UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

DEPARTAMENTO DE ESTATÍSTICA

UTILIZAÇÃO DE ALGORITMOS DO TIPO MACHINE LEARNING SUPERVISIONADO PARA A CARACTERIZAÇÃO DOS RESULTADOS DA COPA DO MUNDO DE FUTEBOL DE 2018

ALUÍSIO MOREIRA NABINGER

Porto Alegre
2018

ALUÍSIO MOREIRA NABINGER

UTILIZAÇÃO DE ALGORITMOS DO TIPO MACHINE LEARNING SUPERVISIONADO PARA A CARACTERIZAÇÃO DOS RESULTADOS DA COPA DO MUNDO DE FUTEBOL DE 2018

Trabalho de Conclusão de Curso submetido como requisito parcial para a obtenção do grau de Bacharel em Estatística.

Orientador

Professor Dr. Hudson da Silva Torrent

Universidade Federal do Rio Grande do Sul (UFRGS)

Porto Alegre
2018

Instituto de Matemática e Estatística

Departamento de Estatística

UTILIZAÇÃO DE ALGORITMOS DO TIPO MACHINE LEARNING SUPERVISIONADO PARA A CARACTERIZAÇÃO DOS RESULTADOS DA COPA DO MUNDO DE FUTEBOL DE 2018

ALUÍSIO MOREIRA NABINGER

Banca examinadora:

Professor Dr. Marcio Valk

Universidade Federal do Rio Grande do Sul (UFRGS)

AGRADECIMENTOS

À UFRGS, pela oportunidade de ensino e crescimento profissional.

Para todos os familiares e amigos que ao relacionarem-se comigo tornaram-se parte do que eu sou hoje.

Aos meus pais por todo o apoio e amor incondicional. Ensinaram-me sempre sobre a importância da educação e esse trabalho conclui mais uma etapa nesse processo. Meu pai, sempre sendo mais do que a definição de prestativo. Minha mãe por todo o conhecimento, não só técnico como também de mim mesmo. Sem ambos esse trabalho não teria sido possível.

Meu irmão por todo o companheirismo em boa parte da vida.

Amanda, por ter sido a pessoa mais sensacional que eu poderia ter desejado em 2018. Obrigado por existir e por ter voltado a existir na minha vida.

Tanise, por ter me impedido de desistir em diversas dificuldades que apareceram durante o trabalho.

Agradeço ao meu orientador Hudson Torrent por toda a ajuda e paciência na execução desse trabalho.

There’s a lot of beauty in ordinary things.

(FISHER, JENNA; THE OFFICE, 2013.)

RESUMO

Entre os torneios de futebol, o mais popular é o torneio da copa do mundo de futebol masculino, que é organizado pela *Fédération Internationale de Futeball Association* (FIFA), que é disputado por 32 seleções mundiais. Por mais que o objetivo do futebol seja sempre um resultado positivo, e que marcar mais gols do que o número de gols sofridos seja a maneira de obter o resultado, não existe uma tática reconhecida como superior a todas as outras. A aprendizagem de máquinas, através da aprendizagem estatística, está cada vez mais presente nas tecnologias e serviços que utilizamos para estudar e prever comportamentos. A aprendizagem de máquinas se apoia na determinação de modelos de comportamento, ou algoritmos, com o objetivo de aplicá-los de forma útil às novas situações. Neste trabalho serão considerados os seguintes modelos, ou métodos: Árvore de Decisão, LASSO e MARS. No modelo de árvore de decisão, para maior precisão, várias árvores serão usadas, por meio de métodos de conjunto, em especial os métodos: ensacamento, floresta aleatória e o *boosting*. O objetivo do presente trabalho é detectar, através da aplicação de métodos de Aprendizagem Supervisionada, de regressão e classificação, quais variáveis foram as mais importantes para que uma seleção conquistasse um resultado positivo, em cada uma das 64 partidas disputadas na Copa do Mundo de 2018, utilizando os dados disponíveis no site da FIFA. Para cada um desses jogos, tem-se 40 variáveis potencialmente explicativas. Modelos de *shrinkage* e redução de dimensionalidade serão considerados para lidar com esse conjunto e é esperado que os métodos convirjam para grupos de variáveis similares. Os métodos em geral tiveram problemas para classificar empates, o que é esperado, conforme trabalhos de assuntos similares. A variável ofensiva chutes no gol e as variáveis defensivas desarmes e chutes bloqueados foram as que apareceram como mais significativas. O modelo MARS, com essas três variáveis, conseguiu 57,8125% de acerto de resultado do jogo.

Palavras-chave: Aprendizado de Máquina. Árvore de Decisão. LASSO. MARS.

ABSTRACT

Among football tournaments, the most popular is the men's world cup tournament, which is organized by the Fédération Internationale de Futeball Association (FIFA), which is disputed by 32 world-class teams. Although the goal of football is always a positive result, and that scoring more goals than the number of goals conceded is the way to obtain the result, there is no tactic recognized as superior to all others. Machine learning through statistical learning is increasingly present in the technologies and services we use to study and predict behavior. Machine learning relies on the determination of behavioral models, or algorithms, in order to apply them in a useful way to new situations. In this work will be considered the following models, or methods: Decision Tree, LASSO and MARS. In the decision tree model, for more accuracy, several trees will be used, by means of assembly methods, especially the methods of *Random Forest* and of *boosting*. The objective of the present study is to detect, through the application of methods of Supervised Learning, regression and classification, which variables were the most important for a selection to achieve a positive result in each of the 64 matches played in the 2018 World Cup, using the data available on the website of the FIFA. For each of these games, there are 40 potentially explanatory variables. Shrinkage and dimensionality reduction models will be considered to deal with this set, and the methods are expected to converge to similar groups of variables. The methods in general had problems to classify draws, which is expected, according to similar works. The offensive variable attempts on goal and the defensive variables disarms and blocked kicks were the ones that appeared as most significant. The MARS model, with these three variables, achieved 57.8125% of correct games result.

Keywords: Machine Learning. Decision Tree. LASSO. MARS.

SUMÁRIO

1. INTRODUÇÃO

Nas últimas décadas, diante do exponencial crescimento tecnológico da humanidade, a Estatística passou a ter um papel cada vez mais relevante em diversos campos, entre esses o esportivo. Esportes como o beisebol foram revolucionados quando times passaram a utilizar a Estatística como base para as suas contratações e definição de seus modelos de jogos e, dessa forma, conquistar bons resultados (THORN; PALMER, 2015), (ALBERT, 2017).

O futebol é outro exemplo importante. Trata-se de um esporte de alcance mundial admirado por pessoas dos mais diversos perfis e poder aquisitivo e, por diversas razões, tem se aliado à ciência e à tecnologia, sendo o uso da estatística uma ferramenta crescente e de destaque. Estudos como os da Universidade de Stanford (TIMMARAJU; PALNITKAR; KHANNA, 2013), (ULMER E HERNANDEZ, 2014), foram realizados para prever os resultados dos jogos no Campeonato inglês (EPL, *English Premier League*), usando algoritmos de inteligência artificial e de aprendizado de máquina (do inglês *Machine Learning*). É oportuno mencionar que Ulmer e Hernandez (2014) citam em seu trabalho o estudo realizado por Joseph et *al*., na década de 90, que utilizaram as redes bayesianas para prever os resultados do time de futebol inglês *Tottenham Hotspur*, no período de 1995-1997.

Entre os torneios de futebol, o mais popular é a Copa do Mundo de futebol masculino, que é organizada pela *Fédération Internationale de Futeball Association* (FIFA), com sede na Suíça. Atualmente, o torneio é composto por 32 seleções, que representam as seis confederações mundiais. A Copa ocorre a cada quadriênio, estando composta por eliminatórias que definem 31 participantes e o 32º participante é a seleção do país que organiza o campeonato. A competição teve sua primeira edição em 1930 e sua 21ª em 2018, não tendo ocorrido apenas em 1942 e 1946, no período da Segunda Guerra Mundial (FIFA, 2018).

Por mais que o objetivo do futebol seja sempre um resultado positivo, e que marcar mais gols do que o número de gols sofridos seja a maneira de obter o resultado, não existe uma tática reconhecida como superior a todas as outras. É interessante notar, também, que as seleções consideradas favoritas têm maneiras tradicionais de jogar, todavia isso não lhes garante bons resultados. As seleções que venceram as copas de 2006, 2010 e 2014, respectivamente a Itália, a Espanha e a Alemanha, acabaram por ser eliminadas na primeira fase da copa seguinte (FIFA, 2018).

A previsão de eventos esportivos no futebol tem sido uma área de destaque, tanto devido à grande popularidade do futebol, como pela movimentação financeira que o esporte traz consigo. Não há dúvida de que a vertente da inteligência artificial indutiva, a aprendizagem de máquinas, está cada vez mais presente nas tecnologias e serviços que utilizamos para estudar e prever comportamentos. A aprendizagem de

máquinas se apoia na determinação de modelos de comportamento, ou algoritmos, com o objetivo de aplicá-los de forma útil às novas situações. Seu uso é frequente em campos muito diversos, com bom emprego na medicina, na engenharia, na agricultura e nas finanças e, também, com boa aplicabilidade no esporte (BELL 2015; MARR, 2018).

Enquanto o reconhecimento de padrões, ou a determinação de um modelo comportamental, tem suas origens na engenharia, o aprendizado de máquina cresceu fora da ciência da computação. No entanto, essas atividades podem ser vistas como duas facetas do mesmo campo, e juntas elas passaram por um desenvolvimento substancial nos últimos anos (adaptado de BISHOP, 2006).

Existem vários algoritmos diferentes empregados na aprendizagem de máquina e, usualmente, o tipo de resposta ou a saída desejada é o que define qual deve ser utilizado. Esses algoritmos usualmente se enquadram em um dos dois tipos de aprendizado: supervisionado ou não supervisionado (HASTIE, TIBSHIRANI, FRIEDMAN, 2009; BELL, 2015). James Le (2016) classifica os algoritmos de aprendizado de máquina em, pelo menos, três categorias: aprendizado supervisionado, não supervisionado e por reforço. Para o autor o aprendizado supervisionado é útil nos casos em que uma propriedade está disponível para um determinado conjunto de dados, mas ausente em outros e precisa ser prevista para essas outras instâncias. Já o não supervisionado é útil nos casos em que o desafio é descobrir relacionamentos implícitos em um determinado conjunto de dados não rotulados. O aprendizado por reforço fica entre esses dois extremos - há alguma forma de realimentação (*feedback)* disponível para cada etapa, ou ação preditiva, mas nenhum rótulo ou mensagem de erro precisa.

