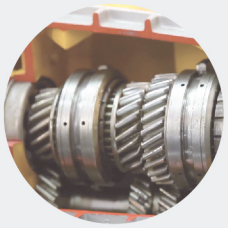
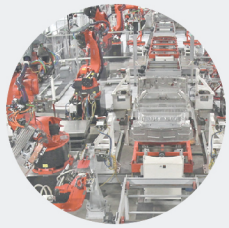
INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA E GESTÃO INDUSTRIAL



Avaliação de Jogadores e Equipas de Basquetebol usando Machine Learning

LUÍS RODOLFO NOGUEIRA E SILVA
setembro de 2022

POLITÉCNICO
DO PORTO

Instituto Superior de Engenharia do Porto

Avaliação de Jogadores e Equipas de Basquetebol usando Machine Learning

Luís Rodolfo Nogueira e Silva

2022

Orientação científica: Carlos Manuel Abreu Gomes Ferreira (CGF)

Instituto Superior de Engenharia do Porto

Departamento de Engenharia Mecânica

P.PORTO


P.PORTO

Instituto Superior de Engenharia do Porto

Avaliação de Jogadores e Equipas de Basquetebol usando Machine Learning

Luís Rodolfo Nogueira e Silva

Estudante nº1171124

Dissertação apresentada ao Instituto Superior de Engenharia do Porto para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia e Gestão Industrial, realizada sob a orientação do Doutor Carlos Manuel Abreu Gomes Ferreira

2022

Instituto Superior de Engenharia do Porto

Departamento de Engenharia Mecânica

P.PORTO


P.PORTO

AGRADECIMENTOS

Quero expressar o meu agradecimento ao professor Carlos Manuel Abreu Gomes Ferreira.

O apoio demonstrado foi fundamental, com as suas sugestões e pontos de vista acerca do desenvolvimento do projeto. Agradeço também toda a disponibilidade e profissionalismo demonstrados.

Por último, especial agradecimento a todos os meus amigos e família, que me ajudaram a ser quem sou hoje e que me acompanharam ao longo de todo o meu percurso académico.

página propositadamente em branco

RESUMO

Ao longo das últimas décadas, o basquetebol passou por uma evolução que transformou o que era um desporto para um negócio com enorme impacto social e financeiro.
Essa transformação, aliada à constante necessidade de obter sucesso desportivo, criou a necessidade de inovação por parte de um clube desportivo de forma a distanciar-se dos seus adversários e de conquistar títulos e alcançar um maior lucro financeiro.

A enorme quantidade de dados impossibilita a análise detalhada por um especialista e requer a utilização de meios computacionais para extrair informação valiosa. Alguns clubes já utilizam *machine learning*, mas este processo ainda se encontra numa fase inicial. Neste sentido, existe um enorme potencial para o tratamento e posterior valorização dos dados.

A maioria das instituições desportivas identificaram a necessidade de capacidade técnica na análise de vastas fontes de informação, de forma que, as decisões tomadas sejam o mais fundamentadas possível, e, consequentemente, haja uma diminuição dos riscos na tomada das mesmas.

Esta dissertação procura resolver este problema, e, com recurso ao *mahine learning*, mais concretamente do auxílio da metodologia CRISP-DM, passa pelo desenvolvimento de modelos de previsão de um *rating* de qualidade de um jogador e de equipa baseado em estatísticas de jogo, como por exemplo o número de pontos por jogo, com recurso a um modelo de previsão das mesmas.

É também desenvolvido um modelo de previsão de resultados de jogos de basquetebol, tendo como base estatísticas de cada equipa, envolvendo diversas variáveis, de forma a tornar o modelo o mais robusto possível, e com uma maior flexibilidade.

PALAVRAS-CHAVE

Machine Learning, CRISP-DM, Rating, Pontos, Random forest, Gradient Boosted, Simple Regression, Numeric Scorer, Accuracy, Previsão

página propositadamente em branco

ABSTRACT

Over the last few decades, basketball has undergone an evolution that has transformed what was a sport into a business with enormous social and financial impact.

This transformation, combined with the constant need to achieve sporting success, created the need for innovation on the part of a sports club in order to distance itself from its opponents and win titles, and achieve greater financial profit.

The huge amount of data makes detailed analysis impossible by an expert and requires the use of computational means to extract valuable information. Some clubs already use machine learning, but this process is still at an early stage. In this sense, there is an enormous potential for the processing and subsequent valorization of the data.

Most sports institutions have identified the need for technical capacity in the analysis of vast sources of information so that the decisions taken are as well-founded as possible, and, consequently, there is a reduction in the risks in making them.

This investigation seeks to solve this problem, and, using machine learning, more specifically with the help of the CRISP-DM methodology, it involves the development of a prediction model of a player and team quality rating based on game statistics, such as the number of points per game, using a prediction model.

A model for predicting the results of basketball games will also be developed, based on statistics for each team, involving several variables, to make the model as robust as possible, and with greater flexibility.

KEYWORDS

Machine Learning, CRISP-DM, Rating, Points, Random forest, Gradient Boosted, Simple Regression, Numeric Scorer, Accuracy, Prediction

página propositadamente em branco

ÍNDICE

página propositadamente em branco

ÍNDICE DE FIGURAS

página propositadamente em branco

ÍNDICE DE TABELAS

página propositadamente em branco

LISTAS DE SIGLAS E SÍMBOLOS

Lista de Siglas

ANN	Aritificial Neural Network
AST	Assistências
DEM	Departamento de Engenharia Mecânica
ESPN	Entertainment and Sports Programming Network
FT	Free-throws
HMM	Hidden Markov Models
IPP	Instituto Politécnico do Porto
ISEP	Insituto Superior de Engenharia do Porto
MAPE	Erro Percentual Absoluto Médio
MEGI	Mestrado em Engenharia e Gestão Industrial
ML	Machine Learning
MVP	Most valuable player
NBA	National Basketball Association
NCAA	National Collegiate Athletic Association
P.Porto	Instituto Politécnico do Porto
PTS	Pontos
REB	Ressaltos
RNAs	Redes Neuronais Artificiais
RNN	Recurrent Neural Network
STL	Roubos de bola
TCI	Tecnologias de comunicação e informação

página propositadamente em branco

1. INTRODUÇÃO

Neste capítulo é efetuado o enquadramento do trabalho, demonstrando a sua pertinência, importância e atualidade. É também enunciada a questão de investigação e são apresentados os objetivos da mesma.

Por fim são apresentadas as opções de metodologias que são usadas ao longo do projeto, bem como o plano de investigação.

1.1. Enquadramento e pertinência

Ao longo das últimas décadas, o basquetebol passou por uma evolução que transformou o que era um desporto para um negócio com enorme impacto social e financeiro. Um exemplo desta mudança é a *NCAA (National Collegiate Athletic Association)*, na qual em 2010, a CBS Sports anunciou um contrato no valor de 10.8 biliões de dólares, ao longo de 14 anos, para transmissão de jogos desta liga (Borghesi, 2018).

Essa transformação, aliada à constante necessidade de obter sucesso desportivo, criou a necessidade de inovação por parte de um clube desportivo de forma a distanciar-se dos seus adversários e de conquistar títulos e alcançar um maior lucro financeiro.

Atualmente são recolhidos muitos dados durante os jogos, nos quais oferecem informações valiosas e que nem sempre estes clubes tiram proveito deles (Tian et al., 2020).

A enorme quantidade de dados impossibilita a análise detalhada por um especialista e requer a utilização de meios computacionais para extrair informação valiosa. Alguns clubes já utilizam *machine learning*, mas este processo ainda se encontra numa fase inicial. Neste sentido, existe um enorme potencial para o tratamento e posterior valorização dos dados (Kaur & Jain, 2018).

Consequentemente, mais e mais comunidades de investigadores da área da análise e exploração de dados (e *machine learning*) aderiram a este movimento, e algumas investigações, como *"Machine Learning and Data Mining for Sports Analytics"* em ECML / PKDD (Pelechrinis et al., 2019) ou *"Large-Scale Sports Analytics" na KDD* (Lucey P, Morgan S, Wiens J, 2016) confirmam a crescente popularidade da análise inteligente de dados para análises desportivas (Brefeld & Zimmermann, 2017).

A análise desportiva reside numa interessante interseção entre a aplicação de problemas reais em diversas áreas (como por exemplo saúde e pesquisa médica / desportiva), as várias indústrias ligadas ao desporto profissional, o interesse comercial significativo, o grande volume de dados, e o enorme interesse do público (Brefeld & Zimmermann, 2017).

Do mesmo modo que a inteligência artificial, os dados desportivos oferecem-se como um ambiente de teste interessante para investigadores da mineração de dados. Contrariamente aos jogos, o universo de dados desportivos não é discreto, e os elementos da individualidade e o erro humano têm um maior impacto / influência, aumentando a dificuldade do processo. No entanto, os desportos são condicionados por diversas regras e lugar (pistas / circuitos na área automóvel,

campos, etc.) onde a ação ocorre, oferecendo repetibilidade que outros cenários em outras áreas do quotidiano não têm (Brefeld & Zimmermann, 2017).

No desporto, principalmente em desportos de alta competição, uma vantagem mínima sobre o adversário é o suficiente para determinar o sucesso ou insucesso de um jogador ou de uma equipa. Mas, para isto, é necessário fornecer ao decisor toda a informação possível, de uma forma limpa, concisa, e com valor (Sarlis et al., 2021).

A maioria das instituições desportivas identificaram a necessidade de capacidade técnica na análise de vastas fontes de informação, de forma que, as decisões tomadas sejam o mais fundamentadas possível, e, consequentemente, haja uma diminuição dos riscos na tomada das mesmas (Sarlis et al., 2021).

Esta dissertação procura resolver este problema, e passa pelo desenvolvimento de um modelo de previsão de um *rating* de qualidade de um jogador baseado em estatísticas de jogo, como por exemplo o número de pontos por jogo, com recurso a um modelo de previsão das mesmas.

Será também desenvolvido um modelo de previsão de resultados de jogos de basquetebol, tendo como base estatísticas de cada equipa, envolvendo diversas variáveis, de forma a tornar o modelo o mais robusto possível, e com uma maior flexibilidade. Os modelos são desenvolvidos com recurso a algoritmos de *machine learning*.

1.2. Questão e objetivos de investigação

É fundamental destacar uma questão de investigação, que será respondida ao longo do desenvolvimento do projeto. Neste sentido, a pergunta na qual o projeto desenvolvido recai é a seguinte:

"Como usar dados passados de jogadores e de equipas, através de um modelo de machine learning, de forma a definir os pontos bons e menos bons de um atleta/equipa, bem como prever resultados, no basquetebol?"

O principal objetivo da dissertação passa pela criação de um modelo de *machine learning*, de forma a determinar o *rating* de qualidade de um jogador baseado em estatísticas de jogo, e também do desenvolvimento de um modelo *machine learning* de previsão de resultados de basquetebol. A ferramenta utilizada para o efeito é o KNIME.

Fundamentalmente, será desenvolvida e disponibilizada uma ferramenta de análise de jogadores e previsão de resultados.

Para alcançar os objetivos pretendidos, é necessário estabelecer metas / objetivos específicos necessários para o desenvolvimento necessário:

- O estudo do estado da arte, de forma a obter um *background* no tema a realizar.

- Procurar a melhor fonte de dados para o modelo a desenvolver. Esta deve possuir uma vasta quantidade de dados, deve ser coesa e possuir várias variáveis.

− Efetuar uma limpeza dos dados a utilizar, de forma a tornar o modelo mais robusto e "sem ruído" causado por outras variáveis. Para isto, é necessário eliminar variáveis desnecessárias, valores repetidos, e *outliers*.

− Efetuar uma normalização e uniformização de variáveis, de forma que seja possível cruzar duas variáveis para tirar conclusões / auxiliar na tomada de decisões.

− Comparar os resultados dos diferentes modelos de previsão (*simple regression, decision tree, random forest, gradient boosted*) e avaliar qual o melhor modelo para responder aos objetivos da investigação.

1.3. Opções metodológicas

Como perspetiva metodológica será abordada uma perspetiva quantitativa. A perspetiva quantitativa "centra-se na análise de factos e fenómenos observáveis e na medição/avaliação de varáveis (…) passíveis de serem medidas, comparadas e/ou relacionadas"(Clara Pereira Coutinho, 2014). A adoção desta perspetiva tem como objetivo testar, verificar e comprovar teorias e hipóteses, recorrendo para isso a grandes amostras, fazendo com que os resultados possam ser generalizados, e, consequentemente, levar ao aumento do conhecimento e da capacidade de prever explicar e controlar fenómenos (Ranjit Kumar, 2019).

Relativamente à classificação, esta pode ser classificada como de natureza analítica. Na investigação analítica, o problema é decomposto nos fatores que o influenciam. É realizada uma análise extensiva de modo a avaliar a influência relativa de cada um dos fatores (Luis Adriano Oliveira, 2011). Procura-se estabelecer e definir relações entre variáveis, e é possível justificar e explicar as causas e os motivos para determinadas ocorrências (Machado et al., 2020).

Também é possível classificar a metodologia da investigação como sintética, pois esta, remete para a reconstituição do todo a partir das suas partes, pois tem como objetivo caraterizar um problema juntando a informação das diferentes partes que o constituem. O conhecimento do todo, permite fazer previsões, entre outros, sendo por isso adequado ao problema da investigação (Luis Adriano Oliveira, 2011).

Para o desenvolvimento do modelo, será usada a metodologia CRISP-DM, que tem como objetivo promover uma metodologia comum para comunicação, auxiliando na conexão das mais variadas ferramentas técnicas, e pessoas com diferentes skills, levando ao desenvolvimento de projetos de um modo eficiente e eficaz. O método CRISP-DM compreende seis fases: *Business Understanding, Data Understanding, Data Preparation,* o *Modeling,* a *Evaluation* e, finalmente, o *Deployment*.
A metodologia CRISP-DM foi concebida para atender aos projetos que estão diretamente envolvidos com o processamento e a análise de um grande volume de dados (Azevedo & Santos, 2008).

Por fim, o método da investigação pode ser classificado como método experimental, pois este consiste na formação de dois grupos de indivíduos, aplicando apenas a um deles o "tratamento" (variável dependente), sendo este o grupo experimental e o restante o denominado de grupo de controlo. Neste método, os grupos são comparados na variável dependente, para avaliar se as diferenças nesta variável são causadas pelo "tratamento" (Machado et al., 2020).

1.4. Estrutura do trabalho

No capítulo 1, é efetuado o enquadramento do trabalho, demonstrando a sua pertinência, importância e atualidade. São também apresentados os objetivos da investigação e opções metodológicas usadas.

No capítulo seguinte, são descritos e especificados os diversos conteúdos necessários para o desenvolvimento da investigação, resultando no estado da arte.

No terceiro capítulo, são apresentados os resultados e são descritos os desenvolvimentos da metodologia utilizada.

Por último, no quarto capítulo são reunidas as principais conclusões e perspetivados futuros desenvolvimentos.

2. REVISÃO BIBLIOGRÁFICA

Neste capítulo, são descritos e especificados os diversos conteúdos necessários para o desenvolvimento da investigação, resultando no estado da arte.

Neste sentido, foi destacada a importância de uma análise de dados, bem como a sua conexão com o *Machine Learning*, nomeadamente os conceitos de *Big Data* e *Data Analysis*.

Posteriormente é descrito o *machine learning* no setor desportivo, bem como alguns dos métodos de previsão do mesmo. De seguida, é especificado o desporto do basketball, pois é o desporto foco do desenvolvimento do modelo do projeto de investigação.

Visto que será utilizada a metodologia CRISP-DM para o desenvolvimento da ferramenta, foram detalhados os passos do processo, desde a compreensão do problema até ao final do desenvolvimento do modelo.

Por fim, são brevemente referidas as utilidades da aplicação KNIME, que será a aplicação principal da realização do estudo.

2.1. Importância da Análise de Dados – Conexão com o Machine Learning

Devido ao avanço exponencial de tecnologias digitais tais como os smartphones, redes sociais, e-commerce, entre outros, os dados encontram-se presentes em todas as organizações. À medida que os recursos analíticos se desenvolvem, termos como inteligência artificial, *Big Data Analytics*, blockchain, surgem e ganham a sua preponderância nas indústrias. Consequentemente, surge um desafio para estas organizações: à luz do mercado extremamente volátil e de um sistema económico global imprevisível, como é que as organizações devem usar a análise de dados para melhorarem as suas práticas de gestão (Araz et al., 2020).

Neste sentido, existe uma preocupação por parte das organizações no aproveitamento e utilização de dados para o apoio de decisões e delineamento de estratégias. Estas procuram efetuar, cada vez mais, um maior investimento num maior processamento de dados, pois a informação retirada deles é tanto mais fidedigna quanto maior a base de dados (Araz et al., 2020).

Apesar de uma maior quantidade de dados fornecer melhores outputs, também se torna difícil trabalhar com os mesmos devido a possíveis limitações de processamento.

Neste sentido, este capítulo pretende explorar o conceito de Big Data e Big Data Analysis.

2.1.1. *Big Data* e *Data Analysis*

O termo *Big Data* pode ser definido como conjuntos de dados cujo volume, velocidade e veracidade são tão grandes que não conseguem ser capturados, armazenados, geridos ou analisados por ferramentas "*TCI*" (Tecnologias de comunicação e informação) (Manyika et al., 2011).

O *Big Data* pode ser classificado em três tipos:

- Estruturado;
- Semi-estruturado;
- Não Estruturado.

Os dados armazenados em bases de dados relacionais como por exemplo o *Oracle* são do tipo estruturado. Por outro lado, os dados disponíveis na *Web* são não estruturados. Estima-se que cerca de 80% dos dados existentes no mundo são não estruturados. Dando como exemplo os tweets na plataforma *Twitter*, estes contêm uma extensa quantidade de "gírias", mistura de idiomas, entre outros, sendo, deste modo, dados não estruturados (Sathi, 2013).

Já o termo *Big Data Analytics* remete para a ciência e tecnologia sobre a organização de enorme quantidade de dados, analisando padrões, visualizando e descrevendo o "conhecimento" para auxílio na tomada de decisões (Sun et al., 2018) .

Os principais componentes da *Big Data Analytics* incluem a análise descritiva de *Big Data* e a análise preditiva, que respondem correspondentemente às três questões do *Big Data*: Quando e o que ocorreu? O que vai acontecer? Qual é a melhor resposta ou escolha sob incerteza? Todas estas questões são frequentemente encontradas na maioria das áreas da ciência, tecnologia, negócios, gestão, organização e indústria (Sun et al., 2018).

O *Machine Learning* é responsável por ter em atenção como um computador se pode adaptar a novas circunstâncias, bem como detetar e extrapolar padrões. O *Machine Learning* tem como objetivo principal a construção/ criação de sistemas que podem desempenhar ou até mesmo exceder a competência de um ser humano a lidar com tarefas complexas ou problemas (Russel & Norvig, 2010).
Os dois termos referidos acima estão diretamente relacionados com o *Machine Learning*. Um modelo de *Machine Learning* pode funcionar perfeitamente com uma base de dados menor. No entanto, quando combinados com uma "*Big Data*", os resultados são proporcionalmente maximizados, como se encontra representado na figura 1 (Russel & Norvig, 2010).

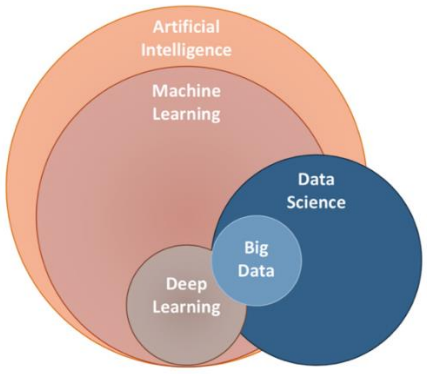


Figura 1 - Esquema de relação entre o Machine Learning e a Big Data (Fonte: *(Thakur, 2020)*)

Um modelo de *Machine Learning* aprende muito mais, e muito mais rapidamente, quando é alimentado por um grande e variado volume de dados e informações. Sendo assim, o *Machine*

Learning pode beneficiar bastante com a conexão com *Big Data*, desvendando possíveis padrões e anomalias, que eventualmente poderão solucionar problemas, ou até mesmo gerar novos insights, permitindo assim uma evolução das organizações (Sun et al., 2018).

Por outro lado, o *Machine Learning* é dos únicos conceitos capazes de dar a devida utilidade à Big Data. Um enorme volume de dados apenas será útil caso estes sejam devidamente analisados e correlacionados, pelo que o *Machine Learning* se torna fundamental. Desta forma, estes conceitos trabalham em sintonia para a criação e desenvolvimento de *Smart Models* (modelos inteligentes), com capacidade para traçarem relações entre variáveis, prever comportamentos e, consequentemente, determinar ações (Salkuti, 2020).

Relacionado também com os conceitos referidos está o conceito de *Data Mining*. O *Data Mining* é um processo de descoberta de vários modelos, resumos, e valores retirados de uma base de dados. Este processo inclui mineração de dados descritiva e preditiva (Salkuti, 2020).

2.2. *Machine Learning* nos Desportos

Cada vez mais o setor desportivo beneficia ao conhecer os processos que sustentam o *Machine Learning* e as suas aplicações na prática, bem como as implicações dentro do campo. O *Machine Learning*, até ao momento, teve impacto em diversas partes do desporto, nomeadamente (Richter et al., 2021):

- As ferramentas usadas para adquirir informação (otimização de leituras de sensores inerciais);
- As informações extraídas dos dados adquiridos pelos dispositivos (cinemática 3D e a previsão via vídeos 2D forças de reação do solo), as trajetórias.
- O processamento de dados adquiridos por dispositivos (modelos de classificação são capazes de dividir de dados em pacotes significativos que anteriormente investigadores do desporto dedicavam um grande tempo no mesmo).
- A utilidade dos dados processados para aprimorar a compreensão do desempenho desportivo e previsões de risco de lesões (modelos de classificação que podem facilitar a decisão objetiva desenvolvendo relação às práticas de reabilitação e prevenção de lesões) (Richter et al., 2021).

É necessário, por isso, aprofundar a utilidade do *Machine Learning* no setor desportivo.

2.2.1. Machine Learning no Desporto - Previsões

Com o crescimento exponencial do desporto, a consciencialização das pessoas sobre a prática desportiva aumentou gradualmente, e os requisitos para uma cultura desportiva têm sido maiores. Contudo, o serviço tradicional relativo a culturas desportivas é único, pelo que é difícil atender às necessidades, cada vez maiores, das pessoas (Keshtkar Langaroudi & Yamaghani, 2019).

Neste sentido, é urgente que este sistema de cultura desportiva consiga acompanhar a evolução, rumo à intelectualização (Keshtkar Langaroudi & Yamaghani, 2019).

No mundo atual, desportos produzem informações estatísticas consideráveis sobre cada jogador, equipa, jogos e temporadas. As organizações perceberam recentemente a extensa e abundante informação disponível, e procuraram tirar proveito, através do uso de técnicas de *Data Mining*. A *Data Mining* nos desportos apoia treinadores nas previsões de resultados, avaliações de desempenho de jogadores, previsões de lesões de jogadores, identificação de talentos desportivos e avaliações de estratégias de jogo, entre outros (Keshtkar Langaroudi & Yamaghani, 2019).

Tendo em conta que as relações entre resultados de desportos e dos diversos elementos de informação são diretamente afetados por diversos fatores referidos anteriormente, vários métodos foram sugeridos para ser efetuada uma previsão de resultados com base nos dados disponíveis (Murphy, 2012).

Mais concretamente, enquanto algumas equipas optam por não utilizar qualquer modelo de previsão, outros dependem fortemente dos especialistas, ou de dados históricos.

As equipas procuram previsões confiáveis, usando-as e tirando proveito das mesmas para auxílio de tomadas de decisão (Murphy, 2012).

No âmbito das previsões usando o *Machine Learning*, surgem vários desafios/incertezas, como por exemplo: Qual o melhor modelo de previsão de dados futuros utilizando dados passados, ou qual o modelo mais adequado para analisar cada tipo de informação, que tipo de análise deve ser efetuada, entre outros (Russel & Norvig, 2010).

