

# Utilização de algoritmos de inteligência artificial na previsão de resultados de partidas de futebol

Olesio Gardenghi Neto<sup>1</sup>; Rodrigo Grassi Martins<sup>2</sup>

**Resumo:** O futebol é uma modalidade esportiva disputada por duas equipes, cujo objetivo é transpor uma bola entre as extremidades – chamadas de baliza - utilizando basicamente toques com o pé. O vencedor da partida será o time que conseguir atingir o objetivo mais vezes no jogo. O interesse de tantas pessoas no mundo pelo futebol gera não apenas telespectadores, mas também muitas movimentações financeiras em torno desse esporte. Sistemas computacionais que trabalham com a previsão de resultados e que auxiliam a minimizar os riscos e maximizar os lucros tornam-se então uma importante ferramenta de trabalho para o dia a dia do futebol. O objetivo deste projeto é testar e comparar o desempenho dos classificadores utilizados na previsão de resultados de partidas de futebol, mais especificamente do Campeonato Inglês, Italiano e Brasileiro, sendo eles Premier League, Serie A e Brasileirão. Os classificadores utilizados foram Regressão Logística, Árvore de Decisão, Random Forest (RF), K-Nearest Neighbors (KNN), Support-Vector Machine (SVM) e Multilayer Perceptron (MLP). Foram utilizados para divisão dos dados os métodos treino e teste e de cross-validation. Como resultado tem-se a acurácia de cada classificador.

**Palavras-chave:** Futebol. Previsão de Resultados. Classificadores. Inteligência Artificial. Aprendizado de Máquina.

## Use of artificial intelligence algorithms in predicting soccer match results

**Abstract:** Soccer is a sport played by two teams, whose goal is to transpose a ball between extremities - called goalposts - basically utilizing foot strokes. The winner of the match will be the team that can reach the goal more often in the game. The interest of so many people in the world for soccer generates not only viewers, but also many financial transactions around this sport. Computer systems that work on predicting results and helping to minimize risks and maximize profits become an important day-to-day work tool for soccer. The objective of this project is to test and compare the performance of the classifiers used to predict soccer match results, specifically the English, Italian and Brazilian Championships, namely Premier League, Serie A and Brasileirão. The classifiers used were Logistic Regression, Decision Tree, Random forest (RF), K-Nearest Neighbors (KNN), Support-Vector Machine (SVM) and Multilayer Perceptron (MLP). Training and testing and cross-validation methods were used for data division. The result is the accuracy of each classifier.

**Keywords:** Soccer. Results prediction. Classifiers. Artificial Intelligence. Machine Learning.

---

<sup>1</sup>Estudante de Ciência da Computação, IFTM, Campus Ituiutaba, netogardenghi@gmail.com

<sup>2</sup>Professor do IFTM, Campus Ituiutaba, MG, rodrigograssi@iftm.edu.br

# 1 INTRODUÇÃO

O futebol é uma modalidade esportiva disputada por duas equipes de 11 jogadores cada um, cujo objetivo é transpor uma bola entre as extremidades – chamadas de baliza - utilizando basicamente toques com o pé. A baliza, mais conhecida como trave, é um retângulo formado por duas traves verticais, perpendiculares ao solo, e uma paralela ao solo. Entre as traves posiciona-se o goleiro o qual é o único jogador que pode colocar as mãos na bola defendendo o gol. O vencedor da partida será o time que conseguir atingir o objetivo mais vezes no jogo (SFEIR, 2011). O fácil entendimento, o baixo custo e a dinâmica empolgante fizeram dessa modalidade uma verdadeira febre mundial. A última copa do mundo jogada na Rússia no ano de 2018 atingiu uma audiência televisiva de mais de 3,5 bilhões de pessoas (FIFA, 2018).

O interesse de tantas pessoas no mundo pelo futebol gera não apenas telespectadores, mas também muitas movimentações financeiras em torno desse esporte. A copa do mundo de 2018 gerou de lucro para a FIFA cerca de 5,35 bilhões de dólares (FIFA, 2019). A movimentação financeira envolve patrocínios a clubes e seleções nacionais, venda de ingressos e produtos licenciados. Além é claro de transações de transferências envolvendo jogadores. Esse fluxo de movimentação financeira atrai uma série de investidores, que sempre visam lucrar. Sistemas computacionais que trabalham com a previsão de resultados e que auxiliam a minimizar os riscos e maximizar os lucros tornam-se então uma importante ferramenta de trabalho para o dia a dia do futebol (PERIN; VUILLEMOT; FEKETE, 2013).

A utilização de técnicas de inteligência artificial para realizar predição e previsão de resultados de partidas de futebol é um problema bastante explorado na literatura (PERIN; VUILLEMOT; FEKETE, 2013). Em Martins et al. (2017) foi realizado um estudo com a predição de resultados de partidas de futebol envolvendo dados do Campeonato Brasileiro, Espanhol e Inglês. A pesquisa demonstrou a eficiência de algoritmos de Inteligência Artificial para demonstrar quais as características mais relevantes na predição de resultados de partidas de futebol. Tal trabalho deixou em aberto a possibilidade de utilização desses algoritmos para realizar a previsão de resultados de partidas de futebol.

O resultado de uma partida de futebol é o personagem central de inúmeros estudos científicos. Existe um esforço muito grande para melhorar as táticas do jogo e as características de uma determinada equipe. Na literatura, existem estudos que se concentram nas previsões de jogos de futebol (CONSTANTINOU; FENTON; NEIL, 2013). A previsão em partidas de futebol é composta por resultado de uma partida (vitória, empate e derrota) que pode ser utilizada para

várias finalidades, incluindo apostas. Muitos esforços foram dedicados para a compreensão do futebol a partir da perspectiva dos resultados preditivos. Prever os resultados é um problema difícil devido ao grande número de fatores que devem ser levados em consideração e que nem sempre podem ser representados em valores quantitativos (HUCALJUK; RAKIPOVIĆ, 2011). Por exemplo, uma equipe pode dominar completamente uma partida sob alguns aspectos, como o número de finalizações certas, o número de passes certos ou a posse de bola e não conseguir marcar um gol a mais do que a equipe adversária para vencer uma partida (BROOKS; KERR; GUTTAG, 2016).

Na literatura é possível encontrar estudos que envolvam predição de resultados de futebol. Ulmer e Fernandez (2013) estudaram as técnicas Baseline, Gaussian Naive Bayes, Hidden Markov Model, Multimodal Naive Bayes, SVM, RBF, One vs All SGD para prever resultados do Campeonato Inglês. Hucaljuk e Rakipović (2011) pesquisaram a previsão de resultados para a UEFA Champions League com os seguintes algoritmos: Naive Bayes (NB), Redes Bayesianas, Logitboost, K-Nearest Neighbours (KNN), RF e Redes Neurais Artificiais. Com o classificador SVM, Igiri (2015) estudou os dados relacionados aos resultados de partidas do Campeonato Inglês.

Owramipur, Eskandarian e Mozneb (2013) analisaram o time de futebol do Barcelona no Campeonato Espanhol. Tax e Joustra (2015) usaram os seguintes algoritmos de classificação no Campeonato Holandês: CHIRP, LogitBoost, DTNB, FURIA, HiperPipes, J48, Naive Bayes, Perceptron e RF. Duarte, Soares e Teixeira (2015) pesquisaram os classificadores: C5.0, JRip, RF, KNN, SVM e NB para prever as partidas do Campeonato Português.

Outros estudos foram desenvolvidos para prever o resultado de partidas de futebol utilizando algoritmos de aprendizagem de máquinas. Esses algoritmos são ferramentas que recebem como entrada um conjunto de características e fornece como saída a previsão do resultados (vitória, empate e derrota).

O objetivo deste trabalho é explorar algoritmos de inteligência artificial como ferramenta de previsão de resultados de partidas de futebol. Como resultado demonstrar e comparar a acurácia obtida pelos classificadores.