Nos problemas de aprendizagem supervisionada, é útil distinguir entre problemas de regressão e classificação (GAMA, 2004). Usualmente os problemas de regressão são quantitativos e os de classificação são qualitativos, embora isto não seja uma regra (JAMES, 2013).

Existem vários métodos de estimação dos resultados e somente a prática dos diferentes modelos, bem como, a comparação dos resultados e o refino dos caminhos percorridos, podem estabelecer as melhores previsões de resultados.

O principal objetivo do presente trabalho é chegar a um modelo, tendo como ponto de partida uma amostra pequena (64 partidas) e muitas variáveis inicialmente (90 variáveis, ou preditores), que auxilie na compreensão de quais destas variáveis foram relevantes para definir o vencedor de cada partida. Ao se determinar quais são as mais importantes, também são reveladas quais não são. Removendo essas, isto é, colocando os seus coeficientes como 0 (zero), através do encolhimento (*shrinkage*) os a seleção de variáveis será feita. As duas técnicas mais conhecidas para encolher os coeficientes de regressão são a regressão Ridge e o LASSO (Least Absolute Shrinkage and Selection Operator). (JAMES *et al.* 2013).

Neste trabalho serão considerados os seguintes modelos, ou métodos: Árvore de Decisão (*Decision Tree*), LASSO e MARS *(Multivariate adaptive regression splines*). No modelo de árvore de decisão, para maior precisão, várias árvores serão usadas, por meio de métodos de conjunto, em especial os métodos: ensacamento (*bagging*) floresta aleatória (*Random Forest*) e o *Boosting*.

Assim, o objetivo do presente trabalho é detectar, através da aplicação de métodos de Aprendizagem Supervisionada, de Regressão e Classificação, quais variáveis foram as mais importantes para que uma seleção conquistasse um resultado positivo, em uma das 64 partidas disputadas na Copa do Mundo de 2018, considerando os dados disponíveis no site da FIFA.

2. REFERENCIAL TEÓRICO

A quantidade de dados disponíveis nos campeonatos, ou torneios de futebol, cresce de modo significativo, aumentando, por consequência, a necessidade do tratamento desses bancos de dados de forma adequada. Neste contexto, o aprendizado de máquinas é uma ferramenta que pode auxiliar satisfatoriamente através da determinação de modelos de comportamento, ou algoritmos.

2.1 Aprendizado De Máquina

Em 1959, Arthur Samuel definiu o Aprendizado de Máquina como:

“Um campo de estudo que dá aos computadores a capacidade de aprender sem serem programados explicitamente” (extraído de BELL, 2015).

Tom Mitchel, no livro *Machine Learning* (McGraw-Hill, 1997), expõe que o aprendizado de máquina ocorre quando: *“Um programa de computador aprende com a experiência **E,** em relação a alguma classe de tarefas **T**, sendo a medida de seu desempenho **P**”.*

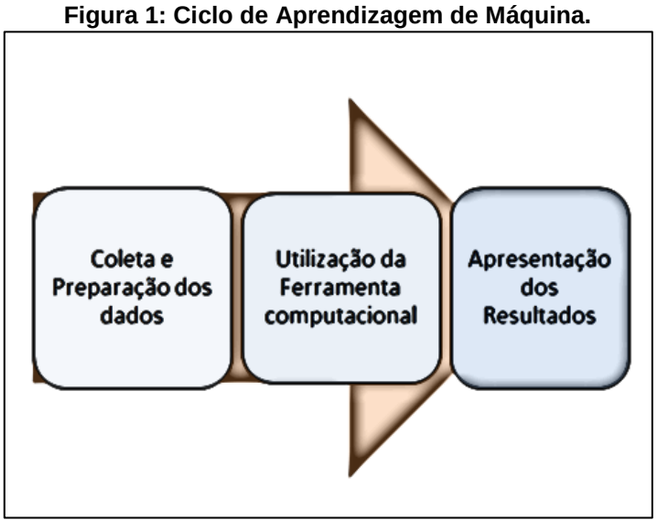
Para Nilsson (1998), o termo Aprendizado de Máquina geralmente é utilizado para se referir a mudanças em sistemas que desempenham tarefas associadas à inteligência artificial (IA) de forma indutiva.

O método é fundamentado no fato de que algumas tarefas são difíceis de serem definidas senão através de exemplos, aceitando que se possam utilizar agentes que podem se modificar, ao invés de agentes explicitamente projetados para uma tarefa específica. Os sistemas utilizados devem possuir a habilidade de se refinarem ao longo do tempo, operando para reduzir os erros e, assim, amortizar a necessidade de intervenções constantes. Dessa maneira o sistema computacional utilizado pode aprender e melhorar com a experiência e, com o tempo, estabelecer um modelo refinado que pode ser usado para prever resultados de perguntas com base na aprendizagem anterior. Assim, em sua essência, o aprendizado de máquina pode ser entendido como uma aplicação matemática computacional que forma previsões baseadas em propriedades conhecidas e aprendidas com dados de treinamento, encontrando padrões nos dados (NILSSON, 1998; HASTIE, TIBSHIRANI, FRIEDMAN, 2009; BELL, 2015).

A utilização do aprendizado de máquinas é vasta, sendo aplicável em desenvolvimento de produtos e serviços, em áreas de ciência e tecnologia (HASTIE, TIBSHIRANI, FRIEDMAN, 2009; BELL, 2015).

Basedo nos estudos de Bell (2015) o ciclo do aprendizado de máquinas pode ser visto reumidamente na Figura 1.

Figura 1: Ciclo de Aprendizagem de Máquina.



Fonte: Autor.

A aprendizagem estatística é uma abordagem do aprendizado de máquina que se refere a um amplo conjunto de ferramentas para entender os dados. Conforme James *et al*.(2013) é uma área da estatística que combina os desenvolvimentos em ciência da computação e o aprendizado de máquina. Um exemplo de aplicação, de aprendizagem estatística, pode ser como estimar o valor de uma função desconhecida, em um novo ponto, dados os valores desta função em um conjunto de pontos de uma amostra. (NILSSON, 1998).

Frequentemente essas ferramentas estatísticas podem ser classificadas como supervisionadas, ou não supervisionadas. Em geral, a aprendizagem estatística supervisionada envolve a construção de um modelo estatístico para a previsão, ou para estimar, uma saída baseada em uma ou mais entradas. Problemas desta natureza ocorrem em campos tão diversos como negócios, medicina, astrofísica e políticas públicas. Com o aprendizado estatístico não supervisionado, há entradas, mas sem saída de supervisão; no entanto, podemos aprender relacionamentos e estruturar tais dados (JAMES *et al.).*

2.1.1 Aprendizado de Máquina Supervisionado

O aprendizado de máquina supervisionado se refere a um conjunto de variáveis identificadas como entradas, que são medidas, ou predefinidas, que têm alguma influência sobre uma ou mais saídas. Assim, o objetivo é usar as entradas para prever os valores das saídas (HASTIE, TIBSHIRANI, FRIEDMAN, 2009).

A aprendizagem supervisionada possui como objetivo a construção de um modelo (classificador) capaz de aprender o mapeamento entre os valores das variáveis independentes, ou preditoras (*features*) e o valor da variável dependente (classes ou *labels*) das instâncias contidas em um conjunto de treinamento, tal que o classificador resultante possua um poder de generalização suficiente para ser utilizado para prognóstico de instâncias não vistas antes (SCHNEIDER, 2018).

Nos problemas de aprendizagem supervisionada, o que distingue um problema de regressão de um problema de classificação é a natureza das variáveis, que podem ser quantitativas, ou qualitativas (categóricas). Sendo as variáveis quantitativas as que assumem normalmente valores numéricos (ex: idade ou valor de uma casa). E as variáveis qualitativas as que adotam, usualmente, categorias diferentes de números (ex: a marca do produto comprado, isto é, A, B ou C). Para os autores, de um modo geral, problemas com uma resposta quantitativa são chamados como problemas de regressão, enquanto que os que envolvem uma resposta qualitativa são frequentemente referidos como problemas de classificação. No entanto, a distinção nem sempre é tão nítida, a regressão linear de mínimos quadrados é usada como uma resposta quantitativa, enquanto que a regressão logística é usada como uma resposta qualitativa (JAMES *et al.* 2013).

Para Matos (2018) a classificação é o processo que ocorre quando se atribui um rótulo a entrada escolhida. Para o autor estes sistemas são usados quando as previsões são de natureza distinta, ou seja, um “sim ou não”, exemplificando temos o mapeamento de uma imagem de uma pessoa e sua classificação como sexo masculino ou feminino. Ainda conforme este autor a categoria de regressão ocorre quando o valor que está sendo previsto difere de um “sim ou não” e segue um espectro contínuo. Sistemas de regressão poderiam ser usados, por exemplo, para responder às perguntas: “Quanto custa?” ou “Quantos existem?” (MATOS, 2018). A regressão é uma técnica que permite explorar e inferir a relação de uma variável dependente (variável de resposta) com as variáveis independentes específicas (variáveis explanatórias). Um dos objetivos da análise de regressão é estimar os parâmetros desconhecidos do modelo (SELAU, 2011; VIDHYA, 2016).

Um modelo de regressão linear assume que a função de regressão E (Y | X) é linear nas entradas X_1,. . . Xp. Modelos lineares são simples e são capazes de fornecer uma descrição adequada e interpretável de como as entradas afetam a saída (HASTIE, TIBSHIRANI, FRIEDMAN, 2009).