Neste sentido, são brevemente referidos abaixo os métodos mais importantes de *Machine Learning*:

Surpervised Learning – tarefa de *Machine Learning* envolvendo uma função que mapeia o *input* num *output* com base em pares de *input-output*, como é possível visualizar na figura 2. Neste método, cada exemplo é um par, consistindo num *input* (normalmente é um vetor) e o desejado *output*, também denominado de sinal de supervisão *(supervisory signal)* (Russel & Norvig, 2010).

As previsões dos dados produzem uma função, que pode ser usada para mapear novos exemplos. Um cenário ideal permitirá que o algoritmo determine corretamente as classes para instâncias não vistas. Este processo requer que o algoritmo generalize, a partir dos dados de treino, identificando padrões que permitam efetuar previsões de uma forma razoável. Essa qualidade estatística de um algoritmo é medida através de uma métrica denominada de erro de generalização. A aprendizagem supervisionada é normalmente utilizada em aplicações onde dados históricos preveem prováveis eventos futuros (Russel & Norvig, 2010).

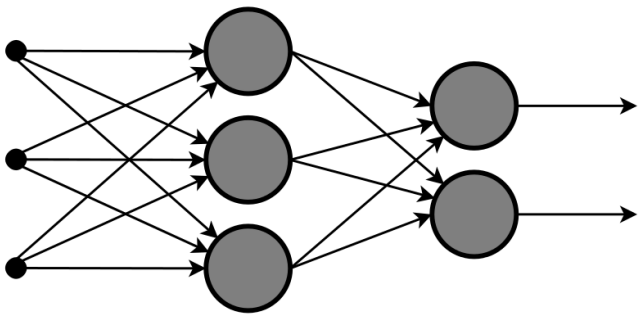


Figura 2 - Rede Supervised Learning (Fonte: (Russel & Norvig, 2010))

Semi-Supervised Learning or Reinforcement Learning – esta é uma abordagem que combina uma pequena quantidade de dados etiquetados com uma grande quantidade de dados não definidos durante a fase de treino. Informação não definida, quando usada em conjunto com uma pequena quantidade de informação etiquetada, pode produzir uma melhoria considerável na precisão do modelo (Russel & Norvig, 2010).

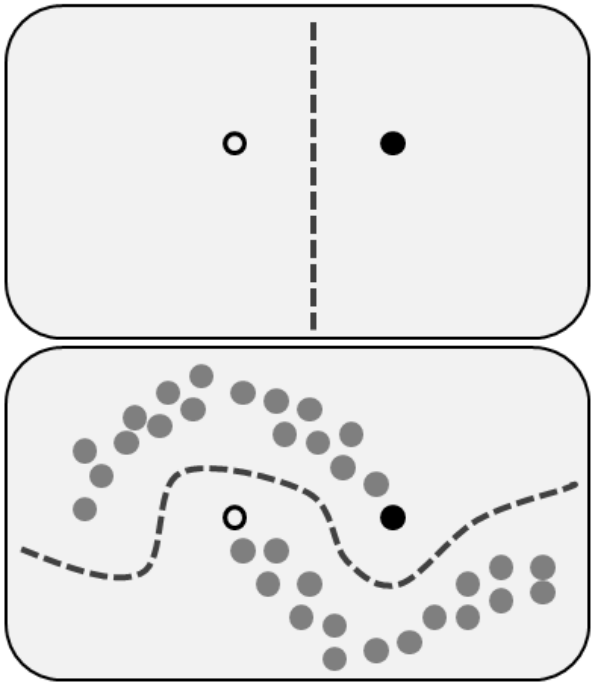


Figura 3 – Ilustração Semi-Supervised Learning (Fonte: (Russel & Norvig, 2010))

O exemplo da figura 3 demonstra a influência de dados não definidos no algoritmo. A primeira imagem mostra um limite de decisão possível de adotar após verificar apenas ume exemplo positivo (círculo branco) e um negativo (círculo preto). Já a imagem inferior mostra um limite de decisão que seria possível adotar se, além dos exemplos definidos, houvesse uma quantidade de dados não definidos (círculos cinzentos). Isto pode ser visto como definição de agrupamentos, e também definir agrupamentos com dados definidos, afastando o limite de decisão das regiões de alta densidade (Russel & Norvig, 2010).

Unsupervised Learning or Clustering – Este tipo de algoritmo aprende padrões de dados não definidos. O objetivo é que, por meio de mímica, que é um modo importante de aprendizagem nas pessoas, a máquina seja forçada a construir uma representação interna e compacta do seu mundo e, consequentemente, gerar conteúdo imaginativo a partir do mesmo. Este método envolve organização autónoma, e captura padrões como densidades de probabilidade, ou combinações de preferências de caraterísticas neuronais (Russel & Norvig, 2010).

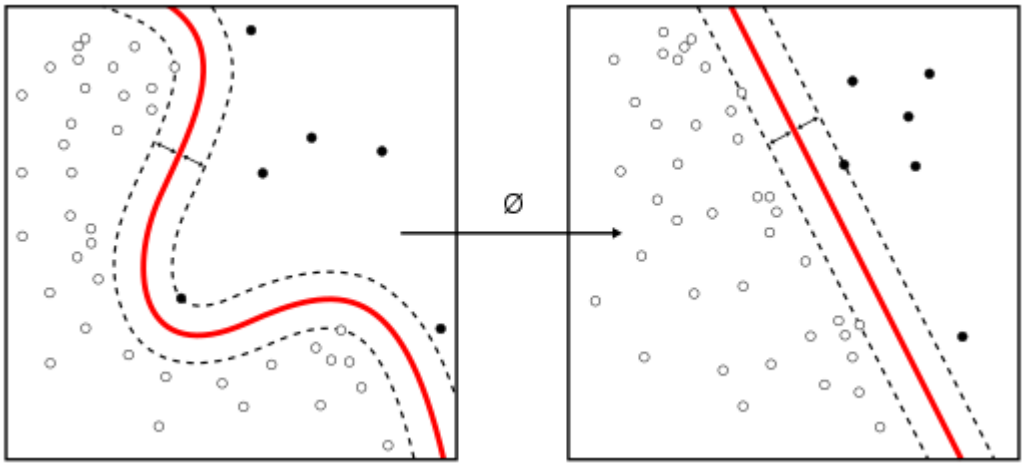


Figura 4 – Ilustração Clustering (Fonte: (Russel & Norvig, 2010))

2.3. Basquetebol – A aplicação do *Machine Learning*

O desporto de basquetebol é conhecido pela grande quantidade de dados coletados para cada jogador, equipa, jogo e temporada. Como resultado, o basquetebol é um setor ideal para trabalhar em diferentes técnicas de análise de dados, e, consequentemente, ganhar insights úteis.

Prever resultados de partidas desportivas é não só interessante para treinadores, como também para fãs e apostadores. Também é interessante como problema de pesquisa, devido à dificuldade inerente: o resultado de um jogo é dependente de vários fatores, como por exemplo, a moral de uma equipa (ou jogador), técnica, estratégia do treinador, etc. Deste modo, mesmo para especialistas, é difícil prever os resultados de jogos. Neste capítulo são referidos alguns trabalhos de investigação da área, de forma a ser efetuado um estudo do estado da arte.

2.3.1. Estado da Arte

Na realização do estudo do estado da arte foram identificadas diversas investigações publicadas na área, aplicando o *machine learning* na previsão de resultados em diferentes variedades de desportos. A aplicação do *Data Mining* no basquetebol foi iniciada na década de 90 pela IBM (Bhandari et al., 1997). O objetivo da ferramenta foi auxiliar uma equipa de gestão da NBA a descobrir padrões escondidos de estatísticas do basquetebol, usando técnicas de *data mining*.

O modelo desenvolvido usou uma técnica de *Data Mining* denominada de *Attribute Focusing*. Esta técnica, comparou a distribuição geral de um atributo com as suas distribuições de diferentes subconjuntos de dados. Posteriormente, se algum subconjunto mostrasse uma distribuição carateristicamente diferente, a combinação de atributos que descreviam o subconjunto eram marcados como interessantes (Bhandari et al., 1997).

Contudo, este modelo seria meramente informativo, alertando apenas o utilizador sobre potenciais distribuições de dados fora do normal, tendo apenas uma explicação ou interpretação limitada sobre estatísticas dos jogadores (Bhandari et al., 1997).

Outro trabalho de investigação notável foi um estudo conhecido como *Hidden Markov Models* (HMMs) (Testolin et al., 2016). Estes modelos foram usados recentemente com o propósito de modelar a progressão de resultados de partidas (vitórias/derrotas) em diferentes épocas / momentos desportivos, aplicando estatísticas avançadas de jogos de NBA como caraterísticas e ter a capacidade de prever o resultado de uma partida.

Estes modelos de previsão revelaram uma precisão bastante promissora para o tema em questão. Os HMMs foram utilizados por outros investigadores como combinações de modelos com os desenvolvidos pelos mesmos, de forma a obterem melhores resultados, não só no basquetebol, como também em outros desportos (Testolin et al., 2016).

Também (Leung & Joseph, 2014), desenvolveram uma ferramenta de *Data Mining* para previsões de resultados de jogos e descoberta de informações úteis nos mesmos. Estes, ao testarem em dados de jogos universitários, obtiveram resultados bastante sorridentes em termos de precisão. A técnica desenvolvida é baseada numa combinação de quatro diferentes medidas de resultados históricos de jogos. O conceito principal era prever o resultado de um jogo entre duas equipas através da análise de um conjunto de equipas que apresentavam similaridades com cada equipa desse jogo, obtendo os resultados dos jogos entre as equipas em cada um dos dois conjuntos, e usando esses resultados como base para previsão.

Existiram vários estudos relacionados com os temas acima referidos, sendo que os mais bem-sucedidos teriam uma eficácia de cerca de 75%. Mas, como é possível verificar, o foco estaria maioritariamente na previsão de resultados de jogos. Não existiria, portanto, um aprofundamento de investigações na avaliação e previsão do desempenho de atletas, que possuem um grande impacto no resultado de um jogo e na sua respetiva equipa (Leung & Joseph, 2014).

2.4. Metodologia CRISP-DM

A metodologia *CRISP-DM* consiste num modelo de processo de *Data Mining*, que detalha abordagens normalmente utilizadas por especialistas em *Data Mining*, para a resolução de problemas. A nomenclatura é proveniente de "*Cross Industry Standard Process for Data Mining*" (Martinez-Plumed et al., 2021).

Deste modo, esta metodologia é usada como referência para construção e desenvolvimento de modelos de *Machine Learning*. A maior vantagem é que promove uma metodologia comum para comunicação, auxiliando na conexão das mais variadas ferramentas técnicas e pessoas com diferentes habilidades e currículos, levando ao desenvolvimento de projetos de um modo eficiente e eficaz. O método *CRISP-DM* envolve seis fases:

2.4.1. Business Understanding

O *Business Understanding* inclui a compreensão do problema, o objetivo da sua modelação, e as caraterísticas específicas do desporto em análise. Tudo isto envolve um conhecimento de como o desporto é jogado e que fatores podem potencialmente influenciar e determinar resultados de jogos. Este conhecimento pode ser obtido através de conhecimento pessoal (adepto do desporto em análise), ou obtido por questionários existentes em artigos publicados, ou até mesmo consultando especialistas do desporto em questão.

É necessária uma certa objetividade quanto ao objetivo do desenvolvimento do modelo, ou seja, o objetivo da análise terá de ser claro e bem definido. Entre vários objetivos, podem servir como exemplo objetivos de competição com previsões de especialistas, competições *online*, ou até mesmo o uso dos resultados do modelo de previsão para apostas desportivas (Martinez-Plumed et al., 2021).

2.4.2. Data Understanding

A *Data Understanding* envolve coletar a informação e iniciar processo de aprendizagem das caraterísticas das mesmas. A informação obtida será uma coletânea de objetos de informação, bem como dos seus atributos. Já um atributo consiste numa propriedade ou caraterística de um objeto, como por exemplo, a cor dos olhos de um indivíduo, a temperatura, entre outros (Huber et al., 2019).

É também necessário identificar problemas de qualidade da informação obtida, obtendo insights, ou detetando subconjuntos interessantes, que possam formar hipóteses sobre informação oculta (Shafique & Qaiser, 2014).

A granularidade/ nível dos dados é algo necessário a ter em consideração. Outros estudos da área possuíam um nível jogo/equipa. É importante salientar que também é possível incluir dados de nível jogador, que consistiria em estatísticas de jogadores que participaram nos jogos em análise (Mahmood et al., 2021).

Para além disso, é preciso ter em conta a definição da variável de classe. Uma extensa quantidade de estudos na área considera os modelos de previsão nos desportos como de classe 2 ou 3, significando equipa da casa vence, equipa de fora vence, ou equipa da casa vence, empate, equipa de fora vence, respetivamente. Também poderá ser considerado um problema de classe numérica, usando técnicas de regressão para previsão de margens de pontos, por exemplo (Mahmood et al., 2021).

2.4.3. Data Preparation

Esta fase inclui todas as atividades requeridas para a construção final da *data set* (data que servirá de alimentação para o modelo de previsão), provenientes dos dados iniciais. Esta tarefa normalmente ocupa 90% do tempo. Este processo é fundamental para produzir modelos válidos e confiáveis. Este processo envolve (Huber et al., 2019):

- Seleção de Informação
- Limpeza de informação
- Construção da informação
- Integração da informação
- Amostragem – principal técnica empregada para a seleção da informação. O princípio chave para uma amostragem eficaz é o seguinte:
    - Usar uma amostra funcionará quase tão bem quanto usar todo o conjunto de dados, se a amostra for representativa.
    - Uma amostra é representativa se tiver aproximadamente as mesmas propriedades (de interesse) que o conjunto original de dados.

É também importante referir os diferentes tipos de amostragem:

- *Simple Random Sampling* – existe uma probabilidade igual de seleção de um determinado item
- *Sampling without replacement* – por cada item selecionado, este é removido da população
- *Sampling with replacement* – objetos não são removidos da população quando selecionados. O mesmo objeto pode ser selecionado mais do que uma vez.
- *Stratified Sampling* – Separação da informação em várias partições, retirar amostras aleatórias de cada partição (Huber et al., 2019).

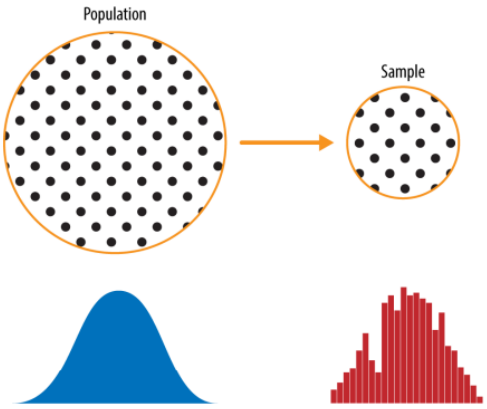


Figura 5 - Sampling – Demonstração (Fonte: (Huber et al., 2019))

Posteriormente, nesta fase é também necessário remover *outliers*, remover campos com nenhuma ou pouca variabilidade, bem como examinar o número de valores de um campo distintos:

- Regra: remover um campo onde quase todos os valores são o mesmo

- Remover campos chave, onde todos os valores são distintos, visto que estes não possuem qualquer semântica associada (Huber et al., 2019).

2.4.4. Modeling

Nesta fase, vários algoritmos de modelação são treinados, para ser escolhido o algoritmo mais qualificado.

Deve ser selecionada e aplicada uma variada gama de técnicas de modelação. Tipicamente, existem várias técnicas que podem ser usadas para resolver o mesmo problema de *data mining*. No entanto, cada técnica possui parâmetros específicos e requisitos sobre a forma de dados. Considerando o referido, é usualmente necessário (Gama et al., 2012):

- Selecionar a técnica de modelação
- Gerar o modelo de teste
- Desenvolver e construir o modelo
- Avaliar o modelo desenvolvido

Como foi explicado acima, existem vários tipos de modelação de dados. Um deles é denominado de modelo preditivo. A modelação preditiva tem como objetivo prever o valor de um atributo específico com base nos valores de outros atributos. O atributo previsto é conhecido como target ou variável dependente e o outro atributo é denominado de variável independente (Gama et al., 2012).

Para além desta, existem também modelos baseados em distâncias, que permitem a utilização de instâncias de dados passados, com *outputs* conhecidos, para previsões de um *output* desconhecido de uma nova instância de dados. São exemplos destes modelos o *Nearest Neighbour* (vizinho mais próximo), *k-Nearest Neighbour* e o *Case-Based Reasoning* (Gama et al., 2012).

Existem também modelos probabilísticos, como por exemplo o *Bayesian Learning, Naive Bayes* e *Bayesian Networks*, nos quais têm como fundamento interpretar as caraterísticas de um *dataset* como variáveis aleatórias e, consequentemente, pensar nos algoritmos como modelos probabilísticos (Gama et al., 2012).

Outro exemplo de um tipo de modelo é a *Decision Tree*. Neste, um novo exemplo é classificado, submetendo-o numa série de testes que determinarão a classe do exemplo. Estes testes são organizados segundo uma estrutura hierárquica chamada de árvore de decisão (Gama et al., 2012).

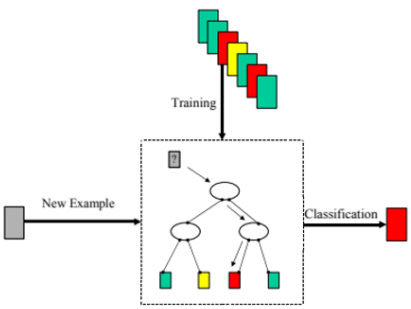


Figura 6 - Decision Tree (Fonte: (Gama et al., 2012))

Os *Optimization-based Methods* são famílias de *machine learning* baseadas em redes neuronais artificiais (RNAs), com múltiplas camadas escondidas. Estas redes são aplicadas em muitas implementações diferentes, com pequenas variações nas suas estruturas, como por exemplo: redes neuronais recorrentes (RNN) e redes neuronais artificiais (ANN) (Lee, 2014).

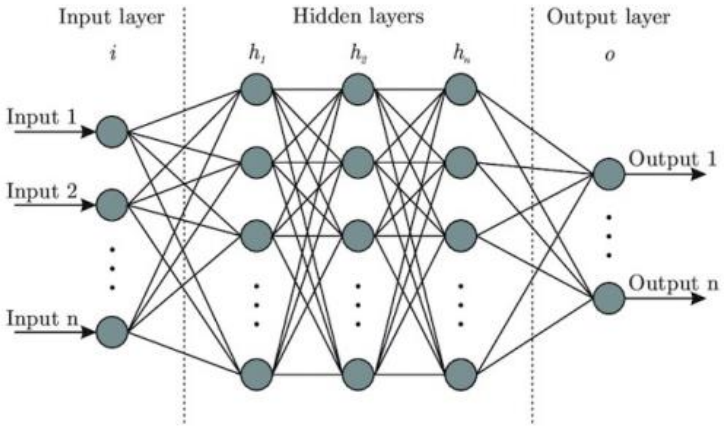


Figura 7 - Multilayer Neural Networks (Fonte: (Lee, 2014))

Redes Neuronais Artificiais (RNAs):

- São inspirados em neurónios biológicos
- Cada neurónio da rede recebe uma ou mais entradas
- Uma função de ativação é aplicada aos inputs, que determina o output do neurónio – chamado de nível de ativação.

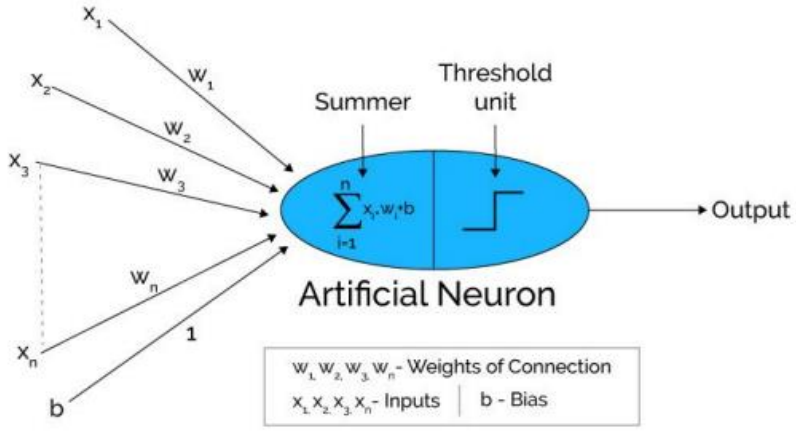


Figura 8 - Artificial Neuron (Fonte: (Lee, 2014))

Vantagens:

- Precisão elevada, mesmo para problemas complexos;
- Processamento distribuído – o conhecimento é distribuído através dos pesos das ligações
- Robusto no controlo, mesmo que contenha erros

- Capacidade para lidar com atributos redundantes.

Desvantagens:

- Difícil de determinar a topologia ótima da rede para um problema
- O espaço de procura (espaço da função erro) tem regiões de mínimos globais. É possível ficar preso no local mínimo ou levar muito tempo para encontrar a solução
- Difícil de usar – tem vários parâmetros para definir, requerendo um longo tempo de treino
- Não fornece um modelo ou explicações de resultados (Lee, 2014).

Redes neuronais recorrentes (RNNs):

As redes neuronais recorrentes são as ferramentas de redes neuronais para problemas relacionados com dados sequenciais.

Nesta recorrência, os nós conectam-se a outros nós, ou a eles mesmos, e o fluxo de informação é multidirecional. Os sistemas nervosos biológicos apresentam altos níveis de recorrência (Lee, 2014).

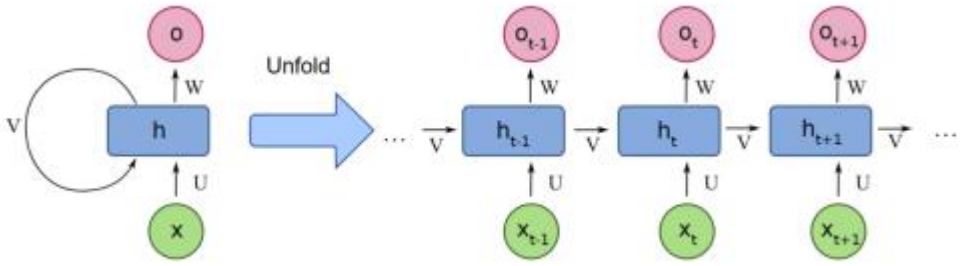


Figura 9 - Recurrent Neural Networks (Fonte: (Lee, 2014))

Problemas:

- Incapacidade de reter informação quando a sequência fornecida é longa;
- Tendem a esquecer informações que foram fornecidas à várias iterações atrás. Isto limita o desempenho de aprendizagem.

Por fim, é possível a combinação de múltiplos algoritmos de *machine learning*, pois não existe um algoritmo que é sempre o mais preciso, em qualquer situação.

Combinando vários algoritmos, como se encontra esquematizado na figura 10, é garantida uma maior precisão, bem como uma maior flexibilidade às diversas situações em estudo (Gama et al., 2012).

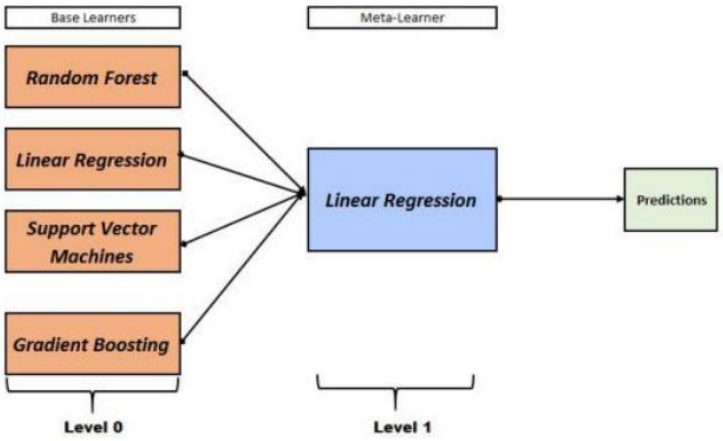


Figura 10 - Combinar múltiplos algoritmos (Fonte: (Gama et al., 2012))

2.4.5. Evaluation

A quinta fase é referente à avaliação do modelo. Para este processo, são efetuados processos com o objetivo de avaliar a performance de um modelo, obtendo estimativas confiáveis e comparar os resultados entre modelos (Gama et al., 2012).