## 2 DESENVOLVIMENTO

### 2.1 METODOLOGIA

A metodologia utilizada é apresentada na figura 1, a seguir, a figura apresenta o fluxograma do sistema proposto.

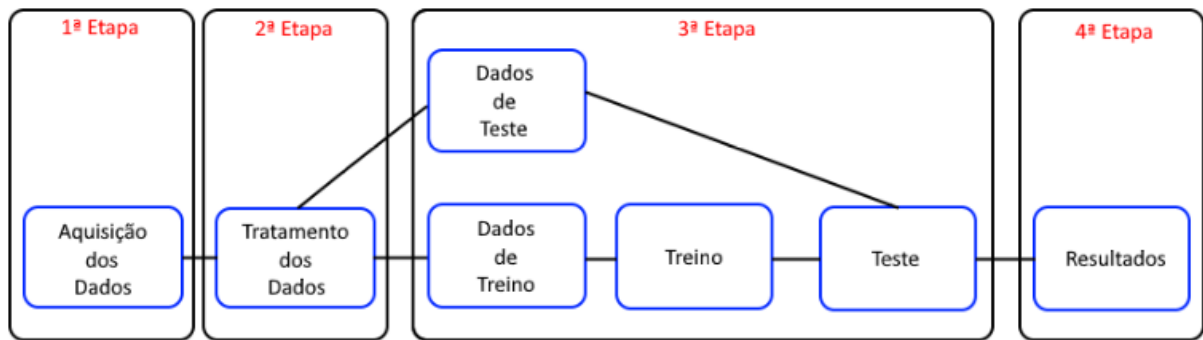


Figura 1: Fluxograma do sistema proposto

#### 2.1.1 1ª Etapa - Aquisição dos dados

A primeira etapa envolve a escolha e a montagem de uma base de dados, algumas bases públicas estão disponíveis em <https://www.football-data.co.uk> e foram objetos de estudo em Ulmer e Fernandez (2013). Foram escolhidas deste mesmo site as seguintes bases de dados:

- Premier League (Temporada 2018/2019);
- Serie A (Temporada 2018/2019).

Foi escolhida também a base de dados do Brasileirão Série A 2012 coletada do UOL Esportes e disponível em <https://github.com/zandree/statsBrazilianLeagueChampionship>. Todas as bases estão no formato CSV.

As bases da Premier League e Serie A foram feitas como cada linha sendo um jogo e esta linha possui os dados do time de casa e o time de fora. O resultado é dado como casa, empate ou fora. A base do Brasileirão tem sua construção diferente. Cada linha também é um jogo porém esta linha só possui os dados de um dos times, ou seja, cada jogo ocorrido terá 2 linhas uma para o time de casa e uma para o de fora. O resultado é dado com vitória, empate ou derrota.

### 2.1.2 2ª Etapa - Tratamento dos dados

Esta etapa envolve todo o pré-processamento de dados. Primeiramente foi criado um Data-Frame utilizando o Pandas para ler as bases de dados. A base da Premier League possui 380 linhas e 62 colunas enquanto a da Serie A possui 380 linhas e 61 colunas e ambas as bases não possuem dados faltantes. A base do Brasileirão possui 760 linhas, 50 colunas e 4 linhas com dados faltantes. As linhas com dados faltantes foram removidas e a base passou a ter 756 linhas.

O próximo passo foi selecionar as características para a previsão. As bases Premier League e Serie A possuem características semelhantes das quais foram selecionadas:

| Característica | Descrição   |
|----------------|---|
| FTR            | Resultado da partida (Casa, Empate, Fora)         |
| HS             | Chutes do time da casa                            |
| AS             | Chutes do time de fora                            |
| HST            | Chutes do time da casa no alvo(gol)               |
| AST            | Chutes do time de fora no alvo(gol)               |
| HF             | Faltas cometidas pelo time da casa                |
| AF             | Faltas cometidas pelo time de fora                |
| HC             | Escanteios do time da casa                        |
| AC             | Escanteios do time de fora                        |
| HY             | Cartões Amarelos do time da casa                  |
| AY             | Cartões Amarelos do time de fora                  |
| HR             | Cartões Vermelhos do time da casa                 |
| AR             | Cartões Vermelhos do time de fora                 |
| HTHG           | Gols marcados pela equipe da casa até o intervalo |
| HTAG           | Gols marcados pela equipe de fora até o intervalo |

Tabela 1: Características das bases Premier League e Serie A.

Na base de dados do Brasileirão foram utilizadas as seguintes características:

| <b>Característica</b> | <b>Descrição</b>          |
|-----------------------|---------------------------|
| assistances           | Assistências              |
| receivedBalls         | Bolas recebidas           |
| recoveredBalls        | Bolas recuperadas         |
| lostBalls             | Bolas perdidas            |
| yellowCards           | Cartões amarelos          |
| redCards              | Cartões vermelhos         |
| receivedCrossBalls    | Cruzamentos recebidos     |
| missedCrossBalls      | Cruzamentos perdidos      |
| defenses              | Defesas                   |
| sucessfulTackles      | Desarmes com sucesso      |
| unsucessfulTackles    | Desarmes sem sucesso      |
| sucessfulDribles      | Dribles com sucesso       |
| unsucessfulDribles    | Dribles sem sucesso       |
| givenCorners          | Escanteios dados          |
| receivedCorners       | Escanteios recebidos      |
| receivedFouls         | Faltas recebidas          |
| committedFouls        | Faltas cometidas          |
| goodFinishes          | Boas finalizações         |
| badFinishes           | Finalizações ruins        |
| ownGols               | Gols marcados             |
| offsides              | Impedimentos              |
| sucessfulLongPasses   | Passes longos com sucesso |
| unsucessfulLongPasses | Passes longos sem sucesso |
| sucessfulPasses       | Passes com sucesso        |
| unsucessfulPasses     | Passes sem sucesso        |
| win                   | Vitória                   |
| draw                  | Empate                    |
| defeat                | Derrota                   |

Tabela 2: Características da base Brasileirão.

Em Premier League e Serie A os dados de FTR são categóricos e como FTR é o alvo da previsão foi preciso transformá-los para numéricos. Casa foi substituído por 2, empate por 1 e fora por 0. No Brasileirão as características win, draw e defeat foram mescladas em uma só, chamada FTR, que também é o alvo da previsão, sendo 2 - vitória, 1 - empate e 0 - derrota.

Com as características definidas o próximo passo foi padronizar os dados. Os dados foram padronizados com o *StandardScaler* que está presente no Scikit-learn. O *StandardScaler* transforma os dados de forma que sua distribuição tenha um valor médio de 0 e o desvio padrão 1. Cada valor no conjunto de dados foi subtraído pela média das amostras e dividido pelo desvio padrão de todo o conjunto de dados (LEARN, 2019). O motivo da padronização foi para que uma característica não sobressaia sobre a outra.

### 2.1.3 3ª Etapa - Treino e teste

Foi definido qual era o alvo da previsão. Em todas as bases o alvo é a característica FTR a qual contém o resultado da partida. As demais características foram usadas para a previsão de FTR. Nesta etapa foram utilizados dois métodos diferentes de divisão de dados.

O primeiro método foi o de divisão de treino e teste. Este método embaralha os dados e os divide em uma porcentagem para treino e outra para teste. Foi dividido 70% para treino e 30% para teste. A divisão dos dados gera quatro novos conjuntos de dados, sendo eles: características para treino, características para teste, alvo para treino e alvo para teste.

O segundo método foi o de validação cruzada (*cross-validation*) *k-fold*. Neste método os dados de entrada são divididos em subconjuntos de dados *k*(chamados de *folds*). O modelo é treinado com *k-1* subconjuntos e depois avaliado no subconjunto que não foi utilizado para treinamento. O procedimento é repetido *k* vezes e em cada vez um subconjunto diferente é reservado para avaliação (SOUZA, 2019). Foram utilizados 10 *folds*.

