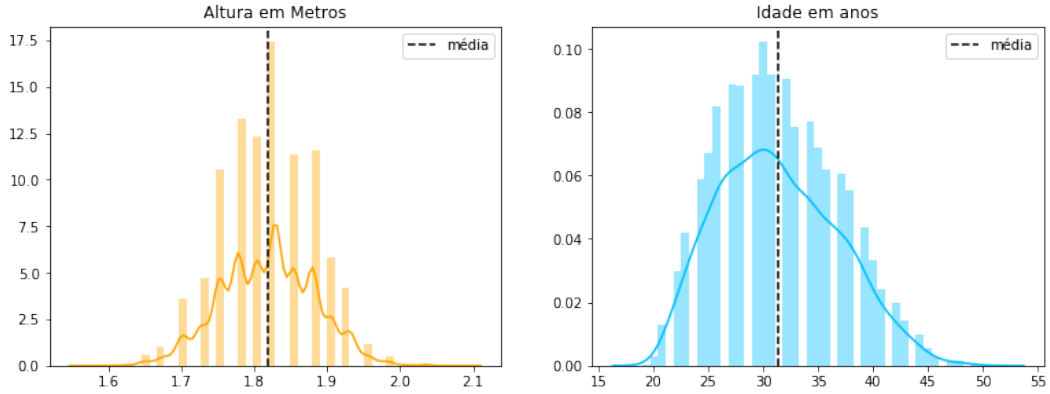
da partida. Como o problema estudado neste projeto possui um alto grau de imprevisibilidade inerentes de esportes coletivos, esta etapa é extremamente importante e exige um conhecimento prévio do modelador para que haja uma seleção adequada de atributos. Dentro da base de dados das partidas podemos apontar alguns eixos de informações que simplificam o estudo e a analise dos atributos, os eixos são: dados da partida, dados dos jogadores e técnicos e dados das casas de apostas. O eixo de atributos da casa de apostas visa diminuir os efeitos que são extra-campo e que influenciam durante a partida como momento politico dos clubes, tradição, momento físico e tático dos jogadores pois as probabilidades são construídas através da soma de atributos históricos e também da analise técnica de especialistas do esporte.

- **Dados da partida**: Dentro desse eixo de informações estão dispostas informações da partida que podem influenciar no momento técnico e físico dos jogadores, uma das variáveis mais importantes aqui é o time que esta como mandante nesta partida sabendo que já foram feitos diversos trabalhos na literatura em que é apontado uma vantagem inerente na possibilidade de vitória para o time mandante conforme apontado em MAHER [1982], outros fatores estudados são importantes e envolvem variáveis da quantidade e assiduidade de jogos, dado que o acumulo de jogos pode influenciar no físico dos jogadores e também a falta de jogos acarreta na falta de ritmo de jogo dos jogadores e configuram-se, portanto, variáveis importantes no que tange o resultado final da partida.



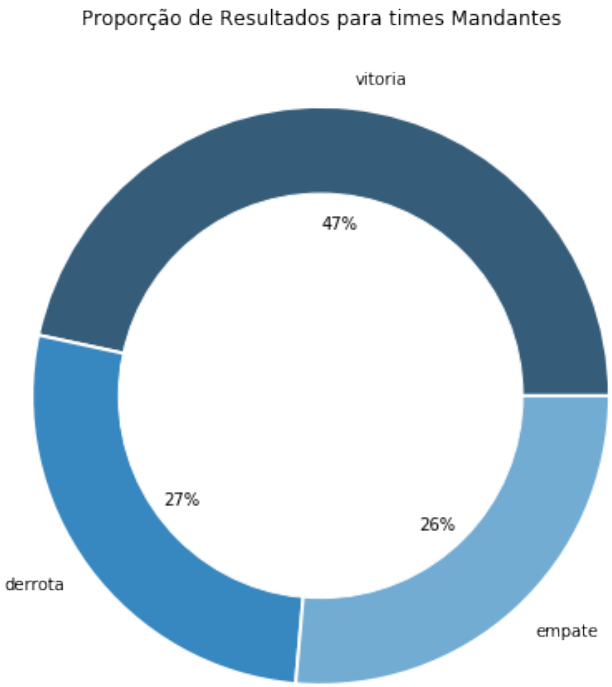
- **Dados dos jogadores**: Dentro desse eixo são fornecidas aos modelos informações sobre os jogadores, essas informações em sua maioria são scouts dos jogadores

e treinadores dos clubes envolvidos na partida, o objetivo desse eixo é generalizar a qualidade técnica dos jogadores envolvidos, momento técnico dos jogadores dado que em boa forma técnica tendem a influenciar mais no resultado final da partida e também do momento psicológico dos jogadores através de informações das últimas partidas dos jogadores, essa informação é importante assim que assumido que jogadores que fizeram gols ou tiveram bons desempenhos em partidas anteriores tendem a influenciar mais no resultado da partida.



