Universidade do Minho
Escola de Engenharia

Carlos Manuel Moreira Rego

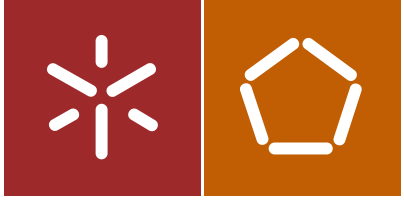
Sistema Inteligente para Análise do Posicionamento de Árbitros de Futebol

outubro de 2015

Sistema Inteligente para Análise do Posicionamento de Árbitros de Futebol

Carlos Manuel Moreira Rego

UMinho | 2015



Universidade do Minho
Escola de Engenharia

Carlos Manuel Moreira Rego

Sistema Inteligente para Análise do Posicionamento de Árbitros de Futebol

Dissertação de Mestrado
Mestrado em Sistemas de Informação

Trabalho efectuado sob a orientação do
Professor Doutor Luís Paulo Reis
Professor Doutor Filipe Meneses

outubro de 2015

Declaração

Nome: Carlos Manuel Moreira Rego

Endereço eletrónico: carlos91moreira@gmail.com

Telemóvel: +351 918 060 341

Número de Bilhete de Identidade: 13917873

Título dissertação: Sistema Inteligente para Análise do Posicionamento de Árbitros de Futebol

Orientador: Professor Doutor Luís Paulo Reis

Coorientador: Professor Doutor Filipe Meneses

Ano de conclusão: 2015

Designação do Mestrado: Mestrado em Sistemas de Informação

Universidade do Minho, __/__/____

Assinatura: _____

O segredo da força está na vontade

Giuseppe Mazzini (1805/1872 - Itália)

À minha querida mãe.

Agradecimentos

Em primeiro lugar, gostaria de expressar a minha gratidão aos meus orientadores, o Professor Doutor Luís Paulo Reis e o Professor Doutor Filipe Meneses pelo constante apoio, paciência, revisões, disponibilidade e pelo imenso conhecimento que me foi transmitido em todas as etapas do trabalho.

Gostaria também de agradecer a todos os meus amigos que me ajudaram, de alguma forma, no desenvolvimento desta dissertação. Nomeadamente ao meu grande amigo Jorge Oliveira pela disponibilidade demonstrada para a análise do jogo e correções à posição dos árbitros.

Por último mas não menos importante, quero agradecer à minha família, principalmente à minha mãe por todo sacrifício ao longo dos anos que me permitiram chegar aqui, a ela lhe devo tudo. À minha irmã por tudo que uma irmã mais nova pode fazer, ao meu pai pelos seus conselhos e à minha tia por todo o apoio e força dada.

A todos eles, o meu mais sincero agradecimento.

Abstract

Soccer is an increasingly competitive sport and clubs always try to achieve a competitive advantage over the opponents. In recent years a great interest for the study of sports analysis performance has emerged in the world of soccer. This analysis falls not only in the performance of the own team but as well in the opposing teams.

As in all sports, in soccer, there are referees whose mission is to ensure that all participants, players and coaching staff, comply with all the conditions laid down in the game's laws. The role of referees and the decisions they take have, in most cases a direct influence on the course of the game. Unfortunately these decisions aren't always correct.

In a large amount of times, these mistakes are due to referee's bad positioning in the soccer field. So it is fundamental to develop a model that may determine the correct position for a soccer referee, in the field, during a match.

For this research work, data with all the game participants' positions in the field during the match has been used as its basis. The data included, players, ball and referees positions gathered by an automatic system, referees with large experience in international competitions (just like the 2006 World Cup final) were used to obtain correct position knowledge.

The achieved results are very positive, both for predicting the assistant referee position, regardless of the team he's accompanying or the match, as to predict the referee position, in controlled game situations, in which the training data is specific for that game situation, such as corner kicks.

In short, despite the limitation on the amount of data available, the models are conclusive and can efficiently determine the correct positioning of the referee regardless of the game situation.

KEYWORDS

Data mining, Position, Data Modelling, Data Cleaning, Refereeing, Soccer.

Resumo

O futebol é um desporto cada vez mais competitivo e os clubes tentam alcançar sempre alguma vantagem competitiva em relação aos seus adversários. Nos últimos anos tem surgido no mundo do futebol um grande interesse sobre a análise da performance desportiva. Esta análise recai não só na própria equipa como também nas equipas adversárias.

Como em todos os desportos, no futebol, existem árbitros que têm como missão garantir que todos os intervenientes deste, jogadores e equipa técnica, cumpram todos os pressupostos estipulados nas leis de jogo. O papel dos árbitros e as decisões que estes tomam têm, em grande maioria dos casos uma influência direta no desenrolar do jogo. Infelizmente estas decisões nem sempre são acertadas.

Estes erros devem-se, numa grande quantidade de vezes, pelo mau posicionamento em terreno do jogo das equipas de arbitragem. É, portanto, fundamental desenvolver um modelo que pode determinar o posicionamento correto dos árbitros em terreno de jogo durante um desafio.

Para este trabalho de pesquisa, foram utilizados dados referentes ao posicionamento de todos os intervenientes de um jogo de futebol. Os dados incluíram o posicionamento dos jogadores, da bola e dos árbitros, obtidos por um sistema automático. Árbitros com vários anos de experiência em competições internacionais (como a final do Mundial de 2006) foram solicitados para obter conhecimento do posicionamento correto.

Os resultados obtidos foram muito positivos, tanto para a previsão do posicionamento do árbitro assistente, independentemente da equipa que este acompanha ou do jogo como para o posicionamento do árbitro em situações de jogo controladas, nas quais os dados de treino eram concretos à situação de jogo, tal como pontapés de canto.

Em suma, e apesar da limitação na quantidade de dados disponíveis, os modelos são conclusivos e conseguem com eficiência determinar o posicionamento correto do árbitro independentemente da situação de jogo.

Palavras-Chave:

Data Mining, Posicionamento, Modelação de dados, Tratamento de dados, Arbitragem, Futebol.

Índice

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

Siglas e Acrónimos

2D – 2 Dimensões

3D – 3 Dimensões

AA – Árbitro(s) Assistente(s)

CRISP-DM - Cross-Industry Standard Process for *Data Mining*

DSRM – Design Science Research Method

FIFA – Fédération Internacionale de Football Association

GPS – Global Position System (Sistema de Posicionamento Global)

IFAB – International Football Association Board

MAE - Mean absolute error

RAE - Relative absolute error

RMSE – Root Mean Square Error (erro da raiz do valor quadrático médio)

RRSE - Root relative squared error

RSE - Relative squared error

RSP - Regression Scatter Plot

SI – Sistemas de Informação

TI – Tecnologias de Informação

UEFA - Union of European Football Associations

1. Introdução

O capítulo introdutório do documento faz o enquadramento do trabalho proposto, enumera os objetivos e resultados esperados, resume os problemas enfrentados no desenvolvimento do trabalho, e indica a estratégia de investigação. É também feita uma breve descrição da estrutura do documento.

1.1. Enquadramento do trabalho proposto

O futebol é um desporto disputado por duas equipas, equipadas de forma distinta, compostas por um máximo de 11 jogadores e com um mínimo de 7 jogadores cada uma, dos quais um é o guarda-redes. Os jogos podem jogar-se em superfícies naturais ou artificiais, com a particularidade de estas terem que ser obrigatoriamente de cor verde, com as dimensões mínimas de 90x45m e máximas de 120x90m com a obrigatoriedade que as linhas laterais sejam maiores do que as linhas de baliza.

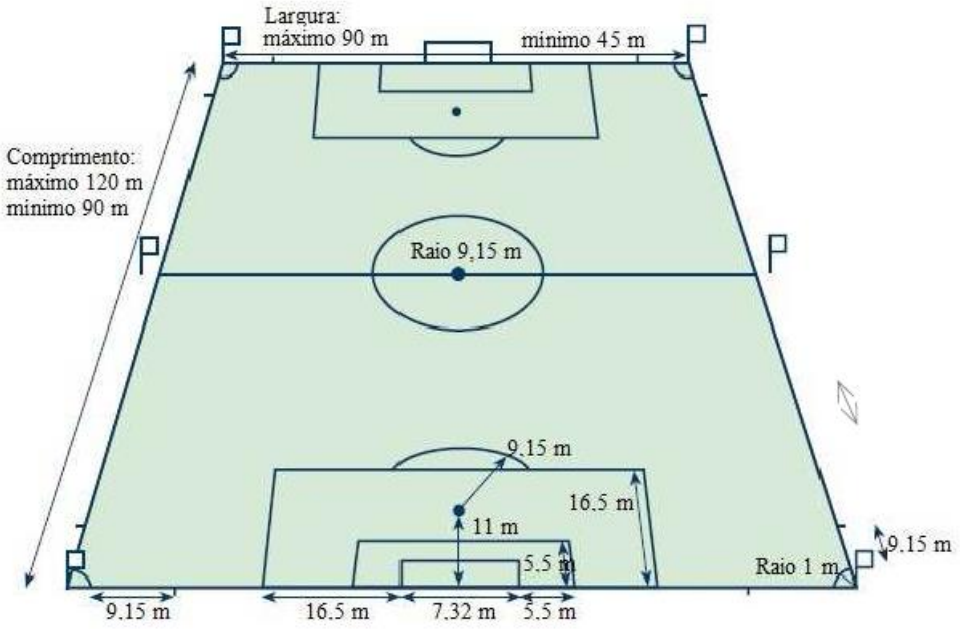


Figura 1 Dimensões e marcações obrigatórias do terreno de jogo. Extraída de "As leis de jogo 2013/2014" (fonte:International Football Association Board (IFAB), 2013)

Tendo em conta as dimensões dos terrenos de jogo de futebol, é fácil compreender que os árbitros têm que percorrer uma grande distância de forma a estarem melhor preparados para assumirem uma correta decisão. A principal missão destes é assegurar que todas as leis de jogo são cumpridas (Internation Football Association Board, 2013). Estudos comprovam que os árbitros correm em média

por jogo, nas altas competições, entre 10 a 12 km, sendo que entre 10 a 15% desta distância é percorrida em alta intensidade (velocidades superiores a 18 km/h). Os valores para os árbitros assistentes (AA) diferem um pouco, correndo estes uma menor distância, entre 6 a 7 km sendo 15 a 20% a alta intensidade (Krustrup et al., 2009).

O tamanho do terreno de jogo, bem como a iluminação artificial e natural, as sombras dos jogadores e do estádio, a deformação de objetos devido aos movimentos rápidos, pessoas que não importam para a análise como por exemplo os adeptos, os treinadores nas áreas técnicas, os jogadores suplentes a aquecerem, entre outros, dificultam a capacidade para localizar a bola, os jogadores e os árbitros no campo (Naidoo & Tapamo, 2006).

1.2. Motivação

O futebol é o desporto mais popular em todo o mundo (Mughal, 2014) e como tal é o responsável pela mobilização de mais massas. O futebol é assim, de uma forma mais ou menos natural, um dos desportos que mais dinheiro gera em torno deste de acordo com o artigo da Forbes (2014) - "As 50 equipas desportivas mais valiosas de 2014" - no qual indica que apesar do enorme crescimento de outros desportos, em especial dos clubes dos Estados Unidos da América, não é possível comparar as equipas de topo do futebol europeu no que respeita a valor e alcance mundial. A lista supracitada é liderada pelo Real Madrid com um valor estimado de 3.44 bilhões de dólares por dois anos consecutivos, seguido pelo rival Barcelona com um valor estimado de 3.2 bilhões de dólares e Manchester United a fechar o pódio com 2.81 bilhões de dólares. Assim como seria expectável o futebol deveria afirmar-se como um desporto em que a adaptação é uma constante em termos tecnológicos. Como é o caso do ténis ou o râguebi, em que em situações de difícil análise a equipa de arbitragem pode-se auxiliar das tecnologias para tomar uma decisão. Devido a esta resistência por parte das instituições que superintendem o futebol, os árbitros de futebol encontram-se constantemente mais expostos à crítica.

Apesar da importância das decisões tomadas pela equipa de arbitragem durante um jogo de futebol, sabe-se muito pouco sobre os fatores que influenciam a qualidade destas (Oudejans et al., 2005). Parece, no entanto, que um posicionamento correto no terreno de jogo seja crucial de acordo com Rontoyannis, Stalikas, Sarros & Vlastaris citados por Javier Mallo, Frutos, Juárez, & Navarro (2012). A distância do árbitro para o lance pode ser justificada pela condição física, o tempo de jogo e o local

da falta (Krustrup et al., 2009). J Mallo, Veiga, López de Subijana, & Navarro (2010) determinaram que durante um jogo internacional um árbitro tem que tomar cerca de 140 decisões assinaláveis, com uma média de 41 faltas por partida. Pese embora a importância de um bom posicionamento, é percetível que caso o árbitro se encontre muito próximo da zona onde a infração ocorreu tal pode comprometer a capacidade de visualizar e analisar o lance e até influenciar a jogada. Todavia, ao encontrar-se demasiado longe aumenta a probabilidade de não obter uma boa visibilidade, intensificando desta forma o risco de errar (Javier Mallo et al., 2012).

Os deveres dos AA, entre outros, focam-se essencialmente em auxiliar o árbitro quando estes têm melhor visão. Ou algum incidente ocorra fora do campo de visão do árbitro e quando um jogador deve ser sancionado por se encontrar na posição de fora de jogo (Internation Football Association Board, 2013). Por esta razão os assistentes devem estar constantemente em linha com o penúltimo adversário ou a bola caso esta esteja mais próxima da linha de baliza. Desta forma o posicionamento destes é fulcral para a tomada correta da decisão em casos de fora de jogo uma vez que tanto a distância para a linha do fora de jogo e o ângulo de visão são aspetos chave (Javier Mallo et al., 2012).

Numa altura que se fala constantemente na implementação das novas tecnologias para auxiliar o trabalho do árbitro em campo e dos seus assistentes, a principal motivação para este trabalho de investigação é a possibilidade de contribuir com um sistema original que seja capaz de ajudar o árbitro na sua colocação no terreno de jogo. Esta colocação será dependente da posição da bola e dos jogadores, nas diferentes situações de jogo, de forma a diminuir a quantidade de erros de avaliação dos lances devido à má colocação dos árbitros.

1.3. Objetivos da dissertação e contributos

O principal objetivo deste trabalho consiste em definir um modelo para o correto posicionamento de árbitros de futebol, nas diferentes situações de jogo, dependendo da posição da bola, dos jogadores das duas equipas intervenientes e da direção da jogada.

Para conseguir atingir este objetivo principal será necessário o cumprimento dos seguintes sub-objetivos:

1. Definir um modelo para o correto posicionamento dos elementos da equipa de arbitragem;

2. Analisar detalhadamente todos os lances dos jogos que exigiram gestão por parte da equipa de arbitragem e pedir auxílio de um perito nos lances onde o árbitro errou e indicar qual o posicionamento correto para que o lance fosse bem ajuizado;
3. Identificar uma correlação entre o posicionamento da equipa de arbitragem e a eficácia das decisões tomadas;
4. Justificar a importância do posicionamento do árbitro no terreno de jogo comparativamente com a bola e com os atletas na tomada correta da decisão.

Após as conclusões atingidas, considera-se que existirá um maior investimento na área, não só da deteção da equipa de arbitragem, inclusive, como também nesta, aqui levantada, problemática da posição ideal do árbitro em terreno de jogo em todas as situações do jogo para a redução de erros.

1.4. Organização do documento

Este documento encontra-se de acordo com as normas da Universidade do Minho para dissertações de Mestrado e Doutoramento. Este foi dividido em nove capítulos.

O primeiro capítulo da dissertação é a introdução. Este capítulo é responsável pelo enquadramento do trabalho proposto, assim como dos objetivos propostos a cumprir ao longo do projeto e por último, é enunciado o problema de investigação, ou seja, o que motivou o desenvolvimento desta solução.

O documento prossegue com a abordagem metodológica. Este é um capítulo fulcral no desenvolvimento da dissertação pois será explicada a estratégia de pesquisa bibliográfica, isto é, qual o processo de seleção dos documentos científicos a analisar, a metodologia de investigação utilizada e por último, como foram equacionadas as questões éticas e tratamento de dados.

O terceiro capítulo do documento é o estado de arte. Este estado de arte é o resultado do trabalho de seleção realizado no ponto anterior. De realçar que o capítulo está dividido em três secções. A primeira centra-se no posicionamento das equipas de arbitragem nas variadas situações de jogo, a segunda nas tecnologias atuais no que toca à deteção dos constituintes do jogo, tais como jogadores, bola e árbitros, e por último, nos pacotes de *software* atualmente existentes no mercado para a análise de futebol.

Data mining figura como o quarto capítulo deste documento. Este capítulo da dissertação centra-se nos fundamentos gerais deste conceito e identifica os principais tipos de problemas de *data mining*. É de seguida elaborada uma discriminação do *data mining* no desporto e que desenvolvimentos existem nesta área, assim como na previsão de localização num espaço.

No quinto capítulo é feito um estudo das possíveis metodologias de desenvolvimento a seguir e ferramentas a utilizar para o desenvolvimento deste trabalho de pesquisa.

O sexto capítulo é o seguimento da metodologia de desenvolvimento selecionada – CRISP-DM. Cada uma das secções correspondem às várias fases da metodologia.

O capítulo seguinte é composto pela discussão dos resultados obtidos e esperados deste modelo e por último é feita uma conclusão a todo o trabalho elaborado e uma perspetiva de trabalhos futuros.

O documento é ainda composto de anexos a respeito dos resultados obtidos da modelação dos dados.

2. Abordagem Metodológica

Neste capítulo será discutida a estratégia de pesquisa bibliográfica para este trabalho e como se procedeu à seleção dos documentos de análise. Procedida da metodologia de investigação e por último das questões éticas e tratamento de dados.

Uma abordagem metodológica adequada e capaz de auxiliar a realização dos objetivos enunciados para o projeto é fundamental em qualquer trabalho de investigação.

2.1. Estratégia de pesquisa bibliográfica

Para a realização desta investigação, será efetuada uma revisão de literatura que incidirá em dois grandes tópicos: (1) Como é que são obtidos os dados para o rastreamento dos jogadores, bola e árbitros num jogo de futebol profissional, o tratamento a que estes dados são sujeitos e os pacotes de *software* existentes atualmente no mercado e (2) a importância do posicionamento da equipa da arbitragem no terreno de jogo para a tomada correta da decisão.