Para previsão dos resultados foram utilizadas diferentes abordagens. Foi feito a previsão com a base completa e com os dados do alvo agrupados de 2 em 2. Nas bases Premier League e Serie A foram feitas previsões com as seguintes abordagens:

- Casa, Empate e Fora;
- Casa e Fora;
- Casa e Empate;
- Fora e Empate.

Na base do Brasileirão as abordagens foram:

- Vitória, Empate e Derrota;
- Vitória e Derrota;
- Vitória e Empate;
- Derrota e Empate.

O próximo passo foi a criação dos modelos de previsão. Os classificadores utilizados foram:

- Regressão logística;

- Árvore de decisão;
- Floresta aleatória;
- K Nearest Neighbours(KNN);
- Support vector machine(SVM);
- Multi Layer Perceptron(MLP).

A regressão logística é uma técnica estatística que tem como objetivo modelar, a partir de um conjunto de observações, a relação “logística” entre uma variável resposta dicotômica e uma série de variáveis explicativas numéricas e/ou categóricas (CABRAL, 2013). Define-se uma variável dependente (Y), ou variável de saída, e procura-se verificar a influência de uma ou mais variáveis ditas variáveis independentes (X’s) sobre esta variável dependente (ZANINI, 2007). A regressão logística busca calcular ou prever a probabilidade da variável dependente assumir um determinado valor em função de outras variáveis.

Árvores de decisão são modelos estatísticos para a classificação e previsão de dados. Formada por um conjunto de nós de decisão uma árvore começa com um nó, que se divide em possíveis resultados. Cada resultado leva a nós adicionais, que se ramificam em outras possibilidades. Uma árvore de decisão é essencialmente uma série de declarações if-elses, que quando aplicados a um registro de uma base de dados, resultam na classificação daquele registro (CARACIOLO, 2009).

Floresta aleatória, do inglês *Random Forest*, é um classificador que consiste em uma combinação de árvores de decisão. Cada árvore depende dos valores de vetores aleatórios amostrados de forma independente e distribuídos igualmente para todas as árvores na floresta (BREIMAN, 2001). Uma floresta aleatória é construída por um simples voto de árvores de decisão usando o algoritmo Bagging. Todas as árvores treinadas são combinadas em uma composição com uma votação simples, usando o erro médio de todas as amostras (DMITRIEVSKY, 2018).

K vizinhos mais próximos, do inglês *K-Nearest Neighbours*, classifica um elemento com base na proximidade de seus k vizinhos. O algoritmo calcula a distância do elemento alvo aos demais e os ordena da menor para o maior. Após ordenadas são selecionados apenas os k mais próximos vizinhos deste elemento alvo. Ele será classificado com as mesmas características do tipo de vizinho que está em maior quantidade dos k selecionados (BOEHMKE, 2019).



Máquina de vetores de suporte, do inglês *Support-Vector Machine*, é um algoritmo cujo objetivo é classificar determinado conjunto de dados que são mapeados para um espaço de características multidimensional usando uma função kernel. Nela, o limite de decisão no espaço de entrada é representado por um hiperplano em dimensão superior no espaço (BOSER; GUYON; VAPNIK, 1992). O que o SVM faz é encontrar um hiperplano entre dados de duas classes buscando maximizar a distância entre os pontos mais próximos em relação a cada uma das classes.

Perceptron multicamadas, do inglês *Multilayer Perceptron*, é uma rede neural com camadas ocultas com um número indeterminado de neurônios. O MLP utiliza métodos derivados do gradiente no ajustes de seus pesos por retropropagação (*backpropagation*). A rede consiste em uma camada de entrada, uma ou mais camadas escondidas e uma ou mais camadas de saída. Um sinal de entrada é propagado da entrada para saída e a saída propaga um sinal em caminho reverso alterando os pesos, para uma nova validação (AMARAL; LIMA; OLIVEIRA, 2014).

O scikit-learn possui todos esses classificadores. Para fazer a previsão primeiro é preciso instanciar o classificador, depois chamar a função de treino e ao final a função de previsão. Os modelos que serão apresentados a seguir valem para as 3 bases de dados. A abordagem a ser tratada é a da base de dados completa, a abordagem de 2 em 2 segue do mesmo raciocínio.

Para os algoritmos de regressão logística, árvore de decisão e MLP foram instanciados os classificadores, depois chamada a função *fit* responsável por treinar o modelo passando como parâmetro as características de treino e o alvo de treino. A função *predict* é responsável pela previsão e nela é passada como parâmetro o alvo de teste, ou seja, com base no treinamento feito com as bases de treino o classificador vai tentar prever uma base ainda não vista, no caso a do alvo de teste. Todo esse procedimento foi feito com as bases do método de divisão 70-30. Para o *cross-validation* é utilizado o *cross\_val\_score* passando como parâmetros o classificador, o vetor de características, o alvo da previsão, e a quantidade de *folds*.

Para os modelos de floresta aleatória e KNN é preciso definir o número de árvores para um e o número de vizinhos para o outro. A fim de encontrar qual o valor que gere a melhor previsão é feito um “método cotovelo”. Este método consiste em fazer previsões alterando os valores dentro de um intervalo. Ao terminar as previsões é feita a média das diferenças entre o valor previsto e o valor real e este valor é salvo em um vetor. O melhor número para utilizar é aquele que representa a menor diferença entre o valor real e o previsto. A floresta aleatória e o KNN seguem do mesmo raciocínio que os demais.

Para o modelo de SVM é utilizado o *GridSearchCV*. O *GridSearchCV* é um método onde é testado diferentes valores para alguns parâmetros do classificador a fim de descobrir qual combinação gera a melhor previsão. Primeiro é definido qual o classificador depois, uma lista com os parâmetros que vão ser testados e seus respectivos valores e a quantidade de *folds*. A previsão segue do mesmo raciocínio que os demais.

#### 2.1.4 4ª Etapa - Resultados

Para avaliar a precisão dos classificadores foi utilizado o método *score* que está contido em cada classificador. Este método compara os valores reais com os valores previstos e retorna a acurácia ou o coeficiente de determinação de cada classificador.

Os resultados obtidos para a base Premier League foram:

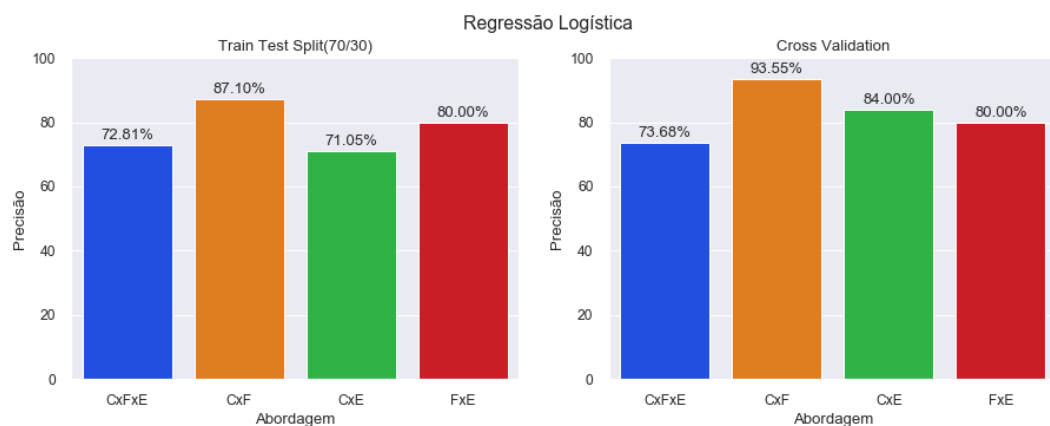


Figura 2: Regressão logística Premier League

Na figura 2 tem-se o resultado da previsão com o classificador regressão logística. O melhor resultado obtido no método 70-30 foi com a abordagem casa x fora com 87.10% de acurácia. No método de validação cruzada o melhor resultado também foi com a abordagem casa x fora com 93.55% de acurácia.