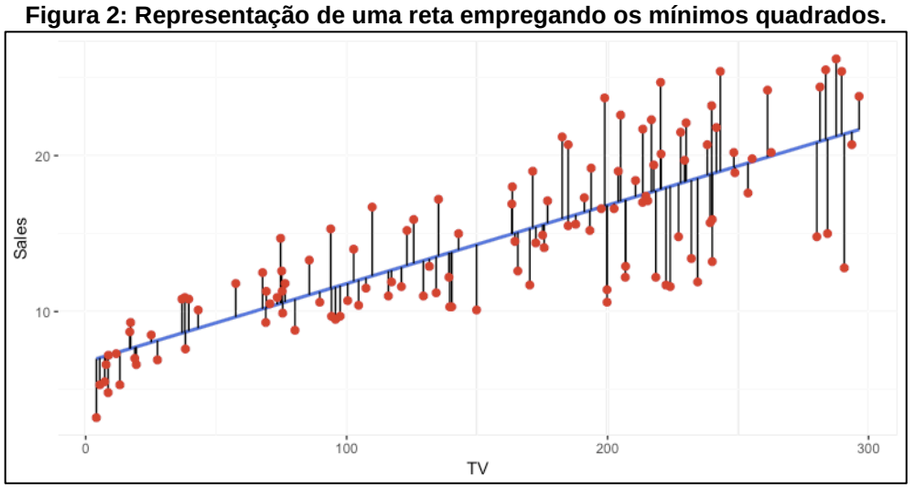
A equação que representa o modelo de regressão linear é uma reta e pode ser representada por:

Equação1:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

A ilustração a seguir mostra a representação de uma reta, obtida pelo meio da linha de “melhor ajuste” encontrada, através da minimização da soma dos erros quadrados (representados pelos segmentos de linhas pretas verticais).

Figura 2: Representação de uma reta empregando os mínimos quadrados.



Fonte: http://uc-r.github.io/public/images/analytics/regression/sq.errors-1.png

A equação que representa o modelo de regressão linear multifatorial pode ser representada por:

Equação 2:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

2.2 LOGIT ORDENADO

A regressão logística ordenada ou modelo *logit* ordenado é um modelo de regressão para variáveis dependentes ordinais, ou qualitativas. Exemplos de várias categorias de respostas ordenadas incluem classificações de títulos, pesquisas de opinião com respostas variando de "concordo totalmente" até "discordo totalmente", níveis de gastos estatais em programas governamentais (alto, médio ou baixo), e no caso do presente trabalho, resultado da partida da Copa do Mundo (derrota, empate ou vitória). Esses resultados são todos os que podem ocorrer em uma partida e são mutuamente excludentes para cada time.

Dados ordinais costumam ser dicotomizados para aplicar métodos estatísticos válidos para comparar duas classes, porém em alguns casos isso pode gerar uma grande de informação (HASTIE, TIBSHIRANI, FRIEDMAN, 2010). No caso do futebol, dicotomizar entre vitórias e não vitórias perde a informação útil que é o empate, que pode ser o objetivo de um time em uma particular partida.

Neste modelo, estima-se a probabilidade de estar em uma categoria ou abaixo dela contra estar nas categorias acima. No caso, a probabilidade de derrota versus empate ou vitória, ou a probabilidade de derrota e empate versus vitória. Modelar vitória não faria sentido, porque sua probabilidade sempre seria 1 (um).

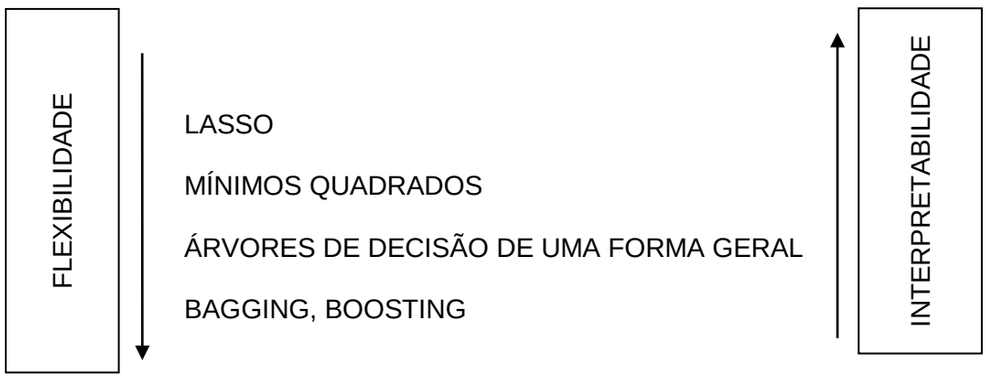
Uma das suposições do *logit* ordenado é que a relação entre cada par de grupos de resultados é a mesma. Assume-se que os coeficientes que descrevem a relação entre derrota e empate são os mesmos que descrevem a relação entre empate

e vitória. Como o relacionamento entre os pares de grupos é o mesmo, existe apenas um conjunto de coeficientes (apenas um modelo).

2.3 MODELOS DE ENCOLHIMENTO

Modelar a variável resposta é um desafio, alguns modelos de aprendizagem estatística se destacam pela flexibilidade, enquanto que outros pela interpretabilidade.

De acordo com James (2013) o modelo LASSO, por exemplo, está baseado em um modelo linear, mas usa um procedimento de ajuste alternativo para estimar os coeficientes $\beta_{0,}\, \beta_1, \ldots \beta_p,$ e define um número deles para exatamente zero. Esse método tem uma abordagem menos flexível do que a regressão linear, porém mais interpretável do que esse modelo. Os métodos não lineares como *bagging*, *boosting*, são flexíveis nas abordagens, porém mais difíceis de interpretar. A seguir uma representação de métodos usados na estatística de máquinas e a sua abordagem quanto à flexibilidade e facilidade de interpretação (JAMES, 2013).



Fonte: Autor, baseado em James (2013).

Dessa forma a combinação de métodos de maior flexibilidade, como os modelos das árvores de decisão com as suas variações, e os modelos de mais fácil interpretação como o LASSO serão os empregados neste trabalho.

Os modelos empregados são:

- LASSO - método de regressão.
- MARS - método de regressão.
- Árvore de Decisão - método de aprendizagem supervisionada de classificação e regressão. No modelo de árvore de decisão, para maior precisão, várias árvores serão usadas, por meio de métodos de conjunto, em especial os métodos: ensacamento, floresta aleatória e o *boosting*.

Os detalhamentos dos três modelos adotados no trabalho e seus desdobramentos estão descritos na sequência.

2.3.1 LASSO

Em Estatística e no Aprendizado de Máquina, LASSO é um método de análise de regressão que executa a seleção e regularização de variáveis para aumentar a precisão e a capacidade de interpretação do modelo estatístico produzido.

Introduzido na literatura geofísica em 1986, mais tarde redescoberto e popularizado por Robert Tibshirani, que criou o termo e forneceu mais *insights* sobre o desempenho observado. Conforme Tibshirani (1996), o LASSO é um método de encolhimento do conjunto de coeficientes, que tem como objetivo estimar um modelo, que determine o conjunto de preditores que melhor expliquem a variável resposta e produzam previsões com pequena variância.

Considerando consistência dos modelos do LASSO, é importante saber se a solução representa bem o modelo, ou seja, conseguir selecionar o subconjunto correto de variáveis relevantes e seus parâmetros que devem ser assintoticamente como dos estimadores Mínimos Quadrados Ordinários (MQO), o que só acontecerá se o subconjunto de variáveis escolhido for o das variáveis relevantes para o modelo (KONZEN, ZIELGELMANN, 2016).

As estimativas são obtidas através da minimização dos quadrados dos resíduos, reduzindo os coeficientes irrelevantes a zero. Conforme James (2013), as duas melhores técnicas de encolhimento para redução dos coeficientes a zero são a regressão Ridge e o LASSO.

A equação que representa o modelo LASSO pode ser representada por:

Equação 3:

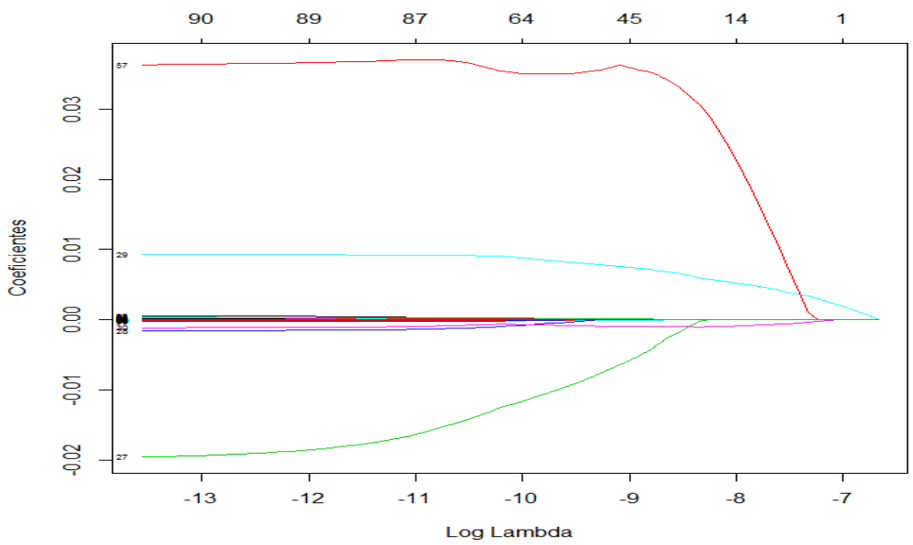
$$\hat{\beta} = argmin_{\hat{\beta}} || Y - X\beta|_2^2 \ + \ \lambda \sum_{j=1}^{p} |\beta_j|$$

Onde β é $um\ vetor\ NX1$

Sendo Y a variável resposta, X uma matriz (pxn) com as variáveis preditoras e λ o parâmetro de encolhimento. Quando λ = 0, as estimativas são as mesmas que os Mínimos Quadrados Ordinários (MQO), sendo o λ é estabelecido por validação cruzada. Essa técnica parte a amostra original em K subamostras de tamanho igual, uma das K subamostras é retirada e o modelo é estimado com K-1 subamostras.

A figura 3 mostra o efeito de λ no eixo x sobre as estimativas dos coeficientes no eixo y, onde cada linha representa o valor do coeficiente de uma variável diferente. No eixo x superior do gráfico temos o número de variáveis. O número de variáveis selecionadas e o tamanho dos coeficientes diminuem à medida que aumenta o λ.

Figura 3: Relação entre λ os coeficientes.



Fonte: Medeiros *et al.* (2016).