Este estado de arte servirá de suporte teórico ao desenvolvimento do trabalho, sendo por isso, fulcral que a estratégia de investigação esteja bem definida para que o trabalho base seja efetuado com o máximo rigor.

O desenvolvimento do estado de arte será feito a partir da definição dos objetivos da pesquisa. Nesta fase serão utilizados alguns motores de busca como o Google Scholar para uma abordagem inicial do tema e após esta etapa, serão utilizadas bases de dados mais voltadas ao tema como o Scopus ou o Web of Science. A definição das *keywords* é o próximo passo para uma correta revisão de literatura. Após a leitura e seleção dos artigos nucleares resultantes da pesquisa e através das citações serão identificados os artigos mais recentes que abordam o tema aqui proposto. Serão estes documentos no qual a revisão de literatura reincidirá.

É necessário definir um critério de seleção para o processo de pesquisa sabendo que as fontes bibliográficas utilizadas contêm informação, nem sempre relevante para o tema aqui desenvolvido.

É neste contexto que algumas fontes bibliográficas foram à priori colocadas de parte, sobretudo na abordagem inicial a este tema. Algumas fontes bibliográficas não possuem a qualidade necessária para desenvolver esta dissertação. De referir a necessidade de filtrar alguns recursos bibliográficos devido aos elevados custos associados à disponibilização integral do conteúdo.

Na fase posterior a esta pré-seleção bibliográfica, foram identificados os pontos fundamentais para o desenvolvimento deste trabalho, entre os quais:

- Autor(es) identificado(s);
- Data de publicação;
- Conteúdo com referências bibliográficas;
- Documento citado por outros artigos;
- Publicado por entidades reconhecidas;
- Elaborado sob supervisão de instituições académicas.

As *keywords* utilizadas até ao momento foram "Football tracking", "Sport analysis", "Soccer vídeo analysis", "Intrusive Systems", "Player detection", "Ball tracking" e "Image processing" para o estado de arte que diz respeito ao rastreamento dos intervenientes dos jogos de futebol. Para a análise do desempenho das equipas de arbitragem foram utilizadas as *keywords* "Referee position", "Referee position model", "Referee decision-making", "Referee position analysis", "Football distance to incidents", "Referee expert performance" e "Assistant referees". Foram enfrentadas algumas dificuldades para encontrar documentos científicos que relatem estudos ao posicionamento do árbitro no que toca a sua posição. Foram no entanto encontrados artigos que faziam esse relacionamento relativamente à distância ao lance, mas não diretamente ao ângulo de visão. Existe uma secção, neste mesmo documento, reservada aos trabalhos relacionados feitos na área.

Deste estado de arte foi também elaborada uma pesquisa sobre *Data Mining* para que a compreensão do conceito fosse mais aprofundada. Esta pesquisa teve essencialmente como objetivo um melhor tratamento de dados após conseguir caracterizar o tipo de problema de *data mining* e consequentemente os melhores algoritmos para se obter o melhor modelo de previsão possível. Para alcançar este objetivo foram feitas pesquisas de artigos científicos pelas seguintes keywords "*Data Mining*", "*Data Mining in sports*", "*Data Mining geospatial*", "*Data Mining cartesian coordinates*".

2.2. Metodologia de Investigação

A metodologia de investigação a utilizar durante este trabalho de pesquisa será a *Design Science Research*, que poderá ser traduzida para português como Projeto Científico de Pesquisa.

O paradigma do projeto científico de pesquisa é muito relevante para os sistemas de informação (SI) porque aborda duas questões chave da área: o papel das Tecnologias de Informação (TI) numa pesquisa de SI e a perceção da falta de relevância profissional da investigação (Klein, 2003).

Esta metodologia requer que primeiro seja desenvolvida uma nova solução e de seguida seja explicado como essa nova solução resolveu o problema. Para que esta medida seja conseguida é necessária uma caracterização do antes e do depois da solução apresentada. O primeiro passo a seguir será a perceção do problema, seguida da sugestão da solução, do desenvolvimento do artefacto e a consequente avaliação à solução, se esta é capaz de satisfazer as necessidades do problema, e por último a conclusão.

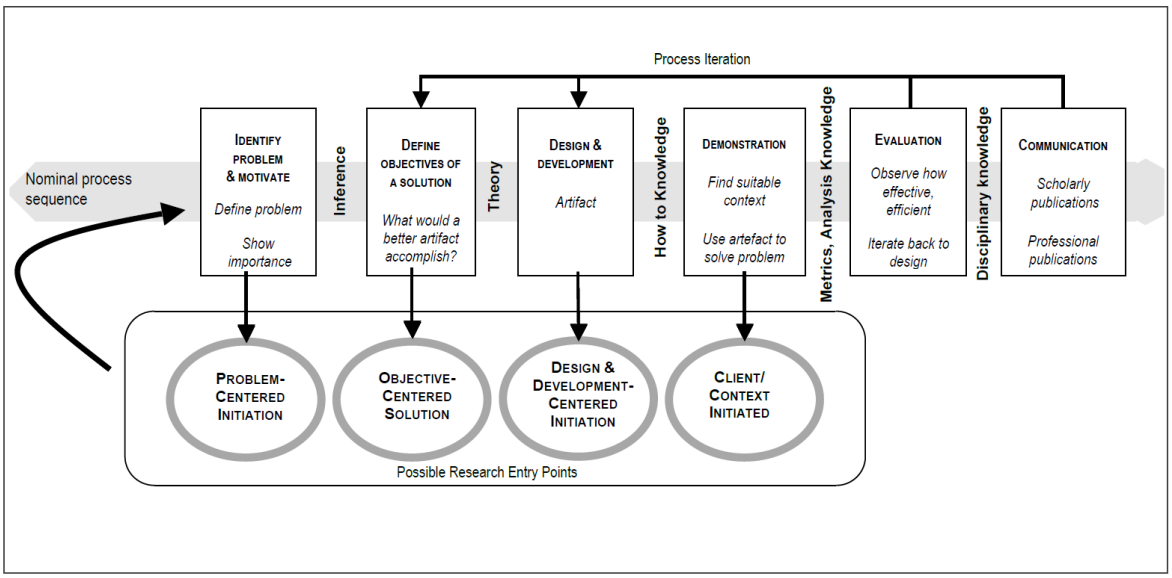


Figura 2 Modelo do Processo de Metodologia do Design Science Research (fonte: Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007)

A Figura 2 representa o modelo de processo de DSRM. Este pode ser dividido em 6 atividades distintas. Este processo apesar de estruturado numa ordem sequencial não há expetativa que os pesquisadores procedam sempre por ordem da atividade 1 à 6. Na verdade podem começar por qualquer fase (Peffers et al., 2007).

Segundo Peffers et al., (2007) a atividade 1 define o problema específico de investigação e justifica o valor de uma solução. A atividade 2 infere os objetivos de uma solução a partir da definição do

problema e do conhecimento do que é possível e viável. A atividade 3 consiste em criar um artefacto. O artefacto pode ser uma construção, um modelo ou um método. A atividade 4 demonstra a utilização do artefacto para resolver uma ou mais instâncias do problema. A atividade 5 comporta observar e medir o quão bem o artefacto suporta a solução do problema e por último durante a atividade 6 comunica-se o problema, o artefacto, a sua utilidade e novidade, o rigor e a sua eficácia à comunidade científica.

Portanto, de acordo com (Hevner & Chatterjee, 2010) esta metodologia suporta um paradigma de pesquisa pragmática que solicita a criação de artefactos inovadores para resolver problemas do mundo real. Assim o projeto científico de pesquisa combina um foco sobre o artefacto de TI com uma alta prioridade de relevância no domínio da aplicação.

2.3. Questões Éticas e Tratamento de Dados

Para o desenvolvimento desta dissertação foi necessário recorrer a dados com informação das posições (cartesianas) dos atletas relativamente ao jogo que opôs a Itália à França no Mundial de 2006 na Alemanha. Estes dados não são do domínio público e consequentemente estes devem-se manter confidenciais.

Estes dados foram obtidos pelo orientador deste trabalho após lhe terem sido fornecidos por uma organização internacional. Estes dados apesar de confidenciais, podem ser utilizados para fins académicos.

Não foi possível obter dados de outros jogos, mesmo de campeonatos amadores, por as leis do jogo não permitirem o uso de sistemas intrusivos e por falta de capacidade de filmar um jogo de futebol de várias perspetivas para depois extrair as posições.

3. Estado da Arte

O estado de arte centra-se nas tecnologias utilizadas para a deteção e rastreamento de jogadores, bola e equipa de arbitragem durante um jogo de futebol profissional e como estes dados são depois processados por pacotes de *software* específico. Outra área de foco desta revisão de literatura é o posicionamento correto dos árbitros no terreno de jogo e a importância desta para uma decisão acertada nos diversos lances do jogo.

3.1. Posicionamento da equipa de arbitragem

Este capítulo incide nas normas e instruções que existem por parte da International Football Association Board (IFAB) para a colocação e movimentação da equipa de arbitragem em lances previstos pelas Leis do jogo e em "bola corrida".

3.1.1. Árbitros

O árbitro de futebol tem como principal dever zelar pela aplicação das Leis do Jogo (Internation Football Association Board, 2013). Todavia este estudo irá focar-se nas faltas e incorreções dos jogadores de futebol. Assim, encontrar-se no local certo, no momento certo, é importantíssimo de forma a obter uma melhor visão do lance e uma avaliação correta do mesmo (Javier Mallo et al., 2012). Obviamente que um bom nível físico do árbitro é também requerido (Oliveira, Orbetelli, & Neto, 2011) uma vez que decisões rápidas, baseadas no julgamento correto do árbitro, e uma ação imediata contribui para o desenvolvimento imparcial dos jogos (Oliveira et al., 2011).

O posicionamento do árbitro com a bola em jogo está já previsto na secção de linhas orientadoras para árbitros no livro das leis do jogo (Internation Football Association Board, 2013). Segundo a Internation Football Association Board, (2013) estes têm as recomendações de:

- o jogo deve desenrolar-se entre o árbitro e o AA mais próximo da jogada;
- O AA mais próximo deve encontrar-se dentro do campo de visão do árbitro e este último deverá utilizar o sistema diagonal[1];

[1] Movimentação do árbitro no terreno de jogo deve ser feita na diagonal desta, de acordo com a figura 3

- Acompanhar o jogo numa posição lateral;
- Deve encontrar-se suficientemente perto da jogada para observar o jogo, mas sem interferir nele e por último;
- "O que é preciso ver" nem sempre acontece nas proximidades da bola e como tal é necessário o árbitro prestar atenção à agressividade das confrontações individuais entre jogadores afastados da bola, às possíveis infrações na zona para onde se dirige o jogo e às infrações que são cometidas depois de a bola ter sido afastada.

A figura 3 mostra qual deve ser o posicionamento dos árbitros durante o jogo, os AA junto às linhas laterais, cada um no seu respetivo meio campo, a movimentação recomendada em diagonal do árbitro e a colocação dos árbitros adicionais (embora a sua presença só se verifique em jogos de competições da UEFA) entre a baliza e a bandeirola de canto do lado dos assistentes. A mesma figura, do lado direito, mostra a área de ação da equipa de arbitragem por zona. Os assistentes assumem a zona lateral e o árbitro a zona central.

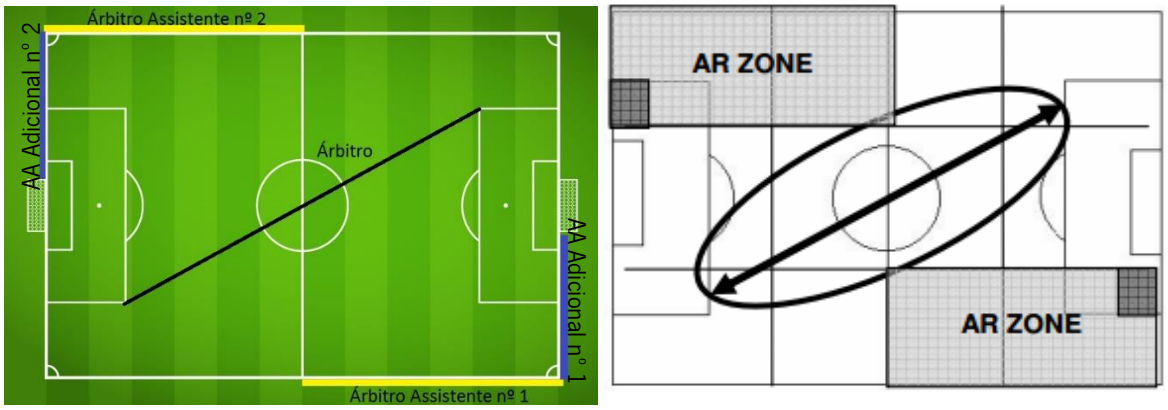


Figura 3 Áreas de ação dos árbitros, AA e árbitros adicionais no terreno de jogo

3.1.2. Árbitros Assistentes

De acordo com a Internation Football Association Board, (2013) os AA ajudam o árbitro a velar pela aplicação das Leis do jogo. Os AA têm como missão, salvo decisão contrária do árbitro, assinalar (1) quando a bola sai completamente do terreno de jogo, (2) a que equipa pertence o lançamento lateral, ou se há pontapé de canto ou de baliza, (3) quando um jogador deve ser sancionado por se encontrar na posição de fora de jogo, (4) quando é solicitado uma substituição, (5) quando um comportamento condenável ou qualquer outro incidente aconteça fora do campo de visão do árbitro, (6) quando forem cometidas infrações em que os AA tenham uma melhor visão que o árbitro e (7) quando nos pontapés

de grande penalidade o guarda-redes se mova para a frente antes que a bolsa seja pontapeada e se a bola transpôs a linha de baliza (Internation Football Association Board, 2013).

Todos estes deveres devem ser cumpridos, destacando-se o fora de jogo pela sua importância fulcral no jogo. Esta lei, apesar de ser das mais pequenas e de fácil entendimento na teoria, é das mais difíceis de colocar em prática no terreno de jogo e origina mais debate.

Os AA são instruídos para estarem constantemente em linha com o penúltimo defensor de uma equipa, ou os dois últimos, ou da bola se esta estiver mais próxima da linha de baliza, de forma a estar o melhor posicionado possível para a sua decisão. A este posicionamento existem algumas exceções, tais como quando o guarda-redes tem a bola controlada nas mãos na sua própria área de grande penalidade, em pontapés de canto, em pontapés de baliza e pontapés de grande penalidade.

3.1.3. Fatores relacionados com a qualidade da tomada de decisão

Um estudo efetuado por Oliveira et al., (2011) enumera alguns motivos que podem estar relacionados com uma correta decisão levada a cabo pela equipa de arbitragem. O primeiro será a visão do árbitro. O sentido que os humanos mais utilizam para obter informação do meio ambiente é a visão. A capacidade do árbitro de interpretar jogadores rápidos, em movimentos simultâneos e sequenciais, depende diretamente da capacidade de visão. Este fator afeta a capacidade de análise em todos os momentos do jogo. Um segundo fator será o aumento de ansiedade, derivado da pressão, por causa das alterações de atenção e concentração. Por fim, outro fator está associado ao posicionamento durante o jogo.

Atualmente todos os estudos realizados pela comunidade científica elaboram apenas uma correlação entre as decisões corretas e erradas com a distância do árbitro à jogada. Esta é uma medida muito mais fácil de avaliar do que a localização do árbitro, no que toca ao ângulo de visão que este tem em relação ao lance, ou se tem jogadores entre ele e onde a infração ocorre, que lhe obstruam a visão.

3.2. Deteção dos constituintes do jogo

Desde há muito tempo verifica-se a necessidade de equipas de futebol profissional se estudarem a si mesmas, bem como os seus adversários procurando obter alguma vantagem competitiva sobre os

mesmos. Esta análise tem sido realizada de forma manual através da visualização das gravações do jogo, por operadores, depois do jogo (C. B. Santiago, Sousa, & Reis, 2012).

A área de deteção automática de jogadores representa um enorme desafio dada a complexidade da análise do desporto em si devido à velocidade e variações de direção dos atletas. Para a obtenção desta informação são usadas tecnologias de deteção e rastreamento que podem ser divididas em dois grupos distintos: Sistemas Intrusivos – nos quais são utilizados *tags* especiais ou sensores nos atletas e sistemas não intrusivos – nos quais não são usados objetos (C. Santiago, 2011).

3.2.1. Jogadores e árbitros

3.2.1.1. Sistemas intrusivos

Os sistemas intrusivos recorrem a sensores sem fios e *tags*. Estas características torna-os sensíveis à degradação e interferências, tanto por outros objetos como por colisões de sinais, entre outros. Nestes sistemas as principais dificuldades parecem estar mais relacionadas com o *hardware* e não tanto com o *software* porque os sinais necessitam de ser fortes o suficiente para serem detetados pelas antenas em boas condições e as *tags* precisam de ser pequenas e leves para não impedir o conforto e a performance dos atletas (C. Santiago, 2011).

É possível destacar algumas tecnologias para o rastreamento dos atletas durante um jogo de futebol, como sendo as mais utilizadas para minimizar o problema da localização dos atletas.

O *Global Positioning System*[2] (GPS) é um sistema de navegação por satélite baseado no espaço que fornece informação da localização e tempo em todas as condições meteorológicas desde que exista uma linha de visão desobstruída para quatro ou mais satélites GPS (National Research Council (U.S.). Committee on the Future of the Global Positioning System, Administration, & System, 1995). O GPS tem sido utilizado nos desportos para localizar os jogadores (C. Santiago, 2011).

Uma outra tecnologia utilizada para a deteção de jogadores de futebol é *Radio Frequency Identification* (RFID). Este é um método de identificação automática sem fios capaz de rastrear os jogadores utilizando ondas de rádio. Esta tecnologia necessita de um recetor e um conjunto de tags que podem

[2] Sistema de Posicionamento Global em português

ser classificadas como passivas (apenas são detetáveis num intervalo inferior a 13 metros do recetor) e ativas (conseguem ser detetadas a 40 metros do recetor, necessitando da sua própria fonte de energia interna) (Abreu, 2010).