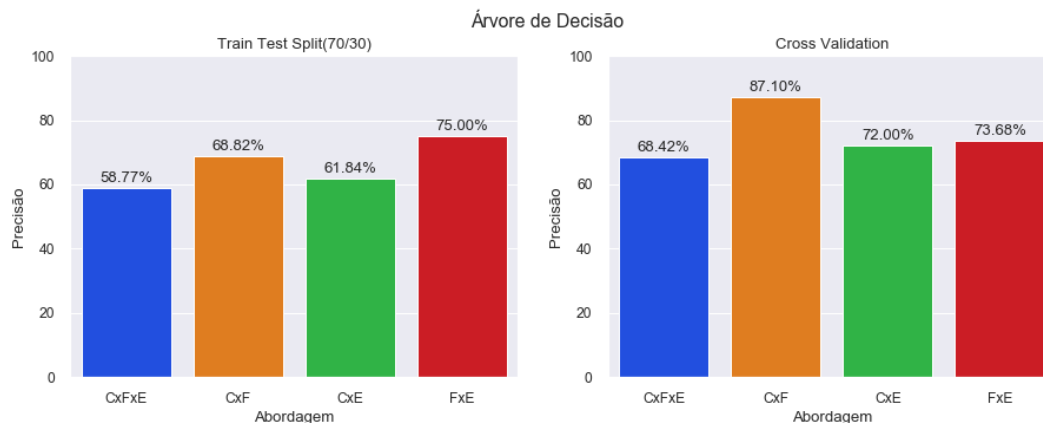


Figura 3: Árvore de decisão Premier League

Na figura 3 tem-se o resultado da previsão com o classificador árvore de decisão. O melhor resultado obtido no método 70-30 foi com a abordagem fora x empate com 75% de acurácia. No método de validação cruzada o melhor resultado foi com a abordagem casa x fora com 87.10% de acurácia.

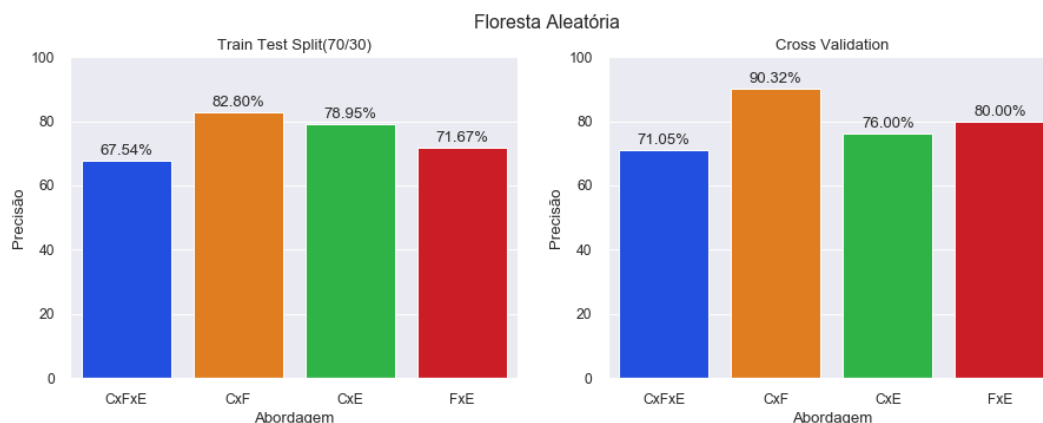


Figura 4: Floresta aleatória Premier League

Na figura 4 tem-se o resultado da previsão com o classificador floresta aleatória. O melhor resultado obtido no método 70-30 foi com a abordagem casa x fora com 82.80% de acurácia. No método de validação cruzada o melhor resultado também foi com a abordagem casa x fora com 90.32% de acurácia.

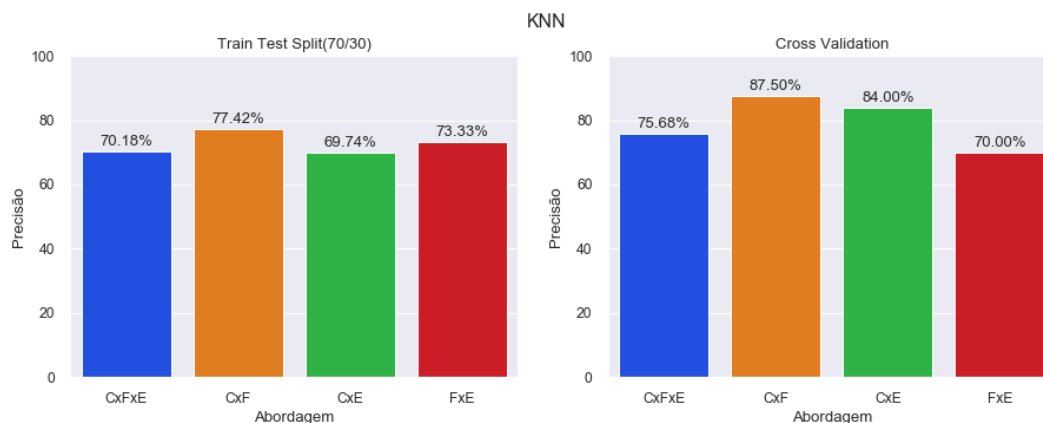


Figura 5: KNN Premier League

Na figura 5 tem-se o resultado da previsão com o classificador KNN. O melhor resultado obtido no método 70-30 foi com a abordagem casa x fora com 77.42% de acurácia. No método de validação cruzada o melhor resultado também foi com a abordagem casa x fora com 87.50% de acurácia.

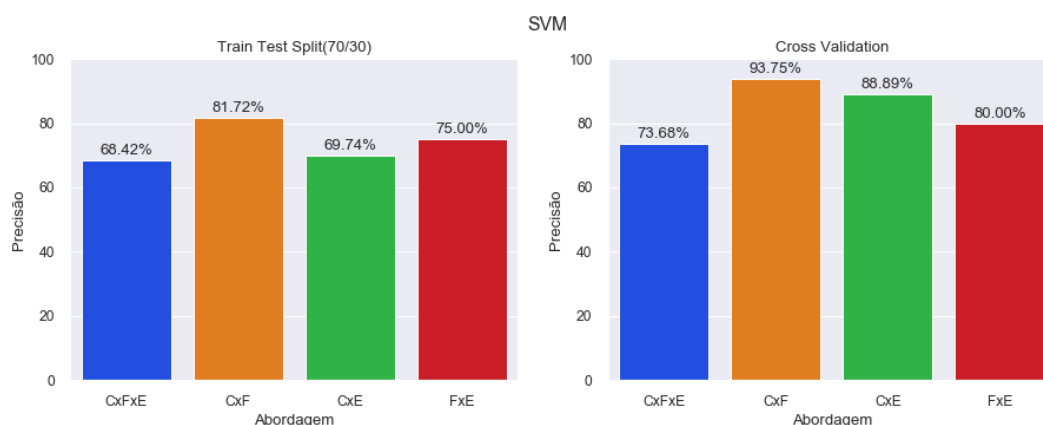


Figura 6: SVM Premier League

Na figura 6 tem-se o resultado da previsão com o classificador SVM. O melhor resultado obtido no método 70-30 foi com a abordagem casa x fora com 81.72% de acurácia. No método de validação cruzada o melhor resultado também foi com a abordagem casa x fora com 93.75% de acurácia.