Para a avaliação do desempenho ser realizada, são necessárias métricas de avaliação de um modelo. A primeira métrica que poder ser usada para o efeito é a "Matriz confusão", que proporcionada uma representação mais detalhada das classificações corretas e incorretas em cada classe (Gama et al., 2012).

		Predicted C	
		Yes	No
Actual	Yes	a (TP)	b (FN)
	No	c (FP)	d (TN)

Figura 11 - Confusion Matrix (Fonte: (Gama et al., 2012)**)**

Outra métrica comum é a precisão, determinando o número de previsões corretas em relação a todas as previsões realizadas (Huber et al., 2019).

Para além destas, existem matrizes de custo, rácio de instâncias positivas (*recall*), rácio de especificidade, rácio de precisão, F1-Score (média entre a precisão e *recall*), e, por fim, a curva ROC. A curva ROC é remetente à caraterística da operação do recetor e é um gráfico. A performance de cada classificador é representada por cada ponto assinalado na curva ROC, no gráfico apresentado abaixo (Huber et al., 2019).

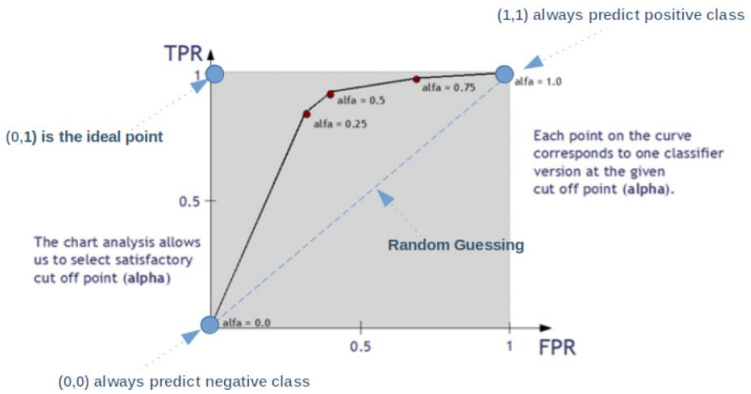


Figura 12 - Curva ROC (Fonte: (Huber et al., 2019))

2.4.6. Deployment

Por fim, a última fase é denominada de desenvolvimento (*deployment*). Esta fase consiste no alinhamento e definição das fases de implementação do projeto de *Data Mining*, tendo em consideração que o modelo resultante da fase de modelação precisa, além de ser flexível às necessidades da organização, ser também interpretável e com capacidade operacional (Bunker & Thabtah, 2019).

É elaborado o relatório final do processo, que apresenta os resultados obtidos e possíveis alternativas de ação no processo aplicado na organização (Bunker & Thabtah, 2019).

2.5. KNIME

A plataforma KNIME Analytics é um software de *No and Low code* para todas as necessidades relacionadas com dados (Fillbrunn et al., 2017).

Esta aplicação contém uma parte significativa das principais técnicas e algoritmos de controlo e gestão de dados e de *Machine Learning*, baseados em programação visual (Fillbrunn et al., 2017).

O KNIME pode ser alimentado por várias fontes de informação, tais como PDF, CSV, JSON, XML, entre outros (Fillbrunn et al., 2017).

Figura 13 - Exemplo de Modelo de Importação de Ficheiros (Fonte: (*KNIME*, n.d.)

Para além disso, é prática em termos de tratamento dos dados em função dos objetivos pretendidos. É possível extrair estatísticas de informação tais como médias ou desvios padrões, e ordenar, agregar, conectar a filtrar dados (Warr, 2012).

Sendo uma aplicação para desenvolvimento de modelos de *Machine Learning*, esta também possui elementos de otimização de performance dos modelos desenvolvidos, assim como métodos de validação dos mesmos (Warr, 2012).

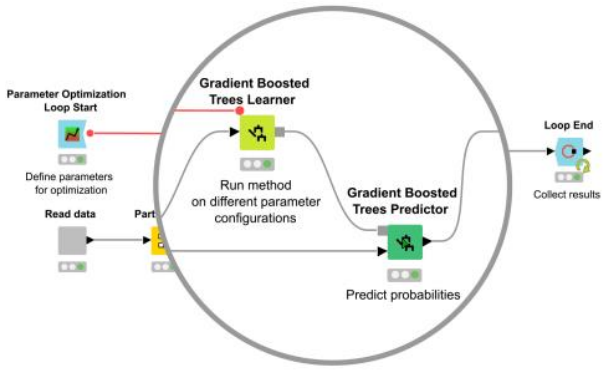


Figura 14 – Exemplo de Otimização de Modelos (Fonte: (*KNIME*, n.d.)

Por último, é possível visualizar dados de uma forma intuitiva e simples, quer sejam gráficos de dispersão, gráficos de barras, ou mesmo gráficos avançados tais como um *heat map* (Warr, 2012).

Figura 15 - Exemplo de Visualização de Dados
(Fonte: (*KNIME*, n.d.)

3. METODOLOGIA E DISCUSSÃO DE RESULTADOS

No presente capítulo é efetuada uma descrição extensiva de cada passo da metodologia CRISP-DM. Começando pelo *Business Understanding*, onde são definidos os objetivos do estudo e do *data mining*. É também efetuada uma avaliação da situação atual e é elaborado um plano do projeto.

Posteriormente é tratado o *Data Understanding*, que consiste numa análise extensiva de várias variáveis da base de dados. De seguida é feita a preparação dos dados (*Data Preparation*), com vista a tornar a base de dados mais robusta. Na fase do *Modeling*, são descritos os modelos de previsão desenvolvidos para a previsão de determinados acontecimentos.

Por fim é efetuada uma reflexão e discussão de resultados obtidos.

3.1. Business Understanding

Determinar os objetivos do estudo

Este estudo tem como finalidade a formulação de conhecimento procurando padrões de dados que suportem hipóteses, usando para isso uma variada gama de métodos e técnicas de *data mining*. Baseado nestes padrões, pretende-se contruir modelos de previsão através do conceito de *machine learning*, permitindo assim visualizar a performance dos atletas e equipas da NBA, com auxílio de estatísticas relativas aos mesmos.
Sendo este um processo que depende intimamente dos dados recolhidos, há fatores que podem colocar a precisão do estudo em causa, nomeadamente o tratamento de dados entre as várias correlações, bem como todas as variáveis envolventes.
Do ponto de vista do estudo, um resultado bem-sucedido seria um em que seja possível tecer conclusões com base nas previsões, assim como através das correlações entre as variáveis de estudo.

Avaliação da situação

Para execução de um estudo aprofundado e fidedigno é necessário recorrer a *datasets* disponibilizados por instituições credíveis. Para este efeito foi usada como fonte de informação o *Website ESPN*. Este estudo será focado no campeonato de basquetebol americana (NBA) em detrimento dos campeonatos de todo o mundo, pois este é considerado o principal campeonato de basquetebol do mundo.
Ao longo deste processo foram-se reunindo alguns *Business Goals* de forma a tornar a análise mais rigorosa possível.
Deste modo, pretendemos a hipótese de que quanto melhores as estatísticas, bem como a eficiência das mesmas, traduzirá em melhores resultados para uma equipa. Para além disso, um melhor elo de entre as duas equipas traduzir-se-á na equipa vencedora do jogo. Também será estudada a hipótese de que uma alta performance de um jogador levará a que o seu ranking seja superior quando comparado aos restantes jogadores.

Estas análises serão alguns dos objetos de estudo nas quais verificaremos a veracidade dos mesmos após o desenvolvimento do modelo acima referido. Em termos do cliente final, os objetivos do negócio passam por:

- Desenvolvimento de modelos capazes de responder às questões que o cliente pretende que sejam respondidas;
- Tornar os modelos flexíveis e robustos para que sejam retornados bons resultados independentemente dos dados a analisar;

Objetivos da data mining

Desenvolver um estudo que poderá ser usada para prever o vencedor de um jogo entre as duas equipas do campeonato americano de basquetebol, e, consequentemente, auxiliará na análise e conclusões acerca de relações entre os diferentes intervenientes de um jogo de basquetebol, como por exemplo, os jogadores. Os objetivos também passam por:

- Ter uma perceção aprofundada das bases de dados utilizadas, efetuando, por exemplo, análises estatísticas;
- Tratamento dos dados e posteriormente a seleção dos mesmos;
- Desenvolver modelos de previsão de forma a responderem aos objetivos do estudo, baseados em *machine learning*.

Plano do projeto

O plano do projeto passa pelo estudo do estado da arte, por obter um conhecimento especializado dos dados e integrá-los. Este conhecimento envolve a recolha, análise e preparação dos dados.

Implementar um modelo assim que este esteja desenvolvido, passando antes disso por uma fase de validação. Caso não seja validada, é necessário obter mais dados, criar novos recursos e executar uma ampla gama de modelos. Por fim, avaliar o modelo usando a precisão e métricas de medida. Para desenvolvimento do tratamento de dados será utilizada a ferramenta *Knime*.

3.2. Data Understanding

Neste capítulo será realizada uma análise de várias variáveis consideradas importantes, não só para a base de uma correta tomada de decisões, como também para uma melhor interpretação dos dados.

3.2.1. Recolha de Dados

A análise foi feita com base nos dados fornecidos pela *ESPN*, das épocas entre 2016 e 2020 do campeonato americano de basquetebol, pois estes forneciam uma maior consistência em termos

de informação, bem como uma informação mais completa em termos da análise pretendida. Foram utilizadas duas bases de dados: uma relativa a estatísticas dos jogadores num determinado jogo, e outra relativa aos jogos entre duas equipas e resultados dos mesmos.

3.2.2. Descrição dos Dados

Inicialmente foi efetuada uma descrição das variáveis da base de dados. Relativamente aos dados relativos aos jogos e resultados:

- **Game_id**: Id relativo a um jogo entre duas equipas, ou seja, é um código único que permite identificar um jogo em específico. Esta variável tem um atributo categórico nominal.
- **Home_team_id**: Código identificador da equipa da casa. Esta variável tem um atributo categórico nominal.
- **Away_team_id**: Código identificador da equipa visitante. Esta variável tem um atributo categórico nominal.
- **Home_team_name**: Nome da equipa da casa. Esta variável tem um atributo categórico nominal.
- **Away_team_name**: Nome da equipa visitante. Esta variável tem um atributo categórico nominal.
- **Date**: Data em que o jogo se realizou. Esta variável tem um atributo categórico ordinal.
- **Home_points:** Pontos marcados pela equipa da casa. É uma variável de atributo numérico discreta. A figura 16 representa o boxplot relativo aos pontos da equipa da casa, centrando-se entre os 100 e 117 pontos por jogo.

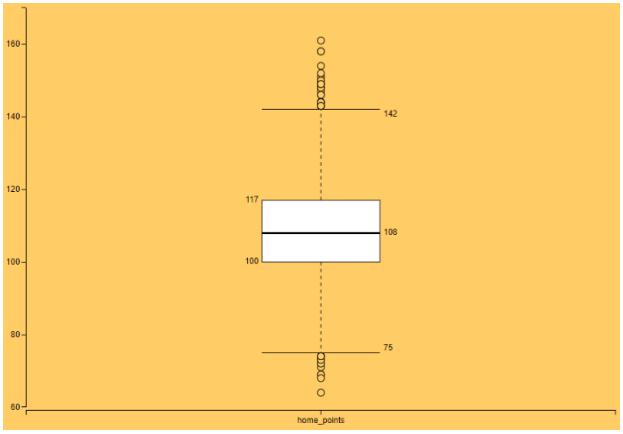


Figura 16 – Boxplot pontos da equipa da casa

- **Away_points:** Pontos marcados pela equipa visitante. É uma variável de atributo numérico discreta. Na figura 17 é possível verificar o boxplot relativo aos pontos da equipa visitante, centrando-se entre os 97 e os 114, valores menores que os da equipa da casa.

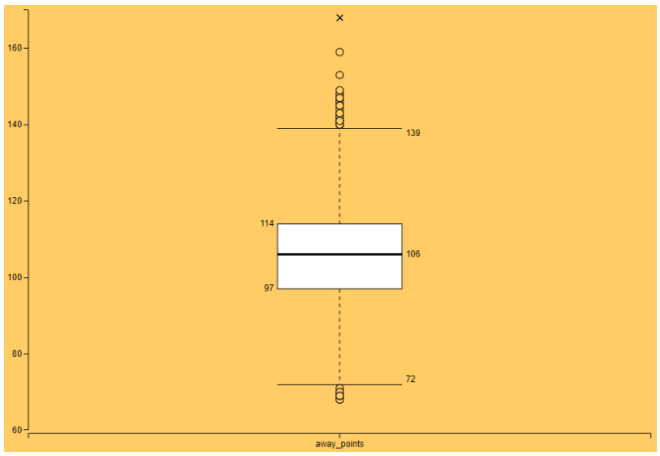


Figura 17 - Boxplot Pontos da Equipa Visitante

- **k:** constante que dita o impacto de um determinado jogo no rating de qualidade da equipa. É uma variável de atributo numérico contínua. Na figura 18 é demonstrado o boxplot da constante "k".

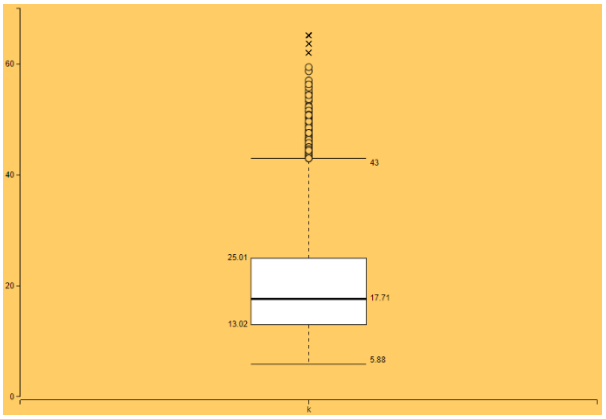


Figura 18 - Boxplot constante K

- **Elo_diff:** diferença entre o rating de qualidade entre as duas equipas. É uma variável de atributo numérico contínua. A figura 19 é ilustrativa do boxplot do diferencial dos ratings elo. Os valores centram-se entre os 53 e os 206.

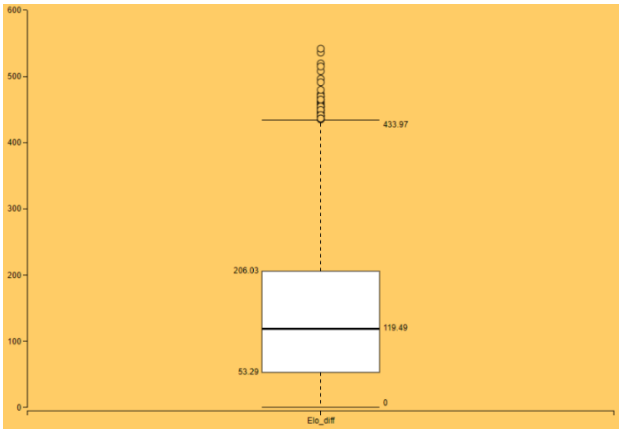


Figura 19 - Boxplot diferencial dos ratings elo

- **MOVwinner:** diferença de pontos entre a equipa vencedora e a equipa derrotada. É uma variável de atributo numérico contínua. Na figura 20 é possível visualizar o boxplot relativo ao MOV, sendo que os valores da margem de vitória centram-se entre os 5 pontos e os 16.

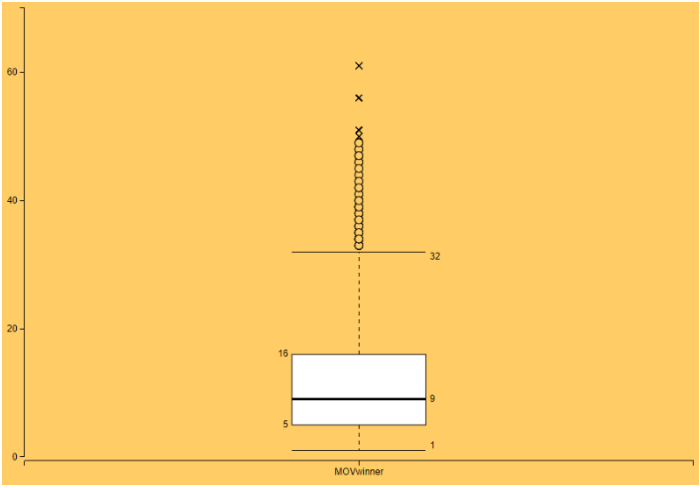


Figura 20 - Boxplot da margem de vitória

- **S home team:** variável que possui o valor 1, caso a equipa da casa tenha vencido o jogo; caso contrário, possui o valor 0. Esta variável tem um atributo categórico binário.
- **S away team:** variável que possui o valor 1, caso a equipa visitante tenha vencido o jogo; caso contrário, possui o valor 0. Esta variável tem um atributo categórico binário.
- **E home Team Values:** probabilidade que a equipa da casa tinha para vencer o jogo. É uma variável de atributo numérico contínua.
- **E away Team Values:** probabilidade que a equipa visitante tinha para vencer o jogo. É uma variável de atributo numérico contínua.
- **R home new:** elo *rating* de qualidade da equipa da casa. É uma variável de atributo numérico contínua. Na figura 21 é possível visualizar o boxplot relativo ao *rating* elo da equipa da casa.

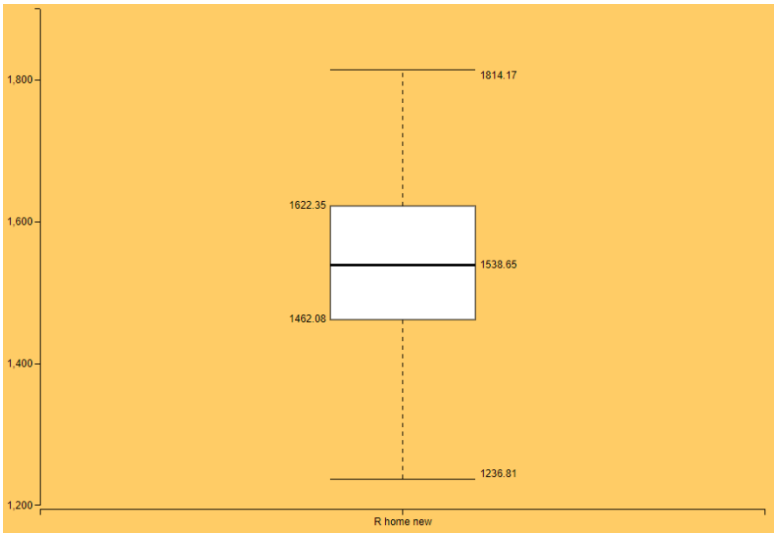


Figura 21 - Boxplot *rating* elo da equipa da casa

- **R away new:** elo *rating* de qualidade da equipa visitante. É uma variável de atributo numérico contínua. A figura 22 é representativa do boxplot do *rating* elo da equipa visitante.

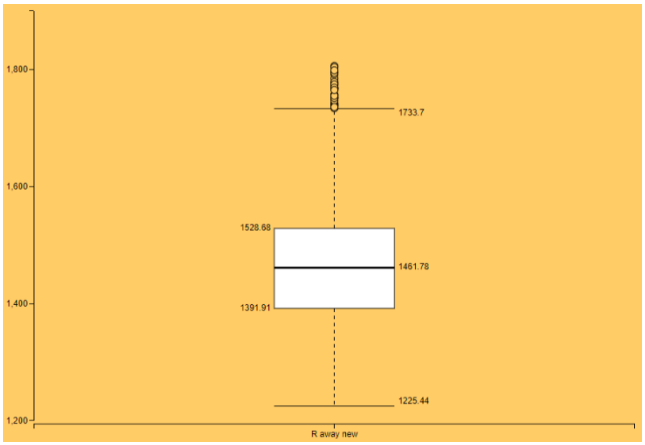


Figura 22 - Boxplot *rating* elo da equipa visitante

Em relação aos dados referentes à base de dados dos jogadores:

- **First_Name:** Primeiro nome do jogador. Esta variável tem um atributo categórico nominal.
- **Last_Name:** Primeiro nome do jogador. Esta variável tem um atributo categórico nominal.
- **Player_id:** Id relativo a um jogador, ou seja, é um código único que permite identificar um jogador em específico. Esta variável tem um atributo categórico nominal.
- **Position (POS):** Posição de um jogador dentro do campo. Esta variável tem um atributo categórico nominal. Os valores desta variável podem ser:
    - *Point-Guard* (PG ou G) (armador ou base) – é responsável por organizar as jogadas ofensivas, criando oportunidades para os colegas de equipa pontuarem. Deverá ser o jogador com o maior tempo com a bola na sua posse.
    - *Center* (C) (poste ou pivô) - é o jogador que fica mais próximo do cesto, e com importantes funções defensivamente.
    - *Forward* (F) (ala de força) – defensivamente, o jogador tem a função de bloquear os adversários e ganhar os ressaltos. Ofensivamente, atua tanto perto do cesto como também realiza lançamentos de média distância.
- **Points (PTS):** Número médio de pontos por jogo. É uma variável de atributo numérico discreta. A figura 23 ilustra o boxplot relativo aos pontos marcados por um jogador num determinado jogo. Os valores centram-se entre os 4 e os 15, mas também é possível verificar registos de jogadores em que marcaram mais de 30 pontos em determinados jogos.

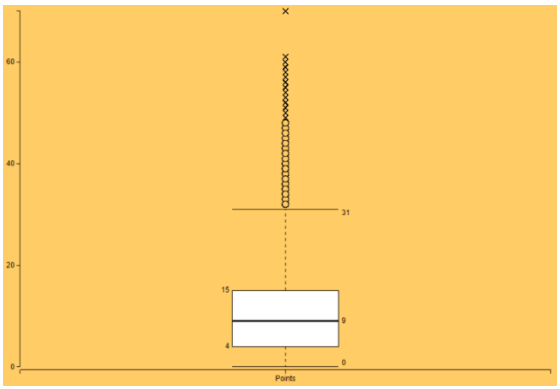


Figura 23 - Boxplot Pontos de um Jogador num determinado jogo

- **Free_Throw_Percent (FT%):** Percentagem de lançamentos livres concretizados em relação aos tentados. É uma variável de atributo numérico contínua.
- **Two_Pt_Percent (2PT%):** Percentagem de lançamentos de dois pontos concretizados em relação aos tentados. É uma variável de atributo numérico contínua.
- **Three_Pt_Percent (3P%):** Percentagem de lançamentos de 3 pontos concretizados em relação aos tentados. É uma variável de atributo numérico contínua.
- **Assists (AST):** Número médio de assistências conseguidas pelo jogador, por jogo. É uma variável de atributo numérico discreta. A figura 24 apresenta o boxplot relativo às assistências efetuadas por um jogador num determinado jogo.

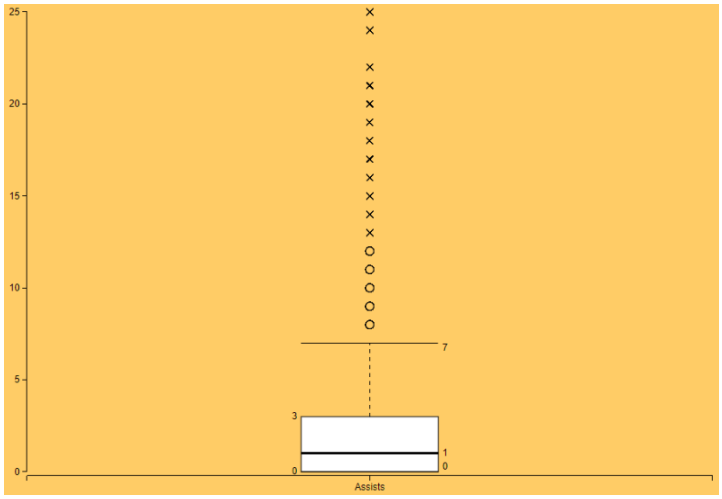


Figura 24 - Boxplot de Assistências

- **Rebounds (REB):** Número médio de ressaltos conseguidos pelo jogador, por jogo. É uma variável de atributo numérico discreta. Na figura 25 é possível visualizar o boxplot relativo aos ressaltos efetuados por um jogador num determinado jogo.