A tecnologia de rede sem fios Wi-Fi usa ondas de rádio para fornecer internet e conexões de rede. Esta pode ser também utilizada para a conceção de um sistema de rastreamento reutilizando a rede de dados sem fios é possível criar um sistema de localização por cima desta infraestrutura (Abreu, 2010). Comparativamente com as duas últimas tecnologias referidas os riscos de oclusão e perda de sinal é muito reduzido, principalmente em ambientes de baixos níveis de concentração de materiais (Mingkhwan, 2006).

O *Bluetooth* é um protocolo sem fios disponível em quase todos os telemóveis móveis do mercado. Embora este protocolo possa ser utilizado num sistema de localização, os consumos das baterias, o pequena área de deteção e o processo de estabelecimento de conexão não transparente fazem desta abordagem inadequada para um sistema de rastreamento eficiente (Abreu, 2010).

A tecnologia *ultra-wideband* é usada para transmitir grandes quantidades de dados num largo *sprectrum* de frequência a um nível de consumo reduzido para pequenas distâncias (USC - Viberti School of Engeneering, 2006). A principal vantagem desta tecnologia é a capacidade de transmitir dados através de portas e outros obstáculos que normalmente refletem os sinais a larguras de banda mais limitadas e a mais consumos (Rouse, 2008).

Local Position Measurement (LPM) é a tecnologia mais precisa usada no desporto para a localização dos seus intervenientes (INMOTIO, 2012). O sistema é baseado na tecnologia RFID aplicando-a numa grande variedade de situações. Os sistemas consiste em antenas (estações base) e de transponders. As estações base são posicionadas em torno do campo e os transponders são usados pelos atletas.

Os cálculos são efetuados em tempo real (as estações calculam a posição dos transponders e consequentemente dos atletas) e por consequente, os dados são analisados durante as medições. A figura 4 facilitará a compreensão deste processo.

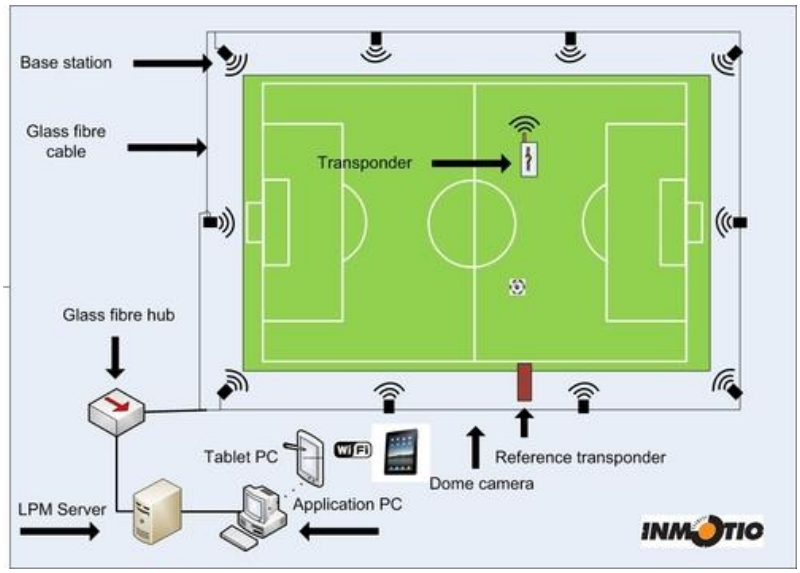


Figura 4 Processo de recolha e tratamento da informação do sistema LPM (fonte: INMOTIO, 2012).

3.2.1.2. Sistemas não intrusivos

De acordo com a lei 4, equipamento dos jogadores, do livro das leis do jogo (Internation Football Association Board, 2013), os atletas não estão autorizados ao uso de qualquer outro dispositivo, considerado joia, para além do equipamento básico definido nas leis. Como tal, o uso de sistemas intrusivos em jogos oficiais torna-se impossível. Como tal o uso de sistemas não intrusivos, sistemas de visão, ganharam utilidade.

Infelizmente, os sistemas de visão, da perspetiva de uma câmara, os vários objetos vistos no terreno de jogo são bastante parecidos, estão em constante movimento, mudam de forma e unem-se muito frequentemente tornando difícil de monitorizar os jogadores de forma individual (C. Santiago, 2011). É possível identificar duas grandes tecnologias utilizadas para superar este problema: transmissões televisivas e sistemas de câmaras dedicadas.

Uma tecnologia importante é a localização, etiquetagem e rastreamento de jogadores. Esta é uma tarefa bastante desafiante dado a muitas dificuldades, tais como oclusão, aparência dos jogadores semelhantes com baixa descriminação, número variável de jogadores, movimento da câmara, variação da silhueta dos jogadores, muito ruído e desfocagem de movimento do vídeo.

O múltiplo rastreamento é uma tarefa muito importante na análise de vídeo. Rastreamento pode ser visto como um problema de associação de dados. O objetivo da associação de dados é recuperar a correspondência entre observações em diferentes *frames.*

São muitos os estudos feitos em análise de vídeo de desporto (Liu et al., 2009). Muitos investigadores procuraram uma solução também para o problema específico da etiquetagem e rastreamento de jogadores em vídeos de transmissão televisiva (Sato & Aggarwal, 2005). Wang, Zeng, Lin, Xu, & Shum, (2004) tentaram extrair modelos de cores do terreno de jogo e dos equipamentos das equipas para a análise semântica. Uma outra tentativa de Sullivan & Carlsson, (2006) um método de trajetória baseada em *clustering* foi proposto para resolver o rastreamento de jogadores em vídeo. Nestes trabalhos a etiquetagem dos jogadores foi conseguida por classificação supervisionada. Por fim Nillius, Sullivan, & Carlsson, (2006) construíram um gráfico de rastreamento e pegaram no problema da localização como inferência numa rede *bayesiana*. Nestes últimos dois trabalhos foram utilizados sistemas de multi câmaras para obter uma visão larga do terreno do jogo em alta definição e estacionária.

Deteção dos jogadores

No trabalho realizado por Liu et al. (2009) a deteção dos jogadores é alcançada através de uma cascata de Haar features em visões globais, que já tinha sido usada previamente por Viola & Jones, (2001). Foram rotulados manualmente seis mil jogadores como amostras positivas. Estas amostras foram então cuidadosamente selecionadas a fim de capturar parcialmente as variações da aparência dos jogadores devido aos movimentos do corpo humano. As amostras negativas são manchas aleatoriamente recortadas de imagens de vídeos para além das retiradas a partir de imagens naturais.

Este esquema melhora a precisão de deteção. Alguns exemplos são visíveis na figura 5. Todas as amostras são adequadamente redimensionadas para uma resolução de 32x64 pixéis. Um detetor cascata é então treinado com este conjunto de amostras. Para este trabalho foi usado de 340 000 Haar features para treinar.

Figura 5 Algumas amostras positivas e negativas utilizadas para treino do sistema. (fonte: (Liu et al., 2009)

Na fase de deteção, a segmentação *playfield* é usada pela primeira vez para filtrar as regiões do fundo. Este processo tem as seguintes vantagens: (1) restringe ainda mais o processamento dentro das regiões candidatas, o que acelera a velocidade de deteção e (2) reduz a possibilidade de falsos alarmes. O detetor é digitalizado através das regiões das imagens filtradas a várias escalas. Várias deteções ocorrem normalmente à volta de cada jogador após a digitalização da imagem. Foram juntados retângulos detetados adjacentes e removidos possíveis respostas falsas através de *clustering* para obter deteções finais com escalas e posições adequadas.

Podem ocorrer deteções falsas devido à desfocagem do vídeo e ajuntamento de jogadores. Estes resultados de deteção imperfeitos podem ser melhorados mais tarde pelo processo de rastreamento. O procedimento de deteção está ilustrado na figura 6.

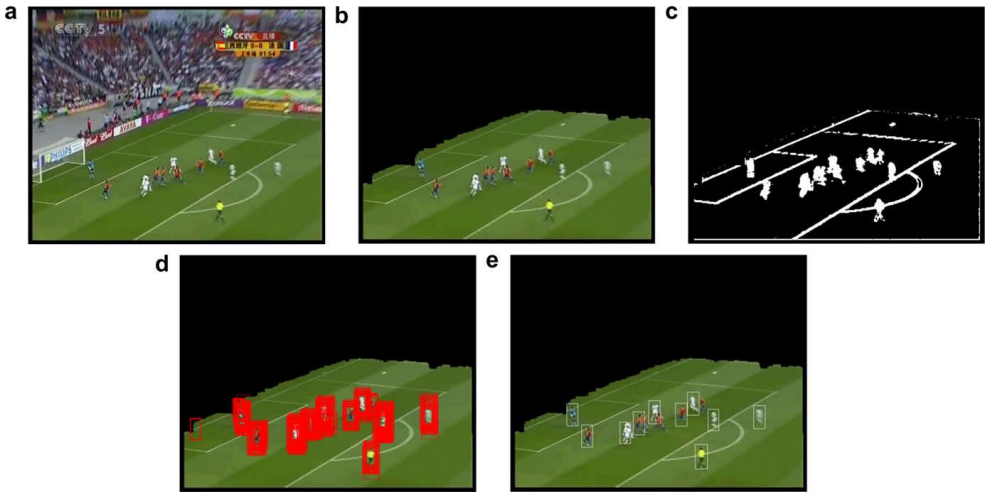


Figura 6 Processo de deteção dos jogadores: (a) imagem original; (b) redução ao jogo; (c) player mask; (d) resposta inicial (retangulos vermelhos); (e) resultado final através de pós-processamento (jogadores e árbitro são delimitados por retangulos brancos). (Fonte: Liu et al., 2009).

Etiquetagem dos jogadores

A finalidade da etiquetagem dos jogadores é distinguir os jogadores da equipa A, equipa B e árbitro. Para conhecer o modelo de aparência do jogador, (Liu et al., 2009) foi executado o detetor de jogadores em todos os 50 *frames* para recolher amostras de aprendizagem do vídeo. Aproximadamente 500 *frames* são processadas e 1500 amostras de jogadores são extraídas.

Foi utilizado um *bag of features* para representar os jogadores. Este processo é ilustrado na figura 7. Para cada amostra de jogador, em primeiro lugar é retirado o fundo usando o modelo de cor dominante e considera-se somente a região do tronco do atleta para a aprendizagem. Os componentes adjacentes com uma distância mais curta que um certo ponto são então unidos. Os componentes unidos são designados *meta-prototypes*.

Juntando todos os pixéis do tronco para o correspondente meta-prototype, cada jogador é então representado por um histograma, de acordo com a figura 7.

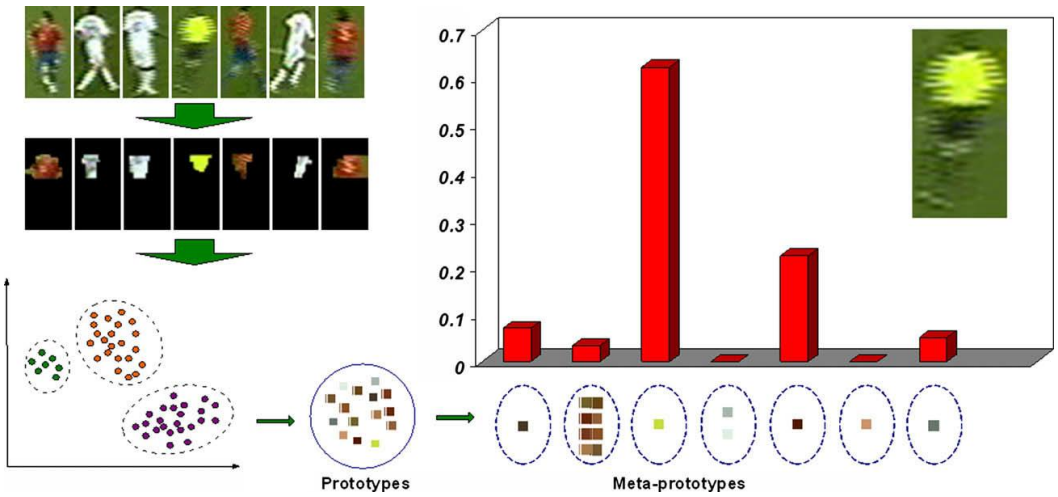


Figura 7 Representação dos jogadores com a utilização de um bag of features. (Fonte: (Liu et al., 2009)

ASPOGAMO

ASPOGAMO foi apresentado em 2006 no RoboCup 2006. Neste evento foram recolhidos dados a partir da cobertura ao vivo dos jogos do Campeonato Mundial de 2006 na Alemanha.

O modo de funcionamento deste sistema de rastreamento visual que determina as coordenadas e trajetórias dos jogadores de futebol a partir das gravações das transmissões televisivas pode ser dividido em duas diferentes distintas: (1) estimativa de parâmetros da câmara e (2) deteção de jogadores (Beetz et al., 2007).

Estimativa de parâmetros da câmara: O sistema é capaz de controlar a câmara, sem se perder em quase todos os cenários testados no Campeonato Mundial de 2006. Usando transmissões televisivas originais sem interrupções da cena, o sistema é capaz de controlar os parâmetros da câmara até 20 minutos sem se perder.

Deteção dos jogadores: A próxima tabela (Tabela I) apresenta as taxas de deteção dos jogadores de alguns jogos do Mundial da Alemanha que foram determinados manualmente pela análise aleatória de 344 frames. Tanto o jogo inicial (Alemanha [ALE] – Costa Rica [CR]) como o jogo entre a Argentina (ARG) – Sérvia e Montenegro (SM) tiveram taxas de deteção acima dos 90% mesmo com os dados de entrada com muito barulho. Por outro lado o jogo entre Portugal (POR) contra o Irão (IRA) a tarefa foi complicada por uma enorme sombra projetada sobre o terreno de jogo dificultando a deteção dos jogadores de branco mesmo para o olho humano. Por último, no trabalho conduzido por Beetz, (2007) foi analisado um pontapé de canto e as deteções que falharam foram causadas maioritariamente por oclusões e agrupamento de jogadores. Os testes procedidos noutros jogos resultaram sempre em taxas de deteção superiores a 90%.

Tabela I Taxa de deteção de jogadores usando ASPOGAMO. (Adaptado de Beetz et al., 2007)

Jogo	Contador de jogadores	Jogadores que não foram detetados	Más classificações	Falsos Positivos
ALE – CR	2353	4.78%	0.38%	2.37%
ARG - SM	1605	5.51%	1.88%	1.23%
POR – IRA	592	21.96%	3.72%	0.34%
Pontapé de Canto	462	17.53%	3.68%	6.06%

As taxas de deteção apresentadas são baseadas em observações que não exploram a informação temporal dada no rastreador utilizado pela equipa de Beetz., (2007). A maioria dos erros fruto das más deteções temporais, falsos positivos ou erros de classificação foram resolvidos quando o rastreador foi incorporado.

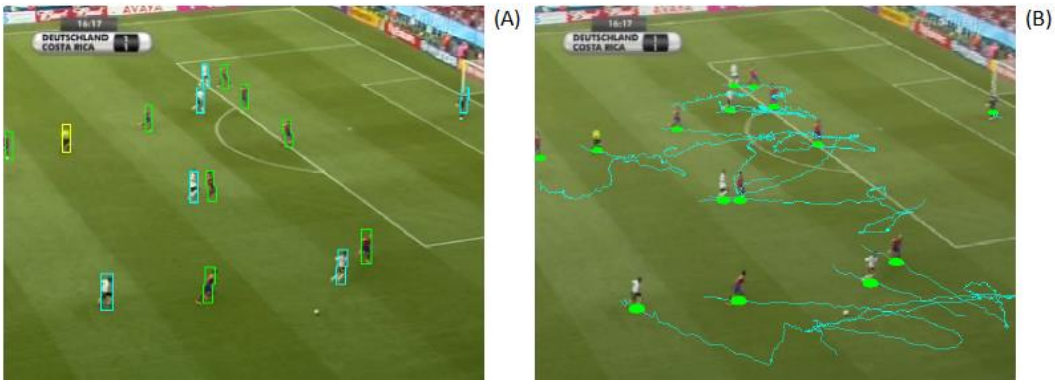


Figura 8 (A) Jogadores a serem detetados pelos sistema; (B) percurso dos atletas a ser rastreado durante esta cena. (Adaptado de (Beetz et al., 2007)).

A figura 8 mostra (a) os jogadores a serem detetados pelo sistema, e (b) o percurso destes nesta pequena filmagem. Estes percursos são definidos com sucesso enquanto o jogador está visível na câmara, normalmente até a filmagem ser interrompida. A exceção acorre quando há jogadores à frente dos painéis de publicidade ou em cenários de ajuntamento de jogadores, como por exemplo, em cantos. ASPOGAMO é capaz de fundir observações de múltiplas câmaras que estejam a gravar a mesma ação de ângulos diferentes.

O sistema ASPOGAMO deteta os jogadores em jogo, segmentando as manchas em campo que não são o terreno verde. Estas manchas são analisadas por meio de restrições de tamanho e modelos de cores. Os jogadores são localizados usando o seu centro de gravidade estimada. Um dos algoritmos utilizados para a identificação dos jogadores em terreno de jogo é um modelo que aproveita as cores distintas dos equipamentos dos jogadores e do árbitro. No entanto, como é facilmente percetível pela figura 9, utilizar segmentação baseada somente na cor está destinada a falhar.

Figura 9 Jogadores aumentados. Os pixéis são desbotados devido à pequena resolução e desfocagem fruto do movimento dos jogadores e da câmara. Os jogadores em frente à publicidade são ainda mais difíceis de detetar (imagem da direita). (fonte: (Beetz et al., 2007)).

Um outro problema que o sistema enfrenta é a qualidade das imagens transmitidas pela televisão, que normalmente apresentam muito ruído, os movimentos rápidos da câmara criam desfocagens e por último, os próprios jogadores ficam escondidos por outros em situações de cantos e faltas. Como resultado disto, formas e cores são misturadas com os seus vizinhos.