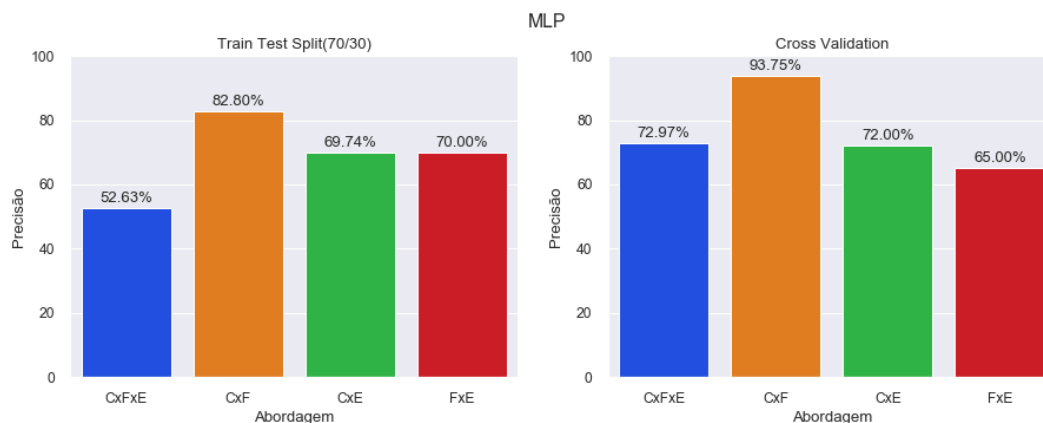


Figura 7: MLP Premier League

Na figura 7 tem-se o resultado da previsão com o classificador MLP. O melhor resultado obtido no método 70-30 foi com a abordagem casa x fora com 82.80% de acurácia. No método de validação cruzada o melhor resultado também foi com a abordagem casa x fora com 93.75% de acurácia.

Os resultados obtidos para a base Serie A foram:

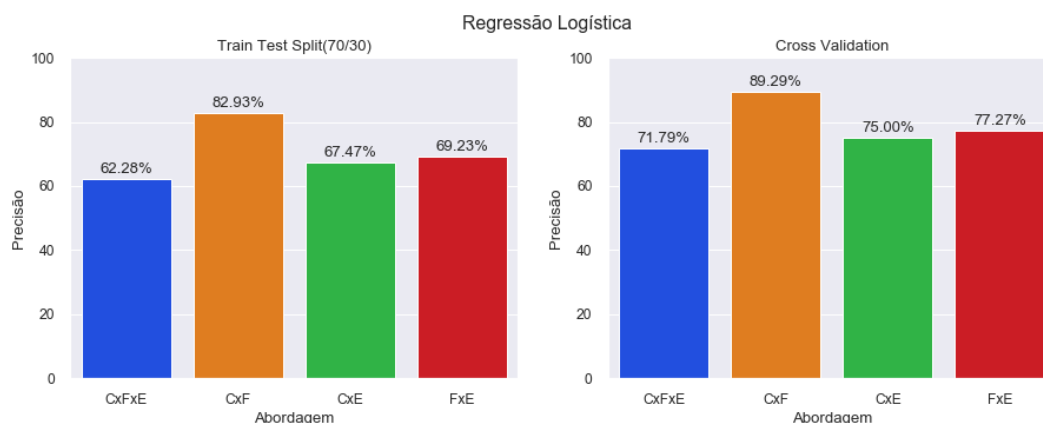


Figura 8: Regressão logística Serie A

Na figura 8 tem-se o resultado da previsão com o classificador regressão logística. O melhor resultado obtido no método 70-30 foi com a abordagem casa x fora com 82.93% de acurácia. No método de validação cruzada o melhor resultado também foi com a abordagem casa x fora com 89.29% de acurácia.

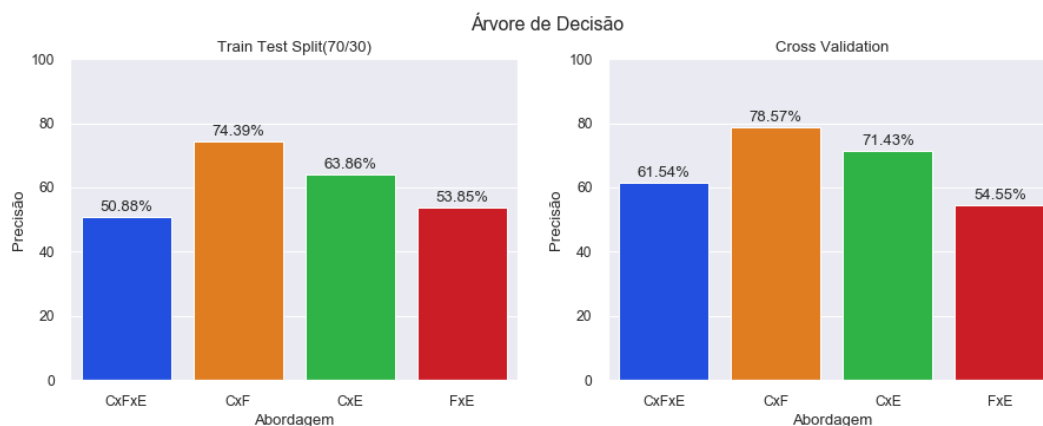


Figura 9: Árvore de decisão Serie A

Na figura 9 tem-se o resultado da previsão com o classificador árvore de decisão. O melhor resultado obtido no método 70-30 foi com a abordagem casa x fora com 74.39% de acurácia. No método de validação cruzada o melhor resultado também foi com a abordagem casa x fora com 78.57% de acurácia.

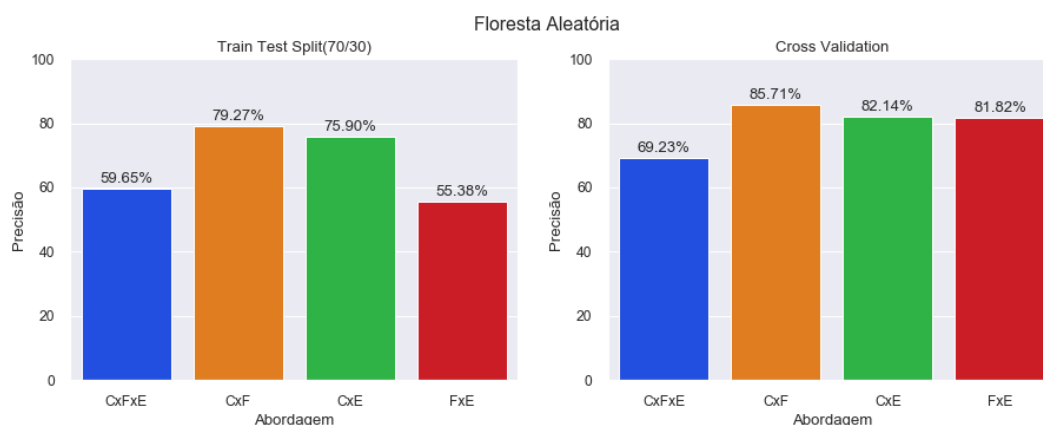


Figura 10: Floresta aleatória Serie A

Na figura 10 tem-se o resultado da previsão com o classificador floresta aleatória. O melhor resultado obtido no método 70-30 foi com a abordagem casa x fora com 79.27% de acurácia. No método de validação cruzada o melhor resultado também foi com a abordagem casa x fora com 85.71% de acurácia.

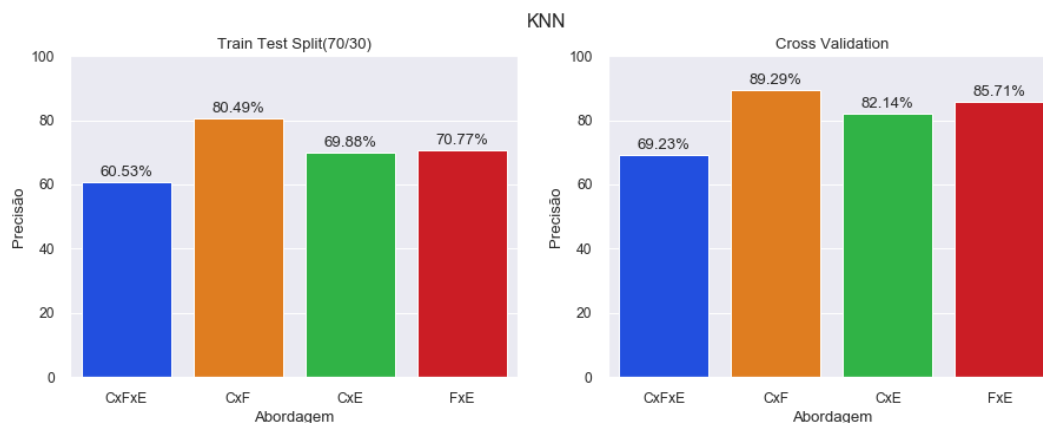


Figura 11: KNN Serie A

Na figura 11 tem-se o resultado da previsão com o classificador KNN. O melhor resultado obtido no método 70-30 foi com a abordagem casa x fora com 80.49% de acurácia. No método de validação cruzada o melhor resultado também foi com a abordagem casa x fora com 89.29% de acurácia.