- **Dados das casas de apostas**: Dentro desse eixo de estudo é fornecido aos modelos dados das odds de aposta, isto é, as probabilidade de vitoria, empate e derrota fornecidas pelas casas de aposta aos apostadores; estas informações são importantes pois são construídas através de analise técnica de especialistas do esporte em conjunto com uma analise histórica do confronto. Estas probabilidades influenciam na quantidade retornada caso a aposta esteja correta, isto é, se as casas de apostas estão apontando maior probabilidade de vitória para o time mandante da partida.

Uma característica muito importante para implantar algoritmos de machine learning de classificação múltipla em grandes bases de dados é que os rótulos da classe alvo estejam balanceados, isto é, que dentro do conjunto de dados possua uma distribuição balanceada da quantidade de vitórias de times mandantes, empates e derrotas. Para resolver este problema durante a etapa de divisão da base de dados em treino e teste é escolhido o atributo "labels"como o divisor, com este atributo setado entre os três rótulos possíveis o algoritmo de cross validation tenta balancear ao máximo possível os dados entre os rótulos da classe alvo. Segue abaixo o balanceamento da base de dados completa com os 25 mil jogos.



No total serão utilizadas 122 atributos para este projeto para aproximadamente 25 mil partidas de ligas de pontos corridos europeias. Este tamanho de base de dados pode diminuir a velocidade dos modelos bem como atribuir alguns atributos que não são capazes de generalizar a influencia no resultado da partida, portanto, para fim de redução de dimensionalidade será utilizado durante a etapa de Pré Processamento o algoritmo PCA (Principal Component Analysis) que mantém na base de dados atributos que são mais importantes na generalização dos eventos.

2.2 Pré-Processamento

Durante a etapa de pré processamento foi feita diversas manipulações da base de dados que estão inicialmente dispostos para compreensão humana e, portanto, devem estar alinhados com as premissas adotadas nos modelos e também coloca-los em formato que a máquina consiga interpretar os dados.

O primeiro objetivo de analise dentro dessa etapa são dados nulos dentro da base de dados, sabendo que ao retirar dados via API das fontes de dados podem vir partidas com dados faltantes por algum motivo extraordinário, existem diversas abordagens para resolver este problema, dado que os modelos adotam a premissa que não hajam dentro do conjunto de treino e teste atributos vazios, a escolha adotada neste projeto foi retirar da analise partidas com dados faltantes; essa escolha é embasada no fato da base de dados possuir poucos dados faltantes no total 2.55 porcento e, portanto, esta abordagem não vai diminuir significativamente a quantidade de observações fornecidas.

Foram criadas ainda alguns outros atributos que não estão inicialmente dispostos nas bases de dados retiradas das fontes referenciadas, estes atributos criados estão em alinhamento com premissas já abordadas na seção de desenvolvimento deste projeto. Os atributos criados foram:

- *saldo-gols-mandante*: Neste atributo foi construído a informação de saldo de gols histórico do clube mandante em todas as temporadas estudadas até o jogo do evento. Este atributo adiciona um peso para clubes que tem historicamente tem melhores desempenhos.

- *saldo-gols-visitante*: Neste atributo foi construído a informação de saldo de gols histórico do clube visitante em todas as temporadas estudadas até o jogo do evento. Este atributo adiciona um peso para clubes que tem historicamente tem melhores desempenhos.

- *confronto-direto-vitoria*: Este atributo adiciona o histórico do confronto direto entre os dois clubes da partida, o valor é calculado através da soma de vitórias do clube mandante nos últimos N=10 jogos do confronto direto até a data da partida observada.Esta informação é importante para entender possíveis vantagens de estilo de jogo ou de jogadores que o time mandante tem sobre o clube visitante.