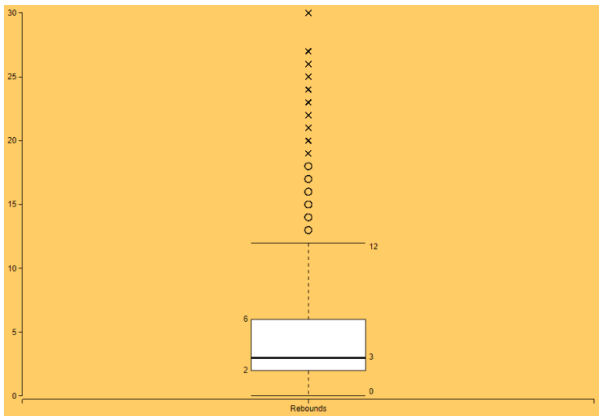


Figura 25 - Boxplot Rebounds

- **Offensive_Rebounds:** Número médio de ressaltos ofensivos conseguidos pelo jogador, por jogo. É uma variável de atributo numérico discreta. A figura 26 representa o boxplot relativo ao número de ressaltos ofensivos efetuados por um jogador num determinado jogo.

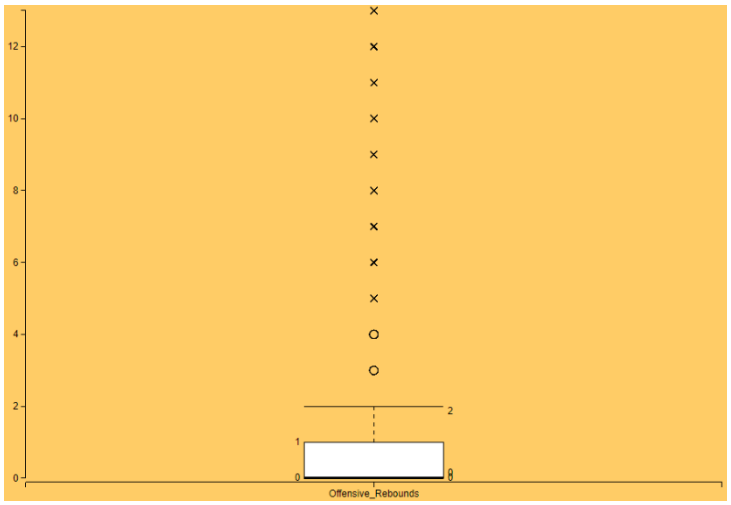


Figura 26 - Boxplot Rebounds Ofensivos

- **Steals (STL):** Número médio de roubos de bola conseguidos pelo jogador, por jogo. É uma variável de atributo numérico discreta. Na figura 27 é identificado o boxplot relativo ao número de roubos de bola efetuados por um jogador num jogo. Como é possível observar, os valores centram-se entre 0 roubos de bola e 1.

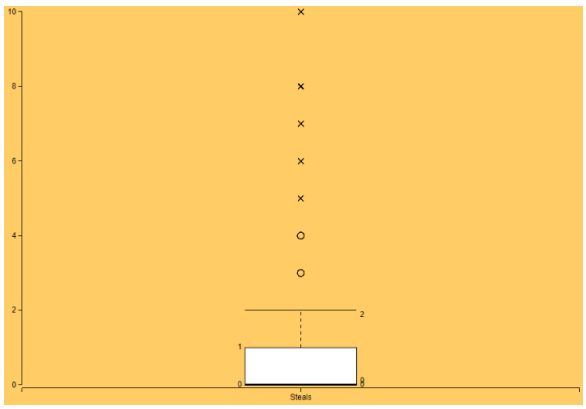


Figura 27 - Boxplot *Steals*

- **Personal_Fouls (PF):** Número médio faltas cometidas pelo jogador, por jogo. É uma variável de atributo numérico discreta. A figura 28 é ilustrativa do boxplot referente ao número de faltas pessoais efetuadas por um jogador num determinado jogo.

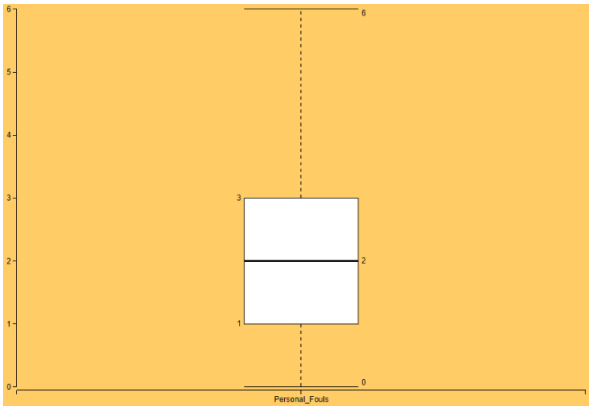


Figura 28 - Boxplot Faltas Pessoais

- **Flagrant_Fouls:** Número médio faltas flagrantes cometidas pelo jogador, por jogo. É uma variável de atributo numérico discreta.
- **Turnovers (TO):** Número médio de perdas de bola do jogador, por jogo. É uma variável de atributo numérico discreta. A figura 29 representa o boxplot relativo ao número de perdas de bola de um jogador num determinado jogo.

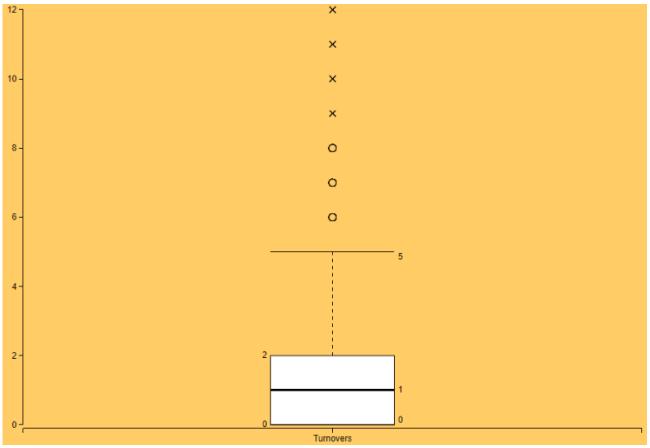


Figura 29 - Boxplot Perdas de Bola

- **Team:** Nome da equipa do jogador. Esta variável tem um atributo categórico nominal.
- **Home_away:** variável que possui o valor 1, caso a equipa do jogador seja a equipa da casa; caso contrário, possui o valor 0. Esta variável tem um atributo categórico binário.
- **Win:** variável que possui o valor 1, caso a equipa do jogador vença o jogo; caso contrário, possui o valor 0. Esta variável tem um atributo categórico binário.
- **Team_points:** Pontos marcados pela equipa do jogador. É uma variável de atributo numérico discreta. Na figura 30 é possível observar o boxplot relativo aos pontos marcados pela equipa do jogador.

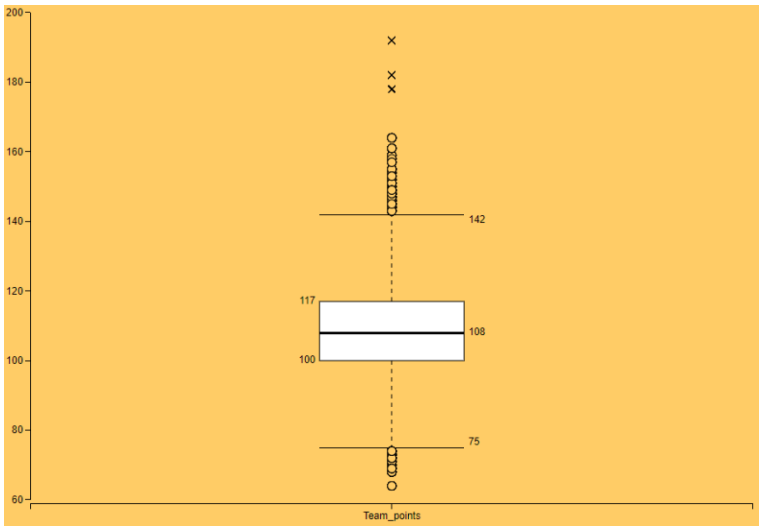


Figura 30 - Boxplot Pontos da Equipa

- **Min_played (MIN):** Número médio de minutos jogados por jogo. É uma variável de atributo numérico contínua. Na figura 31 é visualizado o boxplot referente ao número de minutos jogados por um jogador num jogo.

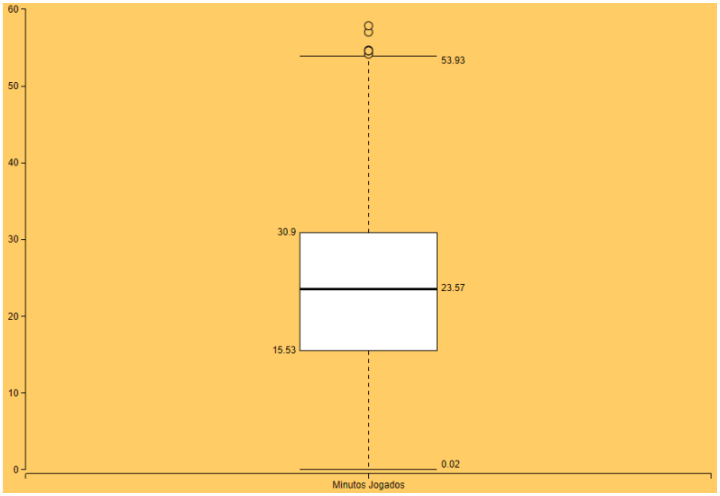


Figura 31 - Boxplot Minutos Jogados

- **Crowd:** Número espectadores por jogo. É uma variável de atributo numérico discreta.
- **Stadium_Cap:** Capacidade máxima do estádio da equipa. É uma variável de atributo numérico discreta.
- **Game_id:** Id relativo a um jogo entre duas equipas, ou seja, é um código único que permite identificar um jogo em específico. Esta variável tem um atributo categórico nominal.
- **Game_date:** Data em que o jogo se realizou. Esta variável tem um atributo categórico ordinal.

Depois de estabelecer o significado e a classificação de cada atributo, calcularam-se algumas estatísticas básicas no KNIME, encontrando-se ilustradas na tabela 1 e na tabela 2.

Tabela 1 - Estatísticas da base de dados dos jogos

	Mean	*Std. deviation*	*Variance*	*Skewness*	*Kurtosis*
home_points	108.6826	12.5924	158.5697	0.1605	0.1719
away_points	106.1136	12.6441	159.8725	0.1243	-0.0026
k	19.7406	9.0457	81.8243	0.9972	0.9678
Elo_diff	139.3775	104.3259	10883.8841	0.7722	-0.0469
MOVwinner	11.3363	8.3355	69.4801	1.3006	2.0468
S home team	0.5799	0.4936	0.2436	-0.3237	-1.8958
S away team	0.4201	0.4936	0.2437	0.3237	-1.8959
E Home Team Values	0.5874	0.1912	0.0366	-0.3443	-0.5985
E away team Values	0.4125	0.1912	0.03657	0.3443	-0.5985
R home new	1536.6609	116.9157	13669.2899	-0.2246	-0.4359
R away new	1464.7898	104.7271	10967.7726	0.3036	-0.1920

Tabela 2 - Estatísticas da base de dados dos jogadores

	Mean	*Std. deviation*	*Variance*	*Skewness*	*Kurtosis*
Points	10.2370	8.2590	68.2116	1.0509	1.2408
Free_Throw_Percent	43.1226	43.2653	1871.8911	0.1971	-1.7099
Two_Pt_Percent	45.6958	30.6382	938.7014	0.0056	-0.8021
Three_Pt_Percent	24.5095	29.4796	869.0494	1.0045	0.1167
Assists	2.2377	2.4913	6.2067	1.7610	4.0503
Rebounds	4.1663	3.4936	12.2057	1.3420	2.4029
Offensive_Rebounds	0.9487	1.3344	1.7808	2.0420	5.6115
Steals	0.7241	0.9655	0.9323	1.5890	3.1351
Personal_Fouls	1.9132	1.4651	2.1467	0.5088	-0.4391
Flagrant_Fouls	0.0048	0.0698	0.0048	14.3090	204.4911
Tech_Fouls	0.0324	0.1843	0.0339	5.8763	36.3118
Turnovers	1.2838	1.3915	1.9364	1.4168	2.6467
Home_Away	0.5060	0.4999	0.2499	-0.0242	-1.9994
win	0.5010	0.5000	0.2500	-0.0040	-2.0000
Team_points	108.6005	12.9095	166.6568	0.1927	0.4968
Minutos Jogados	22.7656	10.2468	104.9974	-0.2636	-0.7168

Foram calculados para todos os atributos os valores da média, desvio padrão, variância e medidas de forma, como *Skewness* and *Kurtosis*.

3.2.3. Exploração dos Dados

Nesta etapa, foi efetuada uma exploração dos dados, para analisar detalhadamente alguns atributos. A primeira ação de exploração dos dados consistiu em comparar as várias posições de acordo com alguns atributos. Nos gráficos apresentados na figura 32, é possível verificar que, a posição *Point-guard* (armador) é a que apresenta uma maior média de pontos marcados por jogo. Por outro lado, os postes são a posição com um maior número médio de ressaltos, devido ao facto de serem jogadores tradicionalmente mais altos, e de, por norma, jogarem perto do cesto.

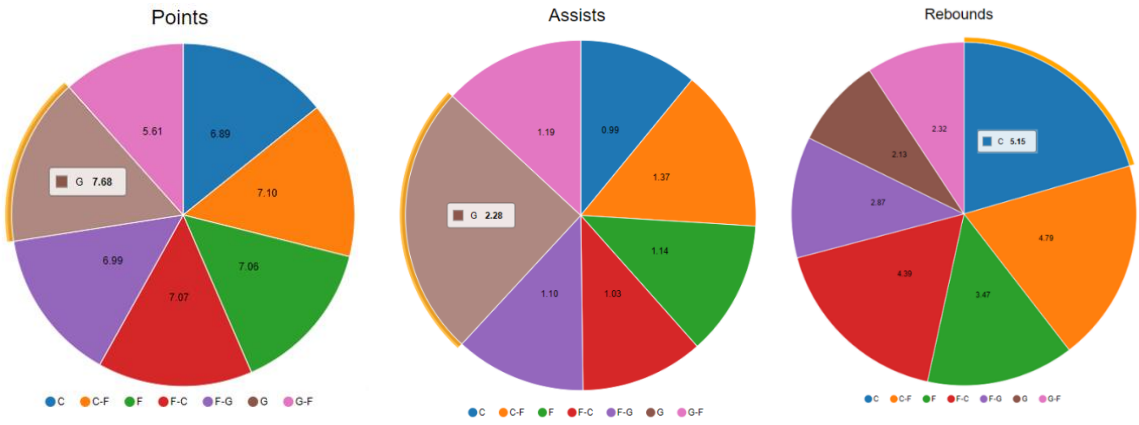


Figura 32 - Pontos, Assists e Rebounds por Posição

Também foi necessário analisar o *rating* de qualidade das equipas, bem como a eficiência dos jogadores. Desta forma, foi importante, para a primeira análise, agrupar os dados por equipa, e para a segunda análise, por jogador. Para isso, foi usado o nó *"Group By"*. Através do gráfico presente na figura 33, é possível observar que as três melhores equipas a jogar em casa, em relação ao indicador de qualidade elo *rating*, são os *Toronto Raptors*, os *Golden State Warriors* e os *San Antonio Spurs*. Por outro lado, as melhores equipas visitantes são, novamente, os *Toronto Raptors* e *Golden State Warriors*, e também os *Houston Rockets*.

Em relação aos jogadores, foi possível verificar na figura 34 que os três jogadores mais eficientes no geral, foram *Russel Westbrook*, *James Harden* e *Giannis Antetokounmpo*. A p*layer effiency* é um indicador criado para avaliar a eficiência de um jogador e será explicada mais detalhadamente no capítulo 3.3.3.

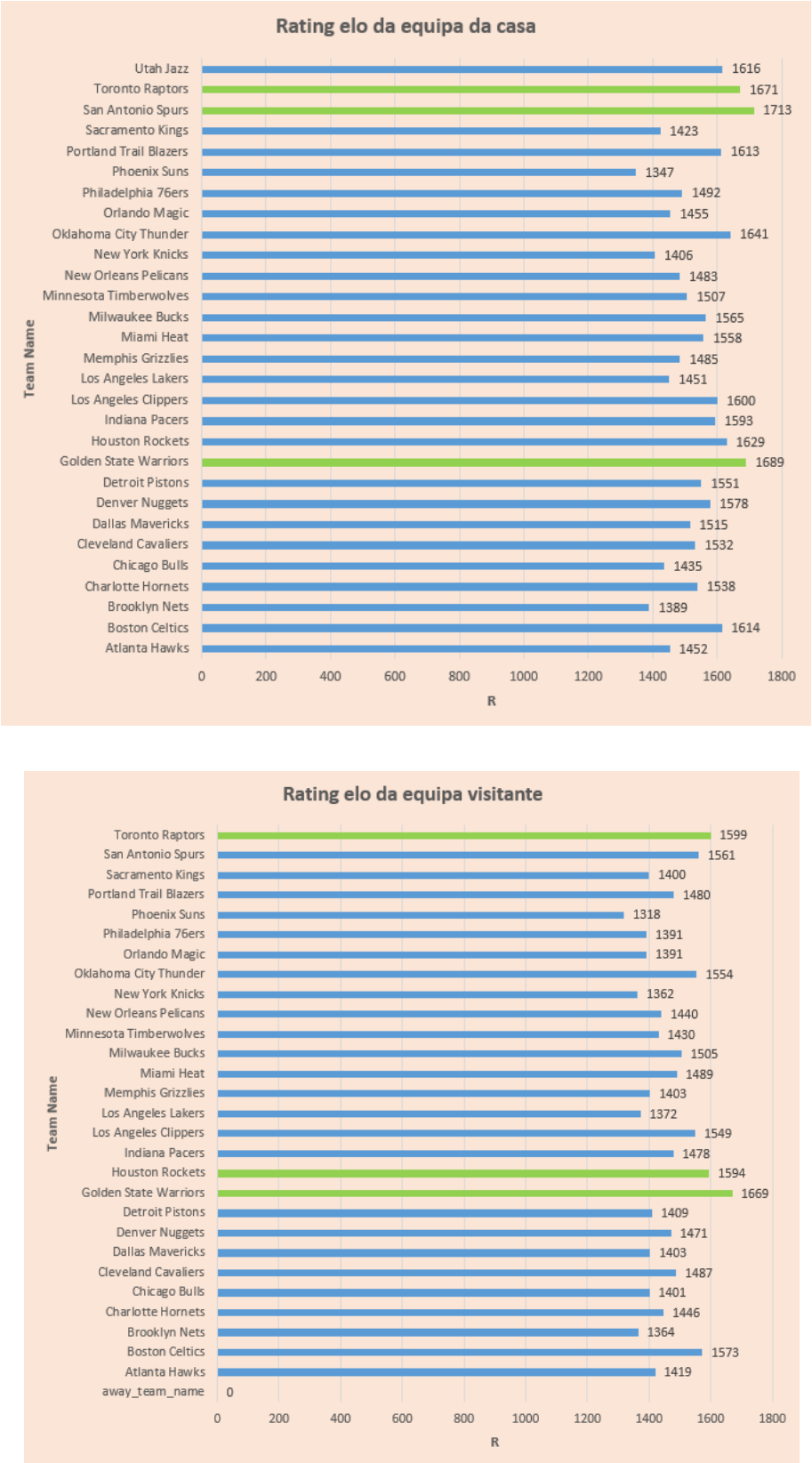


Figura 33 - *Ratings* Elo por Equipa

Figura 34 - Player Efficiency Top 10

Posto isto, foi analisada a evolução do elo *rating* de qualidade das três melhores equipas ao longo do tempo, em casa e fora, representada nas figuras 35 e 36.

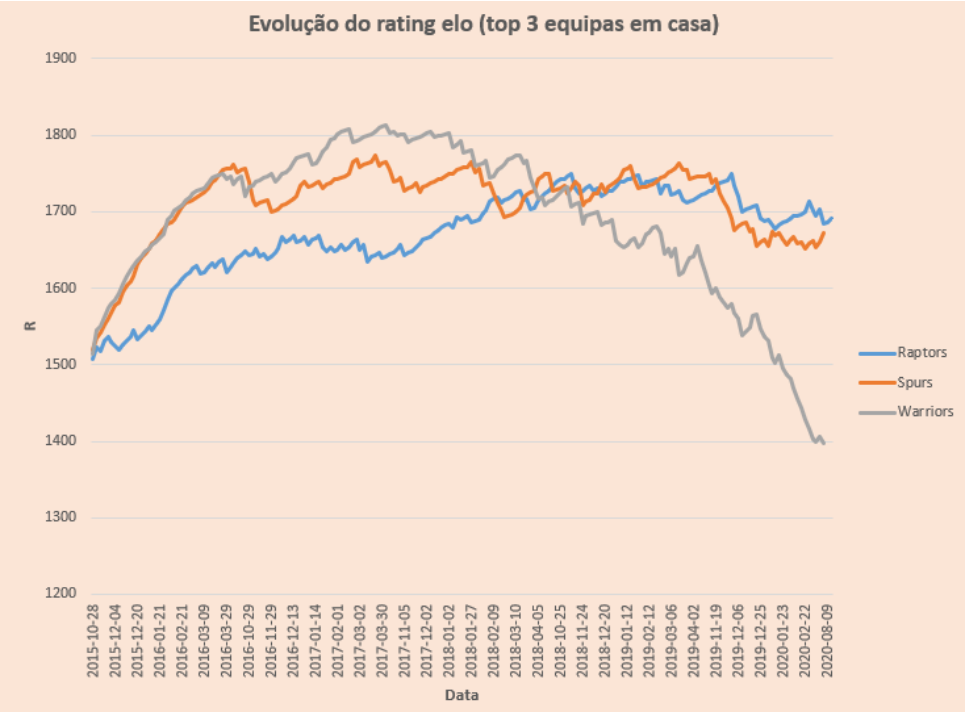


Figura 35 – Evolução do *rating* elo (top 3 equipas em casa)

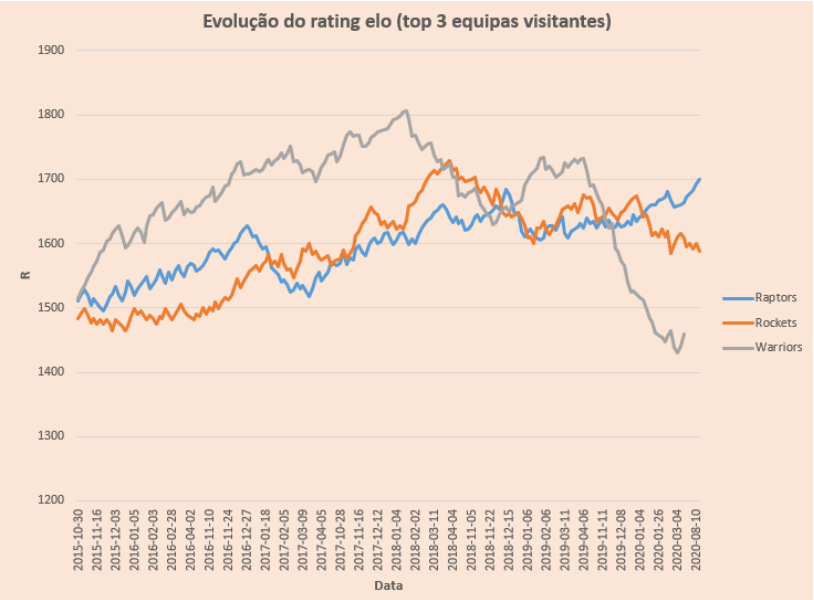


Figura 36 - Evolução do *rating* elo (top 3 equipas visitantes)

Observando as figuras 35 e 36, é possível destacar uma queda dos *Golden State Warriors* entre a época de 2019 e 2020, motivada pelas lesões de jogadores importantes para a equipa. Em relação aos *Toronto Raptors*, observa-se uma subida constante, justificando a vitória do campeonato de NBA na época de 2018-2019.

Para além disso, foi analisado de uma forma macro o impacto de um jogador num determinado jogo. Através da figura 37, verifica-se que os jogadores *James Harden*, *Russel Westbrook* e *Giannis Antetokounmpo* foram os que tiveram um maior impacto nos jogos, em 119 jogos, 142 e 149, respetivamente. O MVP representa o *most valuable Player* de um determinado jogo, e foi uma variável criada para o efeito. A métrica será explicada mais detalhadamente no capítulo 3.3.3.