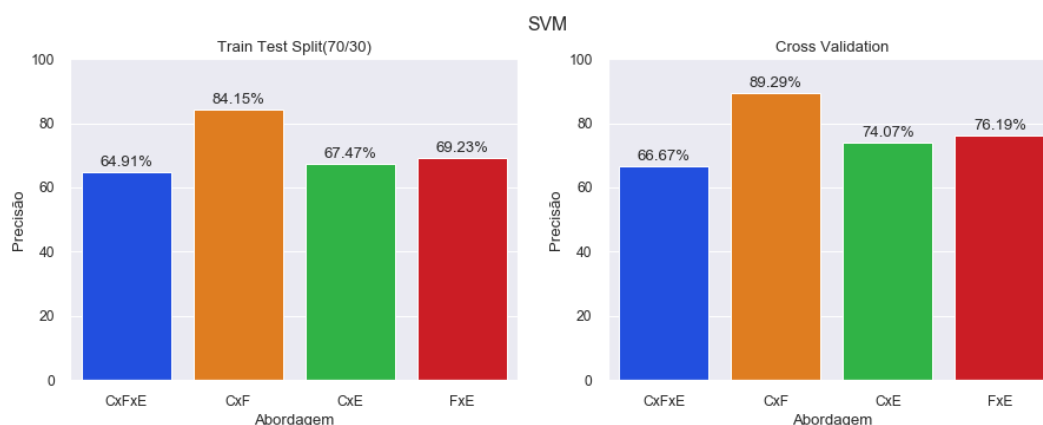


Figura 12: SVM Serie A

Na figura 12 tem-se o resultado da previsão com o classificador SVM. O melhor resultado obtido no método 70-30 foi com a abordagem casa x fora com 84.15% de acurácia. No método de validação cruzada o melhor resultado também foi com a abordagem casa x fora com 89.29% de acurácia.

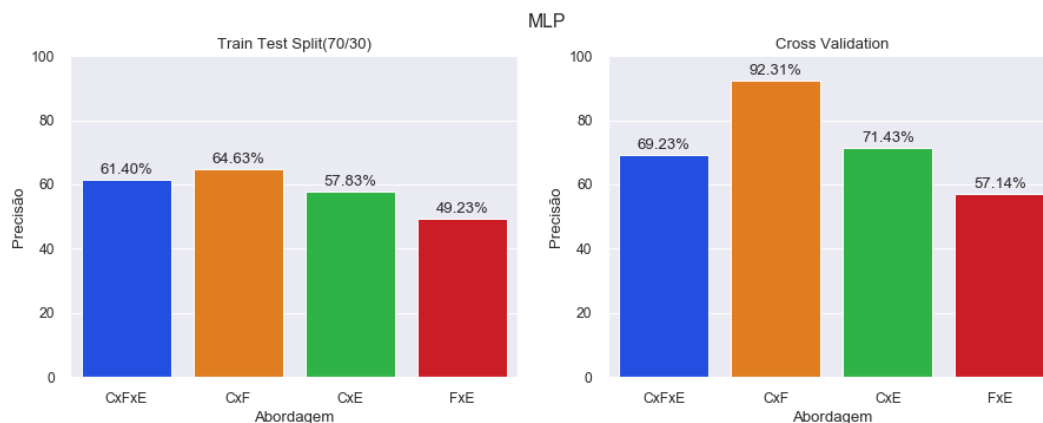


Figura 13: MLP Serie A

Na figura 13 tem-se o resultado da previsão com o classificador MLP. O melhor resultado obtido no método 70-30 foi com a abordagem casa x fora com 64.63% de acurácia. No método de validação cruzada o melhor resultado também foi com a abordagem casa x fora com 92.31% de acurácia.

Os resultados obtidos para a base Brasileirão foram:

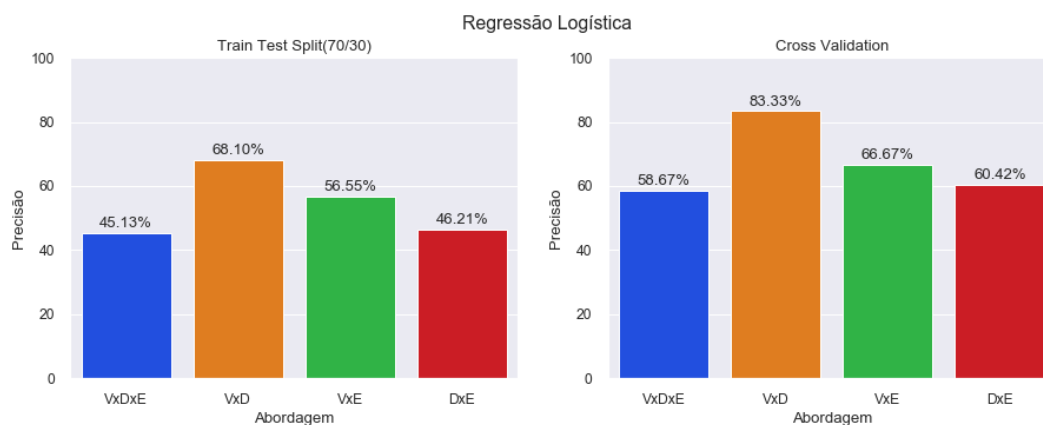


Figura 14: Regressão logística Brasileirão

Na figura 14 tem-se o resultado da previsão com o classificador regressão logística. O melhor resultado obtido no método 70-30 foi com a abordagem vitória x derrota com 68.10% de acurácia. No método de validação cruzada o melhor resultado também foi com a abordagem vitória x derrota com 83.33% de acurácia.



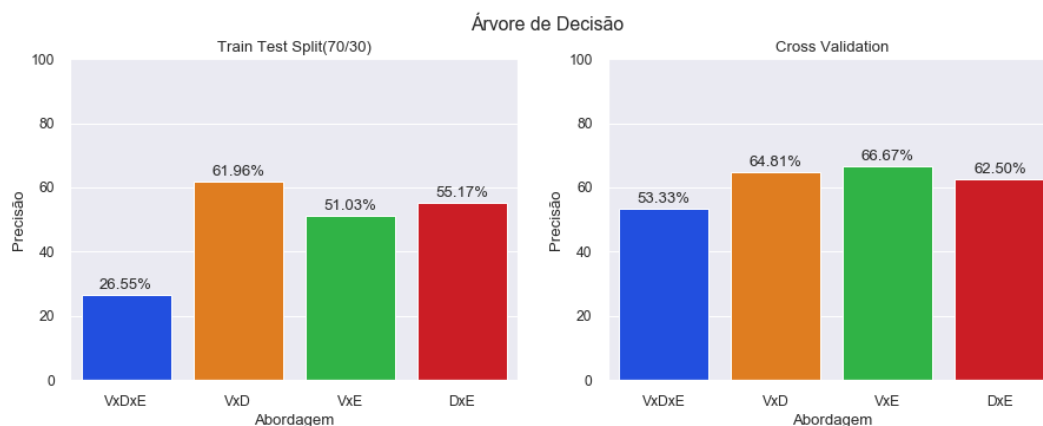


Figura 15: Árvore de decisão Brasileiro

Na figura 15 tem-se o resultado da previsão com o classificador árvore de decisão. O melhor resultado obtido no método 70-30 foi com a abordagem vitória x derrota com 61.96% de acurácia. No método de validação cruzada o melhor resultado foi com a abordagem vitória x empate com 66.67% de acurácia.