- *confronto-direto-derrota*:Este atributo adiciona o histórico do confronto direto

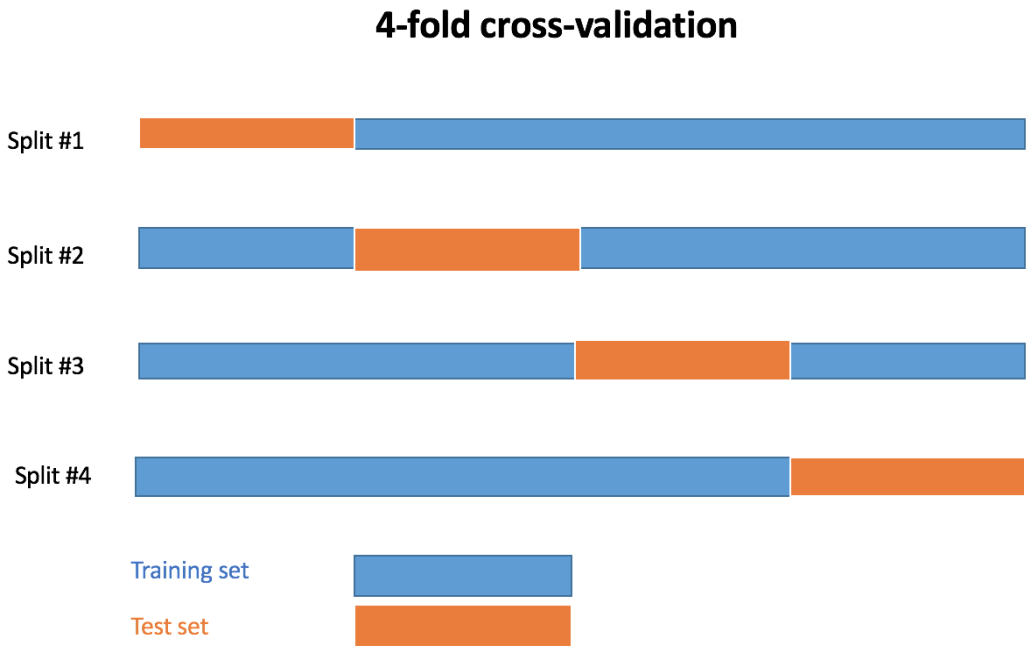
entre os dois clubes da partida, o valor é calculado através da soma de vitórias do clube mandante nos últimos N=10 jogos do confronto direto até a data da partida observada. Esta informação é importante para entender possíveis vantagens de estilo de jogo ou de jogadores que o time visitante tem sobre o clube mandante.

Após a criação de novos atributos é feito a transformação de atributos categóricos em atributos numéricos, outra premissa adotada nos modelos é que todos os atributos de entrada dos modelos são numéricos por que a máquina não tem capacidade de distinguir a diferença entre os atributos categóricos e como eles influenciam na generalização do problema. Para esta etapa foi utilizado o algoritmo Get Dummies, algoritmo nativo da biblioteca Pandas da linguagem Python. O Get Dummies consiste em transformar os atributos categóricos em novas colunas no conjunto de dados onde o valor do atributo é construído através da aderência da partida se comparada à aquele atributo, um exemplo simples é os times envolvidos, serão criadas através desse algoritmo colunas com todos os times da base de dados, caso o time daquela coluna esteja envolvido na partida ele receberá o valor 1 caso contrario 0.

Depois da etapa de transformação de atributos categóricos em numéricos é acrescentado à base de dados muitos atributos e o alto valor de atributos fornecidos nos dados de entrada pode comprometer o tempo de processamento dos modelos e assim diminuir significativamente a capacidade de realizar testes e modificar parâmetros. Para resolver este problema foi utilizado um algoritmo de redução de dimensionalidade chamado PCA, ou Principal Component Analysis, que mantém na base de dados apenas os atributos que possuem maior capacidade de generalizar os dados do problema.

Por fim, a última etapa do pré-processamento dos dados é a separação da base de dados em conjuntos de treino, que serão utilizados para treinar os modelos sugeridos, e de teste, conjunto que tem como função realizar testes dos modelos afim testar os hiperparâmetros e também para mensurar a performance de acerto dos modelos. Utilizar apenas uma divisão não aleatória dos dados de treino e teste pode causar overfitting, em que o modelo se adapta muito bem somente ao conjunto de treino e não generaliza bem o problema e, portanto, não interessa para a predição de partidas que ainda não ocorreram ainda. Para resolver este problema foi escolhido neste projeto adotar a abordagem através do algoritmo de Cross-Validation distribuído também pela biblioteca scikit-learn, que

consiste em realizar a divisão da base de dados de treino e teste em n (atributo n-folds do algoritmo) divisões aleatórias e rodar o modelo de treinamento e teste para cada uma dessas divisões e adotar como avaliação do modelo a média obtida de performance para estas n divisões e iterações utilizadas.



3 Modelagem e Resultados

Neste projeto o objetivo é predizer resultado de partidas de futebol baseado em dados históricos onde a classe alvo tem cardinalidade maior que um, configurando-se um problema clássico de Multiclass Classification onde os modelos podem retornar três tipos de resposta (vitória, empate e derrota). Para este problema iremos testar quatro tipos de algoritmos de aprendizado de máquina, alguns destes propostos por Airback [2016] ,distribuídos pelo pacote scikit-learn são eles: Regressão Logística, Classificador Random Forest, Naive Bayes Gaussiano e Classificador K-Nearest Neighbors.