O LASSO é atualmente uma usual técnica de regularização, trabalhando, essencialmente, nos casos em que existem muitos coeficientes nulos dentro do conjunto de coeficientes a serem estimados, com o objetivo de selecionar as variáveis preditoras que melhor explicam a variável resposta e construir previsões com menor variância (SILVEIRA, 2016).

Zhao e Yu (2006) consideram dois problemas: 1) se existe uma quantidade determinística de regularização que fornece consistência de seleção; 2) se para cada amostra existe uma quantidade correta de regularização que seleciona o melhor modelo. Estes resultados mostram que existe uma condição que denominam por “Condição Irrepresentável” que é quase necessária e suficiente para ambos os tipos de consistência. Estes resultados são válidos para modelos lineares.

Conforme Tibshirani (1996), uma interpretação bayesiana do porque o LASSO estima coeficientes de variáveis como 0 é que este pode ser interpretado como uma regressão linear para a qual os coeficientes têm distribuições à priori de Laplace. A distribuição de Laplace tem sua massa de probabilidade mais concentrada no zero do que a distribuição normal. Isso reflete a tendência do LASSO de zerar coeficientes de variáveis.

2.3.2 MARS

Na estatística, MARS é uma forma de análise de regressão introduzida por Jerome H. Friedman, da Universidade de Stanford, no início da década de 90. É uma técnica de regressão não paramétrica e pode ser vista como uma extensão de modelos lineares que automaticamente modelam a não linearidade e as interações entre variáveis.

De um modo introdutório, no MARS se tem um algoritmo que cria essencialmente um modelo linear por partes que fornece um bloco intuitivo de degrau para a não linearidade. Regressão adaptativa multivariada de splines fornece uma abordagem conveniente para capturar o aspecto de não linearidade da regressão polinomial, avaliando pontos de corte (*nós*) semelhantes às funções de etapa. O procedimento avalia cada ponto de dados para cada preditor como um nó e cria um modelo de regressão linear com o(s) recurso(s) candidato(s) (http://uc-r.github.io/mars).

O termo "MARS" é registrado e licenciado para a Salford Systems (http://www.salfordsystems.com) e pode ser usado como uma abreviação, mas, no entanto, ele não pode ser usado para soluções de software concorrentes. É por isso que o pacote R usa o nome *Earth* (Bell, 2015).

Através do MARS se constrói uma relação a partir de um conjunto de coeficientes e funções de base que são "impulsionados" a partir dos dados de regressão. Em certo sentido, o método é baseado na estratégia "dividir e conquistar", que divide o espaço de entrada em regiões, cada uma com sua própria equação de regressão. Isto torna os MARS particularmente adequados para problemas com maiores dimensões de entrada (isto é, com mais de 2 variáveis), onde a dimensionalidade provavelmente criaria problemas para outras técnicas.

A equação geral do modelo MARS é dada como:

Equação 4:

$$y = f(X) = \beta_0 + \sum_{m-1}^{M} \beta_m \, h_m \, (X)$$

Onde *y* é previsto como uma função das variáveis preditoras *X* (e suas interações). Essa função consiste em um parâmetro de interseção β_0 e a soma ponderada dos pesos $\beta_m \, d$e uma, ou mais funções básicas. Também se pode pensar neste modelo como uma soma ponderada de funções de base h do conjunto que abrange todos os valores de cada preditor (ou seja, esse conjunto consiste em uma função de base h e parâmetro *t*, para cada valor distinto para cada variável preditora). O algoritmo MARS, então, pesquisa o espaço de todas as entradas e valores

preditores (localizações dos nós t), bem como as interações entre as variáveis. A função h será definida com limite em t, ou seja, ou para valores menores do que t, ou para valores maiores. Como ela é 0 para parte do intervalo, pode ser usada para dividir os dados em regiões disjuntas, cada uma das quais pode ser tratada de forma independente. Durante esta busca, um número cada vez maior de funções de base é adicionado ao modelo, para maximizar um critério geral de adequação de mínimos quadrados. Como resultado destas operações, o MARS determina automaticamente as variáveis independentes mais importantes, bem como as interações mais significativas entre elas. (*http://www.statsoft.com/Textbook/Multivariate-Adaptive-Regression-Spline).*

O MARS constrói um modelo em dois estágios: o *forward* e o *backward*. Essa abordagem é a mesma que é usada nas árvores de decisão. O modelo começa com a média da variável de saída e vai adicionando funções base com um algoritmo ganancioso com base na redução máxima de erro residual da soma dos quadrados. O processo continua até que a alteração no erro residual seja muito pequena para continuar, ou até que o número máximo de termos seja atingido. Como o *forward* cria um sobre ajuste, após o mesmo é feito o *backward*, que tem a vantagem de ver o modelo completo, e não o modelo atual adicionando apenas uma variável de cada vez. Os termos são excluídos um a um, sempre retirando o menos eficaz até encontrar o melhor submodelo. (HASTIE, TIBSHIRANI, FRIEDMAN, 2009).

A deficiência dos modelos MARS é a interpretação da variável resposta como uma variável contínua que teve apenas como valores observados -1, 0 e 1 no caso do presente trabalho. Portanto as previsões do MARS não ficam restritas nesses valores e também não podem ser interpretadas como probabilidades diretamente. Esses problemas são contornados considerando-se no ranking das classificações. A classificação é feita para a classe com o maior valor de resposta previsto. (HASTIE, TIBSHIRANI, FRIEDMAN, 2009).

MARS considera as interações entre todas as variáveis disponíveis. Em relação à modelagem do MARS, ela é hierárquica, no sentido de que o efeito de interação só é incluído no modelo se uma das variáveis que estará presente na interação já estiver no modelo. A ideia é que a interação só pode ser significativa se pelo menos uma das variáveis também tivesse um impacto. Essa é uma suposição razoável para que não seja procurado por um número exponencial de alternativas. (HASTIE, TIBSHIRANI, FRIEDMAN, 2009).

Além disso, outra restrição do MARS é que cada variável pode aparecer apenas uma vez na interação. Isso impede que a variável interaja consigo mesma, ou seja, evita que a variável entre no modelo como quadrática ou de ordem superior. (HASTIE, TIBSHIRANI, FRIEDMAN, 2009).

2.3.3 Árvore de Decisão

Uma árvore de decisão é um mapa dos possíveis resultados de uma série de escolhas relacionadas. Uma árvore de decisão pode ser usada para ajudar a criar modelos preditivos automatizados, que têm aplicações em pesquisa de dados e em estatísticas. Esse tipo de árvore é conhecido como uma árvore de classificação, onde cada ramificação contém um conjunto de atributos, ou regras de classificação, associado a um determinado rótulo de classe encontrado na extremidade da ramificação. Essas regras, conhecidas como regras de decisão, podem ser expressas em uma cláusula de "se → então", onde cada decisão, ou valor de dados, forma uma cláusula, de tal forma que, por exemplo, "se as condições forem cumpridas, então o desfecho x será o resultado, com uma certeza de y". Cada pedaço de dados adicional ajuda o modelo a prever, com mais precisão, a qual conjunto finito de valores o sujeito em questão pertence. Esta informação pode então ser usada em um modelo maior de tomada de decisão. Árvores de decisão com resultados contínuos são chamadas de árvores de regressão (LUCIDCHART).

No contexto de problemas de classificação, os algoritmos que geram árvores multivariadas são capazes de explorar várias linguagens de representação usando testes de decisão baseados em uma combinação de atributos (GAMA, 2004).

Árvore tem tendência de sobre ajuste, ou seja, ajusta-se muito bem aos dados, mas não é muito eficaz para fazer previsões. Para contornar esse problema, serão utilizadas duas abordagens. Uma é a poda da árvore e a outra são métodos de conjunto. Na poda, que é feita depois que o treinamento do aprendizado de maquinas é concluído, retiram-se os galhos da árvore, ou seja, remove os nós de decisão a partir do nó da folha, de modo que a precisão geral não seja perturbada. Para maior precisão, às vezes várias árvores são usadas juntas por meio de métodos de conjunto. A seguir exemplos de árvores que podem ser utilizadas para aumentar a precisão:

2.3.3.1 Modelos de árvore de decisão

- **Ensacamento -** se refere à criação de várias árvores para modelar os dados da fonte e, em seguida, a partir dessas chegar a uma árvore de consenso.
- **Floresta aleatória** - é composta por várias árvores concebidas para aumentar a taxa de classificação.
- ***Boosting*** - podem ser utilizadas para árvores de regressão e de classificação (JAMES et al. 2013; LUCIDCHART).

O ensacamento é um procedimento de árvores de decisão frequentemente empregado para reduzir a variância de um método estatístico de aprendizagem. É feito o *bootstrap*, pegando amostras repetidas do conjunto de dados. São gerados conjuntos diferentes e nesses conjuntos o método é treinado. Após, é feita a média das predições. Sendo um conjunto de n observações independentes $Z_1, \ldots, Z_n$, cada uma com variância σ^2, então a variância da média das observações é dada por σ^2/n.

Equação 5 :

$$\hat{f}avg(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^b\,(x)$$

A floresta aleatória é feita de tal maneira que evita que as árvores fiquem correlacionadas, o que ocorria no ensacamento. Em cada divisão da árvore o algoritmo não pode considerar a maioria dos preditores disponíveis. Como não teremos a possibilidade de fazer diversas árvores novas altamente correlacionadas, o que não reduz muito a variância. Dado isso, a média das árvores resultantes é menos variável e portanto, mais confiável.