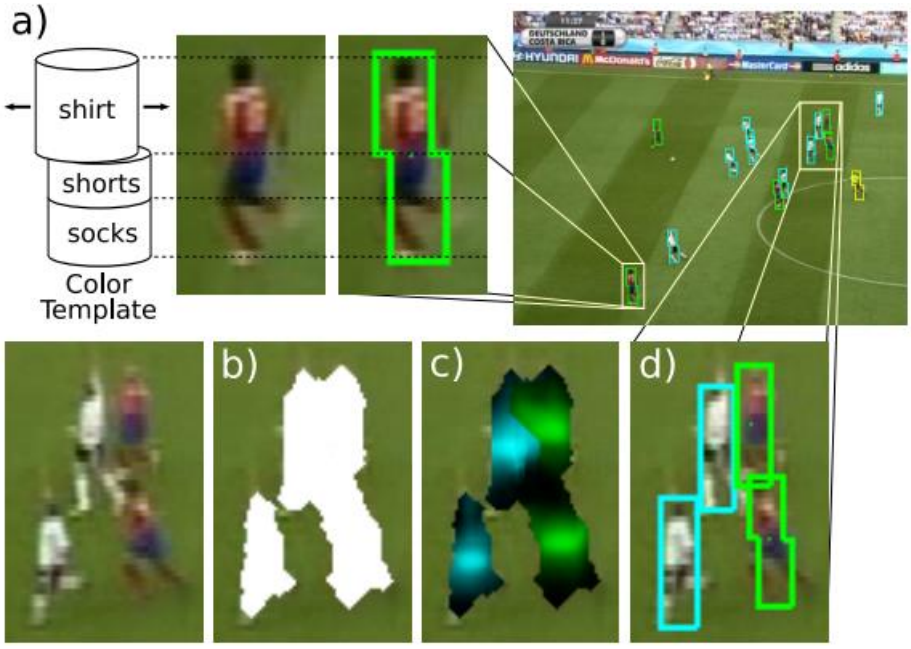


Figura 10 Reconhecimento dos jogadores. (fonte: (Beetz et al., 2007))

De acordo com Beetz., (2007) para resolver este problema do reconhecimento dos jogadores em campo, foram combinados sinais percetuais simples mas poderosos de uma forma probabilística, a fim de alcançar a robustez, velocidade e precisão.

Numa primeira fase todos as potenciais regiões dos jogadores improváveis de pertencer ao terreno de jogo ou às redondezas, são segmentadas (Figura 10B). De seguida é estimado os centros de massa dos jogadores, mais precisamente o centro entre a camisola e os calções pois este é o ponto mais fácil de detetar. Este passo é conseguido através do cálculo de probabilidades do paradeiro do jogador usando várias evidências com base nos modelos de cor, restrições de tamanho e previsões (Figura 10C). As posições dos jogadores são extraídas do mapa pelo centro de gravidade e as suas coordenadas são projetadas para um ponto a meio da altura de jogador sobre o terreno de jogo (Figura 10D).

Uma abordagem alternativa é o uso de várias câmaras fixas ao longo do terreno do jogo. Este método aumenta o campo de visão global, minimiza os efeitos da oclusão dinâmica, fornece estimativas em 3D da localização da bola e melhora a precisão e robustez da estimativa devido à fusão da informação (Xu, Orwell, & Jones, 2004).

O sistema de Xu et al., (2004) utiliza um grupo de oito câmaras fixas dispersas pelas instalações desportivas de acordo com a figura 11.

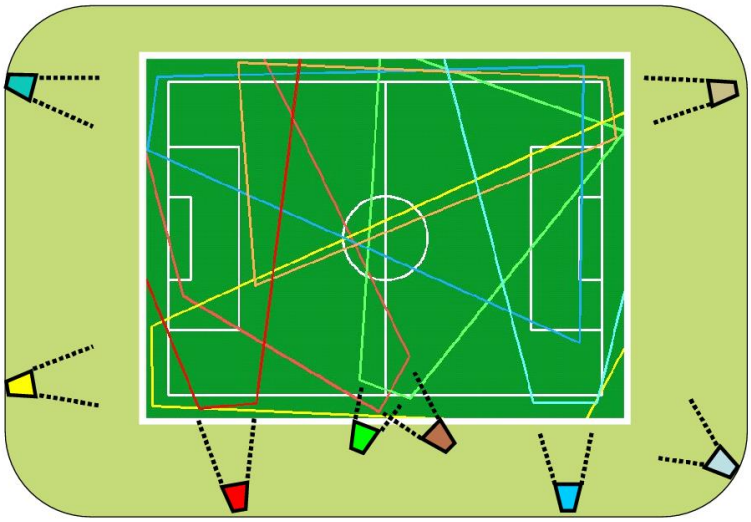


Figura 11 Posicionamento das câmaras e respetivos campos de visão do sistema usado por (Xu et al., 2004)

Para o processo de rastreamento multiview são necessárias três etapas. O primeiro passo é associar medições a trilhos definidos e atualizar estes. O segundo passo é inicializar faixas para as medições

incomparáveis a qualquer trilho. Por último a restrição da população fixa para cada categoria de jogadores (jogadores de campo, guarda-redes e árbitros) é utilizado para reconhecer os membros de cada categoria.

Cada jogador é modelado como um trilho e tem o seu estado atualizado, se possível, por pelo menos uma câmara. Depois de verificar as medições dos trilhos existentes, podem existir medições sem relação. Essas medidas são verificadas de novo, cada uma de câmaras diferentes, para encontrar novos trilhos. Assim, um novo trilho é definido. A medida da probabilidade do jogador é calculada para cada alvo com base na confiança de categoria estimativa, o número de câmaras de apoio, conhecimento de domínio em posições (para guarda-redes e AA), frames de rastreamento bem como a restrição de população fixa.

3.2.2. Bola

A forma de detetar a bola no terreno de jogo é conseguida da mesma forma que são detetados os atletas, no entanto, a alta complexidade associada obriga pesquisadores a desenvolver metodologias e técnicas específicas, uma vez que esta é muito pequena na imagem e de cor branca, tornando difícil de a diferenciar (Seo, Choi, Kim, & Hong, 2005).

O primeiro problema que surge para assegurar o posicionamento da bola são as câmaras utilizadas para transmissões televisivas com baixa taxa de frames e velocidade do obturador muito rápida, conduzindo a imagens desfocadas quando a bola se movimenta a grandes velocidades. Consequentemente é necessário o uso de um sistema de câmaras dedicadas colocadas em locais estratégicos e com algumas características, tais como a taxa de *frames* e a resolução (C. Santiago, 2011).

A bola movimenta-se variadas vezes pelo ar, o que obriga a um rastreamento 3D para um posicionamento mais preciso. São necessárias múltiplas câmaras para estimar e localizar a bola em 3D. Esta tarefa é dividida em duas iterações. Primeiro a bola é detetada e monitorizada numa vista única independente. Em segundo lugar as posições em 2D da bola de várias câmaras são integradas de modo a obter posições 3D.

Todos os objetos em movimento são detetados pelas 8 câmaras dispersas pelo estádio de futebol de acordo com Orwell & Jones, (2004). Todos os candidatos a bola são filtrados por tamanho, forma e

cor. De seguida as posições 3D destas bolas são estimadas e controladas nas coordenadas bases do sistema. Este sistema propõe uma *framework* para a estimativa e localização da bola em 3D, usando múltiplas sequências de imagens e técnicas de reconstrução geométrica.

A Cairos Tecnologies em conjunto com a Adidas desenvolveu um sistema intrusivo para a bola, que é inserido nesta de acordo com a figura 12. Este sistema serve para resolver situações de golo/não golo no qual existem dúvidas se a bola entrou por completo na baliza. Este protótipo foi já utilizado no Mundial de 2014 no Brasil e foi mesmo posto em prática nos jogos da França vs Honduras e Costa Rica vs Itália, nos quais não seria possível à equipa de arbitragem validar os golos dada a difícil análise dos lances.

Figura 12 Bola Inteligente Adidas Teamgeist II

3.2.3. Comparação dos sistemas de deteção

Após a análise efetuada a alguns sistemas que são atualmente utilizados para o rastreamento de jogadores, equipa de arbitragem e bola é possível verificar que todos têm características distintas e consequentemente nem todos podem ser aplicados em todas condições. Os sistemas intrusivos nos atletas por exemplo, embora mais eficazes, não estão viabilizados de serem utilizados em competições oficiais por proibição da IFAB que não autoriza o uso de qualquer objeto para além do equipamento base. Esta mesma organização permitiu no entanto, o uso de sistemas intrusivos nas bolas no Mundial de 2014. A impossibilidade do uso destes sistemas originou o surgimento dos sistemas não intrusivos.

A seguinte tabela (Tabela II) pretende resumir e comparar os variados sistemas utilizados para a deteção da bola, jogadores e árbitros.

Tabela II Resumo dos sistemas intrusivos e não intrusivos para a deteção e rastreamento da bola, atletas e árbitros

	Sistema	Método	Objetivo	Técnicas	Área de cobertura	Tempo de resposta	Precisão
Intrusivo	GPS	Sensor e Antena	Detetar jogadores		Muito alta	Baixo	Baixa
	RFID tags passivas		Detetar jogadores		Baixa	Baixo	Baixa
	RFID tags ativas		Detetar jogadores		Média	Baixo	Média
	Wi-Fi		Detetar jogadores		Alta	Baixo	Alta
	Bluetooth		Detetar jogadores		Baixa	Baixo	Baixa
	Cairos Tecnologies		Golo / não golo	Deteção magnética	Baixa	Baixo	Alta
Não intrusivo	(Liu et al., 2009)	Televisão	Detetar atletas e bola	Subtração do terreno de jogo; Haar features;	Média	Baixa	Alta

	Sistema	Método	Objetivo	Técnicas	Área de cobertura	Tempo de resposta	Precisão
	(Beetz et al., 2007)	Câmaras dedicadas	Detetar atletas e bola	Segmentação da cor; Multiple Hypothesis Tracker (MHT); Subtração do terreno de jogo;	Média	Baixa	Alta (>90%)
	(Xu et al., 2004)		Detetar atletas e bola	Segmentação da cor; Algoritmo Tsai;	Alta	Baixa	Alta
	(Orwell & Jones, 2004)		Detetar a bola em 3D	Segmentação da cor, forma e tamanho;	Alta	Baixa	Alta

Analisando a Tabela II é possível dividir todos os sistemas de rastreamento existentes em dois grupos: (1) Os sistemas intrusivos e (2) os sistemas não intrusivos.

Os primeiros caracterizam-se pela utilização de *tags* e antenas para obter os dados dos atletas em terreno de jogo. Estes sistemas têm o inconveniente de não poderem ser utilizados em jogos oficiais por imposição da IFAB e da FIFA. O sistema Cairos Tecnologies desenvolvido em parceria com a Adidas é um sistema intrusivo que difere dos restantes pois serve apenas para a avaliação de situações em que a bola entra totalmente na baliza. Apesar de também requerer uma *tag* e uma antena para o uso com sucesso, esta *tag* está dentro da bola e tem a particularidade de já ter sido utilizado numa competição oficial, nomeadamente, no Mundial de 2014 no Brasil.

Os sistemas não intrusivos são os mais utilizados e existem diversos produtos e sistemas no mercado sobre análise de performance desportiva. Estes podem ainda ser divididos em dois grupos mais pequenos. Os que utilizam a imagens televisivas para a deteção dos atletas e da bola, e os que utilizam câmaras dedicadas para o efeito dispersas pelo estádio de futebol. Uma técnica em comum com estes sistemas é o reconhecimento dos jogadores através da segmentação da cor dos equipamentos. A bola para além da segmentação da cor, como esta é de pequenas dimensões e branca, é também localizada através da sua forma e tamanho.

3.3. Software de análise de Futebol

Atualmente são muitos os sistemas de análise de performance desportiva capazes de apresentar variadas funcionalidades de deteção de jogadores (Needham, 2003). Os softwares mais utilizados são Prozone, Ascensio Match Expert, Match Vision Studio, Mambo Studio, Sportcode Elite, Performasports, Quintic e Nacsport Elite. Este capítulo centra-se nas características e resumo de alguns destes pacotes de software.

3.3.1. Prozone

A *Prozone* é um sistema que utiliza um conjunto de câmaras espalhadas pelo terreno de jogo e captura os eventos durante a partida, produzindo informação sobre o desempenho de cada jogador e da equipa. Este sistema produz estatísticas e informações durante a partida de futebol, o que é vantajoso para a equipa técnica (Alves, 2011).

Em 2011 a *Amisco* e a *Prozone* juntam forças sobre o nome de *Prozone* com a missão de oferecer perceções com verdadeiro impacto na área (Prozone, 2014). Amisco criou a tecnologia e a Prozone foi responsável por ter fundado a indústria.

O sistema é composto por 6 a 8 câmaras dispostas ao longo do terreno de jogo. Este é capaz de capturar as imagens relativas a um jogo de futebol.

As imagens são processadas manualmente identificando muitas situações do jogo tais como o posicionamento dos jogadores e da bola permitindo uma representação gráfica assim como a identificação dos movimentos dos jogadores. A grande desvantagem deste sistema é necessitar de

operadores a tempo inteiro que consigam identificar manualmente alguns eventos do jogo como faltas e foras de jogo (Abreu, 2010).

A análise do jogo consiste em 3 partes: Modo de Animação, Modo Tático e Modo Físico. Esta animação é feita em 2D e mostra os movimentos dos jogadores durante toda a partida e algumas estatísticas indivíduas. O modo tático permite analisar eventos do jogo, através da representação de áreas, e o desempenho individual (Alves, 2011).

Hoje trabalham com cerca de 250 clientes a nível mundial, entre os quais se destacam clubes como o Manchester United, Bayern Munique e Paris Saint German (Prozone, 2014).

3.3.2. Ascensio Match Expert

Este software é usado para um nível avançado e profissional de análise de performance de jogos de futebol. O software permite o visionamento em 2D e 3D, ver a velocidade e aceleração dos jogadores, a distância entre jogadores, esquema detalhado de remates e informação de eventos que fazem parte do jogo. Estatísticas completas de jogadores tais como a posição deste, alterações das táticas das equipas durante o jogo, as trajetórias dos jogadores durante um período de tempo previamente definido e um esquema visual de densidade nas várias zonas do terreno de jogo das ações dos jogadores.

A informação é representada em tabelas interativas e gráficos estatísticos (PCWorld, 2010).

3.3.3. Mambo Studio

Mambo Studio é um software recente, lançado em 2012 pela empresa Match Analysis, que conta com mais produtos para a mesma área. O mambo studio caracteriza-se por vir revolucionar a edição de vídeo de um jogo de futebol e vangloriasse de reduzir o trabalho que anteriormente durava dias para alguns segundos (Mambo Studio, 2015). O software é capaz de responder a querys avançadas num espaço de tempo muito reduzido (1-2 segundos). Como por exemplo, se algum responsável de um clube de futebol, quiser saber quantas vezes um seu jogador tocou na bola, em apenas 2 segundos tem a resposta e os diversos momentos em que tal evento aconteceu em vídeo. Isto é, num jogo do Barcelona a pergunta: Quantas vezes o Iniesta tocou na bola? O resultado seria 17 vezes, e

seriam listados os 17 momentos, em vídeo, em que o Iniesta toca na bola. Outras perguntas mais complexas tais como: "Quero saber quantas vezes o meu trinco passou a bola a um avançado".

3.4. Análise do Desempenho de Árbitros

O estudo sobre o posicionamento da equipa de arbitragem na qualidade das decisões tomadas é algo pouco aprofundado até ao momento. Existem no entanto vários documentos que analisam o trabalho dos árbitros de futebol quanto à sua condição física, essencial para estar constantemente bem posicionado, a sua resistência às pressões exteriores e a distância destes ao lance que necessita de ser avaliado.

Um dos estudos mais relevantes foi o de Oliveira et al., (2011) que consistiu na avaliação da relação entre a distância do árbitro para o lance onde ocorreu a falta e a consequente decisão tomada. O estudo não encontrou nenhuma associação direta entre a distância e o acerto da decisão, tendo por outro lado detetado um aumento significativo na qualidade da decisão nos últimos 15 minutos do jogo. Não foi encontrada, contudo, uma justificação sólida para tal acontecer.

Outro dos estudos de maior destaque foi elaborado por Javier Mallo et al., (2012) no qual seu objetivo consistia no efeito do posicionamento na qualidade da decisão de árbitros de topo em jogos internacionais. Os jogos em análise são da competição da FIFA, a Taça das Confederações de 2009 no qual 380 faltas e 165 situações de fora de jogo foram revistas. O estudo revelou que o erro médio das equipas de arbitragem encontra-se nos 14% e a percentagem mais reduzida ocorreu no meio campo, local onde a colaboração com os AA é limitada, o jogo mais lento, e o árbitro se encontrava entre 11 a 15 metros do lance. Esta percentagem aumentou para 23% no último quarto de hora do jogo. Por outro lado a percentagem de erro dos AA relativamente ao fora de jogo foi de 13% e foi concluído que a distância do AA para a linha do fora de jogo não tem impacto na qualidade da decisão mas esta (percentagem de erro) é consideravelmente reduzida quando o lance é analisado de ângulo compreendido em 46° e 60°.

O estudo efetuado por Mascarenhas, Dicks, O'Hare, & Button, (2009) refere que usando uma combinação inovadora de vídeo e GPS foi explorado o movimento, os batimentos cardíacos e a tomada de decisão de árbitros de futebol. Este estudo foi feito em jogos do campeonato da Nova Zelândia da

época 2005/06. A percentagem de sucesso no que toca às decisões assumidas foi de 64%, no entanto, não foi identificado uma correlação com as variáveis suprarreferidas.

Outros estudos semelhantes, tais como Elsworthya, Burkeb, & Ben J. Dascombea, (2014), acabam por concluir o mesmo, que a distância do árbitro estava quase sempre compreendida entre 11 e 15 metros e não teve grande impacto na qualidade da decisão.

Os AA no que toca ao posicionamento para o fora do jogo merecem muito mais destaque por parte da comunidade científica. São inúmeros os estudos que indicam que o posicionamento, desde que esteja entre os 0.81 metros para trás e 0.77 metros para a frente, não é a principal causa de erros. A principal causa deve-se à hipótese de flash-lag que pode ser definida como uma ilusão visual quando um flash e um objeto em movimento que aparecem na mesma posição são interpretados como estando deslocados um do outro. Os estudos que suportam esta teoria são Helsen, Gilis, & Weston, (2006), Oudejans et al., (2005) e Catteeuw et al., (2010).