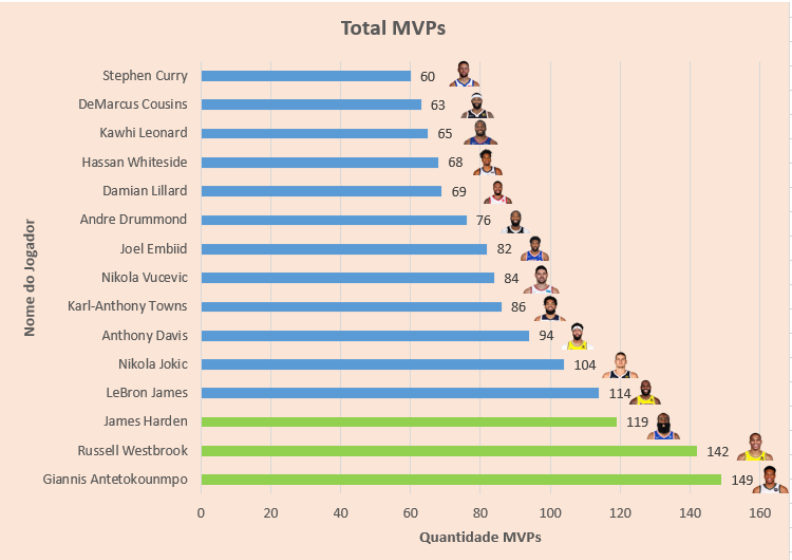


Figura 37 - Total MVPs

Numa fase posterior, achou-se relevante cruzar o número de pontos com a percentagem de lançamentos de dois pontos concretizados, de forma a verificar se existia relação entre as variáveis. Assim, através da análise da figura 38, foi possível comprovar que com o aumento do número de pontos, a percentagem de lançamentos de dois pontos concretizados tem uma tendência a estabilizar. Uma das causas da estabilização será um aumento do volume de lançamentos tentados, sendo cada vez mais difícil manter uma eficiência elevada.

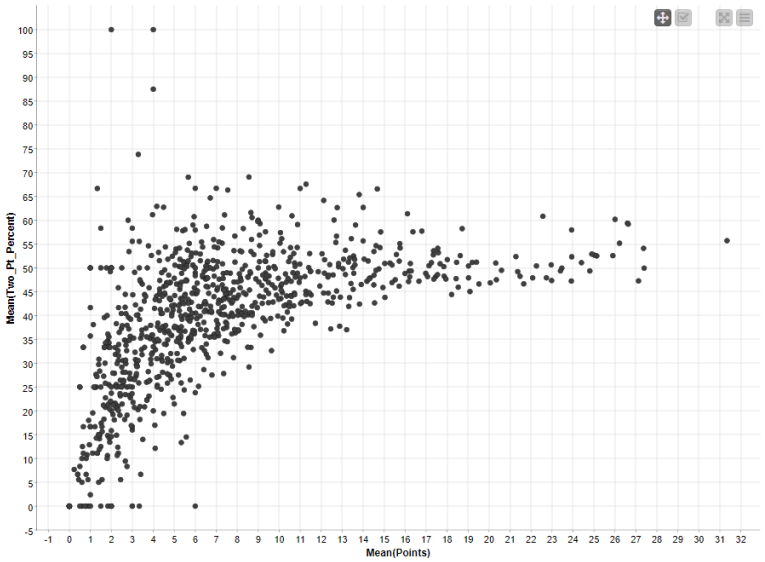


Figura 38 - Correlação entre Pontos e % de Lançamentos de 3 pontos concretizados

Também foi relevante realizar uma análise de correlação entre as variáveis do *dataset*, para tecer conclusões e perceber que informação se pode revelar importante para uma futura previsão de uma variável.

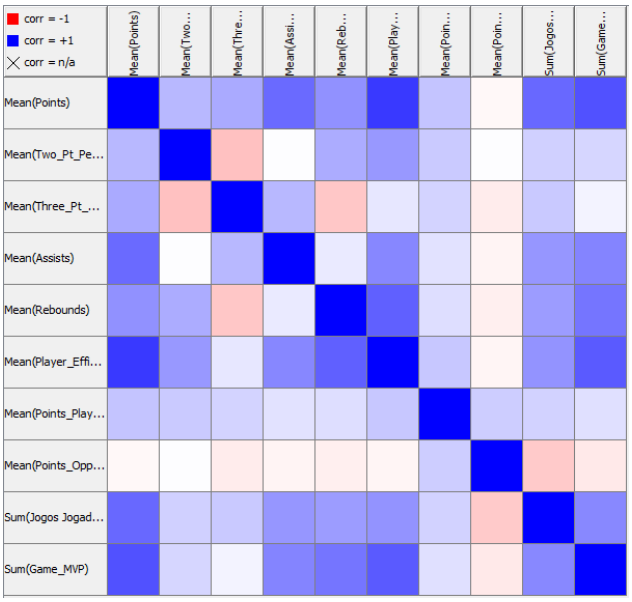


Figura 39 - Correlação entre variáveis na base de dados dos jogadores

Através da figura 39, é possível verificar que uma maior percentagem de lançamentos concretizados (dois pontos e três pontos), conduz a um aumento do número de pontos de um jogador (*Two_Pt_Percent*, *Three_Pt_Percent* com Points, de valor igual a 0.28 e 0.33, respetivamente). Para além disso, um aumento de ressaltos também possui um impacto positivo no número de pontos, pois quanto maior o número de ressaltos maior o número de posses de bola do jogador. Em relação à eficiência de um jogador, esta sofre positivamente com o aumento de todas as restantes variáveis, sendo que a correlação entre esta variável e as restantes é uma correlação fraca (entre 0.1 e 0.25). Já o MVP tem uma correlação forte com a eficiência do jogador, de 0.6485.

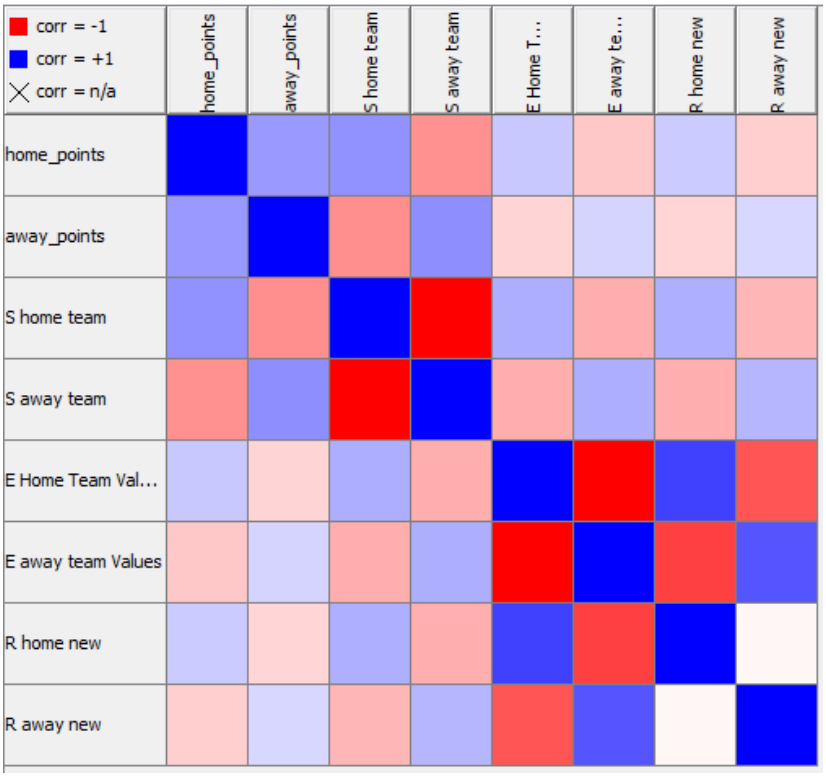


Figura 40 - Correlação entre variáveis na base de dados dos jogos

Observando a figura 40, a probabilidade de uma equipa ganhar o jogo está fortemente correlacionada com o *rating* de qualidade da equipa (0.74). Isto significa que existe uma correlação linear positiva forte entre as duas variáveis referidas.

3.2.4. Verificação da Qualidade dos Dados

Após análise das variáveis do *dataset* foram identificadas algumas irregularidades. Na tabela 3 é possível visualizar problemas encontrados nos dados que irão requerer o devido cuidado na fase de preparação dos dados (*data preparation*). A tabela não terá em conta variáveis que não serão importantes para o estudo.

Tabela 3 - *Missing Values*

Variável	*Missing Values*
home_points	274
away_points	274
k	274
Elo_diff	274
MOVwinner	269
E Home Team Values	274
E away team Values	274
R home new	274
R away new	274

De notar que o *dataset* dos jogos realizados possui aproximadamente 6300 linhas, incluindo as 274 referidas na tabela 3.
Para além das variáveis descritas acima, existem registos de jogos entre equipas que não fazem parte do campeonato oficial da NBA, como é o caso do fim de semana do *All-Star*. O *All-Star Game* da NBA é um jogo de exibição realizado, por norma, no mês de fevereiro e exibe 24 dos melhores jogadores da liga. Apesar deste tipo de jogos estar presente no *dataset*, estes dados não serão tidos em conta para uma posterior análise e previsão.

3.3. Data Preparation

Nesta fase foram efetuadas várias modificações à base de dados original, de forma a ser possível obter uma base de dados mais concisa, robusta e sem ruido de outras variáveis que possam afetar o funcionamento de previsão da máquina.
Assim analisados todos os dados no *Data Understanding*, segue-se a fase de eventuais correções dos mesmos.

3.3.1. Seleção dos Dados

A seleção de dados tem como base selecionar os dados que serão necessários para uma fase posterior. Assim prevê-se a utilização do conjunto total dos dados na fase de modelação.
No entanto, os dados referidos acima serão alterados de forma a permitir o desenvolvimento dos modelos de previsão.
Em termos de eliminação de atributos, foram eliminadas as variáveis que não possuíam qualquer tipo de importância para as posteriores previsões, e que também não permitiam responder aos objetivos definidos nos *Business Goals*. Na tabela 4 é possível verificar os dados eliminados, bem como uma breve descrição dos mesmos e a razão pela qual foram eliminados.

Tabela 4 - Variáveis eliminadas do *dataset*

Variável	Descrição	Razão da eliminação
Column0	Coluna com uma numeração sequencial	Não relevante pois esta coluna não representa qualquer tipo de informação com valor acrescentado
Unnamed: 0	Coluna com uma numeração sequencial	Não relevante pois esta coluna não representa qualquer tipo de informação com valor acrescentado
Offensive Rebounds	Ressaltos ofensivos de um jogador	Não utilizada para as previsões visto não ser uma variável fundamental na consequência da análise
Personal_Fouls	Faltas de um jogador	Não utilizada para as previsões visto não ser uma variável fundamental na consequência da análise
Tech_Fouls	Faltas técnicas de um jogador	Não utilizada para as previsões visto não ser uma variável fundamental na consequência da análise
Flagrant_Fouls	Faltas flagrantes de um jogador	Não utilizada para as previsões visto não ser uma variável fundamental na consequência da análise
Crowd	Atendimento do estádio	Não utilizada para as previsões visto não ser uma variável fundamental na consequência da análise
Stadium	Capacidade do estádio de uma equipa	Não utilizada para as previsões visto não ser uma variável fundamental na consequência da análise
Player_id	Id do jogador	Não utilizada para as previsões visto não ser uma variável fundamental na consequência da análise. Para efeitos de agrupamento será utilizado o nome do jogador

3.3.2. Limpeza dos Dados

Após a seleção dos dados mais importantes para a previsão, é necessário limpar alguns registos da base de dados. Assim, na fase de limpeza de dados foi necessário eliminar as linhas que continham *missing values*, com o auxílio do *nó "Row Filter"* no KNIME. Deste modo, foram eliminados todos os registos da base de dados nos quais as equipas de um determinado jogo não possuíam pontos. Estes registos não possuem esse tipo de informação pois foram jogos agendados para datas de início de pandemia covid-19, na qual a NBA decidiu que esses jogos não se iriam realizar. Deste modo, estes não farão parte da análise.

Outra alteração importante à base de dados deve-se aos registos referidos no capítulo 3.2. Como foi dito anteriormente, existem registos de jogos entre equipas que não fazem parte do campeonato oficial da NBA, como é o caso do fim de semana do *All-Star*. Neste sentido, foram eliminados estes registos da base de dados para não influenciarem negativamente as previsões numa fase posterior.

Foram também removidos registos de estatísticas de jogadores em que jogaram menos de 10 minutos, devido ao facto de, ao serem normalizadas as variáveis, alguns jogadores ficarem com indicadores considerados *"outliers"* e não reais de acordo com a sua performance. A título de exemplo, um jogador que jogou 2 minutos e acertou 1 lançamento em 1 tentado, iria, por jogo, ter uma influência superior à do melhor jogador da própria equipa, não constituindo assim um dado real.

Por último, foram eliminados registos duplicados através do nó *"Duplicate Row Filter"*.

3.3.3. Construção de Dados

Com vista a construir uma base de dados relevante para a fase de modelação, tornou-se fundamental a criação de atributos derivados dos já fornecidos.
Desta forma, foram criadas variáveis com vista a obter resultados concisos, melhorando a disposição dos mesmos e reduzindo o ruído da base de dados.
Primeiramente foram concatenadas as duas colunas relativas ao primeiro e últimos nomes dos jogadores em uma coluna chamada "Nome Completo". Desta forma, é possível filtrar por um jogador em específico apenas com um nó, tornando assim a base de dados mais limpa. Para esta alteração foi utilizado o nó *"String Manipulation"*, onde através de uma fórmula foi concatenado o primeiro nome com o último.
Para além desta, foi alterada a variável dos minutos jogados. Na base de dados original, o tempo jogado estava definido como *string*, sendo, por isso, impossível usar esta variável para médias, previsões ou filtros. Deste modo, foi necessário converter esta variável para um número decimal. Para obter este resultado, foi necessário separar a *string* em duas colunas relativas a minutos e segundos (novamente usando o nó *"String Manipulation"*), convertendo-as em números através do nó *"String to Number"* e juntando-as novamente em apenas uma nova coluna através do nó *"Math Formula"*.
Foi fundamental criar um sistema de rating de qualidade das equipas ao longo do *dataset*, de acordo com os objetivos definidos na dissertação. Para a criação destes atributos, foi utilizada a linguagem de programação visual basic (VBA), onde o algoritmo se irá reger pelos seguintes princípios:

- O *elo rating* de todas as equipas iniciará com o valor de 1500;
- O algoritmo fará iterações registo a registo, por ordem crescente da data em que o jogo foi realizado;
- Cada vez que o algoritmo itera com um registo, é gerado um novo *elo rating* para cada uma das equipas (equipa da casa e equipa visitante).
- A mesma equipa terá um *elo rating* em casa e um fora de casa, visto serem indicadores com condicionantes diferentes.
- O *elo rating* tem como base a seguinte fórmula (Practicallypredictable, 2018).

$$R_{New} = R_{old} + (k * S_{team} - e_{team})$$

(1)

Em que "k" é uma constante que dita o impacto de um determinado jogo no rating de qualidade da equipa; S corresponde à variável binária que indica se a equipa ganhou ou não e, por último, a variável "e" que corresponde à *expected win* da equipa (probabilidade que a equipa tem para vencer um determinado jogo).

A constante "k" é calculada pela seguinte fórmula:

$$k = 20 * \frac{(MOV_{winner} + 3)^{0.8}}{7.5 + 0.006 * ABS(elo_{diff})}$$

(2)

Em que "MOV_winner" representa a margem de vitória da equipa vencedora (diferencial de pontos) e a variável "elo_diff" representa a diferença absoluta entre os elos das duas equipas.

- O atributo "e" – *expected win* é obtido pela seguinte fórmula:

$$e_{team} = \frac{1}{10^{\frac{(r_{away} - r_{home})}{400}}} \quad (3)$$

Para além do *rating* elo, revelou-se importante a criação de uma variável que corresponde à eficiência de um jogador em cada jogo, e que será fundamental para futuras previsões, sendo calculada pela seguinte fórmula:

$$Player\_efficiency = \frac{PTS + REB + AST + STL - FT\% - 2PT\% - 3PT\% - TO}{MP} * 48 \quad (4)$$

Em que:

- PTS – Pontos marcados;
- REB – Ressaltos;
- AST – Assistências
- STL – Roubos de bola;
- FT% - Percentagem de lançamentos livres;
- 2PT% - Percentagem de lançamentos de dois pontos;
- 3PT% - Percentagem de lançamentos de três pontos;
- TO – Turnovers;
- MP – Minutos jogados
- 48 corresponde ao tempo regulamentar de um jogo de basquetebol

Para além da variável referida acima, foi também desenvolvida uma variável para determinar o MVP de um jogo, em que, através de linguagem VBA, o algoritmo percorreu todas as estatísticas de todos os jogadores presentes num determinado jogo, e ditou o MVP (jogador mais valioso) do mesmo.

3.4. Modelação

Na modelação, são utilizados modelos de previsão para a previsão de determinados acontecimentos. Deste modo, os modelos são utilizados para prever indicadores, tais como o vencedor de um jogo entre duas equipas, o número de pontos que um jogador concretiza num determinado jogo, e o ranking de jogadores por posição.

3.4.1. Construção dos modelos de previsão

Neste subcapítulo, são descritos os modelos de previsão utilizados. Para as previsões de variáveis qualitativas, foram utilizados os modelos *Decision Tree*, *Random Forest* e *Gradient Boosted Tree*. Já para as previsões de variáveis quantitativas foram usados os modelos *Simple Regression Tree*, *Random Forest (Regression)*, *Gradient Boosted Tree (Regression)*.

3.4.1.1. Previsão da vitória/derrota da equipa da casa num determinado jogo

O processo metodológico passa pela aprendizagem da previsão por parte do modelo, a partir de um conjunto de dados de treino. Assim, o modelo utilizado consegue efetuar as previsões com o conjunto de dados de teste e, posteriormente, são analisados os indicadores de performance do modelo.

Para o desenvolvimento dos modelos que iriam prever se uma equipa X venceria/perderia o jogo, quando defrontava a equipa Y, foi necessário selecionar jogos de uma determinada equipa em análise. No caso, foram realizadas as previsões relativas às três melhores equipas identificadas na fase do Data Understanding (*Raptors, Warriors e Spurs*).

Para o descrito acima foi utilizado o nó *Row Filter* a par do nó *Group By*, para que o output fosse todos os jogos realizados pela equipa em questão.

Posteriormente foi usado o nó *"Window Slider".* Este nó tem como função a criação de colunas deslizantes a partir de um input. Desta forma, irá permitir a previsão do dia seguinte (T+1). O nó referido foi criado para as variáveis que incorporaram o modelo de previsão.

Posto isto, foi colocado o *nó Joiner* para que fossem agrupadas todas as informações do *Window Slider*.

Assim, é criado o *nó "Partitioning"*, de forma que haja uma divisão entre dois conjuntos: um conjunto de treino e um conjunto de teste. O conjunto para treino constitui 70% dos dados, pelo que o conjunto para teste constitui 30%.

Tanto para as varáveis numéricas e para as variáveis categóricas, foi selecionada a opção *"Draw Randomly"*.

Na tabela 5 são mostradas as variáveis que influenciaram na previsão das variáveis target, para as duas previsões.

Tabela 5 - Variáveis que afetam as previsões da Vitória e Derrota de uma equipa

Previsão	Variáveis que influenciam a *target*
Vitória / Derrota da equipa da casa *(S Home Team)*	*Home Team Name* – Nome da equipa da casa *Away Team* – variável que refere se a equipa joga em casa ou fora *E Home Team* – Probabilidade de vitória da equipa da casa *E Away Team* – probabilidade de vitória da equipa visitante *R Home* – elo *rating* da equipa da casa *R Away* – elo *rating* da equipa visitante

Em relação às parametrizações dos modelos de previsão, foram testados os parâmetros pré-definidos com valores *"default"* no KNIME, originando determinados resultados. Foi efetuada uma abordagem sistemática, através de uma análise iterativa com os parâmetros que impactaram os resultados, com vista a definir os melhores parâmetros dos diferentes modelos e, por sua vez, originar os melhores resultados.

Esta análise foi efetuada através dos *nós "Parameter Optimization Loop Start"* e *"Parameter Optimization Loop End",* em que foram testadas várias combinações de parâmetros até obter os melhores resultados possíveis. É possível visualizar na figura 41 as configurações para um dos exemplos do modelo (*decision tree* na previsão da vitória/derrota da equipa da casa num determinado jogo).

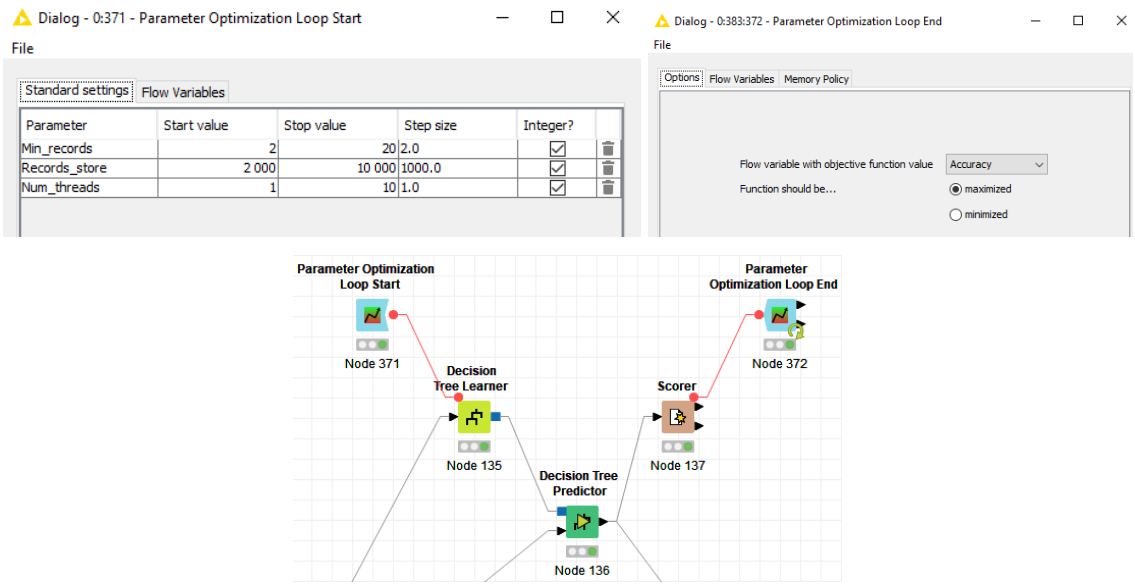


Figura 41 - Configurações de Otimização de Parâmetros

A título de exemplo, através da figura 42 é possível observar os diferentes resultados de algumas das combinações de parâmetros da análise de sensibilidade.

Min_records	Records_store	Num_threads	Objective value
20	2000	2	0.75
18	2000	2	0.7
16	2000	2	0.7
14	2000	2	0.733
12	2000	2	0.7
10	2000	2	0.7
8	2000	2	0.667
6	2000	2	0.667
4	2000	2	0.7
2	2000	2	0.683
20	2000	1	0.75
20	10000	10	0.75
18	10000	10	0.7
16	10000	10	0.7
14	10000	10	0.733
12	10000	10	0.7
10	10000	10	0.7
8	10000	10	0.667
6	10000	10	0.667
4	10000	10	0.7

Figura 42 - Exemplo Análise de Sensibilidade para a *Decision Tree*

Após a realização da análise iterativa, são verificados os melhores parâmetros para o modelo *Decision Tree* na tabela 6, bem como o exemplo de uma *Decision Tree*, para os Toronto Raptors, na figura 43.