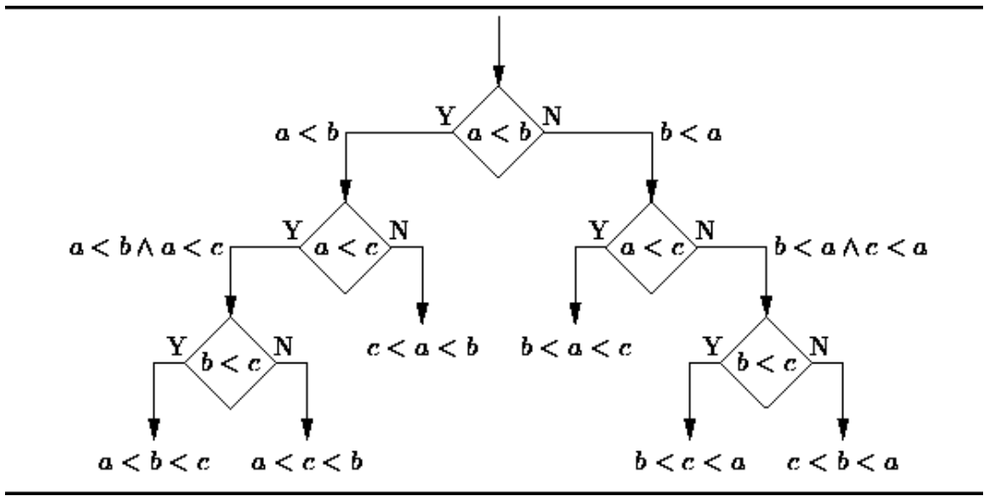
Boosting são similares ao ensacamento, mas com as árvores elaboradas sequencialmente. Cada árvore nova é cultivada usando informação das que foram geradas anteriormente, e em vez de *bootstrap*, elas gradativamente são ajustadas em uma versão modificada do conjunto de dados original.

2.3.3.2 Representação de uma árvore de decisão

Uma árvore de decisão é composta por nós, onde cada nó está associado a uma das entradas das variáveis. As arestas desses nós são os valores possíveis do nó. A folha representa o valor obtido com base nos valores fornecidos a partir da variável de entrada que vai do nó raiz para a folha.

A Figura 4 ilustra um esquema para árvore de decisão.

Figura 4: Árvore de Decisão.



Fonte: https://cdn-images-1.medium.com/max/600/1*dUlGB4eoqXG4MdKsykvFKg.gif.

O algoritmo para cultivar a árvore segue a abordagem padrão de divisão e conquista. Os aspectos mais relevantes são: a regra de divisão, o critério de terminação e o critério de atribuição de folha (GAMA, 2004). O algoritmo padrão para construir árvores univariadas tem duas fases. Na primeira fase, uma grande árvore é construída. Na segunda fase, esta árvore é podada de volta.

2.3.3.3 Possibilidades e limites no emprego de árvores de decisão

De acordo com Bell (2015) existem algumas boas razões para usar árvores de decisão. Por um lado, eles são fáceis de ler e, depois que um modelo é gerado, é fácil relatar a outros sobre como a árvore funciona. Além disso, com árvores de decisão, é possível manipular informações numéricas (árvore de regressão), ou categorizadas (árvore de classificação). A preparação dos dados não é trabalhosa, ainda que se tenha um grande conjunto de dados.

Usar árvores de decisão em aprendizado de máquinas tem vantagens como:

- Funciona para variáveis qualitativas e quantitativas;
- Modela problemas com várias saídas;
- A confiabilidade de uma árvore pode ser testada e quantificada;
- Tende a ser preciso, independentemente da possibilidade de violar os pressupostos de dados de origem (LUCIDCHART).

Mas também tem algumas desvantagens, tais como:

- Ao lidar com dados categóricos com vários níveis, o ganho de informação é tendencioso em favor dos atributos com mais níveis.
- Cálculos podem tornar-se complexos quando se lida com a incerteza e com muito resultados vinculados.

- Conjunções entre nós estão limitadas a “E”, enquanto gráficos de decisão permitem nós ligados por “OU” (LUCIDCHART).

Segundo Bell (2015), uma das principais questões das árvores de decisão é que elas podem criar modelos complexos, dependendo dos dados apresentados no conjunto de treinamento. Para evitar que o algoritmo de Aprendizado de Máquina sobreponha os dados, às vezes é necessário rever os dados de treinamento e podar os valores para as categorias, o que irá produzir um modelo mais refinado e melhor ajustado. Alguns dos conceitos da árvore de decisão podem ser difíceis de aprender porque o modelo não os expressa facilmente. Esta falha às vezes resulta em uma árvore maior do que o desejado.

3 METODOLOGIA

No site da FIFA estão disponíveis todas as estatísticas de cada um dos jogos da Copa de 2018. Cada uma das páginas dos 64 jogos foi aberta e os dados interessantes para o trabalho foram anotados em uma planilha. Foram selecionadas as estatísticas usuais como as de ataque (finalizações), performance (posse de bola, distância percorrida), defesa (desarmes) e disciplina (cartões recebidos). Além de estatísticas típicas, também estavam disponíveis dados de rastreamento sobre as seleções, incluindo posse de bola em diversos setores do campo e por onde foram feitos os ataques de cada uma das equipes. Também se encontram estatísticas sobre o clima e horário do jogo. Dados complementares como situação da partida no intervalo, prorrogação e disputa por pênaltis também são encontrados.

Após o banco estar completo, foi criada uma nova variável, a diferença entre a variável do mandante e do visitante para cada jogo. Deste modo cada variável que está no modelo gerado pelo trabalho envolve ambas as seleções.

Os termos mais utilizados no futebol estão discriminados, no Anexo 3.

3.1 Coletas dos Dados

Os dados foram coletados no site da FIFA, totalizando as 64 partidas que ocorreram na Copa de 2018 (Anexos 1 e 2). O banco é composto inicialmente por 90 variáveis (*inputs*). A Tabela 1 a seguir ilustra os fatos.

Tabela 1: Variáveis de Entrada (*Inputs)* - Situação Inicial.

	Variáveis	Mandante e Visitante
Ataque	2	Gols
	4	Finalizações
	6	Tentativas no alvo
	8	Tentativas fora do alvo
	10	Tentativas bloqueadas
	12	Escanteios
	14	Impedimentos
	16	Tentativas na trave
Performance	18	Posse de bola
	20	Precisão nos Passes
	22	Passes

Continua

Continua

	24	Passes completados
	26	Distancia Percorrida
Defesa	28	Bolas recuperadas
	30	Carrinho
	32	Bolas afastadas
	34	Bloqueios
Disciplina	36	Cartões Amarelos
	38	Cartões Vermelhos Diretos
	40	Cartões Vermelhos Indiretos
	42	Faltas cometidas
Momento do Gol	44	0- 15 minutos
	46	15- 30 minutos
	48	30-45 minutos
	50	45-60 minutos
	52	60-75 minutos
	54	75-90 minutos
Posse de bola - Setor do Campo	56	Defesa - lado direito
	58	Defesa - centro
	60	Defesa - lado esquerdo
	62	Meio de campo - lado direito
	64	Meio de campo - centro
	66	Meio de campo - lado esquerdo
	68	Ataque - pelo lado direito
	70	Ataque - pelo centro
	72	Ataque - pelo lado esquerdo
Setor de Ataque Predominante	74	Direito
	76	Centro
	78	Esquerdo
Bola parada	80	Pênaltis marcados
	82	Pênaltis convertidos
	84	Gols de falta
Clima	85	Tipo de clima – nublado, ensolarado, noite
	86	Temperatura
	87	Umidade
	88	Vento
Situação da partida no intervalo	89	Ganhando no intervalo
Prorrogação	90	Teve prorrogação no jogo

Fonte: https://www.fifa.com/worldcup.

3.2 Elaboração do Banco de Dados

Como fase de preparação dos dados foi testado um Modelo de Regressão com as noventa variáveis iniciais, tendo como resultado 36 variáveis não definidas. Em sequência, como no modelo havia muitas variáveis, foi feita uma discussão para manter aquelas que poderiam ter impacto no modelo final. Além disso, decidiu-se excluir também as variáveis que causariam multicolinearidade e variáveis que dominariam o modelo, como por exemplo, as variáveis que se referem aos gols marcados. As variáveis do setor de ataque predominante, que não possuiam dados para todos os jogos e dependem de qual lado um time utiliza para atacar também foram excluídas do modelo.

Como resultados iniciais estavam considerando algumas variáveis apenas do mandante, ou do visitante o banco de dados foi remodelado. As variáveis de mandante e visitante foram transformadas na diferença entre ambas, visto que um modelo que não considerasse um dos times da partida não faria tanto sentido. Além delas, também foram testados modelos considerando o resultado da partida no intervalo, sob uma ótica de apostas esportivas, que podem ser feitas em qualquer momento da partida, tendo acesso a todas estatísticas que ocorreram no jogo até aquele momento. Essa variável também foi removida para testar a precisão do modelo sem ela e quais variáveis foram ou deixaram de ser escolhidas.

Observa-se que no presente trabalho o mandante é apenas a seleção sorteada pela FIFA, para jogar com o uniforme principal. Dado que todos os jogos da Copa foram realizados todos em um mesmo país, a Rússia, não é esperado que fator local seja algo significativo. Como é possível ver na Tabela 2 a seguir:

Tabela 2: Resultados dos mandantes na Copa.

	Vitória	Empate	Derrota
Mandante	26	13	25

Observa-se, na Tabela 2, que não parece existir diferença entre ser sorteado como mandante ou visitante com relação ao desfecho resultado da partida.

Apos o tratamento dos dados o banco foi composto da diferença das seguintes variáveis, mostradas na Tabela 3.

Tabela 3: *Inputs* - Situação após a eliminação das variáveis.

		Mandante e Visitante
	ATAQUE (7)	Finalizações
		Tentativas no alvo
		Tentativas fora do alvo
		Tentativas bloqueadas
		Escanteios
		Impedimentos
		Tentativas na trave
	PERFORMANCE (5)	Posse de bola
		Precisão de Passes
		Passes
		Passes completados
		Distância Percorrida
	DEFESA (4)	Bolas recuperadas
		Carrinho
		Bolas afastadas
		Bloqueios
	DISCIPLINA (4)	Cartões Amarelos
		Cartões Vermelhos Diretos
		Cartões Vermelhos Indiretos
		Faltas cometidas
	Variáveis por jogo	40
	Número de Jogos	64

Seguindo um modelo multinomial ordenado, a variável dependente teve os seguintes valores atribuídos:

Tabela 4: Variável multiclasse - valores atribuídos ao mandante.

	Valor Atribuído
Vitória do mandante	1
Empate	0
Derrota do mandante	-1

Observação: Na Copa do Mundo o mandante é o time sorteado pela FIFA.

Todos os modelos utilizaram essa variável como resposta.