3.5. Conclusões

A revisão de literatura efetuada para este trabalho de pesquisa apresentada neste capítulo (o posicionamento da equipa da arbitragem e a sua influência na decisão tomada e o rastreamento dos intervenientes do jogo) permitiu concluir que ainda são poucos os estudos efetuados que analisem detalhadamente esta correlação entre o posicionamento (no que toca ao ângulo com que o árbitro vê a jogada e não tanto com a distância com que esta é analisada). Foi encontrado um número maior de documentos que abordam a mesma questão mas direcionada para os AA e qual a relação do ângulo com que o assistente aborda o lance (entre 46° e 60°) e o posicionamento deste comparativamente com a linha do fora do jogo.

No que toca ao rastreamento e localização dos intervenientes do jogo (jogadores, árbitros e bola) existem dois grandes grupos que englobam toda a tecnologia desenvolvida até ao momento – sistemas intrusivos e sistemas não intrusivos. Os primeiros são essencialmente compostos por tags (emissores) e antenas (recetores). Estas tecnologias embora apresentem resultados muito bons têm a grande desvantagem de não poderem ser utilizados em jogos oficiais das competições FIFA por desrespeito às leis do jogo estipuladas. Os sistemas não intrusivos são efetuados recorrendo às transmissões televisivas ou câmaras dedicadas dispersas pelo complexo desportivo. Estes algoritmos

consistem na segmentação por cor essencialmente (todos os jogadores de uma equipa vestem de uma cor, iguais entre si e distintos da equipa adversária, do próprio guarda-redes e da equipa de arbitragem) e a remoção de todo o espaço envolvente da imagem, tal como o terreno de jogo, adeptos e treinadores. A bola merece um tratamento um pouco diferente pois o seu tamanho reduzido e cor branca dificulta a sua localização mesmo para o olho humano. Assim, esta é localizada recorrendo a uma segmentação combinada de cor, tamanho e forma. O facto desta se deslocar maioritariamente das vezes acima do solo, é importante um rastreamento em 3D.

Para finalizar, são inúmeros os pacotes de *software* disponíveis no mercado que oferecem aos clubes uma possibilidade de analisar a performance desportiva das equipas, nos quais se destaca a Prozone.

4. Data Mining

"The goal is to turn data into information, and information into insight." – Carly Fiorina, former executive, president, and chair of Hewlett-Packard Co. In (December 6, 2004)

Este capítulo tem como objetivo fazer uma breve introdução ao conceito de *Data Mining* pois será bastante utilizado ao longo do desenvolvimento desta dissertação. Para além dos fundamentos gerais desta área da inteligência artificial. Será também focado com maior pormenor para o desenvolvimento do *Data Mining* no desporto e na previsão da localização para servir como auxílio para o desenvolvimento descrito nos capítulos seguintes.

4.1. Fundamentos Gerais

Segundo Aggarwal (2015), o *data mining* é o estudo de colecionar, limpar, processar, analisar e obter conhecimento de dados. Existe uma grande variedade de problemas, aplicações, formulações e representações de dados que são encontrados em aplicações reais. Por isso, o conceito "*data mining*" é um termo guarda-chuva que é usado para descrever todos estes aspetos distintos do processamento de dados.

A imensidão de dados é o resultado direto de vantagens na tecnologia e da computorização de todos os aspetos da vida moderna. É, portanto, natural analisar se é possível extrair ideias concisas e possivelmente conhecimentos dos dados disponíveis para objetivos específicos de aplicações. É nestas situações que o *data mining* é necessário. Os dados originais podem ser arbitrários, não estruturados ou até estarem num formato que não é apropriado para o processamento automático. Por exemplo, a coleção manual dos dados pode ser originária de várias fontes e em diferentes formatos. Para resolver este problema é necessário desenvolver um processo em que os dados são colecionados, limpos e transformados num formato único. Este processo origina que a grande maioria do trabalho esteja relacionado com a preparação dos dados.

O processo de *data mining* pode ser descrito pela figura 13 como explica Aggarwal (2015). Este processo vai de encontro às várias fases da metodologia CRISP-DM aprofundada no próximo capítulo.

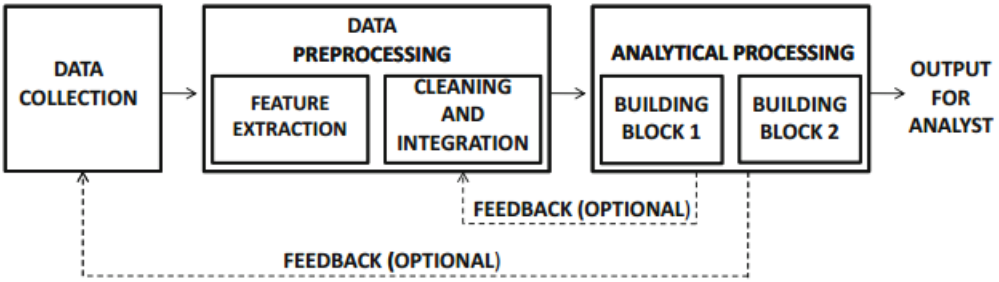


Figura 13 O processo de *data mining*. (fonte: Aggarwal, 2015)

1. **Coleção dos dados**: esta fase é muito específica da aplicação utilizada e extremamente importante pois escolhas corretas nesta etapa pode ter um impacto significativo no processo de *data mining*. Após a coleção, os dados são geralmente guardados em bases de dados ou *data warehouses*.
2. **Extração de características e limpeza dos dados**: os dados quando recolhidos estão frequentemente num formato pouco próprio para processamento. Por exemplo os dados podem estar codificados em *logs* complexos ou em documentos de escrita livre (uma caixa de texto por exemplo em que vários tipos de dados podem estar arbitrariamente misturados). Para tornar os dados prontos para processamento é essencial transformar estes num formato amigável para os algoritmos de *data mining*. É crucial extrair características relevantes para o processo de *data mining*. A fase de extração de características é habitualmente feita em paralelo com a limpeza de dados onde os valores omissos e errados são ou corrigidos ou eliminados.
3. **Processo analítico e algoritmos**: a parte final do processo de *data mining* é a conceção de métodos analíticos eficazes a partir dos dados processados. Em muitos casos não é possível usar diretamente um problema regular, tais como associação de padrões, clustering, classificação ou deteção de outliers. No entanto estes problemas têm uma cobertura tão ampla que em muitos casos pode-se dividir os componentes de forma a utilizar estes para a busca do conhecimento.

Os problemas mais típicos de *data mining* podem ser divididos em duas categorias: previsão e descrição. A previsão é caracterizada por problemas com objetivos específicos, tendo como base conjuntos de dados do passado que servirão de base para o que se pretende prever. A descrição, por outro lado, tem o propósito de detetar informação numa base de dados complexa para aumentar o conhecimento a ser extraído (Almeida, 2009).

A regressão e a classificação são os dois problemas típicos da previsão. A regressão tem como objetivo encontrar uma função que relacione uma variável dependente com uma ou mais variáveis independentes. O resultado da regressão pode ser, por exemplo, a posição ideal de um jogador ou de um árbitro num determinado momento do jogo de futebol (coordenada x e y) tal como esta dissertação se propõe. A classificação tem como objetivo encontrar uma função que associe casos de um determinado domínio a classes pré-determinadas. As técnicas utilizadas na classificação requerem o uso de conjuntos de treino com casos etiquetados, tendo como finalidade a construção

de modelos adequados à descrição das classes. Estes modelos são posteriormente aplicados aos novos casos, não etiquetados, com o objetivo de determinar a classe mais adequada. A resposta para este género de problemas provém de amostras do passado que são analisadas e generalizadas para futuros casos. Quando se quer saber se a decisão de um árbitro foi correta ou não comparativamente com a sua posição em terreno de jogo, depara-se com um problema típico de classificação (Almeida, 2009).

Os principais tipos de padrões de dados que são descobertos a partir dos conjuntos de dados através de algoritmos de *data mining* são apresentados nesta secção do documento.

4.1.1. Classificação e previsão de padrões

Classificação e padrões de previsão capturam relações entre variáveis de atributos, $x_1,...x_n$, e variáveis alvo, $y_1,...,y_n$, que são fornecidas por um determinado conjunto de dados, $(x_1,...,x_n,y_1,...,y_n)$. Classificação e previsões de padrões permitem classificar ou prever valores de variáveis a partir dos valores das variáveis de atributos.

> **IF**(Atributo$_1$ = Portugal) **OR** (Atributo$_2$ = Verde **AND** Atributo$_3$ = Branco) **THEN** Target = Verdadeiro; **ELSE** Target = Falso

Esta relação permite classificar o valor de uma variável a partir do valor dos seus atributos.

Padrões de classificação e previsão, que capturam a relação entre atributos, x_1, ..., x_n, com variáveis alvo, y_1, ..., y_n, podem ser representadas na forma geral de **y = F(x)**. Os padrões de classificação para F tanto podem assumir a forma de regras de decisão como de modelo linear.

Geralmente o termo "padrões de classificação" é utilizado, se a variável alvo é uma variável categórica, e o termo "padrões de previsão" é utilizado se a variável alvo é uma variável numérica.

4.1.2. Cluster e associação de padrões

Padrões de cluster e associação normalmente só envolvem variáveis de atributos, x_1, ..., x_n. Padrões de cluster identificam grupos de registos de dados semelhantes de maneira que os registos num grupo são semelhantes mas têm maior diferença dos dados de outros grupos. Ou seja, padrões de cluster revelam padrões de semelhanças e diferenças entre os dados. Os padrões de associação são

estabelecidos baseados em coocorrências de itens nos registos de dados. Há situações em que as variáveis alvo, y1, ..., yn, são também utilizadas no clustering mas são tratadas da mesma forma que os atributos.

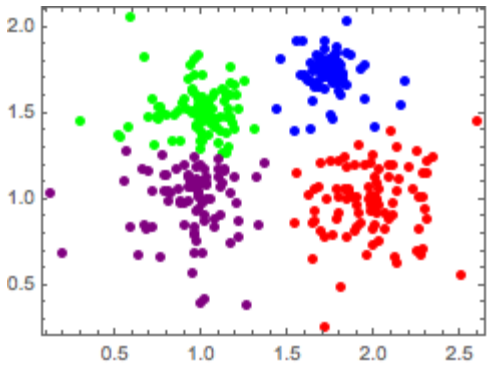


Figura 14 Exemplo de um gráfico de cluster. (fonte: Wolfram, 2014)

4.1.3. Padrões de redução de dados

Este tipo de algoritmos relacionados com padrões de redução de dados procuram um pequeno número de variáveis que podem ser utilizadas para representar um conjunto de dados com um número muito maior de variáveis. Assim uma variável indica uma dimensão de dados, os padrões de redução de dados permitem que um conjunto de dados num espaço de alta dimensão de dados para ser representado num espaço de baixa dimensão. Por exemplo, a figura 15 assinala 10 pontos num espaço de duas dimensões (*x,y*), sendo *y=2x* e *x=1,2,...,10*. Este conjunto de dados de duas dimensões pode ser representado como sendo uma dimensão com o *z* como eixo, e o *z* relacionado com as variáveis originais, *x* e *y*.

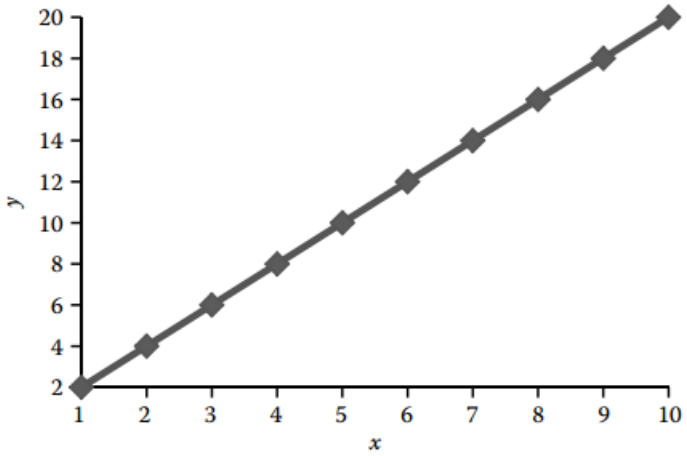


Figura 15 Redução do conjunto de dados de duas dimensões para um de uma dimensão. (fonte: Ye, 2014)

4.1.4. Outliers e padrões de anomalias

Outliers e anomalias são registos de dados que diferem muito dos restantes, ou seja, da norma. A norma pode ser definida de muitas formas. Por exemplo, a norma pode ser definida pelo intervalo de valores da maioria dos dados e qualquer valor fora desse intervalo é considerado um outlier (Ye, 2014).

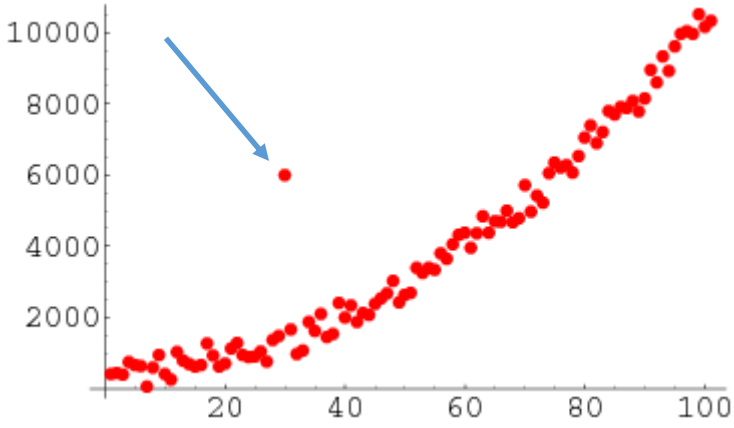


Figura 16 Exemplo de um outlier, indicado na imagem pela seta. (fonte: Weisstein, 2015)

4.1.5. Padrões sequenciais e temporais

Este tipo de padrões revelam modelos numa sequência de pontos. Se a sequência é definida no momento durante o qual são observados os dados, designa-se a sequência de pontos como uma série temporal.

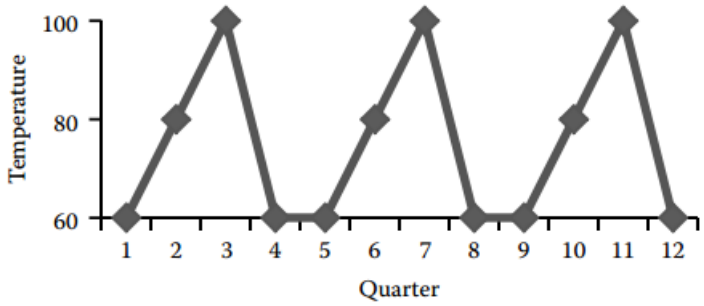


Figura 17 Exemplo de sequência temporal. (fonte: Ye, 2014)

A figura 17 mostra uma série temporal de valores de temperatura para uma cidade para cada trimestre durante 3 anos. É facilmente percetível um padrão de 60, 80, 100 e 60 que se repete todos os anos.

4.2. Data Mining no Desporto

Existe uma quantidade enorme de dados em todos os desportos. Estes dados podem ser relativos à forma individual de cada atleta, dos treinos e decisões táticas, eventos relacionados com o jogo ou o funcionamento da equipa em jogo. A tarefa não é como obter os dados mas sim que dados devem ser obtidos e como fazer uso destes. Esta abordagem de procurar conhecimento pode ser utilizada para melhorar a performance da equipa, para olheiros através de análise estatística e técnicas de projeção para identificar que talento terá maior impacto.

Data mining tem sido usada nos desportos em gerais desde a fundação de organizações internacionais que se dedicam à recolha de dados tais como a International Association on Computer Science in Sports (IACSS), fundada em 1997 para melhorar a cooperação entre investigadores internacionais interessados em aplicar técnicas de ciência de computação e tecnologias para os desafios relacionados com o desporto (International Association of Computer Science in Sport, n.d.) e a International Association for Sports Information (IASI) fundada em 1960 com o objetivo de uniformizar e arquivar as bibliotecas desportivas a nível mundial. A IASI é uma rede mundial de especialistas em desporto, bibliotecários e repositório de documentos (IASI, 2013).

Muitos tipos distintos de análises estatísticas podem ser aplicados a dados de desportos como o basebol e basquetebol e embora as técnicas e indicadores mudem de desporto para desporto o foco da questão, as estatísticas, são identificáveis independentemente da atividade, mesmo quando os seus indicadores não podem ser diretamente comparados (Schumaker, Solieman, & Chen, 2010a).

Uma contribuição importante é a capacidade de prever quando um jogador pode estar a ter quebras físicas através da previsão de lesões. O AC Milan, clube de futebol italiano, monitoriza os treinos dos seus atletas (Flinders, 2002). Este *software* compara o desempenho a performance do treino de um atleta e quaisquer indícios de sub-rendimento pode indicar que o jogador está lesionado.

Segundo o trabalho desenvolvido por Almeida (2009) utilizando os registos de log de vários jogos de futebol robótico é possível identificar qual a melhor formação tática que uma equipa deve adotar quando defronta uma outra equipa. Para isso foram utilizados vários testes a estes registos de jogos e utilizados vários algoritmos de classificação.

Em suma o *data mining* no desporto é utilizado com o intuito de criar alguma vantagem competitiva sobre o adversário que facilite a vencer o desafio.

4.3. *Data Mining* na previsão de localização

Os requisitos para fazer *data mining* em dados geoespaciais são diferentes dos realizados nas clássicas bases de dados relacionais. A razão deve-se às propriedades especiais deste tipo de dados: alta dimensionalidade, auto correlação espacial, a heterogeneidade, a complexidade, dados mal estruturados e dependência da escala (Demšar, 2006).

De acordo com Demšar (2006) os dados geoespaciais são espacialmente dependentes. O que significa que os atributos de uma localização no espaço têm tendência a estarem relacionados. Esta característica de fenómenos geográficos que dita que objetos semelhantes aglomeram-se no espaço, é tão fundamental que os geógrafos definiram como a primeira lei da geografia (Lei de Tobler): "Tudo está relacionado com tudo o resto, mas objetos próximos estão mais relacionadas do que objetos distantes".