Para identificar os melhores parâmetros que podem ser utilizados para generalizar o problema a biblioteca scikit learn também contém uma função chamada Grid Search, em que é apontado listas de hiper-parâmetros distintos para cada uma das variáveis de entrada do modelo desenhado pelo pacote, esta função realizada todas as combinações possíveis de hiper-parâmetros e retorna a combinação na qual o modelo configurado retorna a melhor performance possível.

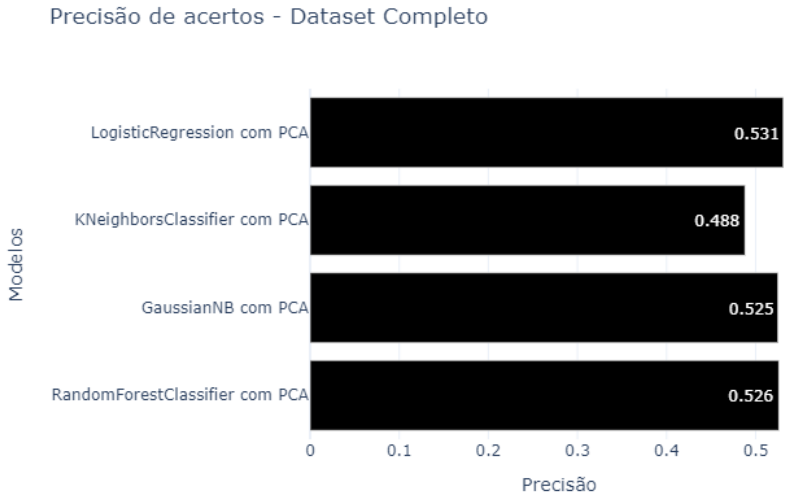
Para testar a performance dos modelos adotados foi feita simulações com bases de dados de treino e teste distintas para que ao final do projeto seja escolhido o melhor modelo e os melhores hiper-parâmetros que generalizam bem o problema envolvido no problema de classificação multilabel.

Para conseguirmos comparar a performance do modelo em questão de ganho de precisão neste projeto foi escolhido definir um baseline ou modelo de base, para que baseado em um modelo extremamente simples tenha-se uma perspectiva real do ganho de informação com a utilização de técnicas de machine learning, dado que este problema envolve um esporte coletivo altamente imprevisível, um modelo simples é sempre apostar que o time mandante irá ganhar a partida. Portanto, estamos interessados em modelos que possam atingir uma precisão maior que o baseline dado que se a performance for menor o algoritmo não traz ganho de informação na predição dos resultados, sendo assim o baseline é importante para adotar uma linha de corte dos modelos.

Base Completa

A primeira base de dados utilizada foi a totalidade de todos os 25 mil jogos da base de dados, quem compreendem jogos dos campeonatos de pontos corridos, as ligas,

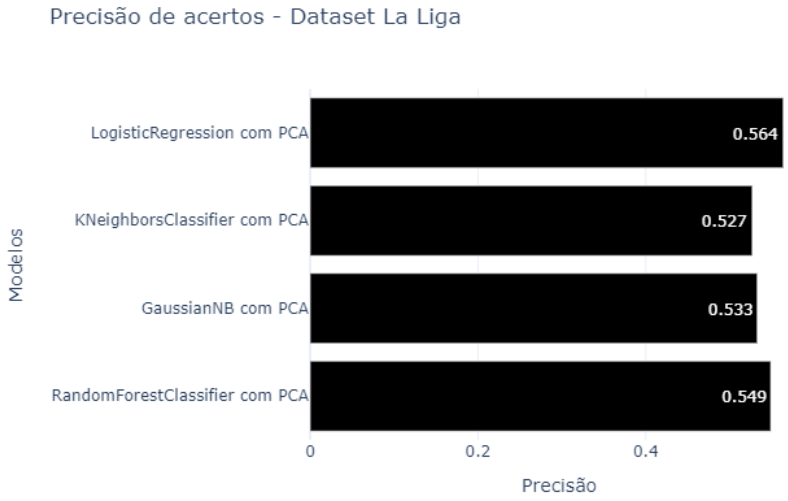
de 11 países diferentes retirando somente as partidas com dados nulos já apontadas na seção de pré-processamento. As temporadas envolvidas nessa base de dados são partidas da temporada 2008/2009 até 2015/2016 na totalidade das partidas dessas temporadas.