3.3 Modelagem / Metodologia Usada

- LASSO - O modelo Lasso foi implementado com o pacote *glmnetcr*.
- MARS - Para fazer o modelo MARS utilizou-se o pacote *earth* do R. Utilizou-se grau de interação 2, para que fossem consideradas as interações entre as variáveis preditoras. O método de seleção foi o *backward,* visto que o “forward” não pode ser feito com modelos de respostas múltiplas.
- Árvore de Decisão - Para fazer a árvore foi usado o pacote *tree*. Foi feita uma árvore de classificação considerando que a saída era vitória, empate e derrota. Após a árvore ser feita, ela foi podada considerando o critério de classificações incorretas. Para fazer o Ensacamento e a Floresta Aleatória foi utilizado o pacote R*andom Fores*t e para o *Boosting* o pacote *gbm* foi utilizado.

4 RESULTADOS

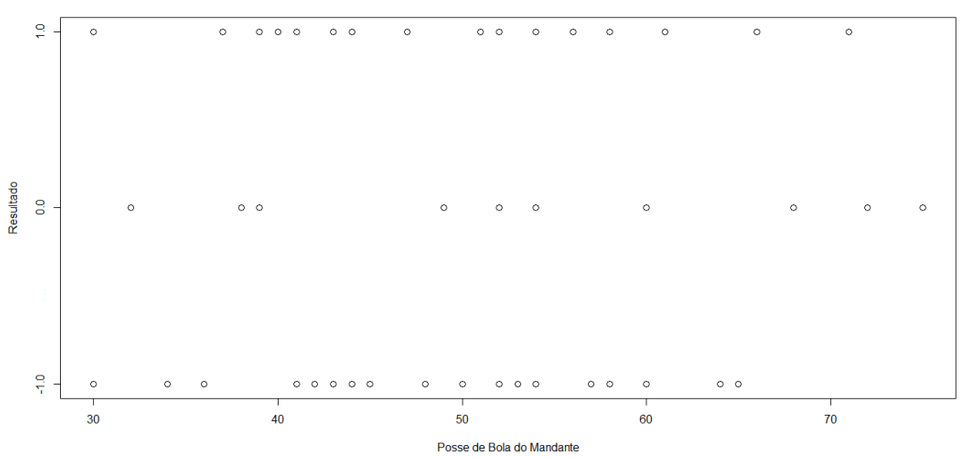
4.1 Posse de Bola *x* Resultado Obtido

Uma das estatísticas mais clássicas para expressar domínio de jogo é a posse de bola e é comumente o foco de algumas seleções, como a da Espanha. Na Copa de 2018, a Espanha teve 1 vitória e 3 empates. Em todos os jogos a Espanha teve bastante superioridade na posse de bola em relação ao adversário, mas só conseguiu uma vitória. A Rússia, uma seleção tecnicamente inferior à Espanha, teve menos posse de bola que seus adversários nos cinco jogos que disputou. Isso não a impediu de conseguir um dos melhores resultados da Copa, 5x0 contra a Arábia Saudita e ainda, eliminar a Espanha nas Oitavas de Final.

A seguir temos um gráfico que representa a posse de bola nos 64 jogos. A posse de bola está representada no eixo **x** e o resultado da partida no eixo **y**.

Pode-se ver que não existe nenhuma relação aparente entre as duas variáveis, visto que times venceram com posses de bola 28(54,9%) e perderam 23(45,1%) das partidas que tiveram um vencedor. No jogo entre Espanha e Rússia, a Espanha teve 75% de posse de bola e o resultado final foi empate. Uma regressão entre as duas variáveis não é significativa (p-valor = 0,444). Pode-se concluir que posse de bola isoladamente não é um fator significativo para determinar o vencedor de uma partida.

Figura 5: posse de bola X resultado do jogo.



4.2 Árvore de Decisão

4.2.1 Árvore de classificação

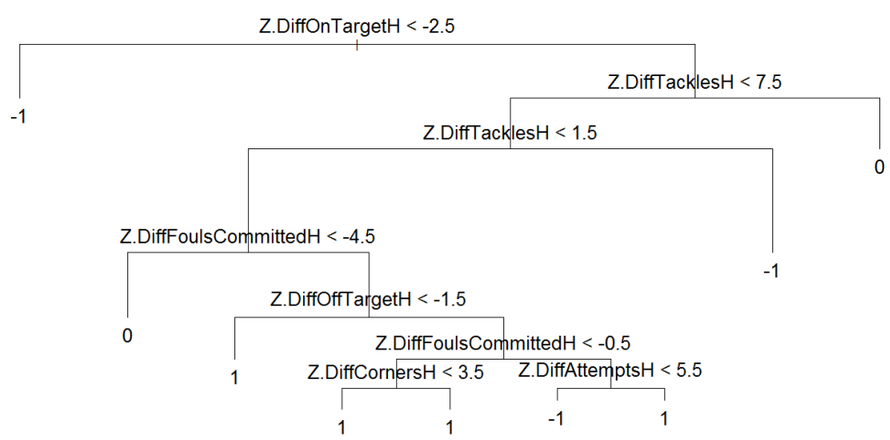
A árvore de decisão não utilizou a variável do resultado da partida no intervalo no modelo. Por isso a árvore não será refeita sem essa variável no banco de dados.

No modelo de árvore de decisão como uma árvore de classificação as variáveis utilizadas foram:

- "DiffOnTarget" (diferença de chutes no alvo),
- "DiffTackles" (diferença de desarmes),
- "DiffFoulsCommitted" (diferença de faltas cometidas),
- "DiffOffTarget" (diferença de chutes fora do alvo) ,
- "DiffCorners" (diferença de escanteios) e
- "DiffAttempts" (tentativas).

Como árvore inicial do modelo, a diferença de chutes no gol foi a variável mais importante para determinar o vencedor de uma partida na Copa, de modo que quando essa diferença foi superior a 2,5, ou seja, o visitante chutou pelo menos 3 vezes mais, o algoritmo colocou o mandante como perdedor. Esta árvore teve 9 nós e errou 16(25%) das 64 classificações. Como estamos com uma árvore de classificação, assume-se que cada observação pertence à classe mais comum de ocorrência. Pode-se ver que a diferença entre escanteios não é de fato utilizada, dado que independente do valor da mesma a árvore dirá que o time ganha o jogo. Vale ressaltar que a variável diferença de escanteios não está levando a decisões diferentes.

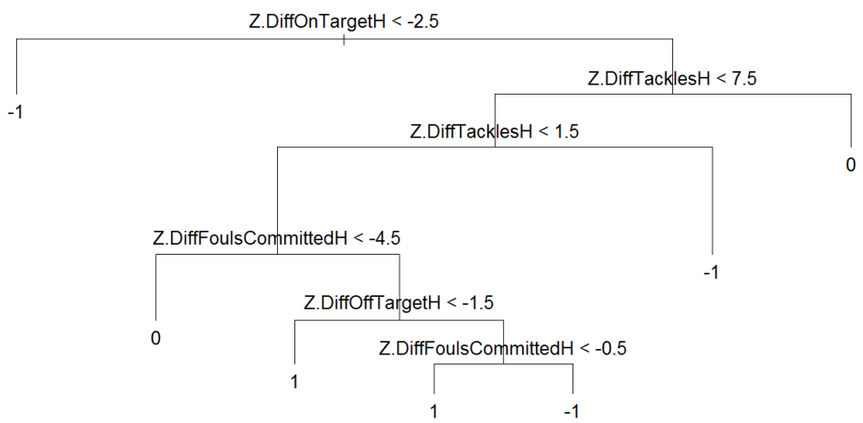
Figura 6: Árvore de classificação.



Após ser podada, a árvore resultante teve 7 nós e 17(26,56%) de classificações erradas. Dois nós a menos do que a árvore inicial e um acerto de resultado a menos do que a mesma. Além disso, essa árvore utilizou a diferença entre 4 variáveis ao invés de 6. As variáveis selecionadas foram:

- DiffOnTarget,
- DiffTackles
- DiffFoulsCommitted e
- DiffOffTarget.

Figura 7: Árvore de classificação podada.



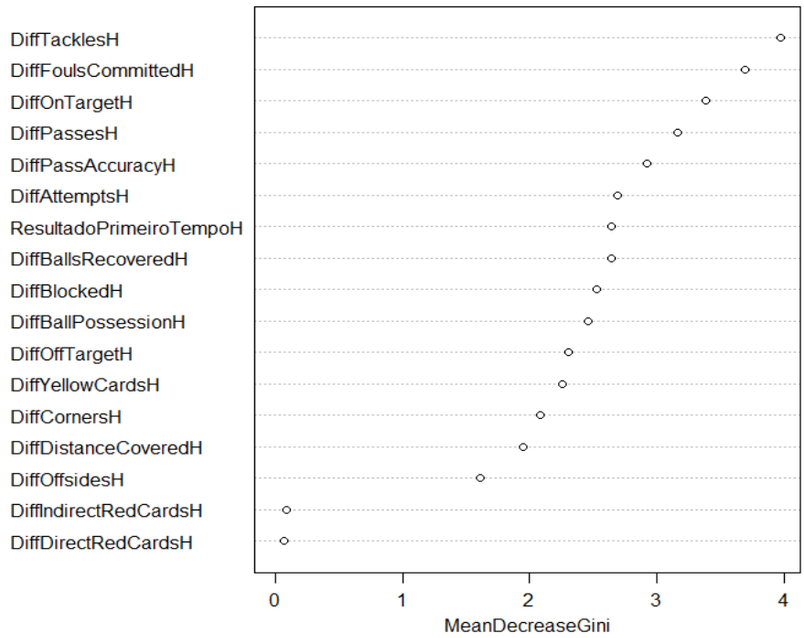
4.2.2 Ensacamento e Floresta Aleatória

Ensacamento foi feito juntamente com floresta aleatória, visto que variância das árvores geradas era grande em decorrência do domínio que a variável chutes no alvo exerce. O algoritmo de floresta aleatória acertou 28(43,75%) dos resultados finais dos jogos. Foram previstas 36 vitórias de mandantes, 4 empates e 24 derrotas. Em particular a dificuldade de prever empates foi muito grande e só 4 foram previstos, 3 incorretamente.

4.2.3 Importância de variáveis

Utilizou-se o critério de Gini para definir quais foram as variáveis mais importantes. O critério de Gini mede a impureza dos nós da árvore. Através do critério MeanDecreaseGini se tem que as diferenças entre desarmes, faltas cometidas e chutes no alvo como variáveis mais importantes respectivamente.