Recentemente e fruto dos avanços tecnológicos, principalmente ao nível dos smartphones e a grande utilização por parte dos consumidores resultou num grande volume e variedade de tipo de dados

móveis (Gomes, Phua, & Krishnaswamy, 2013). Como tal a previsão da localização é uma tarefa importante para as operadoras de telemóveis e administradores de cidades inteligentes, de forma a oferecerem melhores serviços e recomendações (Keles, Ozer, Toroslu, & Karagoz, 2015). Todavia devido à natureza incerta dos dispositivos móveis e das limitações de sistemas de posicionamento, a localização de um dispositivo é desconhecida por largos períodos de tempo. Nestes casos é necessário um método para prever as possíveis próximas localizações de um objeto em movimento (Monreale, Pinelli, Trasarti, & Giannotti, 2009).

Muitos estudos limitam-se aos movimentos históricos de um objeto para adivinhar a futura posição, utilizando os dados espaciais e temporais como dados de treino. Contudo, pode também ser utilizado o movimento de todos os objetos numa determinada área para aprender uma classificação. Para isso assume-se que as pessoas frequentemente seguem a multidão – os indivíduos tendem a seguir caminhos comuns. Por exemplo, as pessoas vão para o trabalho todos os dias por caminhos parecidos. Assim, se existir dados suficientes para modelar comportamentos semelhantes, é possível, utilizando tal conhecimento, de prever os movimentos futuros (Monreale et al., 2009).

Através dos padrões de movimento extraídos pode ser utilizado o modelo matemático "Trajectory Pattern" (Giannotti, Nanni, & Pedreschi, 2006). Este algoritmo interpreta os padrões de movimento como sequências de regiões onde períodos de viagem típicos são frequentemente seguidos.

Atualmente na comunidade científica ainda não existem estudos sobre a previsão da localização de atletas dentro do terreno de jogo, no entanto, é possível retirar algumas conclusões dos trabalhos existentes na área de previsão de localização, como por exemplo: as pessoas utilizam caminhos comuns e que estas seguem a multidão. Assim é fácil de perceber que um jogador numa determinada posição da respetiva formação tática vai utilizar movimentos repetidos ao longo do jogo e como tal é possível determinar qual será o seu comportamento para uma determinada situação. O mesmo princípio poderá ser aplicado às equipas de arbitragem. Trabalho esse que irá ser desenvolvido nesta dissertação. Assim, na existência de dados suficientes para modelar comportamentos de equipas de arbitragem de topo, será possível determinar qual o posicionamento ideal da equipa de arbitragem num determinado jogo.

4.4. Conclusões

Data mining é um complexo processo dividido em várias fases. Sendo estas fases as de recolha de dados, pré-processamento e análise. A fase de pré-processamento é muito específica da aplicação porque diferentes formatos de dados requerem algoritmos diferentes a ser aplicados (Aggarwal, 2015a).

Esta tecnologia está cada vez a ser mais utilizada pois a quantidade de dados disponíveis é cada vez maior. Como tal é cada vez mais fácil obter conhecimento de tarefas do dia-a-dia ao contrário do que acontecia antigamente que apenas grandes organizações tinham tal poder.

O desporto é atualmente uma das áreas na qual o *data mining* está a ser utilizado, desde a previsão de resultados de jogos ou até mesmo se um jogador irá estar lesionado nos próximos dias ou se a compra de um jogador é justificado por parte de um clube. Para isso são utilizados dados existentes de resultados históricos e confrontos entre equipas, assim como de atletas e do seu estado clínico.

Outras áreas como a previsão de localizações estão também atualmente a ser investigadas à procura de conhecimento e para isso contribuiu o grande consumismo de dispositivos móveis como os smartphones que possuem GSP.

No entanto, ainda não existem estudos que conciliem estas duas áreas no que toca a previsão da localização de atletas em terreno de jogo para além de todo o estudo realizado à volta das formações das equipas de futebol.

5. Metodologia e Ferramentas de desenvolvimento

5.1. Metodologia de desenvolvimento

5.1.1. CRISP-DM

A metodologia Cross-Industry Standard Process for *Data Mining* (CRISP-DM) é um modelo de processos hierárquicos, compreendendo quatro níveis de abstração (em ordem descendente): fases, tarefas genéricas, tarefas específicas e instâncias de processos.

De acordo com o descrito por Wirth & Hipp, (2000) no nível superior, o processo de *data mining* é organizado num número de fases, sendo que cada uma consiste em várias tarefas genéricas (assim designado por cobrir todas as situações possíveis de *data mining*). O terceiro nível, o nível das tarefas específicas, é o lugar para descrever como as ações nas tarefas genéricas devem ser realizadas em situações específicas. Por último, o nível de instâncias de processos é um registo de ações, decisões e resultados de uma real mineração de dados. Uma instância do processo é organizada de acordo com as tarefas definidas nos níveis anteriores mas representa o que de facto aconteceu num trabalho específico, ao invés do que acontece na generalidade.

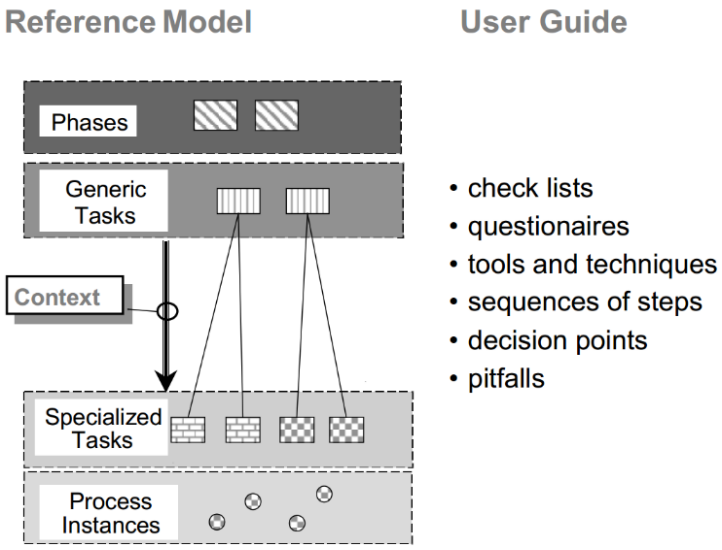


Figura 18 Os quatro níveis da metodologia de CRISP-DM para *Data Mining*. (fonte: Wirth & Hipp 2000)

A própria metodologia faz a distinção entre *Reference Model* e *User Guide*. Considerando que o primeiro apresenta uma visão geral das fases, das tarefas e dos *outputs*, descrevendo o que fazer num projeto de *data mining*, o User Guide dá dicas mais detalhadas e dicas para cada fase e cada tarefa de uma fase retratando como fazer um projeto de *data mining* (Wirth & Hipp, 2000).

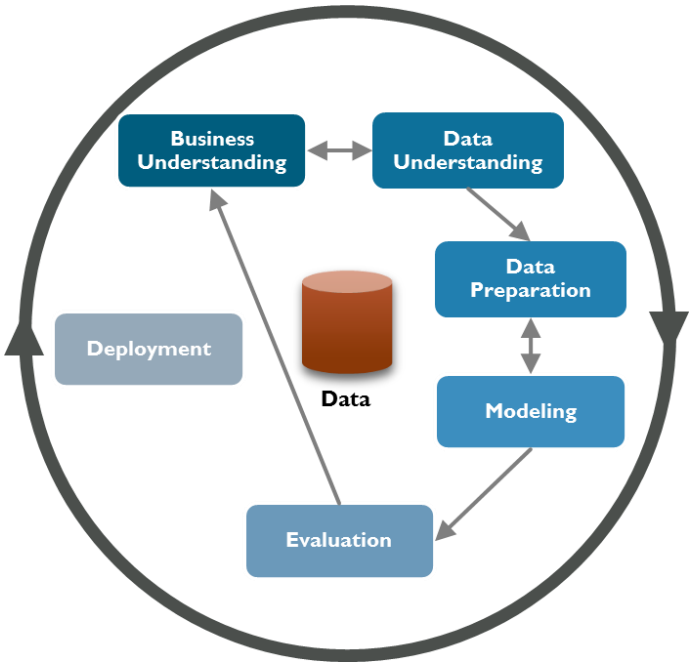


Figura 19 Fases do CRISP-DM Process Model for *Data Mining*. (fonte: Decisive Facts 2015)

A descrição das fases e tarefas como passos discretos executados numa ordem específica representa uma sequência idealizada de eventos. Na prática, muitas das tarefas podem ser executadas numa ordem diferente e, muitas vezes, ser necessário recuar para tarefas anteriores e repetir certas ações. O modelo de processo CRISP-DM não tenta capturar todos esses percursos possíveis através do processo de *data mining*, porque isso exige um modelo de processo excessivamente complexo e os benefícios esperados seriam muito reduzidos.

5.1.2. SEMMA

A metodologia SEMMA é composta por cinco fases que lhe dão o nome: Sample (Amostra), Explore (Exploração), Modify (Modificação), Model (Modelação) e Assess (Avaliação). Esta foi desenvolvido pelo SAS Institute como o processo para orientar um projeto de *data mining*.

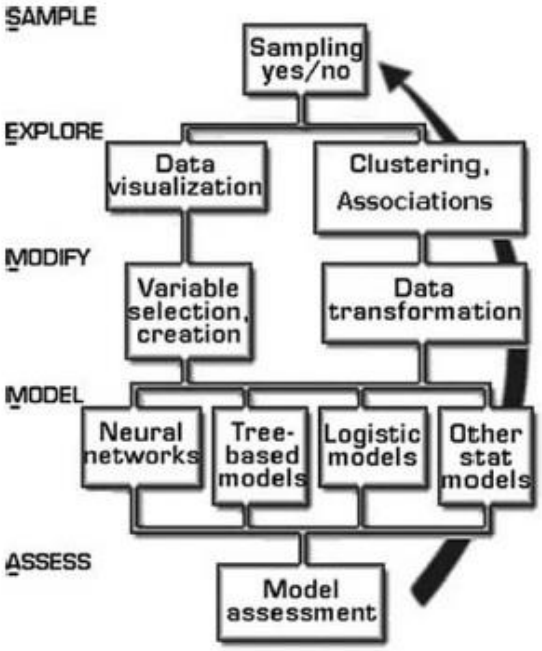


Figura 20 Fases do SEMMA. (fonte: "Metodologia SEMMA", 2010)

De acordo com o website da SAS, (2015) SEMMA não é uma metodologia de *data mining* mas sim uma organização lógica da ferramenta funcional do SAS Enterprise Miner para a realização das tarefas essenciais de mineração de dados. Enterprise Miner pode ser usado como parte de qualquer metodologia de *data mining* iterativa adotada. Obviamente há passos tais como a formulação bem definida de um negócio, ou problemas de pesquisa e implementação.

Naturalmente as fases tais como a formulação de um negócio bem definido ou problemas de pesquisa e a seleção de fontes de qualidade de dados representativos, são críticas para o sucesso geral de qualquer projeto de *data mining*. SEMMA é focado em aspetos ligados ao desenvolvimento do modelo de mineração de dados.

SEMMA, apesar de estar ligado ao SAS Enterprise Miner Software este, é independente da ferramenta de *data mining* escolhida e pretende guiar o utilizador numa implementação de uma aplicação de *data mining*. Este oferece um processo fácil de compreender, permitindo um desenvolvimento organizado, adequado e a manutenção de projetos de *data mining*. Atribui uma estrutura para a conceção, criação e evolução. Ajuda a apresentar soluções para problemas de negócio assim como a definir objetivos de *data mining* (Azevedo & Santos, 2008).

5.2. Ferramentas para a análise e visualização de dados

Neste capítulo será feita uma análise às ferramentas que foram equacionadas para a resolução deste problema, os seus prós e contras, assim como uma justificação do porquê de ter optado por uma em particular, em detrimento das restantes.

Esta pré-seleção de ferramentas deve satisfazer uma série de critérios que auxiliem na resolução do problema proposto.

5.2.1. Matchflow

No desenvolvimento deste modelo para o posicionamento da equipa de arbitragem foi selecionada uma ferramenta gráfica utilizada no futebol robótico. Esta ferramenta designada *Matchflow,* é capaz de interpretar várias formações táticas das equipas de futebol posicionando os jogadores ao longo do terreno de jogo, assim como de identificar qual o jogador com a posse de bola e a quem este deve passar a bola, de acordo com vários fatores, tais como a distância e a existência (ou não) de jogadores adversários próximos e capazes de intercetar o passe.

De acordo com o trabalho realizado por Marques, (2010) para que estas funcionalidades sejam possíveis, a ferramenta utiliza o algoritmo *Delaunay Triangulation and Linear Interpolation* como mecanismo de posicionamento baseado em triangulação. O valor de *input* do mecanismo é um ponto focal no campo de futebol, normalmente a posição da bola. Os valores de *output* são posições estratégicas dos jogadores de acordo com o valor de entrada. O campo de futebol é dividido em vários triângulos de acordo com determinados dados. A triangulação de *Delaunay* é usada para encontrar o triângulo para cada posição da bola. Em seguida, um algoritmo de interpolação linear é utilizado para calcular a posição do jogador.

Inicialmente, ao definir uma nova formação, é necessário um ponto para criar o triângulo inicial. Em cada vértice do triângulo podemos definir as posições estratégicas do jogador para aquele ponto. Cada vez que um ponto é adicionado, é definido o triângulo na triangulação que o rodeia e, a partir deste ponto, são definidas as arestas para os vértices do seu triângulo cercado na triangulação. Se coincidir com um vértice já existente, são definidos vértices novos para as arestas opostas dos dois triângulos. Na figura 18 são percetíveis os dois casos quando um vértice **Pr** é adicionado.

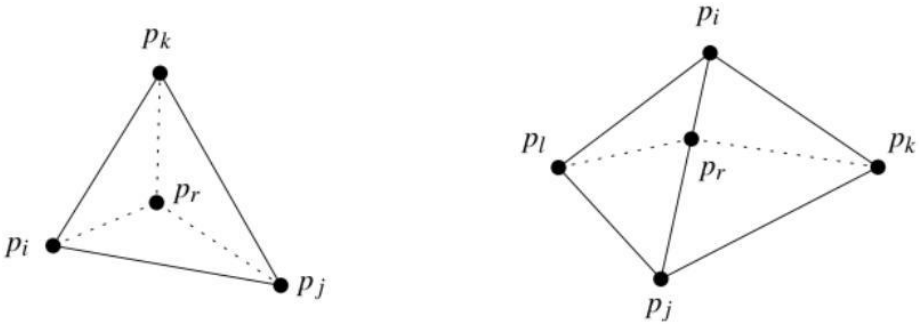


Figura 21 Vértice adicionado dentro e numa aresta de um triângulo. (fonte: Marques, 2010)

A triangulação de Delaunay tende a evitar triângulos com ângulos internos muito pequenos maximizando o menor ângulo de todos os triângulos na triangulação.

A forma mais simples e eficiente de computar a triangulação de Delaunay é adicionar um vértice de cada vez, triangulando novamente as partes afetadas do grafo a cada adição. Quando um vértice v é adicionado dividimos em três o triângulo que o contém, então aplicamos o algoritmo flip (este algoritmo permite rodar uma das arestas do triângulo no caso de este ser um triângulo não-Delaunay) (Marques, 2010).

Assim que os triângulos estejam definidos, é usado um algoritmo de interpolação linear para calcular a posição estratégica dos jogadores na formação.

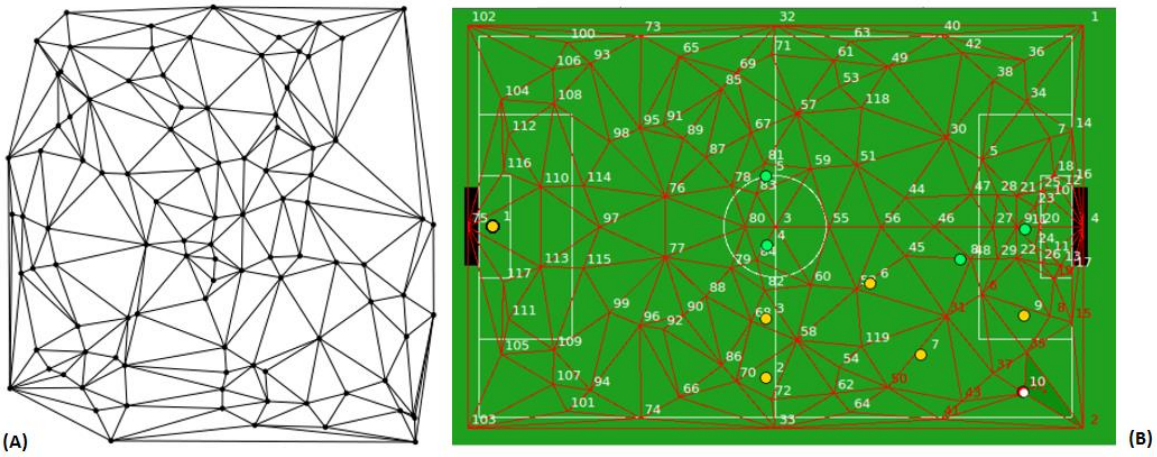


Figura 22 (A) Triangulação de Delaunay de um conjunto de 100 pontos aleatórios no plano. (B) Redefinição da posição dos jogadadores quando o jogador número 10 tem a posse da bola. (fonte: (A) Wikipedia, 2015 e (B) Marques, 2010)

Adaptação da ferramenta

Esta ferramenta será adaptada para que seja capaz de incluir também a figura do árbitro dentro do terreno de jogo e dos dois AA ao longo das correspondentes linhas laterais. Esta nova equipa deverá movimentar-se no terreno de jogo, não só de acordo com o que as normas e regulamentos para árbitros de futebol descritas nas leis do jogo, mas também de forma a estarem o melhor posicionados possível para serem capazes de tomar decisões corretas.

5.2.2. R, Rattle

O R é um *software* livre para computação estatística e de gráficos fornecendo uma ampla variedade de técnicas. A linguagem pode facilmente ser estendida com pacotes disponíveis no CRAN[3] (Zhao, 2012). Um destes pacotes é o *Rattle*.

Rattle GUI é um pacote gratuito (GNU GPL v2) que oferece um *graphical user interface* (GUI) para *data mining* usando a linguagem R.