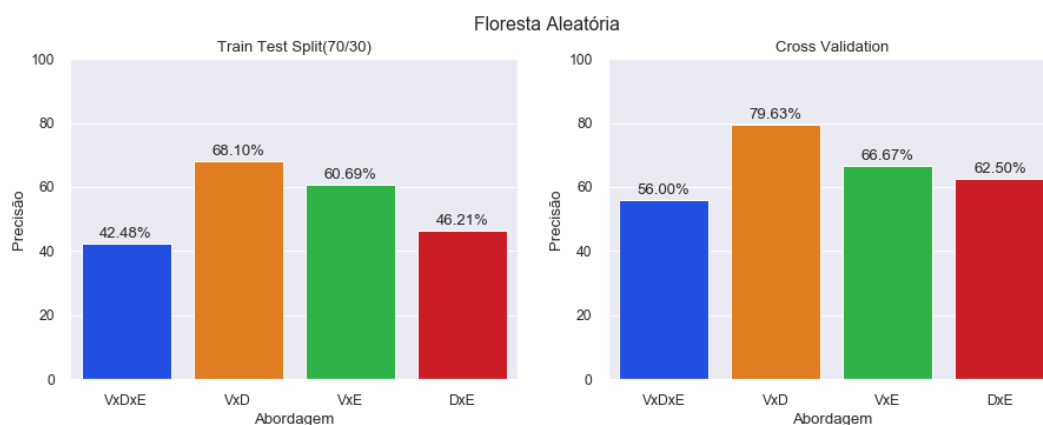


Figura 16: Floresta aleatória Brasileiro

Na figura 16 tem-se o resultado da previsão com o classificador floresta aleatória. O melhor resultado obtido no método 70-30 foi com a abordagem vitória x derrota com 68.10% de acurácia. No método de validação cruzada o melhor resultado também foi com a abordagem vitória x derrota com 79.63% de acurácia.

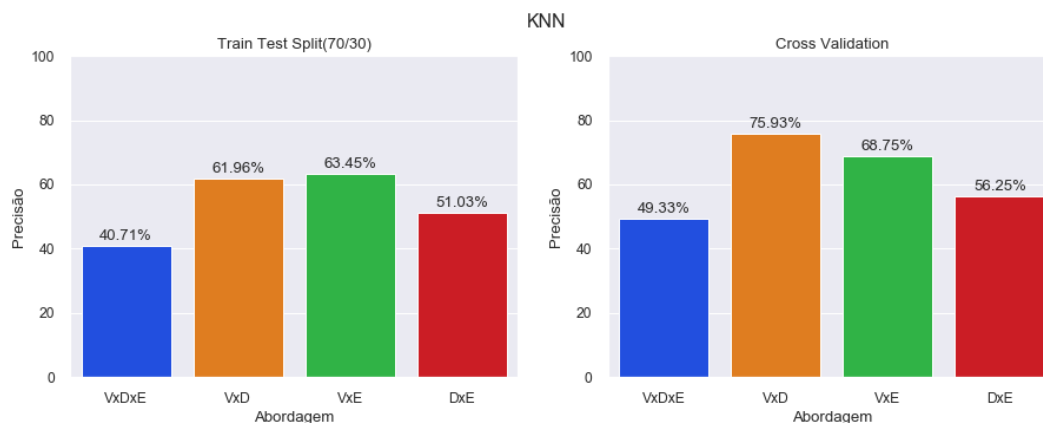


Figura 17: KNN Brasileiro

Na figura 17 tem-se o resultado da previsão com o classificador KNN. O melhor resultado obtido no método 70-30 foi com a abordagem vitória x empate com 63.45% de acurácia. No método de validação cruzada o melhor resultado foi com a abordagem vitória x derrota com 75.93% de acurácia.

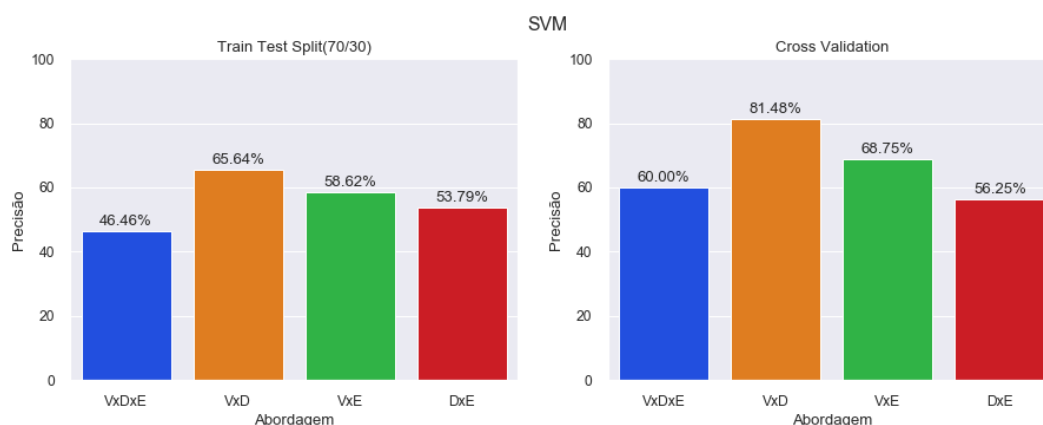


Figura 18: SVM Brasileiro

Na figura 18 tem-se o resultado da previsão com o classificador SVM. O melhor resultado obtido no método 70-30 foi com a abordagem vitória x derrota com 65.64% de acurácia. No método de validação cruzada o melhor resultado também foi com a abordagem vitória x derrota com 81.48% de acurácia.

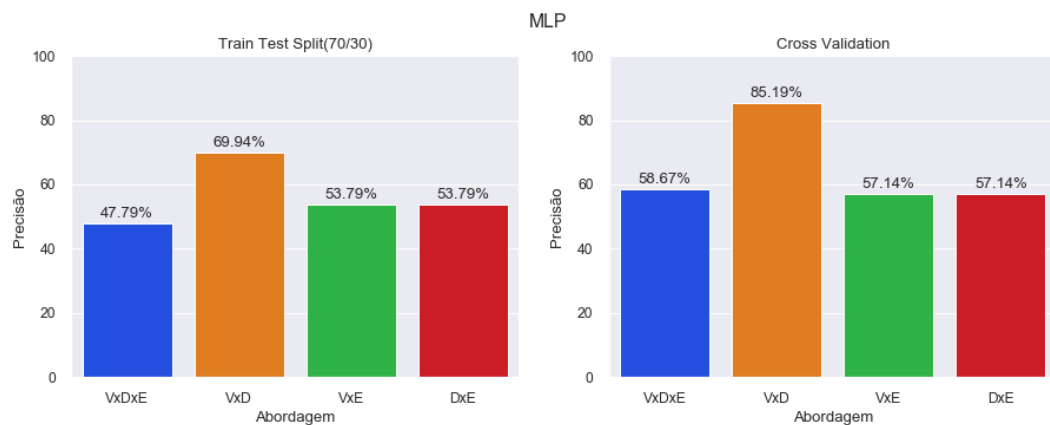


Figura 19: MLP Brasileiro

Na figura 19 tem-se o resultado da previsão com o classificador MLP. O melhor resultado obtido no método 70-30 foi com a abordagem vitória x derrota com 69.94% de acurácia. No método de validação cruzada o melhor resultado também foi com a abordagem vitória x derrota com 85.19% de acurácia.

### 3 DISCUSSÃO E CONCLUSÃO

Conclui-se que em questão de dados para os modelos é mais difícil prever o resultado de Casa x Fora x Empate e Vitória x Empate x Derrota. Os modelos performam melhor quando o alvo da previsão é binário (2 classes). Dentre as bases com 2 classes Fora x Empate e Derrota x Empate foram as que tiveram menor performance e Casa x Fora e Vitória x Derrota foram as que tiveram melhor performance. Podemos concluir então que os modelos performam melhor em prever resultados sem empate.