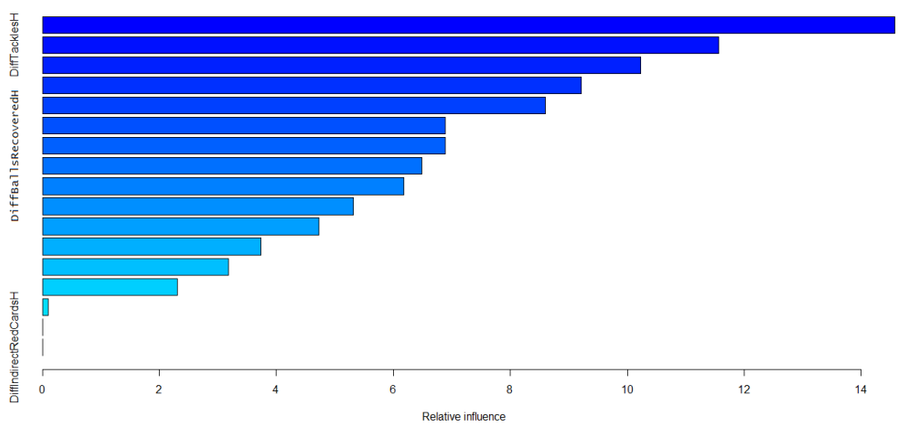
Figura 8: Importância das variáveis no *Random Forest*.



4.2.4 Boosting

Após ser estimado o número ideal de 3 interações de *boosting* foi feito o modelo. Para o modelo de *boosting* duas diferenças entre variáveis tiveram influência não nula: bolas recuperadas e desarmes, duas variáveis defensivas.

Figura 9: Importância das variáveis no boosting:



As previsões de *boosting* foram bem melhores do que as de floresta aleatória, porém com uma ressalva que o modelo previu somente um empate no jogo Suíça e Costa Rica. Acertou 36(56,25%) dos jogos.

Variáveis do modelo de *boosting* e sua relevância: diferença entre desarmes, com relevância de 54,32% e diferença entre bolas recuperadas, com 45,68%.

Note que são duas variáveis defensivas.

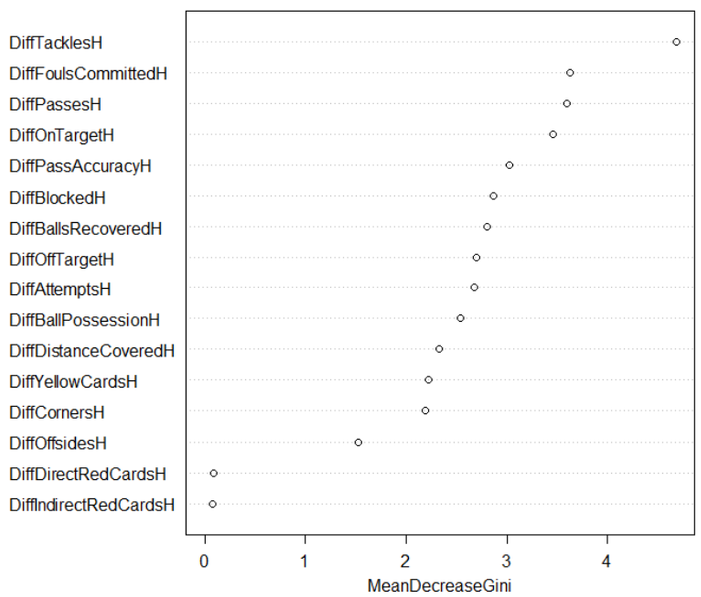
4.2.5 Floresta Aleatória sem o resultado do primeiro tempo:

Como a árvore inicial não tinha utilizado o resultado do primeiro tempo como variável significativa, tirar essa variável do modelo não modifica a mesma. *Boosting* é o mesmo caso. Já para floresta aleatória existem diferenças, já que o resultado do primeiro tempo era o mais importante.

Os resultados da floresta aleatória sem o resultado do primeiro tempo foram os piores em termos de predição do resultado. 24(37,5%) classificações corretas nos 64 jogos da Copa.

As diferenças entre desarmes, faltas cometidas e passes foram as variáveis mais importantes. É interessante notar que as variáveis faltas cometidas e passes, que estão como mais importantes nesse modelo não foram escolhidas em outros.

Figura 10: Índice de Gini para Random Forest



4.3 LASSO

LASSO força diversos coeficientes para 0, tirando essas variáveis do modelo e o tornando mais simples. Das variáveis originalmente disponíveis, o LASSO utilizou apenas a diferença de tentativas no alvo. Quando a variável do resultado no intervalo esteve disponível ela foi utilizada e melhorou um pouco os resultados do modelo. Foram 38(59,38%) acertos com a variável resultado do intervalo disponível e 32(50%) acertos sem ela.

O LASSO não classificou nenhum jogo como empate.

4.4 MARS

A grande vantagem do MARS é incluir a interação de variáveis. O MARS apresentou os melhores resultados de acerto de resultado e também foi quem mais acertou previsões de empate, ainda que tenha previsto apenas 3 e acertado os 3. Isso só ocorreu quando o resultado do intervalo não estava disponível, ou seja, mesmo com resultado de empate no intervalo o modelo ainda assim não colocava empate como previsão do resultado final, assim como outros modelos.

Quando o resultado do intervalo estava no banco de dados o MARS utilizou apenas a diferença entre tentativas no alvo do banco de dados e acertou 39(60,94%) dos resultados das partidas. Sem ela, as variáveis utilizadas foram: diferença de tentativas no alvo, diferença de bloqueios e diferença de desarmes. Neste modelo existiu a interação entre as variáveis bloqueios e desarmes.

Temos como funções h do modelo as pmax. Pmax é uma função indicadora entre os máximos das duas variáveis. Por exemplo, quando a diferença de tentativas no alvo ultrapassa -1, a mesma passa a ser multiplicado pelo coeficiente na linha da respectiva saída. No caso as funções h são pmax(0, DiffTentativasnoAlvo +1) e pmax(0, ResultadoPrimeiroTempo) no primeiro modelo e pmax(0, DiffTentativasnoAlvo +1) e pmax(0, DiffBloqueios – 5) * DiffDesarmes no segundo modelo. No caso do termo de interação do segundo modelo, quando a diferença entre os bloqueios do mandante ultrapassa 5, o valor é multiplicado, além de pelo coeficiente, também pela diferença entre desarmes.

Nas Tabelas 5 e 6 estão os dois modelos MARS gerados. Substituindo-se os valores das variáveis selecionadas temos as 3 saídas e a classificação é feita para a classe com o maior valor de resposta previsto.

Tabela 5: Modelo MARS

Saída	Intercepto	pmax(0, DiffTentativasnoAlvo	pmax(0,

		+1)	ResultadoPrimeiroTempo)
-1	0,63	-0,055	-0,36
0	0,22	-0,0044	-0,021
1	0,15	0,06	0,38

Tabela 6: Modelo MARS sem resultado do primeiro tempo

Saída	Intercepto	pmax(0, DiffTentativasnoAlvo +1)	pmax(0, DiffBloqueios – 5) * DiffDesarmes
-1	0,54	-0,049	0,0088
0	0,22	-0,019	-0,027
1	0,24	0,068	0,018

Dos modelos de regressão, o MARS teve o melhor desempenho de taxa de acerto, tanto com quanto sem o resultado do primeiro tempo. A Tabela 7 exibe como foram as previsões e os resultados dos jogos do MARS para o modelo sem o resultado do primeiro tempo. O modelo previu 35 derrotas dos mandantes, sendo que ocorreram 25 e previu 26 vitórias de mandantes, acertando 61,54% desse tipo de previsão.O modelo acertou todos os empates que previu, mas somente colocou 3 dos 64 jogos como empates, sendo que o número de empates na Copa foi 13.

Tabela 7: Previsões do modelo MARS sem resultado do primeiro tempo

	-1	0	1
-1	18	0	7
0	7	3	3
1	10	0	16

A Tabela 8 mostra uma comparação do desempenho de todos os modelos executados no trabalho em relação a cada previsão e a taxa de acerto. É interessante notar o número baixo de previsões de empate, tanto certas quanto erradas, com exceção da árvore.

Tabela 8: Desempenho dos modelos:

Modelo	Derrota previsão certa	Derrota previsão errada	Empate previsão certa	Empate previsão errada	Vitória previsão certa	Vitória previsão errada	Taxa acerto
Árvore Podada	22	7	9	7	16	3	73,44%
LASSO c	18	10	0	0	20	16	59,38%
LASSO s	14	14	0	0	18	18	50%
MARS c	20	13	0	0	19	12	60,94%
MARS s	18	17	3	0	16	10	57,81%
Floresta Aleatória c	10	14	1	3	17	19	43,75%
Floresta Aleatória s	11	23	1	3	12	16	37,50%
Boosting c	16	15	1	0	19	13	56,25%
Boosting s	16	15	1	0	19	13	56,25%

A Tabela 9 tem as variáveis selecionadas por cada um dos modelos finais do trabalho.

Tabela 9: Variáveis selecionadas por cada modelo:

Modelo	Variáveis
Árvore Podada	DiffTentativasnoAlvo, DiffDesarmes, DiffFaltasCometidas, DiffTentativasforadoAlvo
LASSO c	DiffTentativasnoAlvo, Resultado intervalo
LASSO s	DiffTentativasnoAlvo
MARS c	Resultado intervalo, DiffTentativasnoAlvo
MARS s	DiffTentativasnoAlvo, DiffBloqueios, DiffDesarmes
Random Forest c	Resultado intervalo, DiffOnTarget
Random Forest s	DiffDesarmes, DiffPasses, DiffFaltasCometidas
Boosting c	DiffDesarmes, DiffBolasRecuperadas
Boosting s	DiffDesarmes, DiffBolasRecuperadas

4 CONCLUSÃO

Entre os modelos a árvore de classificação foi o que conseguiu um maior número de acertos, utilizando-se de diversas variáveis. Depois de podada ela conseguiu, considerando 4 variáveis, acertar 47 classificações. No caso da árvore as variáveis importantes foram: diferença de tentativas no alvo, diferença de desarmes, diferença de faltas cometidas e diferença de tentativas fora do alvo. *Boosting* ficou com diferença de bolas recuperadas e diferença de desarmes.