[3] The Comprehensive R Archive Network – repositório de pacotes para R

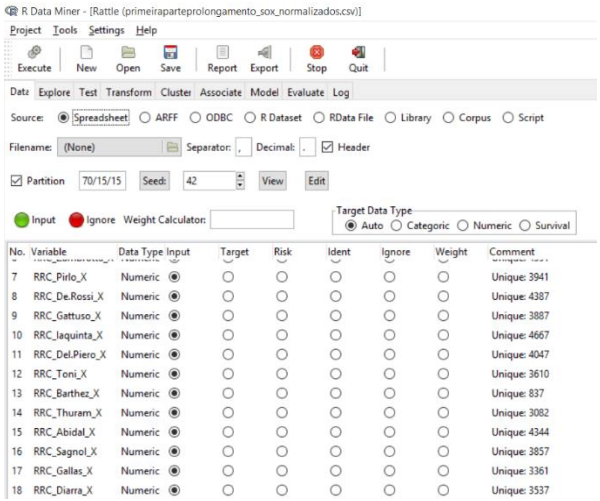


Figura 23 Exemplo do Rattle GUI

Rattle fornece muitas funcionalidades de *data mining* explorando o poder do *R Statistical Software*. Para além desta funcionalidade este pode ser utilizado como uma ferramenta para aprender a linguagem R. Existe uma *tab* designada de "*Log*" que replica o código R para qualquer atividade executada na GUI, que pode ser copiado e colado.

O Rattle pode ser utilizado para análise estatística ou para construção de modelos, permitindo o utilizador dividir os dados em três tipos de partições: treino, validação e teste (Togaware Pty Ltd, 2015).

5.2.3. Weka

De acordo com o (Machine Learning Group at the University of Waikato, 2015) WEKA é uma coleção de algoritmos de *machine learning* para tarefas de *data mining*. O *software* é munido de ferramentas para pré-processamento de dados, classificação, regressão, *clustering*, regras de associação e visualização.

Weka foi desenvolvido pela Universidade de Waikato na Nova Zelândia, e o seu foco é principalmente a comunidade académica, como uma ferramenta de *data mining* (Schumaker, Solieman, & Chen, 2010b). Este *software* é gratuito e licenciado ao abrigo *da GNU General Public License*.

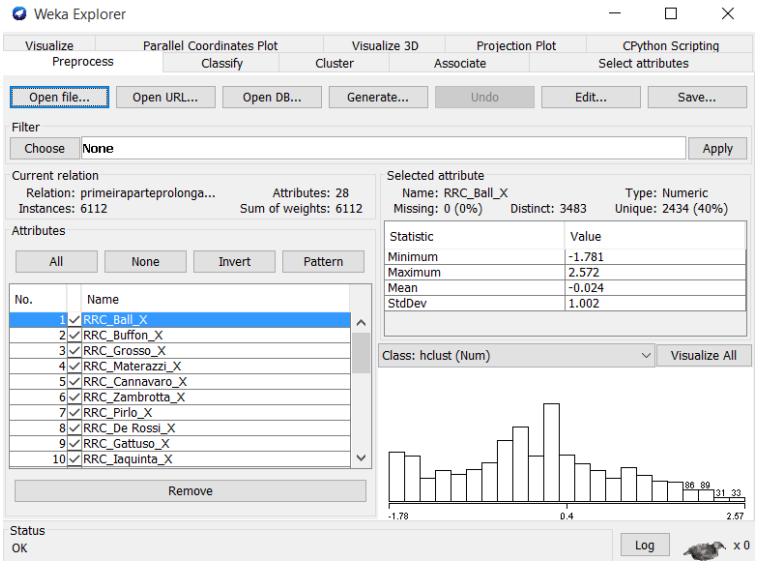


Figura 24 Exemplo da ferramenta Weka

Todas as técnicas são baseadas no pressuposto de que os dados estão disponíveis num único ficheiro ou relação, onde cada ponto de dados é descrito por um número fixo de atributos (normalmente numéricos ou nominais). Weka oferece acesso a bases de dados SQL processando o resultado retornado por uma query. Uma das funcionalidades que falta a este *software* é não ser capaz de fazer *data mining* de dados multi-relacionais. Para contornar esta restrição é necessário ligar todos os dados de várias tabelas num único ficheiro (Reutemann, Pfahringer, & Frank, 2004).

5.3. Conclusões

Foi realizada uma comparação das duas metodologias, através da correspondência entre as duas metodologias e uma ligação entre as diferentes fases de cada uma delas e a sua homóloga na metodologia contrária, conforme a Tabela III.

Tabela III Sumário da correspondência das distintas fases das duas metodologias de desenvolvimento

SEMMA	CRISP-DM
—	Estudo do negócio
Amostra	Estudo dos dados
Exploração	
Modificação	Preparação dos dados

Modelo	Modelação
Avaliação	Avaliação
—	Implementação

Após a comparação, a metodologia a ser seguida para este projeto de desenvolvimento é a CRISP-DM pois esta é mais flexível do que a criada pelo SAS Institute.

No que respeita às ferramentas analisadas, serão aproveitadas as funcionalidades de cada uma delas e o R, estendido com o pacote Rattle, será utilizado essencialmente para análise e tratamento de dados, tal como eliminações de registos de valores omissos e normalização. O Weka servirá como complemento para a fase de *data mining* pois esta ferramenta é munida de uma maior variedade de algoritmos. É de salientar que o MS Excel também será utilizado como apoio para a gestão dos dados pois oferece uma visualização mais "crua" dos dados, ao contrário das ferramentas supramencionadas. Estas ferramentas foram as selecionadas para este trabalho por serem de uso livre.

A ferramenta Matchflow era a que apresentava maior potencial, contudo esta foi descartada devido a diversos problemas de compilação do código em Windows mais modernos.

6. Previsão da posição da equipa de arbitragem

Este capítulo apresenta a descrição de todas as experiências elaboradas ao longo do trabalho. Como referido, o desenvolvimento foi pensado com a metodologia CRISP-DM como base.

6.1. Estudo do negócio

Esta fase inicial incide na compreensão dos objetivos do projeto e requisitos numa perspetiva de negócio. Com estes conhecimentos deve-se, então, transformar os objetivos e os requisitos num problema de *data mining* elaborando um plano de projeto preliminar para atingir os objetivos (Wirth & Hipp, 2000).

O objetivo deste trabalho é identificar a posição ideal para a equipa da arbitragem num determinado momento do jogo em relação à posição e direção da bola, assim como dos jogadores atacantes e defensores.

Para se atingir tal objetivo será feito um estudo a jogos de futebol como base de análise para a previsão da posição ideal da equipa de arbitragem em terreno de jogo. Estes jogos de futebol serão de grande referência e com equipas de arbitragem muito experientes. Com o auxílio das imagens televisivas de lances que possam ter sido mal decididos, será pedido auxílio a árbitros e observadores da primeira liga portuguesa para indicar qual a posição correta para o árbitro naquele instante de tempo que lhe teria proporcionado um juízo correto do lance, tornando a amostra de dados de aprendizagem o mais eficiente possível.

Pretende-se construir um modelo que sirva como referência para árbitros de futebol amadores e profissionais, no que respeita ao melhor posicionamento dentro de campo.

6.2. Estudo dos dados

A fase de estudo dos dados inicia-se com a recolha dos dados e a familiarização destes para identificar problemas de qualidade de dados, descobrir primeiras perceções dos dados ou detetar subconjuntos interessantes para formar hipóteses de informação oculta. Esta fase está próxima da anterior pois a formulação do problema de *data mining* e o plano do projeto requer pelo menos algum conhecimento dos dados existentes (Wirth & Hipp, 2000).

A recolha dos dados não fez parte do trabalho efetuado, como já referido, estes foram fornecidos por uma organização ao orientador desta dissertação, para o desenvolvimento do projeto. As situações com lances mal decididos pela equipa de arbitragem são exceção. Nessas situações será feita uma consulta a especialistas na área que indicarão onde o árbitro e os assistentes deveriam estar colocados para que tivessem melhores condições para ajuizar melhor a jogada. Nesses casos serão feitas alterações aos dados para que a amostra de dados permita obter o melhor modelo possível.

Ball	Buffon	Grosso	Materazzi	Cannavaro	Zambrotta	Camoranesi	Perrotta	Pirlo	De Rossi	Gattuso	Totti	Iaquinta	Del Piero	Toni
52.36,32.64,0.00	4.98,34.00	35.84,18.77	32.97,30.00	32.97,36.00	36.02,47.14	-65000.00,-65000.00	-65000.00,-65000.00	42.29,32.97	43.00,29.84	43.04,37.25	-65000.00,-65000.00	50.97,47.00	44.37,24.92	52.36,24.36
52.36,32.64,0.00	4.96,34.00	35.81,18.72	32.97,30.00	32.97,36.00	36.04,47.18	-65000.00,-65000.00	-65000.00,-65000.00	42.31,32.97	43.00,29.81	43.04,37.29	-65000.00,-65000.00	50.97,47.00	44.45,24.89	52.40,24.43
52.36,32.64,0.00	4.96,34.00	35.79,18.69	32.97,30.00	32.97,36.00	36.04,47.20	-65000.00,-65000.00	-65000.00,-65000.00	42.36,32.97	43.00,29.78	43.05,37.30	-65000.00,-65000.00	50.97,47.00	44.50,24.87	52.46,24.46
52.36,32.64,0.00	4.96,34.00	35.77,18.67	32.97,30.00	32.97,36.00	36.05,47.25	-65000.00,-65000.00	-65000.00,-65000.00	42.40,32.97	43.00,29.76	43.06,37.36	-65000.00,-65000.00	50.97,47.02	44.54,24.85	52.53,24.52
52.36,32.64,0.00	4.94,34.00	35.72,18.61	32.97,29.97	32.97,36.00	36.06,47.27	-65000.00,-65000.00	-65000.00,-65000.00	42.45,32.97	43.00,29.72	43.09,37.38	-65000.00,-65000.00	50.97,47.02	44.61,24.85	52.61,24.60
52.25,33.06,0.00	4.94,34.00	35.70,18.59	32.97,29.97	32.97,36.00	36.09,47.30	-65000.00,-65000.00	-65000.00,-65000.00	42.50,32.97	43.00,29.68	43.11,37.43	-65000.00,-65000.00	50.97,47.02	44.70,24.80	52.69,24.67
52.15,33.48,0.01	4.94,34.00	35.65,18.53	32.97,29.97	32.97,36.00	36.11,47.36	-65000.00,-65000.00	-65000.00,-65000.00	42.54,32.97	43.00,29.67	43.12,37.46	-65000.00,-65000.00	50.97,47.02	44.77,24.78	52.78,24.72
52.05,33.90,0.01	4.92,34.00	35.63,18.51	32.97,29.97	32.97,36.00	36.13,47.38	-65000.00,-65000.00	-65000.00,-65000.00	42.61,32.97	43.00,29.61	43.14,37.52	-65000.00,-65000.00	50.97,47.03	44.84,24.76	52.87,24.79
51.93,34.24,0.00	4.90,34.00	35.59,18.45	32.97,29.96	32.97,36.00	36.14,47.43	-65000.00,-65000.00	-65000.00,-65000.00	42.65,32.97	43.00,29.59	43.18,37.54	-65000.00,-65000.00	50.97,47.04	44.90,24.71	52.97,24.86
51.95,34.21,0.00	4.90,34.00	35.54,18.42	32.97,29.96	32.97,36.00	36.18,47.46	-65000.00,-65000.00	-65000.00,-65000.00	42.70,32.97	43.00,29.53	43.20,37.59	-65000.00,-65000.00	50.97,47.05	45.00,24.68	53.09,24.94
51.94,34.27,0.00	4.88,34.00	35.50,18.36	32.97,29.95	32.97,36.00	36.20,47.52	-65000.00,-65000.00	-65000.00,-65000.00	42.77,32.97	43.00,29.51	43.22,37.62	-65000.00,-65000.00	50.97,47.09	45.06,24.63	53.18,25.02
51.93,34.25,0.00	4.85,34.00	35.45,18.34	32.97,29.95	32.97,36.00	36.21,47.54	-65000.00,-65000.00	-65000.00,-65000.00	42.81,32.97	43.00,29.45	43.27,37.65	-65000.00,-65000.00	50.97,47.11	45.15,24.60	53.30,25.10
51.93,34.20,0.00	4.84,34.00	35.43,18.28	32.97,29.94	32.97,36.00	36.25,47.59	-65000.00,-65000.00	-65000.00,-65000.00	42.87,32.97	43.00,29.42	43.29,37.70	-65000.00,-65000.00	50.96,47.11	45.25,24.54	53.43,25.18
52.20,33.84,0.00	4.82,34.00	35.38,18.25	32.97,29.93	32.97,36.00	36.27,47.62	-65000.00,-65000.00	-65000.00,-65000.00	42.93,32.97	43.00,29.36	43.34,37.72	-65000.00,-65000.00	50.96,47.12	45.34,24.51	53.54,25.26
52.48,33.50,0.22	4.80,34.00	35.34,18.19	32.97,29.92	32.97,36.00	36.29,47.65	-65000.00,-65000.00	-65000.00,-65000.00	43.00,32.97	43.00,29.34	43.36,37.77	-65000.00,-65000.00	50.95,47.14	45.43,24.45	53.68,25.35
52.77,33.15,0.44	4.79,34.00	35.30,18.17	32.97,29.92	32.97,36.00	36.34,47.70	-65000.00,-65000.00	-65000.00,-65000.00	43.04,32.97	43.00,29.28	43.38,37.79	-65000.00,-65000.00	50.95,47.18	45.52,24.39	53.79,25.43
53.04,32.80,0.64	4.77,34.00	35.27,18.11	32.97,29.89	32.97,36.00	36.36,47.72	-65000.00,-65000.00	-65000.00,-65000.00	43.11,32.97	43.00,29.25	43.43,37.80	-65000.00,-65000.00	50.95,47.20	45.59,24.35	53.93,25.51
53.34,32.46,0.84	4.76,34.00	35.25,18.09	32.97,29.87	32.97,36.00	36.38,47.77	-65000.00,-65000.00	-65000.00,-65000.00	43.15,32.97	43.00,29.19	43.45,37.84	-65000.00,-65000.00	50.95,47.21	45.70,24.28	54.04,25.59
53.61,32.11,1.02	4.75,34.00	35.20,18.03	32.97,29.85	32.97,36.00	36.39,47.78	-65000.00,-65000.00	-65000.00,-65000.00	43.20,32.97	43.00,29.17	43.47,37.86	-65000.00,-65000.00	50.93,47.25	45.77,24.22	54.18,25.68
53.89,31.76,1.20	4.75,34.00	35.18,18.01	32.97,29.84	32.97,36.00	36.43,47.81	-65000.00,-65000.00	-65000.00,-65000.00	43.27,32.97	43.00,29.11	43.50,37.86	-65000.00,-65000.00	50.90,47.27	45.86,24.18	54.30,25.76
54.18,31.42,1.36	4.73,34.00	35.14,17.95	32.97,29.81	32.97,36.00	36.45,47.84	-65000.00,-65000.00	-65000.00,-65000.00	43.31,32.97	43.00,29.09	43.52,37.87	-65000.00,-65000.00	50.89,47.27	45.95,24.10	54.43,25.84
54.45,31.07,1.52	4.73,34.00	35.13,17.93	32.97,29.78	32.97,36.00	36.45,47.86	-65000.00,-65000.00	-65000.00,-65000.00	43.36,32.97	43.00,29.03	43.54,37.88	-65000.00,-65000.00	50.87,47.29	46.04,24.03	54.56,25.92
54.75,30.73,1.66	4.71,34.00	35.11,17.87	32.97,29.76	32.97,36.00	36.46,47.88	-65000.00,-65000.00	-65000.00,-65000.00	43.43,32.97	43.00,29.01	43.54,37.88	-65000.00,-65000.00	50.86,47.31	46.14,23.96	54.69,26.00
55.02,30.38,1.80	4.71,34.00	35.09,17.85	32.97,29.72	32.97,36.00	36.46,47.89	-65000.00,-65000.00	-65000.00,-65000.00	43.47,32.97	43.00,28.95	43.55,37.89	-65000.00,-65000.00	50.84,47.34	46.25,23.89	54.84,26.05
55.30,30.04,1.92	4.71,34.00	35.06,17.79	32.97,29.68	32.97,36.00	36.46,47.93	-65000.00,-65000.00	-65000.00,-65000.00	43.54,32.97	43.00,28.93	43.56,37.88	-65000.00,-65000.00	50.81,47.36	46.34,23.84	54.95,26.13
55.59,29.69,2.04	4.71,34.00	35.04,17.77	32.97,29.67	32.97,36.00	36.46,47.93	-65000.00,-65000.00	-65000.00,-65000.00	43.59,32.97	43.00,28.87	43.56,37.88	-65000.00,-65000.00	50.78,47.36	46.43,23.76	55.09,26.20
55.86,29.34,2.14	4.71,34.00	35.04,17.71	32.97,29.61	32.97,36.00	36.46,47.95	-65000.00,-65000.00	-65000.00,-65000.00	43.61,32.97	43.00,28.84	43.55,37.87	-65000.00,-65000.00	50.77,47.36	46.52,23.68	55.21,26.27
56.15,29.01,2.24	4.73,34.00	35.04,17.68	32.97,29.59	32.97,36.00	36.45,47.95	-65000.00,-65000.00	-65000.00,-65000.00	43.68,32.97	43.00,28.79	43.54,37.86	-65000.00,-65000.00	50.72,47.36	46.61,23.60	55.34,26.34
56.43,28.65,2.32	4.73,34.00	35.04,17.62	32.97,29.53	32.97,36.00	36.45,47.95	-65000.00,-65000.00	-65000.00,-65000.00	43.70,32.97	43.00,28.76	43.54,37.86	-65000.00,-65000.00	50.70,47.36	46.71,23.52	55.45,26.42
56.71,28.31,2.40	4.75,33.97	35.02,17.59	32.97,29.50	32.97,36.00	36.43,47.96	-65000.00,-65000.00	-65000.00,-65000.00	43.77,32.97	43.00,28.70	43.52,37.84	-65000.00,-65000.00	50.65,47.36	46.80,23.43	55.59,26.45
57.00,27.96,2.46	4.75,33.97	35.02,17.53	32.97,29.44	32.97,36.00	36.39,47.96	-65000.00,-65000.00	-65000.00,-65000.00	43.79,32.96	43.00,28.67	43.50,37.80	-65000.00,-65000.00	50.62,47.36	46.90,23.35	55.69,26.51

Figura 25 Exemplo dos dados recebidos. As posições cartesianas (x,y) de todos os jogadores do jogo da Final do Mundial de 2006 entre a Itália e a França

Os dados são referentes à posição cartesiana dos jogadores durante o jogo, jogo este que se estendeu ao prolongamento, ou seja, 120 minutos, sendo que existem 24 registos por segundo, relativos às 24 frames por segundo de capacidade das câmaras de televisão.