Premier League A Premier League é a liga nacional do Reino Unido, apontada como a liga mais disputada e importante do planeta a Premier League é a liga esportiva mais assistida do mundo, transmitida em 212 territórios para 643 milhões de lares e uma audiência potencial de TV de 4,7 bilhões de pessoas Ebner [2013]. Foram realizados treino e teste em todas as partidas da Premier League da temporada 2008/2009 até 2015/2016.

LIGA BBVA : A LIGA BBVA ou comumente chamada de La Liga é a liga

nacional da Espanha, organizada pela Liga Nacional de Fútbol Profesional, essa liga é muito importante no contexto mundial, um exemplo disso é que os últimos 10 prêmios de melhor jogador (The Best FIFA Football Awards) do ano concedido pela FIFA foi de jogadores que atuam nesta liga FIFA [2020] . Foram realizados treino e teste em todas as partidas da LIGA BBVA da temporada 2008/2009 até 2015/2016.



Serie A : A Serie A é a liga nacional da Itália,organizada pela , essa liga é muito relevante no contexto mundial, principalmente pelo contexto histórico da liga que sempre teve grandes times na sua liga, atualmente a sua importância esta mais ligada à tradição do que propriamente a qualidade dos clubes atuais. Foram realizados treino e teste em todas as partidas da Serie A da temporada 2008/2009 até 2015/2016.

4 Conclusão

Para concluir este projeto conseguiu atingir métricas de precisão superiores ao baseline construído através da intuição de que clubes mandantes tendem a vencer as partidas, utilizando mais de 25 mil partidas dentro do contexto do futebol europeu os modelos propostos por Airback [2016] atingiram as seguintes precisões para os resultados das 5 mil partidas incluídas no conjunto de teste:

Tabela de Precisão dos Modelos :

Modelos	Dados Completos	Premier League	La Liga	Serie A
Baseline	46	46	49	47
LogisticRegression	53.1	53.4	56.4	53.7
RandomForestClassifier	52.6	51.4	54.9	52.1
GaussianNB	52.5	50.9	53.2	52.8
KNeighborsClassifier	48.8	47.5	52.7	47.8

Sendo assim o modelo com maior precisão dentre os testados foi a Regressão Logística atingindo 53.1 porcento de precisão nos palpites gerados dentro do conjunto de dados completo, este valor esta 8 porcento acima do modelo de base demonstrando uma performance positiva para as ligas nacionais europeias. Para as demais ligas testadas o algoritmo Gaussian Naive Bayes atingiu métricas positivas também, superando o baseline em média de 6 porcento e para suavizar os efeitos inerentes aos estilos de jogos dos clubes, que são distintos entre as ligas, foi testado também neste projeto os modelos para as três maiores ligas do futebol europeu, dentro desse contexto é possível interpretar que a La Liga é o campeonato mais previsível dentre estes, pois os modelos em geral tiveram uma assertividade muito maior para essa liga, uma das hipóteses é que existe uma disparidade econômica gigantesca dos clubes na La Liga e consequentemente os clubes com mais recursos financeiros tem maior capacidade para investir em bons jogadores, com elencos mais fortes os clubes top da La Liga tendem a vencer mais, afirmação que esta completamente alinhada com os resultados obtidos.

Referências

Airback. Match outcome prediction in football. "https://www.kaggle.com/airback/match-outcome-prediction-in-football", 2016.

Football Data. Data matches and bookmaker. "http://www.football-data.co.uk/", 2016. Accessed: 2018-12-06.

Sarah Ebner. "history and time are key to power of football", says premier league chief. *The Times*, 2013. Accessed: 10.08.2020.

Enetscores. Football data api. "http://football-data.mx-api.enetscores.com/", 2016. Accessed: 2018-12-06.

FIFA. Fifa activity report 2017. "https://resources.fifa.com/image/upload/big-count-stats-
2017. Accessed: 2020-08-06.

FIFA. The best fifa football awards. "https://www.fifa.com/the-best-fifa-football-awards/",
2020. Accessed: 2020-08-06.

M. J. MAHER. Modelling association football scores. *Statistica Neerlandica*, 36(3):109–118, 1982.

Hugo Mathien. European soccer database. "https://www.kaggle.com/hugomathien/soccer", 2016.

Pandas. Pandas pydata. "https://pandas.pydata.org/docs/user$_guide/index.html$", 2020.

Scikit-Learn. Scikit-learn. "https://scikit-learn.org/stable/", 2020.

Sofifa. So fifa player data. "https://sofifa.com/", 2016. Accessed: 2018-12-06.