Levando em consideração os métodos de divisão das bases de dados concluímos que o método de validação cruzada gera uma performance melhor do que o método de divisão 70% treino 30% teste.

A tabela abaixo mostra quais classificadores obtiveram uma melhor performance em cada método de divisão e em cada abordagem. Para a base brasileiro considere C como vitória e F como derrota.

| 70-30             |                        |                        |                                      |
|-------------------|------------------------|------------------------|--------------------------------------|
| Abordagem         | Premier League         | Serie A                | Brasileirão                          |
| CxFxE             | Regressão Logística    | SVM                    | MLP                                  |
| CxF               | Regressão Logística    | SVM                    | MLP                                  |
| CxE               | Floresta Aleatória     | Floresta Aleatória     | KNN                                  |
| FxE               | Regressão Logística    | KNN                    | Árvore de Decisão                    |
| Validação Cruzada |                        |                        |                                      |
| CxFxE             | KNN                    | Regressão Logística    | SVM                                  |
| CxF               | SVM                    | MLP                    | MLP                                  |
| CxE               | SVM                    | Floresta Aleatória/KNN | KNN/SVM                              |
| FxE               | Regressão/Floresta/SVM | KNN                    | Árvore de Decisão/Floresta Aleatória |

Tabela 3: Classificadores com melhor performance em cada método de divisão e abordagem.

Podemos concluir que os algoritmos que obtiveram a melhor performance mais vezes foram nesta ordem: SVM, KNN, regressão logística, floresta aleatória, MLP e árvore de decisão.

Todos os algoritmos abordados neste trabalho foram implementados em Python e encontram-se disponíveis em <https://github.com/olesiogn/Football-prediction>.

## Referências

- AMARAL, C. do; LIMA, J. M. C. de; OLIVEIRA, L. E. S. de. *Rede Neural Supervisionada Multicamadas - Árvore de decisão plataforma WEKA*. 2014. Disponível em: <https://pt.slideshare.net/cesardoamatal/a-rede-neural-supervisionada-chamada-perceptron-multicamadas>. Acesso em: 26 nov 2019.
- BOEHMKE, B. M. W. S. U. G. B. *Hands-On Machine Learning with R*. Taylor & Francis Ltd, 2019. ISBN 1138495689. Disponível em: [https://www.ebook.de/de/product/38343783/bradley/\\_boehmke/\\_brandon/\\_m/\\_wright/\\_state/\\_university/\\_greenwell/\\_hands/\\_on/\\_machine/\\_learning/\\_with/\\_r.html](https://www.ebook.de/de/product/38343783/bradley/_boehmke/_brandon/_m/_wright/_state/_university/_greenwell/_hands/_on/_machine/_learning/_with/_r.html).
- BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A training algorithm for optimal margin classifiers. In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. New York, NY, USA: ACM, 1992. (COLT '92), p. 144–152. ISBN 0-89791-497-X. Disponível em: <http://doi.acm.org/10.1145/130385.130401>.
- BREIMAN, L. *Machine Learning*, Springer Nature, v. 45, n. 1, p. 5–32, 2001.
- BROOKS, J.; KERR, M.; GUTTAG, J. Using machine learning to draw inferences from pass location data in soccer. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, Wiley, v. 9, n. 5, p. 338–349, jun 2016.
- CABRAL, C. I. S. *Aplicação do Modelo de Regressão Logística num Estudo de Mercado*. Dissertação (Mestrado) — Universidade de Lisboa, 2013. Disponível em: [https://repositorio.ul.pt/bitstream/10451/10671/1/ulfc106455\\_tm\\_Cleidy\\_Cabral.pdf](https://repositorio.ul.pt/bitstream/10451/10671/1/ulfc106455_tm_Cleidy_Cabral.pdf).
- CARACIOLO, M. P. *Introdução a Árvores de decisão para classificação e mineração de dados*. 2009. Disponível em: <http://aimotion.blogspot.com/2009/04/artigo-introducao-arvores-de-decisao.html>. Acesso em: 24 nov 2019.
- CONSTANTINO, A. C.; FENTON, N. E.; NEIL, M. Profiting from an inefficient association football gambling market: Prediction, risk and uncertainty using bayesian networks. *Knowledge-Based Systems*, Elsevier BV, v. 50, p. 60–86, sep 2013.
- DMITRIEVSKY, M. *Floresta de decisão aleatória na aprendizagem por reforço*. 2018. Disponível em: <https://www.mql5.com/pt/articles/3856>. Acesso em: 24 nov 2019.
- DUARTE, L. M. da S.; SOARES, C.; TEIXEIRA, J. *Previsão de resultados de jogos de futebol*. Dissertação (Mestrado) — Faculdade da Engenharia da Universidade do Porto, 2015.
- FIFA. *More than half the world watched record-breaking 2018 World Cup*. 2018. Disponível em: <https://www.fifa.com/worldcup/news/more-than-half-the-world-watched-record-breaking-2018-world-cup>. Acesso em: 15 jun 2019.
- FIFA. *Financial Report 2018*. 2019. Disponível em: <https://resources.fifa.com/image/upload/xzshoe2ayttyquuxhq0.pdf>. Acesso em: 15 jun 2019.
- HUCALJUK, J.; RAKIPOVIĆ, A. Predicting football scores using machine learning techniques. *2011 Proceedings of the 34th International Convention MIPRO*, may 2011.

IGIRI, C. P. Support vector machine–based prediction system for a football match result. *IOSR Journal of Computer Engineering*, v. 17, n. 3, p. 21–26, jun 2015.

LEARN scikit. *sklearn.preprocessing.StandardScaler*. 2019. Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>. Acesso em: 24 nov 2019.

MARTINS, R. G. et al. Exploring polynomial classifier to predict match results in football championships. *Expert Systems with Applications*, Elsevier BV, v. 83, p. 79–93, oct 2017.

OWRAMIPUR, F.; ESKANDARIAN, P.; MOZNEB, F. S. Football result prediction with bayesian network in spanish league-barcelona team. *International Journal of Computer Theory and Engineering*, IACSIT Press, p. 812–815, 2013.

PENDHARKAR, P. C.; KHOSROWPOUR, M.; RODGER, J. A. Application of bayesian network classifiers and data envelopment analysis for mining breast cancer patterns. *Journal of Computer Information Systems*, v. 40, n. 4, p. 127–132, 2000.

PERIN, C.; VUILLEMOT, R.; FEKETE, J.-D. SoccerStories: A kick-off for visual soccer analysis. *IEEE Transactions on Visualization and Computer Graphics*, Institute of Electrical and Electronics Engineers (IEEE), v. 19, n. 12, p. 2506–2515, dec 2013.

SFEIR, M. N. Laws of the game (adapted from fifa 2010-11). *World Literature Today*, v. 85, n. 3, p. 38–39, may 2011.

SOUZA, A. *Validação Cruzada: Conceito e Exemplo em R*. 2019. Disponível em: <https://pessoalex.wordpress.com/2019/04/16/validacao-cruzada-conceito-e-exemplo-em-r/>. Acesso em: 24 nov 2019.

TAX, N.; JOUSTRA, Y. Predicting the dutch football competition using public data: A machine learning approach. Unpublished, 2015.

ULMER, B.; FERNANDEZ, M. *Predicting Soccer Match Results in the English Premier League*. Tese (Doutorado) — Stanford University, 2013.

ZANINI, A. *Regressão Logística e Redes Neurais Artificiais: Um Problema de Estrutura de Preferência do Consumidor e Classificação de Perfis de Consumo*. Dissertação (Mestrado) — Faculdade de Economia e Administração da Universidade Federal de Juiz de Fora, 2007. Disponível em: [http://www.ufjf.br/poseconomia/files/2010/01/td\\_007\\_20071.pdf](http://www.ufjf.br/poseconomia/files/2010/01/td_007_20071.pdf).