Tabela 6 - Melhores Parâmetros para *Decision Tree*

Equipa	Número Mínimo de Registos por Nó	Número de Registos Armazenados	Número de Processadores	Precisão
Toronto Raptors	20	2000	1	0.75
Golden State Warriors	16	2000	1	0.814
San Antonio Spurs	4	2000	1	0.814

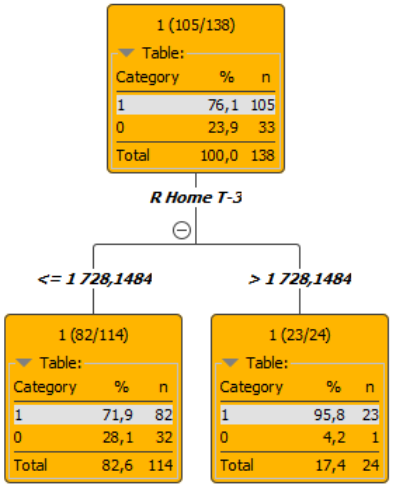


Figura 43 - *Decision Tree* para Toronto Raptors

Foi efetuado o mesmo processo para o modelo *Random Forest*, originando os seguintes resultados, apresentados na tabela 7. Este parâmetro foi testado para valores entre 1 e 1000.

Tabela 7 - Melhores Parâmetros para *Random Forest*

Equipa	Número de Modelos	Precisão
Toronto Raptors	3	0.783
Golden State Warriors	10	0.78
San Antonio Spurs	8	0.746

Realizou-se a mesma análise para o modelo *Gradient Boosted Tree Learner,* apresentada na tabela 8.

Tabela 8 - Melhores Parâmetros para *Gradient Boosted Tree Learner*

Equipa	Profundidade da Árvore	Número de Modelos	Taxa de Aprendizagem	Precisão
Toronto Raptors	2	200	0.1	0.75
Golden State Warriors	5	200	0.1	0.729
San Antonio Spurs	1	1000	0.2	0.729

3.4.1.2. Previsão dos pontos marcados por um jogador

Para o desenvolvimento dos modelos que iriam prever o número de pontos marcados por um jogador num determinado jogo, foi necessário normalizar alguns dados da base de dados. Na base de dados representativa dos jogos entre duas equipas, indicadores como os pontos da equipa, a probabilidade de a equipa ganhar, entre outros, eram dados baseados na equipa da casa ou da equipa visitante, sendo por isso registos independentes entre eles.

Posto isto, revelou-se fundamental a conversão destes indicadores para serem baseados por jogador, em que o sistema vai verificar registo a registo os dados do jogador, e ajustando as variáveis em função dos mesmos, convertendo o conceito de equipa da casa e equipa visitante para equipa do jogador e equipa oponente. Esta conversão foi efetuada através de fórmulas, construídas com o auxílio do nó *"Column Expressions"*.

Para os modelos desenvolvidos, foram realizadas as previsões relativas aos cinco melhores jogadores retirados do *Data Understanding* (*Giannis Antetokounmpo*, *Russel Westbrook*, *James Harden*, *Lebron James* e *Nikola Jokic*).

Para o descrito acima foi utilizado o *nó Row Filter*, para que o *output* fosse todos os registos de um determinado jogador.

Em relação ao *Window Slider* e ao *Partitioning*, foram configurados de igual forma ao descrito no capítulo 3.4.1.

Na tabela 9 são mostradas as variáveis que influenciaram na previsão da variável *target*.

Tabela 9 - Variáveis que afetam as previsões dos pontos marcados por um jogador

Previsão	Variáveis que influenciam a *target*
Pontos Marcados por um Jogador (*Points*)	*2PT%* - Percentagem de lançamentos de 2 pontos concretizados; *3PT%* - Percentagem de lançamentos de 3 pontos concretizados; *Min_Jogados* – Minutos jogados pelo jogador; *Player_efficiency* – Eficiência do jogador; *Points_Player_Team* – Pontos marcados pela equipa do jogador; *Game_MVP* – Foi o melhor jogador de um determinado jogo; *Win* – Equipa do jogador venceu o jogo;

Posto isto, foram novamente otimizados os parâmetros para os modelos *Random Forest Learner (Regression)* e *Gradient Boosted Trees Predictor (Regression)*. Como foi referido no subcapítulo anterior, para o *Simple Regression Tree Learner* não foi efetuada qualquer otimização de parâmetros.
Na tabela 10 é possível observar os melhores parâmetros para o *Random Forest*, nos cinco jogadores.

Tabela 10 - Melhores Parâmetros para *Random Forest*

Jogador	Número de Modelos	Erro Percentual Médio Absoluto
Giannis Antetokounmpo	100	34.6%
Russel Westbrook	600	40.1%
James Harden	800	35.9%
Lebron James	100	24.6%
Nikola Jokic	700	52.8%

É possível verificar que para diferentes jogadores, o número de modelos ideal varia. Isto dever-se-á à discrepância das estatísticas entre os mesmos.

De seguida são verificados os melhores parâmetros no modelo *Gradient Boosted Tree Learner*.

Tabela 11 - Melhores Parâmetros para *Gradient Boosted Tree Learner*

Jogador	Profundidade da Árvore	Número de Modelos	Taxa de Aprendizagem	Erro Percentual Médio Absoluto
Giannis Antetokounmpo	1	50	0.1	35%
Russel Westbrook	1	50	0.8	38.1%
James Harden	1	50	0.1	35.5%
Lebron James	1	50	0.1	24.3%
Nikola Jokic	1	50	0.2	52.4%

Através da análise da tabela, e possível verificar que para 4 dos 5 jogadores, a aprendizagem do modelo é lenta (entre 0.1 e 0.2), e o número de modelos ideal é 50. Para além disso, a profundidade da árvore ideal é de 1.

3.4.1.3. Previsão do ranking de jogadores por posição

A previsão do *ranking* de jogadores por posição é essencialmente um Top de jogadores de acordo com a sua qualidade.
Para esta previsão, os jogadores, por posição, foram classificados pela sua performance através do nó "*Rank*", que consiste em calcular classificações para os grupos selecionados com base nos atributos de classificação selecionados e também no modo de classificação. Posteriormente foram divididos em *bins* para que os dados sejam distribuídos em grupos de 5, através do nó *"Numeric Binner".* Na figura 44 é possível verificar as diferentes classes, como por exemplo a classe 5-10, que significa que o jogador estará entre o *rank* 5 e o *rank* 10 quando comparado com os restantes jogadores do grupo testado.

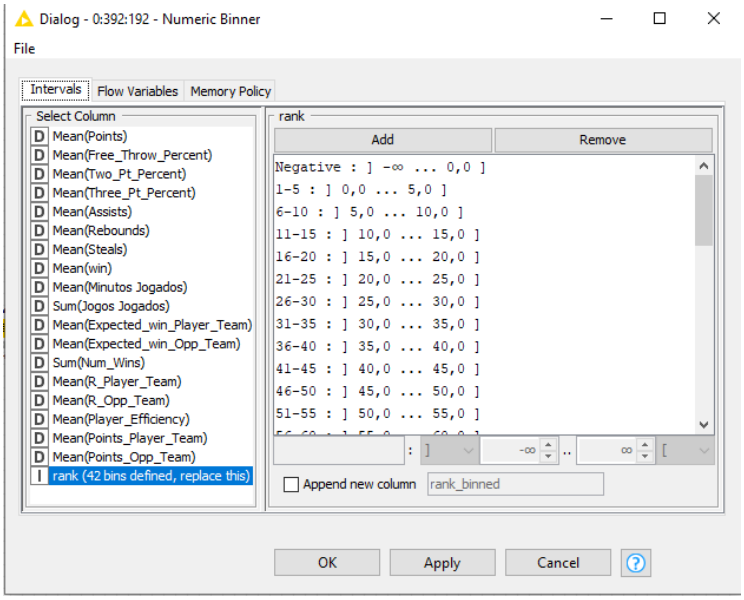


Figura 44 - Configuração Numeric Binner

Para os modelos desenvolvidos, foram realizadas as previsões relativas as três posições principais na equipa: C (*Center* / poste) , G (*Guard*/ Armador), F (*Forward*/ Ala Pivô)

Para o descrito acima foi utilizado o *nó Row Filter*, para que o *output* fosse todos os registos de jogadores de uma determinada posição.
Em relação ao *Window Slider* foi configurado de igual forma ao descrito no 3.4.

Na tabela 11 são mostradas as variáveis que influenciaram na previsão da variável *target*.

Tabela 11 - Variáveis que afetam as previsões do ranking de jogadores por posição

Previsão	Variáveis que influenciam a *target*
Ranking de jogadores por posição - *Rank*	*Points* – Pontos marcados pelo jogador

Foram otimizados os parâmetros para os modelos *Random Forest Learner* e *Decision Tree Learner e Gradient Boosted Tree Learner*.

Na tabela 12 é possível observar os melhores parâmetros para o *Random Forest*, nas três posições.

Tabela 12 - Melhores Parâmetros para *Random Forest*

Posição	Número de Modelos	Precisão
C (Center)	100	0.438
G (Guard)	300	0.483
F (Forward)	100	0.317

De seguida são verificados os parâmetros para o modelo *Decision Tree*, representados na tabela 13.

Tabela 13 - Melhores Parâmetros para *Decision Tree*

Posição	Número Mínimo de Registos por Nó	Número de Registos Armazenados	Número de Processadores	Precisão
C (Center)	2	2000	2	0.688
G (Guard)	2	2000	2	0.667
F (Forward)	2	2000	2	0.55

Por último são demonstradas as melhores parametrizações para o modelo *Gradient Boosted*.

Tabela 14 - Melhores Parâmetros para *Gradient Boosted*

Posição	*Tree Depth*	Número de Modelos	*Learning Rate*	Precisão
C (Center)	2	200	0.2	0.75
G (Guard)	2	200	0.2	0.633
F (Forward)	2	200	0.2	0.55

3.4.1.4. Previsão do rating de qualidade da equipa da casa

Essencialmente, nesta previsão são agrupados os jogos realizados por uma determinada equipa, e, com base em determinadas variáveis, é efetuada a previsão do *rating* elo para o dia seguinte.
Para a previsão do *rating* de qualidade da equipa da casa foi efetuado um filtro pelos jogos de uma determinada equipa em análise. No caso, foram realizadas as previsões relativas às três melhores equipas identificadas na fase do *Data Understanding* (*Raptors, Warriors e Spurs*).
Para o descrito acima foi utilizado o nó *Row Filter* a par do nó *Group By*, para que o output fosse todos os jogos realizados pela equipa em questão.

Foi novamente usado o *nó "Window Slider".* Desta forma, irá permitir a previsão do dia seguinte (T+1). O nó referido foi criado para as variáveis que incorporaram o modelo de previsão.

Posto isto, foi colocado o *nó Joiner* para que fossem agrupadas todas as informações do *Window Slider*.

Foi criado o *nó "Partitioning"*, de forma que haja uma divisão entre dois conjuntos: um conjunto de treino e um conjunto de teste. O conjunto para treino constitui 70% dos dados, pelo que o conjunto para teste constitui 30%.
Na tabela 15 são mostradas as variáveis que influenciaram na previsão da variável *target*.

Tabela 15 - Variáveis que afetam as previsões do ranking de jogadores por posição

Previsão	Variáveis que influenciam a *target*
Rating de qualidade da equipa da casa (*R Home*)	*Home Team Name* – Nome da equipa da casa *Away Team* – variável que refere se a equipa joga em casa ou fora *S Home* Team – Equipa da casa ganhou / não ganhou *S Away* Team – Equipa visitante ganhou/ não ganhou *E Home Team* – Probabilidade de vitória da equipa da casa *E Away Team* – probabilidade de vitória da equipa visitante

Em relação às parametrizações dos modelos de previsão, foram testados os parâmetros pré-definidos com valores *"default"* no KNIME, originando determinados resultados. Em função dos resultados, foi efetuada uma análise iterativa com os parâmetros que impactaram os resultados, com vista a definir as melhores parametrizações dos diferentes modelos e, por sua vez, originar os melhores resultados.
Esta análise foi efetuada através dos *nós "Parameter Optimization Loop Start"* e *"Parameter Optimization Loop End",* em que foram testadas várias combinações de parâmetros até obter os melhores resultados possíveis.
São mostrados os resultados do *Random Forest Learner* e *Gradient Boosted Tree Learner*, respetivamente, nas tabelas 16 e 17.

Tabela 16 - Melhores Parâmetros para *Random Forest*

Equipa	Número de Modelos	Erro Percentual Médio Absoluto
Toronto Raptors	500	1.8%
Golden State Warriors	1000	2.7%
San Antonio Spurs	100	1.4%

Tabela 17 - Melhores Parâmetros para *Gradient Boosted Tree*

Equipa	Profundidade da Árvore	Número de Modelos	Taxa de Aprendizagem	Erro Percentual Médio Absoluto
Toronto Raptors	1	400	0.1	0.3%
Golden State Warriors	1	400	0.1	0.4%
San Antonio Spurs	1	400	0.1	0.2%

3.4.2. Avaliação dos Modelos

Após efetuada a análise de sensibilidade dos parâmetros a serem utilizados nos diferentes modelos, são analisados os resultados dos mesmos.

Para esta avaliação serão tidas em conta variáveis provenientes dos nós "*Numeric Scorer*" para as previsões de variáveis quantitativas e do "*Scorer*" para as previsões de variáveis qualitativas.

No caso das informações geradas no *Numeric Scorer*, será dada especial atenção ao Erro Percentual Médio Absoluto, tendo também em conta as restantes variáveis na decisão final. No entanto, serão também analisadas outras métricas, como o R^2, o Desvio (MSD) e o erro quadrático médio (EQM). Por outro lado, no *Scorer* será analisada com maior importância a precisão (*Accuracy*).

Aquando das previsões, foi também criado um algoritmo, denominado de algoritmo de persistência, que servirá como referência para uma comparação com os dados obtidos nos modelos de previsão. Este algoritmo consiste no facto de que o valor da previsão da variável do dia seguinte irá ser o valor dessa mesma variável no dia anterior. Conclui-se que, caso o modelo de previsão obtenha melhores resultados que o algoritmo de persistência, é um modelo favorável/bem-sucedido.

Foram avaliados os mesmos modelos para diferentes "*Window Sliders*", ou seja, foram executados os mesmos modelos para colunas deslizantes em que T {2, 3, 4}.

3.4.2.1. Previsão da vitória/derrota da equipa da casa num determinado jogo

Na tabela 18 são mostrados os resultados dos modelos de previsão, para os diferentes valores de T (tamanhos do *Window Slider*), ou seja, para as previsões com base nos dois dias anteriores, nos três dias anteriores e nos quatro dias anteriores.

Tabela 18 - Resultados Previsão Vitória/Derrota da equipa da casa

Modelo	Precisão T=2	Precisão T=3	Precisão T=4
Decision Tree - Raptors	80%	75%	76.667%
Decision Tree - Warriors	68.333%	64.407%	79.661%
Decision Tree - Spurs	78.333%	74.576%	83.051%
Random Forest - Raptors	78.333%	70%	71.667%
Random Forest - Warriors	71.667%	74.576%	79.661%
Random Forest - Spurs	75%	74.576%	81.356%
Gradient Boosted Trees - Raptors	66.667%	68.333%	65%
Gradient Boosted Trees - Warriors	75%	72.881%	66.102%
Gradient Boosted Trees - Spurs	66.67%	71.186%	67.797%
Raptors - (Alg. Persistência)	58.333%	60%	66.667%

Warriors - (Alg. Persistência)	61.667%	69.492%	72.881%
Spurs - (Alg. Persistência)	68.333%	64.407%	69.492%

Como é possível verificar, para um valor do *Window Slider* igual a 3, todos os modelos apresentam uma maior precisão quando comparados com o algoritmo de persistência, excetuando o modelo *Decision Tree* para a equipa *dos Golden State Warriors*. No entanto, o *Random Forest* apresenta os resultados mais consistentes e com uma maior precisão comparativamente às restantes, com uma média de precisão de aproximadamente 73% (referentes à média dos registos de *Random Forest* das três equipas). Com um *Window Slider* igual a 2, os resultados da previsão mostram-se ligeiramente superiores ao anterior, com uma média de previsão a rondar os 75%, tanto para o modelo *Decision Tree* como para o *Random Forest*.

Por fim, com um *Window Slider* igual a 4, os resultados da previsão apresentam-se melhores em comparação com as duas verificações anteriores, sendo que a média da previsão para o método *Decision Tree* é de aproximadamente 79,8%, e no *Random Forest* de 77,6%. Para o caso da *Decision Tree*, uma precisão de 79.8% significa que 79.8 em cada 100 jogos o resultado previsto foi o que realmente aconteceu.

Posto isto, conclui-se que os melhores modelos a aplicar na previsão da vitória ou derrota da equipa da casa num determinado jogo são o *Decision Tree* o *Random Forest*, com um valor de colunas deslizantes de 4.

3.4.2.2. Previsão dos pontos marcados por um jogador

Nas tabelas 19, 20 e 21, 22 e 23 são demonstrados os resultados dos modelos de previsão dos pontos marcados por um jogador (Giannis Antetokounmpo, Russel Westbrook, James Harden, Lebron James e Nikola Jokic), para os diferentes valores de T (tamanhos do *Window Slider*), ou seja, para as previsões com base nos dois dias anteriores, nos três dias anteriores e nos quatro dias anteriores.

Tabela 19 - Previsão para os Pontos marcados pelo Giannis Antetokounmpo

Window Slider	Jogador	R^2	Erro quadrático médio (EQM)	Desvio (MSD)	Erro Percentual Médio Absoluto (MAPE)
T = 2	*Simple Regression Tree – Giannis Antetokounmpo*	-0.783	89.024	0.286	0.328
	Random Forest - Giannis Antetokounmpo	-0.05	52.405	-0.286	0.275
	Gradient Boosted Trees - Giannis Antetokounmpo	-1.601	129.844	-0.087	0.402
	Giannis Antetokounmpo - (Alg. Persistência)	-1.648	132.19	-0.214	0.415

T = 3	*Simple Regression Tree – Giannis Antetokounmpo*	-0.996	147	1.747	0.483
	Random Forest - Giannis Antetokounmpo	-0.002	73.771	-0.614	0.355
	Gradient Boosted Trees - Giannis Antetokounmpo	-0.784	131.419	2.4	0.45
	Giannis Antetokounmpo - (Alg. Persistência)	-0.575	116	-0.048	0.414
T = 4	*Simple Regression Tree – Giannis Antetokounmpo*	-0.587	107.024	-0.518	0.334
	Random Forest - Giannis Antetokounmpo	-0.018	68.699	-1.229	0.273
	Gradient Boosted Trees - Giannis Antetokounmpo	-0.841	124.203	-1.482	0.347
	Giannis Antetokounmpo - (Alg. Persistência)	-0.856	125.217	-0.398	0.379

Tabela 20 - Previsão para os Pontos marcados pelo Russel Westbrook

Window Slider	Jogador	R^2	Erro quadrático médio (EQM)	Desvio (MSD)	Erro Percentual Médio Absoluto (MAPE)
T = 2	*Simple Regression Tree – Russel Westbrook*	-0.131	133.282	-0.923	0.463
	Random Forest - Russel Westbrook	0.044	112.654	-0.372	0.453
	Gradient Boosted Trees - Russel Westbrook	-0.351	159.205	-1.801	0.537
	Russel Westbrook - (Alg. Persistência)	-0.607	189.372	0.192	0.519
T = 3	*Simple Regression Tree – Russel Westbrook*	-0.619	184.667	-3.359	0.475
	Random Forest - Russel Westbrook	0.055	107.833	-1.038	0.405
	Gradient Boosted Trees - Russel Westbrook	-0.709	194.878	-0.047	0.547
	Russel Westbrook - (Alg. Persistência)	-0.597	182.167	-0.449	0.517
T = 4	*Simple Regression Tree – Russel Westbrook*	-0.725	167.857	-2.429	0.421
	Random Forest - Russel Westbrook	0.055	91.935	-1.442	0.359
	Gradient Boosted Trees - Russel Westbrook	-1.006	195.228	-3.061	0.443

	Russel Westbrook - (Alg. Persistência)	-0.597	155.442	-1.078	0.409

Tabela 21 - Previsão para os Pontos marcados pelo James Harden

Window Slider	Jogador	R^2	Erro quadrático médio (EQM)	Desvio (MSD)	Erro Percentual Médio Absoluto (MAPE)
T = 2	*Simple Regression Tree - James Harden*	-0.519	147.783	2.05	0.423
	Random Forest - James Harden	-0.141	111.067	3	0.37
	Gradient Boosted Trees - James Harden	-0.833	178.389	4.205	0.456
	James Harden - (Alg. Persistência)	-0.907	185.583	1.583	0.453
T = 3	*Simple Regression Tree - James Harden*	-1.154	214.883	1.283	0.491
	Random Forest - James Harden	-0.071	106.783	1.017	0.36
	Gradient Boosted Trees - James Harden	-0.74	173.505	1.119	0.431
	James Harden - (Alg. Persistência)	-0.874	186.9	1.367	0.438
T = 4	*Simple Regression Tree - James Harden*	-1.51	202.283	-3.617	0.351
	Random Forest - James Harden	-0.1	88.667	-1.233	0.256
	Gradient Boosted Trees - James Harden	-1.532	204.054	1.255	0.41
	James Harden - (Alg. Persistência)	-0.927	155.283	-0.917	0.333

Tabela 22 - Previsão para os Pontos marcados pelo Lebron James

Window Slider	Jogador	R^2	Erro quadrático médio (EQM)	Desvio (MSD)	Erro Percentual Médio Absoluto (MAPE)
T = 2	*Simple Regression Tree - Lebron James*	-1.646	120.037	1.963	0.374
	Random Forest - Lebron James	-0.144	51.926	1.309	0.275
	Gradient Boosted Trees - Lebron James	-1.208	100.196	2.936	0.358

	Lebron James - (Alg. Persistência)	-1.646	120.037	1.963	0.374
T = 3	*Simple Regression Tree - Lebron James*	-1.333	98.16	-0.802	0.31
	Random Forest - Lebron James	-0.152	48.444	0.42	0.248
	Gradient Boosted Trees - Lebron James	-1.306	97.014	0.656	0.315
	Lebron James - (Alg. Persistência)	-1.333	98.16	-0.802	0.31
T = 4	*Simple Regression Tree - Lebron James*	-2.338	133.407	0.642	0.367
	Random Forest - Lebron James	-0.209	48.333	-0.235	0.234
	Gradient Boosted Trees - Lebron James	-1.904	116.086	-0.654	0.334
	Lebron James - (Alg. Persistência)	-2.338	133.407	0.642	0.367

Tabela 23 - Previsão para os Pontos marcados pelo Nikola Jokic

Window Slider	Jogador	R^2	Erro quadrático médio (EQM)	Desvio (MSD)	Erro Percentual Médio Absoluto (MAPE)
T = 2	*Simple Regression Tree - Nikola Jokic*	-0.731	107.174	-0.035	0.542
	Random Forest - Nikola Jokic	0.058	58.326	0.349	0.546
	Gradient Boosted Trees - Nikola Jokic	-0.784	110.416	-2.602	0.503
	Nikola Jokic - (Alg. Persistência)	-0.194	73.93	-0.907	0.494
T = 3	*Simple Regression Tree - Nikola Jokic*	-0.684	128.477	0.453	0.677
	Random Forest - Nikola Jokic	0.056	72.012	0.64	0.535
	Gradient Boosted Trees - Nikola Jokic	-0.692	129.091	0.002	0.604
	Nikola Jokic - (Alg. Persistência)	-0.299	99.058	0.384	0.523
T = 4	*Simple Regression Tree - Nikola Jokic*	-1.53	144.553	1.024	0.698
	Random Forest - Nikola Jokic	0	57.129	0.776	0.506

	Gradient Boosted Trees - Nikola Jokic	-1.493	142.439	1.168	0.673
	Nikola Jokic - (Alg. Persistência)	-0.753	100.165	-0.047	0.578

Analisando as tabelas 19, 20 e 21, 22 e 23 pode-se comprovar que a variação da variável dependente não é explicada pela variável independente, devido ao facto dos valores do r^2 ajustado serem inferiores a 0 (negativos). Assim como a previsão analisada anteriormente, isto acontece para todos os valores de *Window Slider*.

Em relação ao indicador do erro quadrático médio, apenas o modelo *Random Forest* se revela com resultados superiores para as três equipas comparativamente com os algoritmos de persistência para as mesmas. Todos os modelos, no entanto, apresentam um erro quadrático médio elevado, o que significa que o estimador (usado para deduzir o valor de um parâmetro desconhecido em um modelo) prevê observações distantes do real.