Quanto aos modelos de regressão, MARS com diferença de tentativas no alvo, diferença de bloqueios e diferença de desarmes e LASSO com a diferença de tentativas no alvo.

É interessante ressaltar que o modelo MARS considerava a interação entre diferença de bloqueios e diferença de desarmes.

Em relação ao desempenho de cada um dos métodos para acertar o que ocorreu, floresta aleatória teve o pior aproveitamento. Acertou 43,75% dos resultados considerando a variável do resultado do primeiro tempo e 37,5% sem ela. LASSO teve 59,375% de aproveitamento quando o resultado da metade da partida estava na previsão e 50% sem ele. *Boosting* teve 56,25% e o melhor dos modelos com menos variáveis foi o MARS, com 60,9375% e 57,8125% de acerto. Conclui-se do que foi apresentado que duas variáveis defensivas, bloqueios e desarmes e a variável ofensiva de tentativas na direção do gol foram as mais impactantes. Nenhum dos melhores modelos considerou as variáveis de performance como significativas estatisticamente, o que vai ao encontro aos resultados de algumas seleções como a Espanha que só ganhou um jogo e da França, que foi campeã vencendo Argentina, Bélgica e Croácia tendo menos posse de bola nos 3 jogos.

REFERÊNCIAS

ALBERT Jim. **A. Baseball Statistics Course**. https://amstat.tandfonline.com/doi/abs/10.1080/10691898.2002.11910663?fbclid=IwAR3btazQSU9mgrDtrkE4VqpoWE0uD9WzfGArA9HCUu1ummRCmKI0xpdjYOA&#.W_m1LzEnbIU.[Acesso em 20-novembro-2018].

BELL, Jason. **Machine Learning: Hands-On for Developers and Technical Professionals.** Published by John Wiley & Sons, Inc. Indianapolis, Manufactured in the United States of America. Copyright © 2015 by John Wiley & Sons, Inc., Indianapolis, Indiana. 2015. 407 p.

BISHOP, Christopher. **Pattern Recognition and Machine Learning**, New York: Springer, 2006. 758 p.

FIFA. https://www.fifa.com/ [Acesso em 19-setembro-2018].

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. **The Elements of Statistical Learning**: Data Mining, Inference, and Prediction. 2. Ed. New York: Springer, 2009. 745 p.

KONZEN, Evandro; ZIEGELMANN, Flávio. **LASSO-Type Penalties for Covariate Selection and Forecasting in Time Series.** November 2016.Pages 592-612. Journal of Forecasting, 2016.

LE, James. **The 10 Algorithms Machine Learning Engineers Need to Know**. 2016. https://www.kdnuggets.com/2016/08/10-algorithms-machine-learning-engineers.html. [Acesso em 25-novembro-2018].

LUCIDCHART. https://www.lucidchart.com/pages/pt/o-que-e-arvore-de-decisao.. . [Acesso em 23-novembro-2018].

GAMA, João. **Functional Trees**. Machine Learning, 55, 219–250, 2004. Kluwer Academic Publishers. Manufactured in The Netherlands. https://link.springer.com/content/pdf/10.1023%2FB%3AMACH.0000027782.67192.13.pdf. [Acesso em 20-novembro-2018].

JAMES, Gareth *et al*. **An Introduction to Statistical Learning**: with Applications in R. 1. Ed. New York: Springer, 2013. 426 p.

MARR, Bernard. **Big Data And AI: 30 Amazing (And Free) Public Data Sources For 2018.** https://www.forbes.com/sites/bernardmarr/2018/02/26/big-data-and-ai-30-amazing-and-free-public-data-sources-for-2018/#78ce82885f8a. [Acesso em 20-novembro-2018].

MATOS, David; **Conceitos Fundamentais de Machine Learning**. http://www.cienciaedados.com/conceitos-fundamentais-de-machine-learning. [Acesso em 21-novembro-2018].

MEDEIROS, M. C. *et al*. **Forecasting Brazilian Inflation with High Dimensional Models.** November 2016. v. 36, n. 2, pp. 223-254. Brazilian Review of Econometrics, 2016.

NILSSON, Nils. **Introduction to Machine Learning**. 1998. Department of Computer Science .Stanford: Stanford University. 188 p.

SCHNEIDER, Cristian. **Machine Learning aplicado na previsão de resultados de partidas de futebol:** um estudo de caso para comparação de diferentes classificadores. / 92 f. Trabalho de conclusão de Curso – UFRGS, Escola de Engenharia, Curso de Engenharia Elétrica, Porto Alegre, 2018.

SELAU, Lisiane Priscila Roldão; RIBEIRO, José Luis Duarte. A systematic approach to construct credit risk forecast models. **Pesqui. Oper**, Rio de Janeiro, v. 31,n. 1,p. 41-56, Apr. 2011 . disponível em <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0101-74382011000100004&lng=en&nrm=iso>. pdf. [Acesso em 22-novembro-2018].

SILVA, Aline M. L. et al. **Descoberta de conhecimento através de métodos de aprendizagem de máquina supervisionados aplicados ao SIGAA**/ Universidade Federal do Piauí. Revista de Sistemas e Computação, Salvador, v. 7, n. 1, p. 68-78, jan./jun. 2017.

SILVEIRA, Raiane. **Comparação das penalizações do tipo Lasso para previsão das taxas de crescimento dos índices de inflação IPCA e IGP-M.** Porto Alegre: Departamento de Estatística/ UFRGS, 2016. Nº de páginas: 34. Trabalho de conclusão de Curso – UFRGS, Instituto de Matemática e Estatística, Bacharelado em Estatística, Porto Alegre, 2016.

TIBSHIRANI, Robert. **Encolhimento regressão e seleção através do lasso.** Jornal da Royal Statistical Society. Série B (metodológica), v. 58, 267-288, 1996.

TIMMARAJU, Aditya.; PALNITKAR Aditya; & KHANNA, Vikesh. **Game ON! Predicting English Premier League Match Outcomes**, 2013, disponível em: http://cs229.stanford.edu/proj2013/TimmarajuPalnitkarKhanna-GameON!PredictionOfEPLMatchOutcomes.pdf. [Acesso em 23-novembro-2018].

THORN John; PALMER Pete; **The hidden Game of Baseball: A Revolutionary Approach to Baseball and Its Statistics.** Chicago, Estados Unidos. E. University of Chicago Press. Ano: 2015. 400 p**.**

ULMER Ben; FERNANDEZ Matthew. **Predicting Soccer Match Results in the English Premier League**: http://cs229.stanford.edu/proj2014/Ben%20Ulmer,%20Matt%20Fernandez,%20Predicting%20Soccer%20Results%20in%20the%20English%20Premier%20League.pdf. [Acesso em 23-novembro-2018].

VIDHYA. **A Complete Tutorial on Tree Based Modeling from Scratch** (in R & Python). https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/, Ano: 2016 [Acesso em 24-novembro-2018].

WAQUIL, Alice Paul**. Mando de Campo e Gol Qualificado – uma Análise da Vantagem na Copa do Brasil.** 84 p. 2018.

ANEXO 1

Seleções participantes da copa do mundo de 2018 – Fase de grupos: 48 jogos.

		Vitória	Empate	Derrota	Pontos
Grupo A	Uruguai	3	0	0	9
	Rússia	2	0	1	6
	Arábia Saudita	1	0	2	3
	Egito	0	0	3	0
Grupo B	Espanha	1	2	0	5
	Portugal	1	2	0	5
	Iran	1	1	1	4
	Marrocos	0	1	2	1
Grupo C	França	2	1	0	7
	Dinamarca	1	2	0	5
	Peru	1	0	2	3
	Austrália	0	1	2	1
Grupo D	Croácia	3	0	0	9
	Argentina	1	1	1	4
	Nigéria	1	0	2	3
	Islândia	0	1	2	1
Grupo E	Brasil	2	1	0	7
	Suíça	1	2	0	5
	Sérvia	1	0	2	3
	Costa Rica	0	1	2	1
Grupo F	Suécia	2	0	1	6
	México	2	0	1	6
	Coreia	1	0	2	3
	Alemanha	1	0	2	3
Grupo G	Bélgica	3	0	0	9
	Inglaterra	2	0	1	6
	Tunísia	1	0	2	3
	Panamá	0	0	3	0
Grupo H	Colômbia	2	0	1	6
	Japão	1	1	1	4
	Senegal	1	1	1	4
	Polônia	1	0	2	3
	TOTAL	39(81,25%)	9(18,75%)	39(81,25%)	

ANEXO 2

Seleções participantes da copa do mundo de 2018 –

Fase de grupos até a final.

Grupo	Fase de grupos	Oitavas	Quartas	Semifinais	Resultado
	Uruguai	Uruguai	Uruguai		
A	Rússia	Rússia	Rússia		
	Arábia Saudita				
	Egito				
	Espanha	Espanha			
B	Portugal	Portugal			
	Iran				
	Marrocos				
	França	França	França	França	1º França
C	Dinamarca	Dinamarca			
	Peru				
	Austrália				
	Croácia	Croácia	Croácia	Croácia	2ºCroácia
D	Argentina	Argentina			
	Nigéria				
	Islândia				
	Brasil	Brasil	Brasil		
E	Suíça	Suíça			
	Sérvia				
	Costa Rica				
	Suécia	Suécia	Suécia		
F	México	México			
	Coreia				
	Alemanha				
	Bélgica	Bélgica	Bélgica	Bélgica	3º Bélgica
G	Inglaterra	Inglaterra	Inglaterra	Inglaterra	4º Inglaterra
	Tunísia				
	Panamá				
	Colômbia	Colômbia			
H	Japão	Japão			
	Senegal				
	Polônia				

ANEXO 3

Glossário de termos empregados neste trabalho relacionados ao futebol.

Inglês	**Português**
Attempts	Tentativas
Ball Possession	Posse de Bola
Balls Recovered	Bolas Recuperadas
Blocked	Bloqueio
Corner	Escanteio
Distance Covered	Distancia percorrida
Fouls Committed	Faltas cometidas
Off Target	Chutes fora do alvo
Offside	Impedimento
On Targets	Chutes no alvo
Pass Accuracy	Precisão no passe
Red Card	Cartão vermelho
Tackles	Desarmes
Yellow Card	Cartão Amarelo