Quanto aos desvios, todos os modelos possuem um maior desvio médio quando comparados com os algoritmos de persistência, exceto o modelo *Gradient Boosted Tree* para um *Windows Slider* de 3. Verifica-se também que o valor dos desvios médios, para todos os modelos, encontra-se próximo do 0, o que significa que há uma baixa variabilidade do conjunto de dados.

Por último, é possível concluir que, em análise ao Erro Percentual Médio Absoluto, o modelo *Random Forest* apresenta melhores resultados em relação aos algoritmos de persistência, sendo o único modelo nessas condições. Os valores do MAPE variam entre os 23% e os 55% para os diferentes modelos, sendo que o *Random Forest* com melhores resultados apresenta uma média de 32.56%, ou seja, em média, a previsão está incorreta em 32.56%, para um valor de colunas deslizantes igual a 4.

3.4.2.3. Previsão do ranking de jogadores por posição

Na tabela 24 são demonstrados os resultados dos modelos de previsão do ranking de jogadores nas diferentes posições (C – Center, G- Guard e F- Forward), para os diferentes valores de T (tamanhos do *Window Slider*), ou seja, para as previsões com base nos dois dias anteriores, nos três dias anteriores e nos quatro dias anteriores.

Tabela 24 - Resultados do *ranking* de jogadores por posição

Modelo	Precisão T = 2	Precisão T = 3	Precisão T = 4
Decision Tree - C (Center)	81.25%	68.8%	50%
Decision Tree - G (Guard)	3.333%	66.7%	6.667%
Decision Tree - F (Forward)	3.333%	55%	5%
Random Forest - C (Center)	37.5%	43.8%	37.5%
Random Forest - G (Guard)	28.333%	48.3%	35%
Random Forest - F (Forward)	20%	31.7%	36.667%

Gradient Boosted Trees - C (Center)	56.25%	75%	31.25%
Gradient Boosted Trees - G (Guard)	53.333%	63.33%	38.333%
Gradient Boosted Trees - F (Forward)	43.333%	55%	38.333%
C (Center) - (Alg. Persistência)	81.25%	87.5%	87.5%
G (Guard) - (Alg. Persistência)	81.667%	76.667%	78.333%
F (Forward) - (Alg. Persistência)	76.667%	75%	76.667%

Através dos dados presentes na tabela 30, todos os modelos apresentam uma menor precisão quando comparados com o algoritmo de persistência. No entanto, o *Decision Tree* apresenta os resultados mais consistentes e com uma maior precisão comparativamente às restantes, com uma média de precisão de aproximadamente 63.5% e apenas para um valor de colunas deslizantes igual a 3 (referentes à média dos registos de *Decision Tree* das três posições). Isto significa que aproximadamente, que 63 jogadores em cada 100, o *ranking* resultante da previsão corresponde ao *ranking* correto.

Visto que os modelos de previsão apresentam piores resultados em relação ao algoritmo de persistência, considera-se que o modelo analisado não estará apto para efetuar uma boa previsão das variáveis em questão.

3.4.2.4. Previsão do rating de qualidade da equipa da casa

Nas tabelas 25, 26 e 27 são demonstrados os resultados dos modelos de previsão do *rating* de qualidade da equipa da casa, para os diferentes valores de T (tamanhos do *Window Slider*), ou seja, para as previsões com base nos dois dias anteriores, nos três dias anteriores e nos quatro dias anteriores.

Tabela 25 - Resultados do *rating* de qualidade da equipa da casa para T=2

Modelo T = 2	R^2	Erro quadrático médio (EQM)	Desvio (MSD)	Erro Percentual Médio Absoluto (MAPE)
Simple Regression Tree - Raptors	0.887	345.656	-12.274	0.009
Simple Regression Tree - Warriors	0.938	639.313	-11.577	0.012
Simple Regression Tree - Spurs	0.856	168.626	-5.238	0.005
Random Forest - Raptors	0.088	1107.814	-8.356	0.016
Random Forest - Warriors	-0.052	3671.511	-23.937	0.023
Random Forest - Spurs	-0.168	606.824	-7.804	0.01
Gradient Boosted Trees - Raptors	0.234	1550.41	-19.082	0.013
Gradient Boosted Trees - Warriors	0.892	938.758	-13.363	0.013
Gradient Boosted Trees - Spurs	0.75	270.764	-5.495	0.005
Raptors - (Alg. Persistência)	0.982	66.665	1.627	0.004
Warriors - (Alg. Persistência)	0.991	95.094	-1.497	0.004

Spurs - (Alg. Persistência)	0.897	168.626	5.238	0.005

Tabela 26 - Resultados do *rating* de qualidade da equipa da casa para T=3

Modelo T = 3	R^2	Erro quadrático médio (EQM)	Desvio (MSD)	Erro Percentual Médio Absoluto (MAPE)
Simple Regression Tree - Raptors	0.961	158.972	-4.98	0.006
Simple Regression Tree - Warriors	0.921	607.628	-14.219	0.012
Simple Regression Tree - Spurs	0.712	905.427	-2.48	0.01
Random Forest - Raptors	-0.865	1814.408	-22.421	0.018
Random Forest - Warriors	-0.555	3882.559	-36.433	0.027
Random Forest - Spurs	-1.081	1128.828	-20.752	0.014
Gradient Boosted Trees - Raptors	0.385	1137.312	-17.714	0.012
Gradient Boosted Trees - Warriors	0.503	2832.873	-21.123	0.017
Gradient Boosted Trees - Spurs	-1.618	3002.284	-26.609	0.02
Raptors - (Alg. Persistência)	0.987	54.846	0.827	0.004
Warriors - (Alg. Persistência)	0.987	105.753	-1.451	0.005
Spurs - (Alg. Persistência)	0.976	53.518	1.64	0.003

Tabela 27 - Resultados do *rating* de qualidade da equipa da casa para T=4

Modelo T = 4	R^2	Erro quadrático médio (EQM)	Desvio (MSD)	Erro Percentual Médio Absoluto (MAPE)
Simple Regression Tree - Raptors	0.738	562.577	-16.124	0.011
Simple Regression Tree - Warriors	0.729	2142.131	-22.515	0.018
Simple Regression Tree - Spurs	0.822	264.311	-5.888	0.007
Random Forest - Raptors	-5.733	2645.974	-31.648	0.022
Random Forest - Warriors	-4.131	8410.655	-55.768	0.038
Random Forest - Spurs	-2.303	1199.32	-18.859	0.014
Gradient Boosted Trees - Raptors	0.903	297.644	-10.549	0.009
Gradient Boosted Trees - Warriors	0.754	1934.745	-26.656	0.02
Gradient Boosted Trees - Spurs	0.21	986.77	-12.692	0.011
Raptors - (Alg. Persistência)	0.984	48.211	1.284	0.003
Warriors - (Alg. Persistência)	0.995	61.765	1.148	0.004
Spurs - (Alg. Persistência)	0.962	79.401	0	0.004

Através dos resultados presentes nas tabelas 25, 26 e 27, é possível verificar que, em relação ao r^2, uma percentagem da variação da variável dependente é explicada pela variável independente, excetuando o *modelo Random Forest*, pois neste os valores do r^2 ajustado são inferiores a 0 (negativos). Isto acontece independentemente do valor atribuído no *Window Slider*.

Por outro lado, relativamente ao erro quadrático médio, nenhum dos modelos revela resultados superiores para as três equipas comparativamente com os algoritmos de persistência para as mesmas.

Quanto aos desvios, todos os modelos possuem um maior desvio médio quando comparados com os algoritmos de persistência. Mesmo assim, o valor dos desvios médios, para todos os modelos, encontra-se próximo do 0, o que significa que há uma baixa variabilidade do conjunto de dados.
Por último, é possível concluir que, em análise ao Erro Percentual Médio Absoluto, o modelo *Simple Regression* apresenta melhores resultados. Os valores do MAPE variam entre os 0.5% e os 1.8% para os diferentes modelos e valores de *Window Slider*, sendo que o *Simple Regression* apresenta uma média de aproximadamente 0.9% para um *Window Slider* de 3, ou seja, em média, a previsão está incorreta em 0.9% das vezes. Já para um Window Slider de 2 a média encontra-se nos 0.86% e para um Window Slider de 4 a média encontra-se nos 1.2%. Considera-se o valor 2 como o melhor valor a utilizar no número de colunas deslizantes.

3.5. Discussão de resultados

De uma forma geral verifica-se que os resultados obtidos no modelo *Random Forest* apresentam melhores valores relativamente aos valores de MAPE, r^2 e erros quadráticos médios. Isto deve-se ao facto do Random Forest ser um modelo muito completo, e flexível mediante diferentes objetivos de previsão. O modelo resolve problemas tanto de regressão, quanto de classificação, apresentando bons resultados em ambos. Será provável que o modelo *Random Forest* apresente melhores resultados que a *Decision Tree* na maior parte dos casos, pois tem as suas origens em árvores de decisão (seleciona subconjuntos de *features* e monta mini árvores de decisão).

Relativamente aos dias a considerar na previsão, os resultados obtidos dependem do tipo de previsão efetuada. Para a previsão da vitória/ derrota de uma equipa e dos pontos marcados por um jogador foram obtidos melhores resultados considerando apenas os quatro dias anteriores à previsão. Por outro lado, para a previsão do ranking de jogadores por posição, foram obtidos melhores resultados considerando apenas os três dias anteriores. Por último, para a previsão do rating de qualidade da equipa da casa, obteve-se melhor desempenho apenas considerando os dois dias anteriores.

Em relação aos desvios, na maioria das previsões efetuadas os desvios mostraram-se próximos de 0, o que significa que há uma baixa variabilidade do conjunto de dados.

Por sua vez, os valores de r^2 apresentaram-se negativos, exceto na previsão do rating de qualidade da equipa da casa. Sendo negativos, significa que a variação da variável a prever não é explicada pelas variáveis independentes. No caso do *rating* de qualidade, o r^2 apresentou valores próximos de 1, o que significa que a variável a prever é explicada pelas restantes variáveis independentes.

Relativamente aos erros quadráticos médios, para a previsão dos pontos marcados por um jogador, bem como a previsão do rating de qualidade de uma equipa, os valores apresentados são mais elevados, significando que os valores previstos apresentam uma discrepância em relação aos valores corretos.

Relativamente às métricas de precisão da previsão, os valores apresentados foram satisfatórios (aproximadamente uma média de 77% na previsão da vitória ou derrota de uma equipa, e aproximadamente 63,5% na previsão do ranking de jogadores por posição).

Por último, os valores do MAPE (Erro Percentual Médio Absoluto) dependem do tipo de previsão efetuada. Para a previsão dos pontos marcados por jogador, a previsão estará incorreta em cerca de 32.5%. Quanto à previsão do *rating* de qualidade da equipa, foram apresentados valores próximos do 0, mais concretamente, 0.86%, o que significa que a previsão estará incorreta em 0.86%, ou seja, considera-se uma previsão quase perfeita.

Analisando os resultados, não é possível concluir qual a base temporal definitiva que permite obter a melhor performance, pois varia consoante as diferentes previsões.

Para os resultados obtidos, os modelos usavam o *dataset* contemplando diversas variáveis, baseadas nos resultados da *data understanding*, pelo que não será possível concluir que variáveis possuem um maior impacto na previsão.

4. CONCLUSÃO

O principal objetivo da dissertação passou pela criação de um modelo de *machine learning*, de forma a determinar o *rating* de qualidade de um jogador e equipa baseado em estatísticas de jogo, e também do desenvolvimento de um modelo *machine learning* de previsão de resultados de basquetebol, utilizando a ferramenta KNIME. Este reconhecimento foi conseguido com a realização de diversas previsões de variáveis, tendo por base a metodologia CRISP-DM.

Tendo em conta o referido na discussão dos resultados obtidos, pode-se concluir que de facto foram encontrados modelos capazes de prever se uma equipa vai ganhar/perder um determinado jogo, dos pontos marcados por um jogador em determinado jogo. Por outro lado, o desempenho dos modelos previsão do *ranking* de um jogador por posição, bem como do *rating* de qualidade de uma equipa, não foi satisfatório, pois estes apresentavam piores resultados quando comparados ao algoritmo de referência.

Foram descritos ao longo do trabalho os mecanismos necessários para o desenvolvimento dos modelos de previsão, bem como otimizações de parametrizações de indicadores envolvidos nas mesmas, contribuindo para obter melhores resultados em relação à precisão dos modelos e aos erros associados.

Foi também definido um método consistente independentemente das previsões, mas concluiu-se que cada previsão deve ser tratada de forma independente, analisando as correlações entre variáveis realizadas no *Data Understanding*, e tomando decisões relativas às variáveis que irão/ deverão afetar a previsão da variável alvo.

4.1. Limitações e investigação futura

As limitações do projeto passaram por um complexo tratamento de dados de forma que as previsões a fazer se tornassem realmente úteis para o cliente final. Para além disso, as bases de dados utilizadas poderiam demonstrar outros indicadores de forma que fosse possível fornecer ao cliente final uma vasta gama de previsões e indicadores.

Relativamente a possíveis melhorias numa fase futura, seria interessante do ponto de vista do cliente final explorar previsões com redes neuronais, como por exemplo RNN, pois estas possuem elevada precisão, mesmo em problemas complexos. Este tipo de modelos é também robusto em termos de controlo. Ademais, seria possível explorar a previsão de probabilidades, como por exemplo a probabilidade de uma equipa ganhar um determinado jogo, através de modelos bayesianos.

Para além disso, será importante também uma análise mais extensiva e pormenorizada relativamente a quantas e quais variáveis devem afetar a variável *target* em cada uma das previsões, de forma a obterem-se os melhores resultados possíveis nas previsões.

REFERÊNCIAS BIBLIOGRÁFICAS

Araz, O. M., Choi, T. M., Olson, D. L., & Salman, F. S. (2020). Data Analytics for Operational Risk Management. *Decision Sciences*, *51*(6), 1316–1319. https://doi.org/10.1111/deci.12443

Azevedo, A., & Santos, M. F. (2008). KDD , SEMMA AND CRISP-DM : A PARALLEL OVERVIEW Ana Azevedo and M . F . Santos. *IADIS European Conference Data Mining*, 182–185. http://recipp.ipp.pt/handle/10400.22/136%0Ahttp://recipp.ipp.pt/bitstream/10400.22/136/3/KDD-CRISP-SEMMA.pdf

Bhandari, I., Colet, E., Parker, J., Pines, Z., Pratap, R., & Ramanujam, K. K. (1997). Advanced scout: Data mining and knowledge discovery in NBA data. *Data Mining and Knowledge Discovery*, *1*(1), 121–125. https://doi.org/10.1023/A:1009782106822

Borghesi, R. (2018). The Financial and Competitive Value of NCAA Basketball Recruits. *Journal of Sports Economics*, *19*(1), 31–49. https://doi.org/10.1177/1527002515617510

Brefeld, U., & Zimmermann, A. (2017). Guest editorial: Special issue on sports analytics. *Data Mining and Knowledge Discovery*, *31*(6), 1577–1579. https://doi.org/10.1007/s10618-017-0530-1

Bunker, R. P., & Thabtah, F. (2019). A machine learning framework for sport result prediction. *Applied Computing and Informatics*, *15*(1), 27–33. https://doi.org/10.1016/j.aci.2017.09.005

Clara Pereira Coutinho. (2014). *Metodologia de Investigação em Ciências Sociais e Humanas*. https://books.google.pt/books?hl=pt-PT&lr=&id=uFmaAwAAQBAJ&oi=fnd&pg=PT3&dq=Metodologia+de+Investigação+em+Ciências+Sociais+e+Humanas&ots=GheC1DdZS7&sig=1NsusHQQxgzmWbMYXCbRHQti_Fg&redir_esc=y#v=onepage&q&f=false

Fillbrunn, A., Dietz, C., Pfeuffer, J., Rahn, R., Landrum, G. A., & Berthold, M. R. (2017). KNIME for reproducible cross-domain analysis of life science data. *Journal of Biotechnology*, *261*(July), 149–156. https://doi.org/10.1016/j.jbiotec.2017.07.028

Gama, J., Lorena, A. C., Faceli, K., Oliveira, M., & Ponce, A. (2012). *Extração de Conhecimento de Dados*. https://doi.org/ISBN: 978-972-618-698-4

Huber, S., Wiemer, H., Schneider, D., & Ihlenfeldt, S. (2019). DMME: Data mining methodology for engineering applications - A holistic extension to the CRISP-DM model. *Procedia CIRP*, *79*, 403–408. https://doi.org/10.1016/j.procir.2019.02.106

Kaur, H., & Jain, S. (2018). Machine learning approaches to predict basketball game outcome. *Proceedings - 2017 3rd International Conference on Advances in Computing, Communication and Automation (Fall), ICACCA 2017*, *2018-Janua*, 1–7. https://doi.org/10.1109/ICACCAF.2017.8344688

Keshtkar Langaroudi, M., & Yamaghani, M. (2019). Journal of Advances in Computer Engineering and Technology. *Science and Research Branch,Islamic Azad University*, *5*(1), 27–36. http://jacet.srbiau.ac.ir/article_13599.html

KNIME. (n.d.). https://www.knime.com/knime-analytics-platform

Lee, S. A. (2014). *Business Intelligence and Analytics Ramesh Sharda Dursun Delen Efraim Turban Tenth Edition*. https://d1wqtxts1xzle7.cloudfront.net/62386890/business_intelligence20200316-50198-128rsmo-with-cover-page-v2.pdf?Expires=1634850341&Signature=djbiCXos7VKx1C4HAXEH9TKOhagpgAGIwuwMKipJ2

~StilXONZPAdrh92h-PrC4ET8ndQQ7c3NRY0TXisGoj5LYXBvso4iVTO42qJfdajKnXKBuX

Leung, C. K., & Joseph, K. W. (2014). Sports data mining: Predicting results for the college football games. *Procedia Computer Science*, *35*(C), 710–719. https://doi.org/10.1016/j.procs.2014.08.153

Lucey P, Morgan S, Wiens J, Y. Y. (2016). (2016). *KDD workshop on large-scale sports analytics*. http://www.large-scale-sports-analytics.org/

Luis Adriano Oliveira. (2011). *Dissertação e Tese em Ciência e Tecnologia segundo Bolonha*.

Machado, C. F., Fernandes, J., & Amaral, L. (2020). *Methodology Used for Determination of Critical Success Factors in Adopting the New General Data Protection Regulation in Higher Education Institutions.* 71–109. https://doi.org/10.1007/978-3-030-40896-1_4

Mahmood, Z., Daud, A., & Abbasi, R. A. (2021). Using machine learning techniques for rising star prediction in basketball. *Knowledge-Based Systems*, *211*, 106506. https://doi.org/10.1016/j.knosys.2020.106506

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity. Retrieved from McKinsey Global Institute:* https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M. J., & Flach, P. (2021). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*, *33*(8), 3048–3061. https://doi.org/10.1109/TKDE.2019.2962680

Murphy, K. P. (2012). Machine Learning - A Probabilistic Perspective - Table-of-Contents. *The MIT Press*, 1049.

Pelechrinis, K., Winston, W., Sagarin, J., & Cabot, V. (2019). Machine Learning and Data Mining for Sports Analytics. *Machine Learning and Data Mining for Sports Analytics*, *11330 LNAI*(September), 106–117.

Practicallypredictable. (2018). *Elo Ratings for NBA Teams*. http://practicallypredictable.com/2018/04/15/elo-ratings-for-nba-teams/

Ranjit Kumar. (2019). *Research methodology: A step-by-step guide for beginners*. https://books.google.pt/books?hl=pt-PT&lr=&id=J2J7DwAAQBAJ&oi=fnd&pg=PP1&dq=Research+methodology:+A+step-by-step+guide+for+beginners.&ots=cvpiJGOCfg&sig=HIvpDmdQZ5MWhJatap3MyiSrJwU&redir_esc=y#v=onepage&q&f=false

Richter, C., O'Reilly, M., & Delahunt, E. (2021). Machine learning in sports science: challenges and opportunities. *Sports Biomechanics*, *00*(00), 1–7. https://doi.org/10.1080/14763141.2021.1910334

Russel, S., & Norvig, P. (2010). *Artificial Intelligence - A Modern Approach*.

Salkuti, S. R. (2020). A survey of big data and machine learning. *International Journal of Electrical and Computer Engineering*, *10*(1), 575–580. https://doi.org/10.11591/ijece.v10i1.pp575-580

Sarlis, V., Chatziilias, V., Tjortjis, C., & Mandalidis, D. (2021). A Data Science approach analysing the Impact of Injuries on Basketball Player and Team Performance. *Information Systems*, *99*, 101750. https://doi.org/10.1016/j.is.2021.101750

Sathi, D. A. (2013). Big data analytics: Disruptive technologies for changing the game. *Mc Press*, 96.

Shafique, U., & Qaiser, H. (2014). A Comparative Study of Data Mining Process Models (KDD , CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research*, *12*(1), 217–222. http://www.ijisr.issr-journals.org/

Sun, Z., Strang, K., & Li, R. (2018). Big data with ten big characteristics. *ACM International Conference Proceeding Series*, 56–61. https://doi.org/10.1145/3291801.3291822

Testolin, A., Stoianov, I., Sperduti, A., & Zorzi, M. (2016). Learning Orthographic Structure With Sequential Generative Neural Networks. *Cognitive Science*, *40*(3), 579–606. https://doi.org/10.1111/cogs.12258

Thakur, N. (2020). *The differences between Data Science, Artificial Intelligence, Machine Learning, and Deep Learning*. https://ai.plainenglish.io/data-science-vs-artificial-intelligence-vs-machine-learning-vs-deep-learning-50d3718d51e5

Tian, C., De Silva, V., Caine, M., & Swanson, S. (2020). Use of machine learning to automate the identification of basketball strategies using whole team player tracking data. *Applied Sciences (Switzerland)*, *10*(1). https://doi.org/10.3390/app10010024

Warr, W. A. (2012). Scientific workflow systems: Pipeline Pilot and KNIME. *Journal of Computer-Aided Molecular Design*, *26*(7), 801–804. https://doi.org/10.1007/s10822-012-9577-7

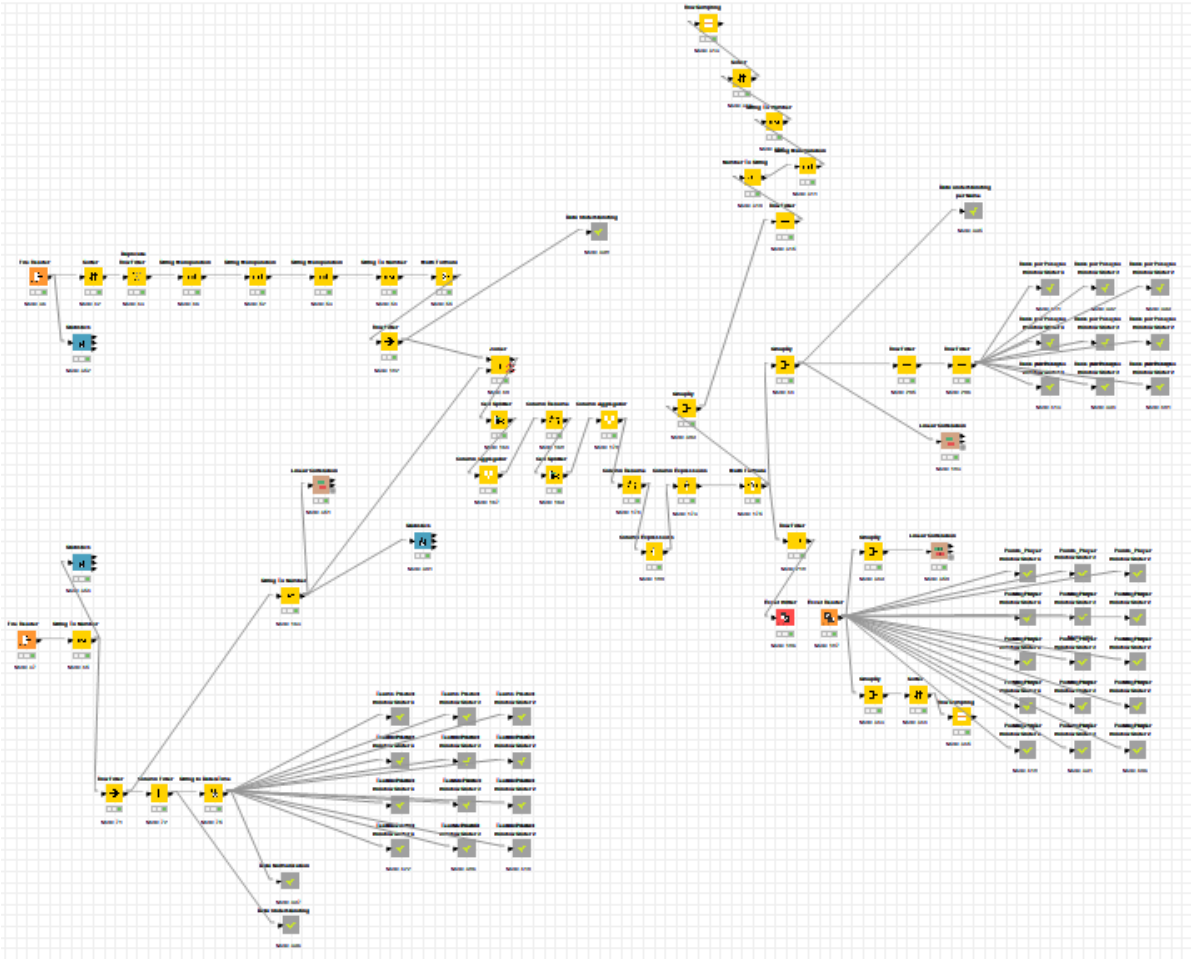
ANEXO A

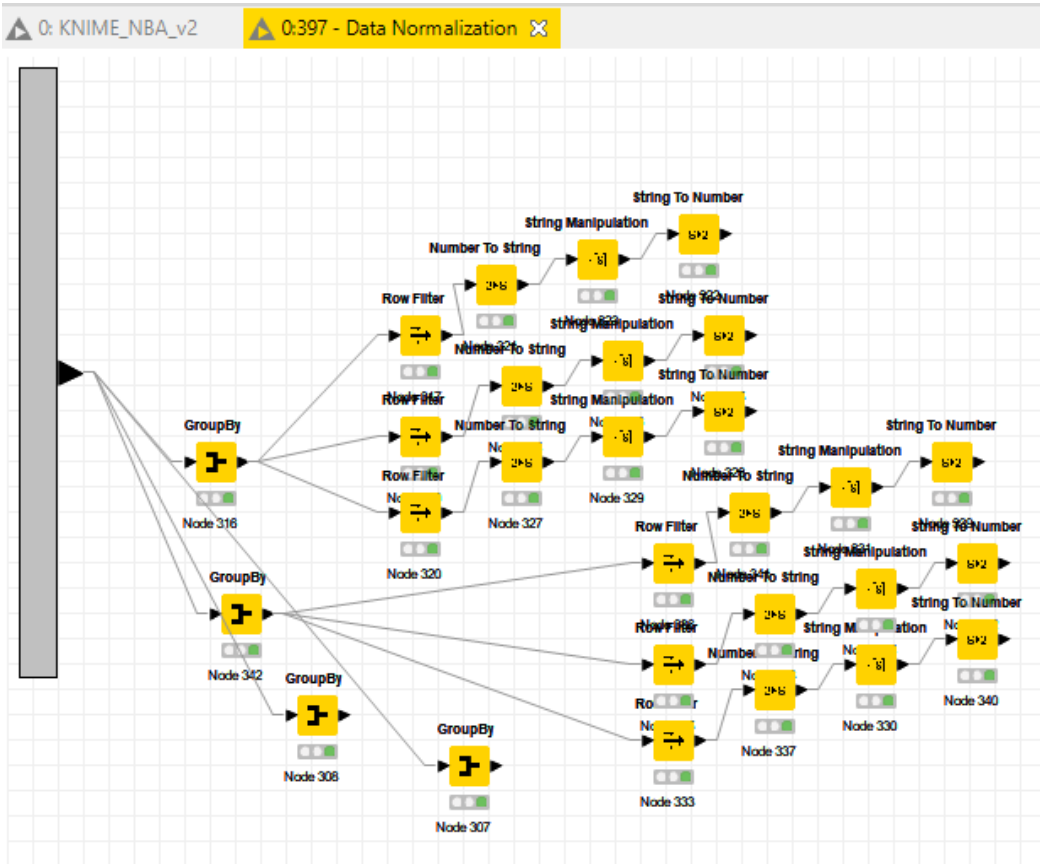
```
Sub funcao()

Dim maior As Double
Dim efficiency As Double
Dim i As Double
Dim linha As Double


lRow = Cells(Rows.Count, 1).End(xlUp).Row
game = ThisWorkbook.Sheets(1).Cells(2, "Z").Value
efficiency = ThisWorkbook.Sheets(1).Cells(2, "BD").Value
Z = 0

For i = 2 To lRow

    If (i <> 2) Then
        game = ThisWorkbook.Sheets(1).Cells(i - 1, "Z").Value
    End If

    If (ThisWorkbook.Sheets(1).Cells(i, "Z").Value = game) Then
        If (ThisWorkbook.Sheets(1).Cells(i, "BD").Value > maior) Then
            maior = ThisWorkbook.Sheets(1).Cells(i, "BD").Value
            Z = linha
            linha = i
            ThisWorkbook.Sheets(1).Cells(i, "BH").Value = "0"
            If (Z > 0) Then
                ThisWorkbook.Sheets(1).Cells(Z, "BH").Value = "0"
            End If
        Else
            ThisWorkbook.Sheets(1).Cells(i, "BH").Value = "0"
        End If
        If (linha > 0) Then
            ThisWorkbook.Sheets(1).Cells(linha, "BH").Value = "1"
        End If
    Else

        ThisWorkbook.Sheets(1).Cells(i, "BH").Value = "1"
        maior = ThisWorkbook.Sheets(1).Cells(i, "BD").Value
        linha = i
    End If

Next i

End Sub


-------------------------------

Sub function()

Dim matrix As Variant

Dim k As Double
Dim elo_diff As Double
```

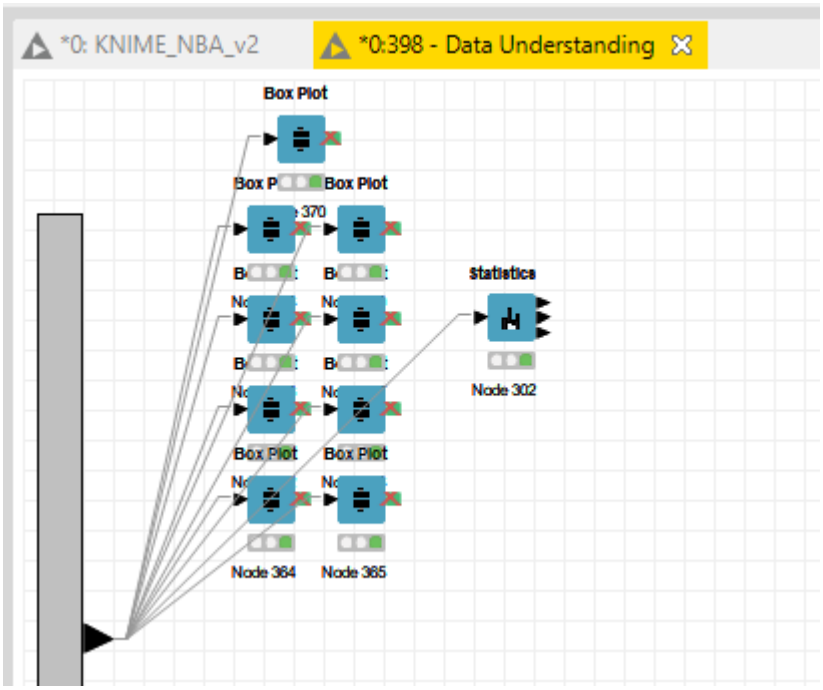
```
Dim e_home As Double
Dim e_away As Double
Dim r_home As Double
Dim r_away As Double

ReDim matrix(1 To 31, 1 To 5) As Variant
matrix(1, 1) = "Atlanta Hawks"
matrix(2, 1) = "Boston Celtics"
matrix(3, 1) = "Brooklyn Nets"
matrix(4, 1) = "Charlotte Hornets"
matrix(5, 1) = "Chicago Bulls"
matrix(6, 1) = "Cleveland Cavaliers"
matrix(7, 1) = "Dallas Mavericks"
matrix(8, 1) = "Denver Nuggets"
matrix(9, 1) = "Detroit Pistons"
matrix(10, 1) = "Golden State Warriors"
matrix(11, 1) = "Houston Rockets"
matrix(12, 1) = "Indiana Pacers"
matrix(13, 1) = "Los Angeles Clippers"
matrix(14, 1) = "Los Angeles Lakers"
matrix(15, 1) = "Memphis Grizzlies"
matrix(16, 1) = "Miami Heat"
matrix(17, 1) = "Milwaukee Bucks"
matrix(18, 1) = "Minnesota Timberwolves"
matrix(19, 1) = "New Orleans Pelicans"
matrix(20, 1) = "New York Knicks"
matrix(21, 1) = "Oklahoma City Thunder"
matrix(22, 1) = "Orlando Magic"
matrix(23, 1) = "Philadelphia 76ers"
matrix(24, 1) = "Phoenix Suns"
matrix(25, 1) = "Portland Trail Blazers"
matrix(26, 1) = "Sacramento Kings"
matrix(27, 1) = "San Antonio Spurs"
matrix(28, 1) = "Toronto Raptors"
matrix(29, 1) = "Utah Jazz"
matrix(30, 1) = "Washington Wizards"

For batata = 1 To 31
    For arroz = 4 To 5
        matrix(batata, arroz) = 1500
    Next arroz
Next batata

lRow = Cells(Rows.Count, 1).End(xlUp).Row

For i = 2 To lRow
If IsEmpty(ThisWorkbook.Sheets(1).Cells(i, "H").Value) = True Then
'If ThisWorkbook.Sheets(1).Cells(i, "H").Value = " " Then
GoTo line1
End If

    x = ThisWorkbook.Sheets(1).Cells(i, "E").Value
    y = ThisWorkbook.Sheets(1).Cells(i, "F").Value



    For Z = 1 To 31
```

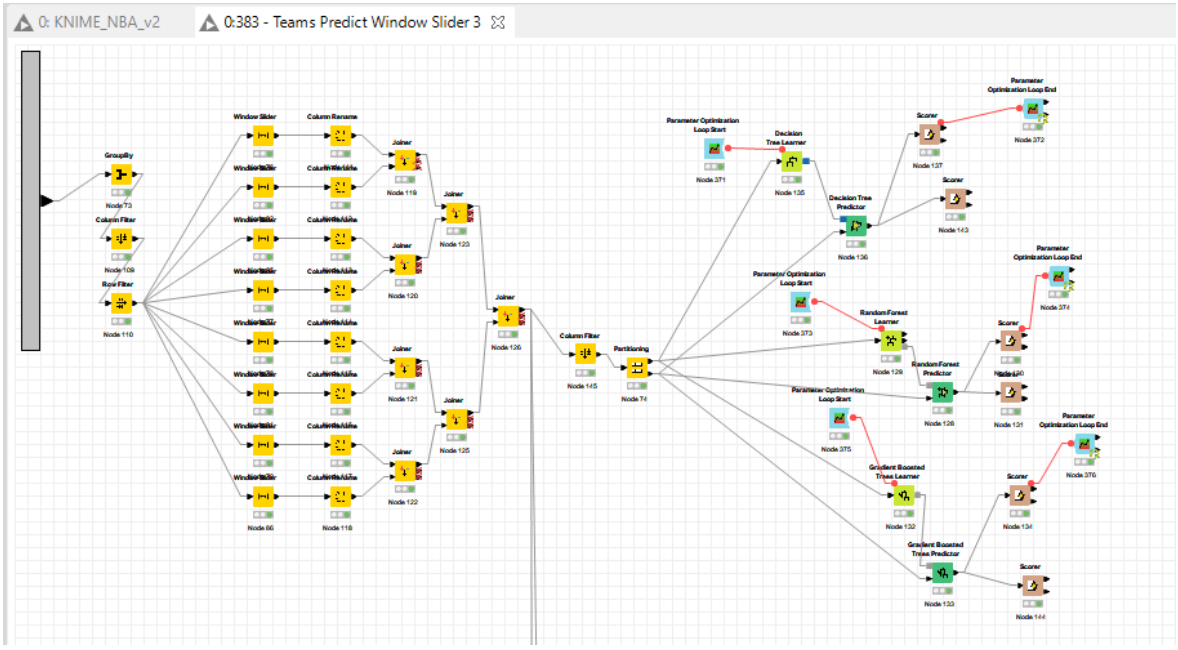
```
        If x = matrix(Z, 1) Then
            r_home_use = matrix(Z, 4)
            massa1 = Z
        End If
        If y = matrix(Z, 1) Then
            r_away_use = matrix(Z, 5)
            massa2 = Z
        End If
    Next Z

         elo_diff = Abs(r_home_use - r_away_use)
         e_home = 1 / (1 + 10 ^ ((r_away_use - r_home_use) / 400))
         e_away = 1 / (1 + 10 ^ ((r_home_use - r_away_use) / 400))
         k = 20 * (((ThisWorkbook.Sheets(1).Cells(i, "L") + 3) ^ 0.8) /
(7.5 + 0.006 * (Abs(elo_diff))))
         r_home = r_home_use + (k * (ThisWorkbook.Sheets(1).Cells(i, "M")
- e_home))
         r_away = r_away_use + (k * (ThisWorkbook.Sheets(1).Cells(i, "N")
- e_away))

         ThisWorkbook.Sheets(1).Cells(i, "O").Value = e_home
         ThisWorkbook.Sheets(1).Cells(i, "P").Value = e_away
         ThisWorkbook.Sheets(1).Cells(i, "J").Value = k
         ThisWorkbook.Sheets(1).Cells(i, "Q").Value = r_home
         ThisWorkbook.Sheets(1).Cells(i, "R").Value = r_away
         ThisWorkbook.Sheets(1).Cells(i, "K").Value = elo_diff

         matrix(massa1, 4) = r_home
         matrix(massa2, 5) = r_away

line1:

Next i

End Sub
```

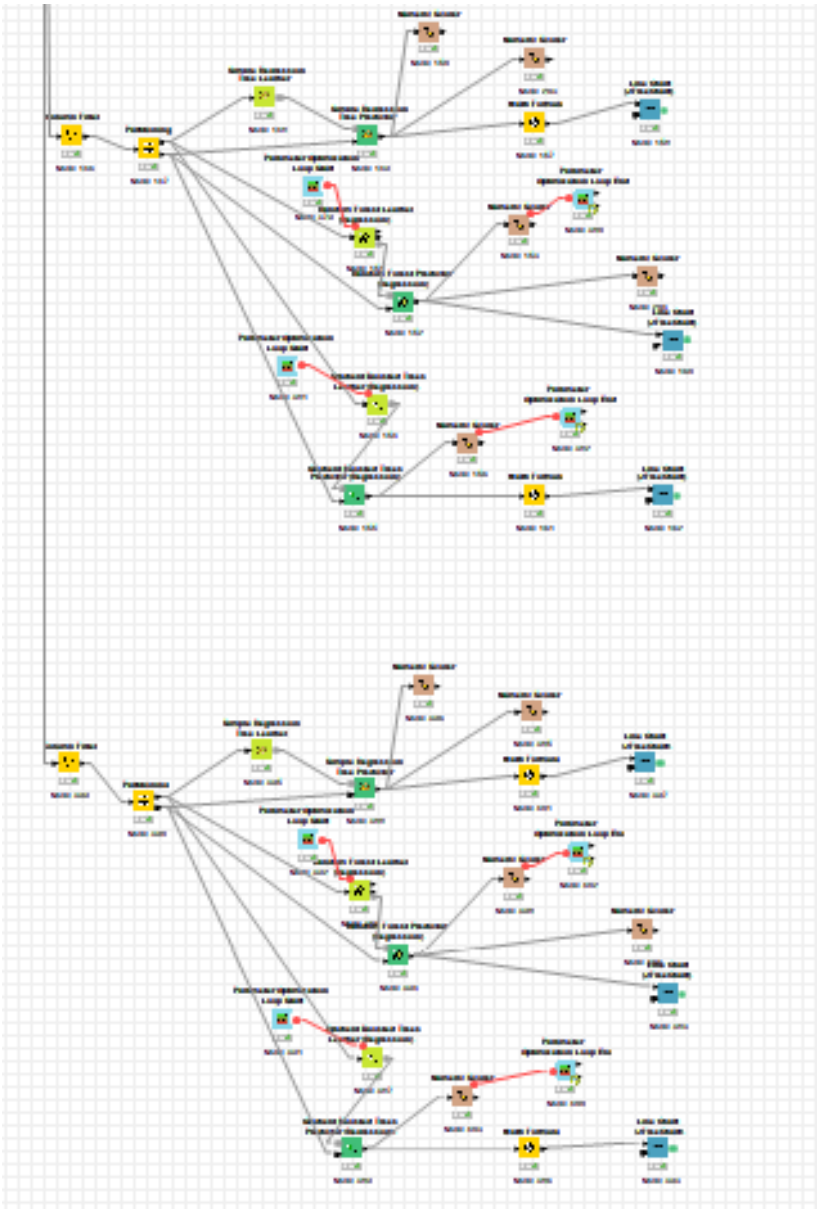
ANEXO B

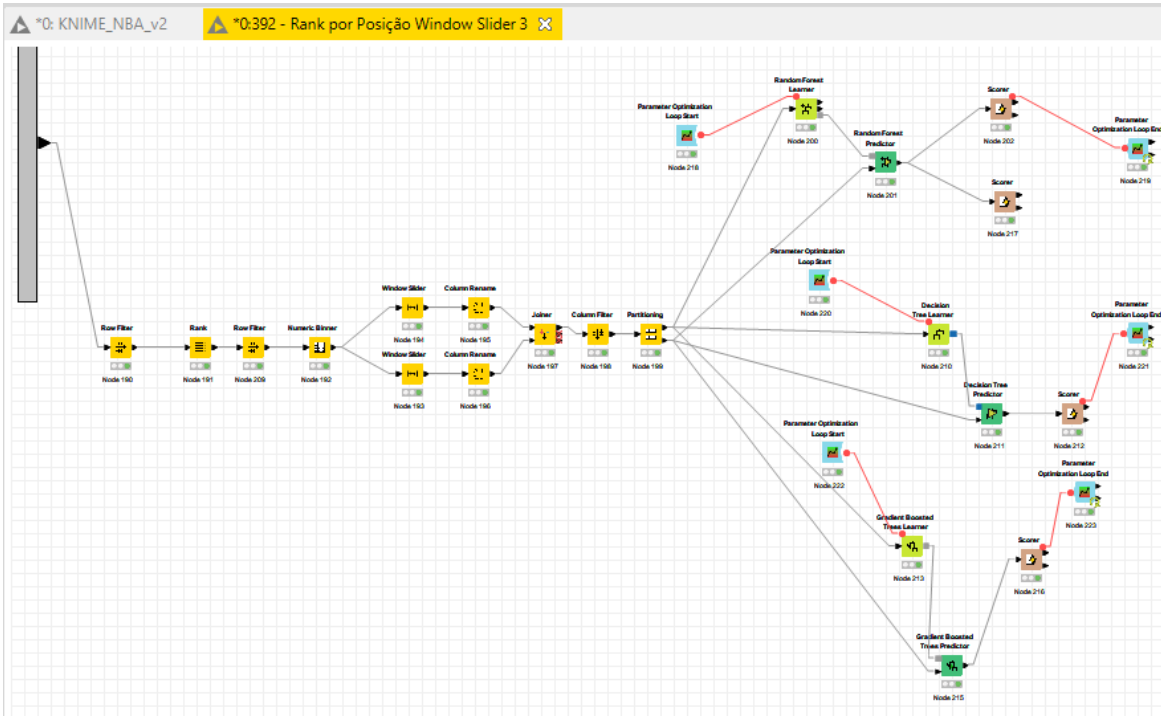


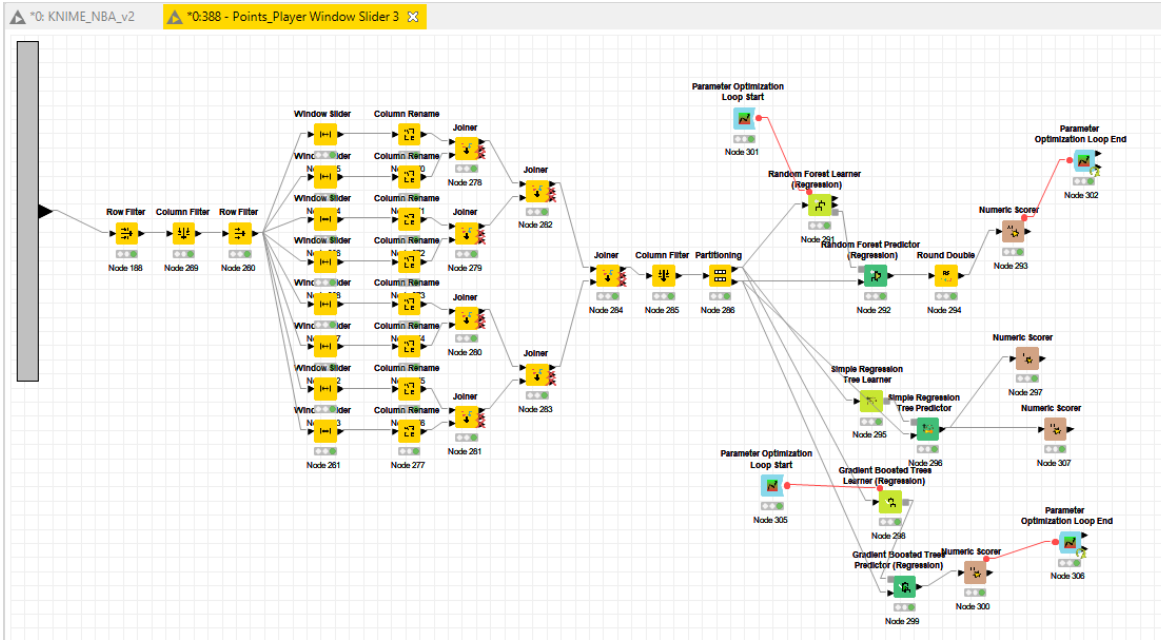

0: KNIME_NBA_v2
0:397 - Data Normalization
GroupBy
Node 316
GroupBy
Node 342
GroupBy
Node 308
GroupBy
Node 307
Row Filter
Number To String
String Manipulation
String To Number
Node 320
Node 327
Node 329
Node 333
Node 337
Node 330
Node 340


*0: KNIME_NBA_v2
*0:398 - Data Understanding
Box Plot
Box Plot
Box Plot
Box Plot
Box Plot
Statistics
Node 302
Node 364
Node 365


0: KNIME_NBA_v2
0:383 - Teams Predict Window Slider 3




*0: KNIME_NBA_v2
*0:392 - Rank por Posição Window Slider 3
Row Filter
Rank
Row Filter
Numeric Binner
Window Slider
Column Rename
Window Slider
Column Rename
Joiner
Column Filter
Partitioning
Parameter Optimization Loop Start
Random Forest Learner
Random Forest Predictor
Scorer
Parameter Optimization Loop End
Scorer
Parameter Optimization Loop Start
Decision Tree Learner
Decision Tree Predictor
Scorer
Parameter Optimization Loop End
Parameter Optimization Loop Start
Gradient Boosted Trees Learner
Gradient Boosted Trees Predictor
Scorer
Parameter Optimization Loop End


*0: KNIME_NBA_v2
*0:388 - Points_Player Window Slider 3
Row Filter
Column Filter
Row Filter
Window Slider
Column Rename
Joiner
Joiner
Joiner
Column Filter
Partitioning
Parameter Optimization Loop Start
Random Forest Learner (Regression)
Random Forest Predictor (Regression)
Round Double
Numeric Scorer
Parameter Optimization Loop End
Simple Regression Tree Learner
Simple Regression Tree Predictor
Numeric Scorer
Numeric Scorer
Parameter Optimization Loop Start
Gradient Boosted Trees Learner (Regression)
Gradient Boosted Trees Predictor (Regression)
Numeric Scorer
Parameter Optimization Loop End