Os jogadores têm os valores no eixo das abcissas e das ordenadas e a bola tem a particularidade de ter também valores para o eixo das cotas.

(A)

Ball_X	Ball_Y	Ball_Z	Zidane_X	Zidane_Y
52.04	33.13	0	61.34	32

(B)

Figura 26 Exemplo das coordenadas da (a) bola no formato (x,y,z) para os eixos das abcissas, ordenadas e cotas respetivamente e dos (b) jogadores no formato (x,y) para os eixos das abcissas e das ordenadas respetivamente.

O jogo em questão foi realizado no Estádio Olímpico, *Olympiastadion*, em Berlim. Este estádio tem uma lotação de 76 243 pessoas e as medidas do campo de 105x68m (ZEROZERO, 2015).

A partir da informação das dimensões do terreno do jogo, é possível analisar os dados e criar alguns pontos de referência. Como tal, é possível criar alguns pontos de referência.

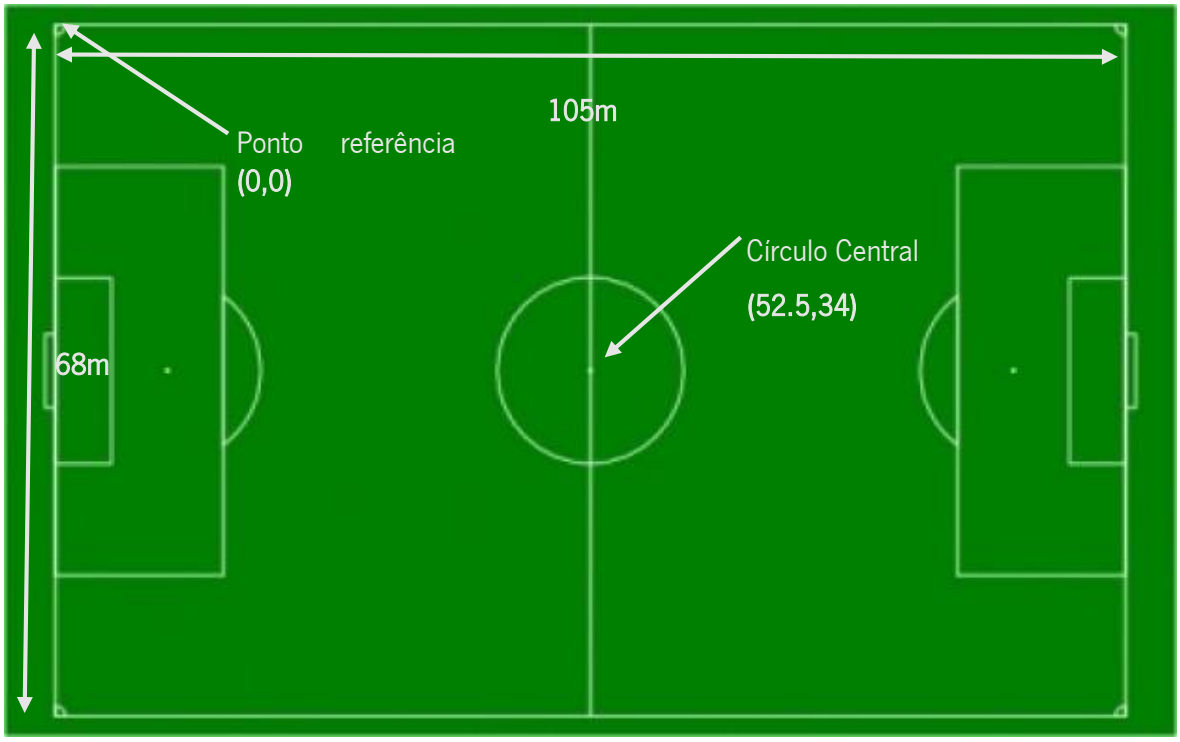


Figura 27 Terreno de jogo com as dimensões do Estádio Olympiastadion e as posições cartesianas de alguns pontos de referência do campo.

Ao comparar alguns dos pontos representados na figura 27 é possível tirar algumas ilações quanto à margem de erro do sistema de obtenção das posições cartesianas por vídeo da bola e jogadores. No instante inicial a bola encontra-se no ponto central, no entanto os registos apontam para a posição (52.04,33.13,0) (ver figura 26 A) quando era expectável que a posição fosse (52.5,34,0), tal como é indicado na figura 28. Existe portanto uma diferença de 46 cm para o eixo das abcissas, sendo esta uma margem de erro muito reduzida, e uma diferença de 86 cm para o eixo das ordenadas. Quanto ao eixo das cotas, a medida vai ao encontro do expectável.

O momento inicial do jogo, no qual todos os jogadores estão nas posições iniciais à espera do apito inicial, é um bom momento para esta análise relativa às posições que são indicadas nos registos e as reais posições de acordo com a figura 28.

Figura 28 Instante inicial da final do Mundial de 2006 entre a Itália e a França. (Adaptada SKY MONDIALE 1, 2014)

Tabela IV Comparação entre a localização real e a localização registada.

ID	Objeto	Localização em Registo	Localização Real	Motivo da expetativa da localização real
1	Bola	(52.04,33.13,0)	(52.5,34,0)	Está no círculo central, que fica a meio dos dois eixos
2	Toni	(52,33.04)	(52,33.5)	Está com a bola no pé para dar o pontapé de saída
3	Henry	(52.97,42.97)	(52.5,43.15)	Deverá estar a 9,15m da bola
4	Zidane	(61.34,32)	(61.65,32)	Deverá estar a 9,15m da bola
5	Cannavaro	(28.2,39)	Diferença para o AA Otero	Penúltimos defensores da Itália
6	Materazzi	(28.97,30)		
7	Thuram	(68.58,27.97)	Diferença para o AA Garcia	Penúltimo defensor da França
8	Garcia	(70.04,68)	Diferença para o penúltimo defensor da França	Está em linha com o penúltimo defensor da França

9	Otero	(25.97,0)	Diferença para os penúltimos defensores italianos	Está em linha com o penúltimo defensor da Itália

A Tabela IV serve para perceber a qualidade dos dados uma vez não fez parte do âmbito deste projeto a recolha destes. Infelizmente, por razões óbvias, apenas foi feita uma análise superficial tendo em conta pontos de referência do terreno do jogo, tais como as dimensões do terreno do jogo e das várias áreas integrantes, nomeadamente o ponto central, o círculo central e as áreas de grande penalidade.

Após esta análise é percetível que a margem de erro é reduzida dada as dimensões do terreno de jogo. À exceção da diferença de valores entre os penúltimos defensores de ambas as equipas para os assistentes, onde se registaram diferenças de aproximadamente 2 metros, a margem de erro é inferior a 1 metro.

As variáveis de saída para este modelo são as posições dos três árbitros: árbitro (Elizondo), AA número um (Otero) e o AA número dois (Garcia).

6.3. Preparação dos dados

Esta fase engloba todas as atividades que conduzem à construção do conjunto de dados finais (os dados que servirão para alimentar as ferramentas de modelação a partir dos dados iniciais). É provável que as tarefas da preparação dos dados sejam feitas mais do que uma vez e sem qualquer ordem definida inicialmente. Fazem parte desta fase tarefas como incluir tabelas, registos e seleção de atributos, limpeza de dados, construção de novos atributos e transformação de dados para melhor se adequarem às ferramentas de modelação (Wirth & Hipp, 2000).

A preparação dos dados foi feita de acordo com o conhecimento que se pretendia obter em cada momento. Assim, os dados foram preparados em dois principais momentos. Para (1) elaborar o modelo de previsão da localização ideal dos AA e (2) elaborar o modelo de previsão da localização ideal do árbitro.

Para o primeiro problema foram removidos todos os valores referentes à coordenada *y* uma vez que esta não tem qualquer influência (ou tem uma influência tão baixa que para a construção deste

modelo foi considerada desprezável) sobre o posicionamento do AA, uma vez que a posição deste é relativa à segunda coordenada de *x* mais baixa (ou segunda mais alta dependendo do lado do ataque que acompanha). Estes resultados foram também confirmados pelas experiências demonstradas na Tabela VIII.

Para o AA apenas foram observados com detalhe o posicionamento destes em foras de jogo mal assinalados. Durante os 120 minutos apenas foi registado um fora de jogo que suscita dúvidas. No entanto, não se procedeu a qualquer correção quanto ao posicionamento do AA em questão (Garcia) pois este encontra-se bem colocado e o lance é muito duvidoso, mesmo com as várias repetições na televisão. No benefício da dúvida, considera-se que o Garcia esteve bem. Se errou foi por motivos alheios ao posicionamento tais como a habilidade do árbitro conseguir interpretar jogadores rápidos, em movimentos simultâneos e sequenciais, de acordo com o referido no capítulo 3.1.3.

Para o segundo problema, modelo de previsão da localização ideal do árbitro, as alterações aos dados foram um pouco diferentes. Numa primeira instância, foram analisadas várias situações de jogo que exigiram uma maior intervenção do árbitro. Dessa série de decisões tomadas ao longo do jogo, foram selecionadas (ver Tabela V) as que o árbitro errou por motivo do seu mau posicionamento.

Tabela V Momentos do jogo no qual o árbitro errou por falha no posicionamento

ID	Tempo	Descrição da jogada	Localização real	Localização correta
1	05:09	O jogador n°7 branco (Malouda) entra dentro de área pelo lado direito e assim que o jogador n° 23 azul (Materazzi) tenta cortar a bola. Existe um pequeno toque no pé do atacante e a grande penalidade é justificada. Esta, no entanto, foi mais pelo aparato da queda, uma vez que o árbitro encontra-se mal colocado.	(27.5,13.84)	(17.5,12)
2	31:53	Perrota é pisado por Ribery. O árbitro perto e sem ninguém a impedir a visualização nada assinala. Este erro não é fruto do mau posicionamento	(58,22.85)	(58,22.85)
3	52:41	Grande penalidade a favor da França. O árbitro longe do lance nada assinala.	Fora de imagem	(15.5,14.84)

ID	Tempo	Descrição da jogada	Localização real	Localização correta
4	65:55	Iaquinta domina a bola com o braço. Árbitro longe e com um atleta a obstruir a visão nada assinala.	(43,18.7)	(37.5,13.84)
5	71:37	O árbitro mal colocado, está numa linha de passe da Itália. Este é surpreendido com a bola.	(66.25,21.85)	(71.75,20.85)
6	79:14	Cannavaro salta de forma negligente sobre Zidane. Nada assinalado.	(72,54.16)	(74.75,54.16)

Após esta análise e correção à posição do árbitro, isto é, onde o árbitro deveria estar colocado para que a sua tomada de decisão fosse outra, procedeu-se às alterações dos valores nos dados para que o conhecimento a adquirir fosse o mais correto possível. Esta correção foi feita de acordo com o feedback recebido pelo perito na área, Jorge Oliveira, AA da 1ª liga do futebol português, Liga NOS.

Figura 29 Exemplos de situações no qual o árbitro teve situações de análise difícil e qual o posicionamento correto

A Figura 29 apresenta 6 situações nas quais o posicionamento do árbitro não era o mais correto para avaliar os lances com os quais se deparou e como tal, errou. Estas imagens são também expostas no apêndice A. A roda verde clara representa o local onde o árbitro deveria estar colocado para tomar a melhor decisão. De referir que a imagem no canto superior direita indica que o árbitro está bem posicionado. No entanto, este erra por outros fatores alheios ao posicionamento.

Por defeito os valores aparecem com um valor de "-65000" quando o objeto (árbitro, jogador ou bola) não está visível para nenhuma das câmaras com as quais os dados foram capturados. Este valor foi considerado como um *outlier* e portanto foram removidos todos os registos que continham um destes valores. Exceção feita para os jogadores que foram substituídos ao longo do jogo. Os valores referentes às posições destes jogadores após abandonarem o terreno de jogo foram substituídos pelos dos jogadores que entraram para os seus lugares. Uma outra ação aplicada aos dados foi a eliminação de todos os registos nos quais houvesse algum elemento em terreno de jogo sem qualquer valor relativo à sua posição.

6.4. Modelação

Nesta fase são selecionadas e aplicadas várias técnicas de modelação sendo os seus parâmetros calibrados para valores ótimos. Tipicamente há várias técnicas para o mesmo tipo de problema de *data mining* e algumas dessas técnicas exigem formatos de dados específicos. Existe uma linha muito ténue a separar a fase da modelação da fase de preparação de dados pois é recorrente identificar problemas nos dados quando se está a modelar ou ter novas ideias para construir novos dados (Wirth & Hipp, 2000).

Para obter um modelo o mais exato possível foram efetuadas várias experiências e exploração dos dados do jogo de futebol. Essas experiências consistiram:

- Para os AA – Garcia e Otero
    - Uso de subconjuntos de dados de um terço, dois terços e a totalidades destes
    - Restrição à coordenada X
    - Restrição à equipa que estavam a acompanhar, bola e árbitro
    - Com os dados não normalizados e normalizados
    - Uso dos 90 minutos de jogo para treino e dos 30 minutos de prolongamento para testes
    - Uso dos primeiros 15 minutos de jogo para treino e dos primeiros 15 minutos da segunda parte para testes
    - Trocar a ordem dos dados (random)
- Para o árbitro
    - Uso de subconjuntos de 45 minutos, 90 minutos e 120 minutos de jogo.
    - Restrição à coordenada X, Y e ambas
    - Com os dados não normalizados e normalizados
    - Uso dos 90 minutos de jogo para treino e dos 30 minutos de prolongamento para testes
    - Dados de apenas uma equipa e de ambas as equipas
    - Uso dos primeiros 15 minutos de jogo para treino e dos primeiros 15 minutos da segunda parte para testes
    - Usar situações controlados do jogo – como só pontapés de canto por exemplo
    - Trocar a ordem dos dados (random)

Estas experiências foram pensadas com o propósito de selecionar o algoritmo de regressão mais adequado para a previsão da melhor posição dos elementos da equipa de arbitragem em fundamento da posição da bola e dos jogadores.

Nesta aplicação a classe das variáveis é numérica. Neste caso **o objetivo é minimizar o erro da raiz do valor quadrático médio (RMSE)**[4] ou valor eficaz da predição da variável.

Após uma análise a vários algoritmos de regressão foram selecionados os modelos matemáticos que apresentaram melhores resultados para o primeiro problema definidos na documentação da ferramenta Weka (Weka Documentation, 2015) da seguinte forma:

- **ZeroR** – Classe para a construção e utilização de um classificador 0-R. Prevê a média (para uma classe numérica) ou a moda (para uma classe nominal).
- **M5P** – Implementa rotinas de base para a geração de modelos de árvores e regras M5.
- **LinearRegression** – Classe para o uso de regressão linear para a predição. Usa o critério Akaike para a seleção de modelo e é capaz de lidar com casos ponderados.
- **DecisionTable** – Classe para a construção e utilização de uma simples tabela de decisão maioritariamente classificadora.
- **REPTree** – É uma árvore de decisão rápida. Esta classe constrói uma árvore de decisão/regressão usando informação adquirida e aprimorada.
- **SimpleLinearRegression** – Aprende um modelo de regressão linear simples. Escolhe o atributo que resulta no menor erro quadrado. Apenas consegue trabalhar com valores numéricos e não se adequa a valores omissos.
- **PaceRegression** – Classe para a construção de modelos de regressão linear com cadência e usa-os para a previsão. Sob condições de regularidade, este modelo é comprovadamente ideal quando o número de coeficientes tende para infinito. Consiste num grupo de avaliadores que são ótimos sob certas condições ou na totalidade.
- **MultilayerPerceptron** – Classificador que utiliza retropropagação para classificar instâncias. Esta rede pode ser construída à mão, criada por um algoritmo ou ambas. Esta também pode ser monitorizada e modificada durante o treino. Os nós nesta rede são todos sigmóides (exceto quando a classe é numérica, caso em que os nós de saída se tornam no limite unidades lineares).
- **IBK (Nearest Neighbor) –** k vizinhos mais próximos. Pode selecionar o valor apropriado de K com base na validação cruzada. Pode também fazer distância ponderada.
- **AdditiveRegression –** Meta classificador que melhora o desempenho de um classificador com base de regressão. Cada iteração encaixa um modelo para os resíduos deixados pelo classificador na iteração anterior. A previsão é conseguida através da adição às previsões de cada classificador. A redução do encolhimento de parâmetros (taxa de aprendizagem) ajuda a evitar overfitting e tem um efeito de alisamento, mas aumenta o tempo de aprendizagem.
- **RandomSubSpace –** Este método constrói um classificador baseado numa árvore de decisão que mantém a mais alta precisão em dados de treino e melhora a precisão à medida que aumenta a complexidade. O classificador é constituído por várias árvores construídas sistematicamente por seleção pseudo aleatória de subconjuntos de componentes do vetor, isto é, árvores construídas em subespaços escolhidos aleatoriamente.

[4] Do inglês: Root mean squared error (RMSE)

- **LeastMedSq** – Implementa uma mediana quadrada mínima de regressão linear utilizando a classe de regressão linear Weka já existente para formar previsões.
- **Bagging** – Classe para ensacar um classificador para reduzir a variância. Pode fazer a classificação e regressão em função da aprendizagem base.
- **IsotonicRegression** – Aprende um modelo de regressão isotónica. Escolhe um atributo que resulta no menor erro quadrado. Não permite valores omissos e só consegue trabalhar com valores numéricos.

Os resultados do erro da raiz do valor quadrático médio gerados pelo Weka Explorer estão representados na Tabela VII para as várias experiências pensadas. Esta análise foi limitada apenas a dados normalizados pois, após uma comparação de vários algoritmos de regressão, conclui-se que todos os valores do RMSE são consideravelmente maiores quando os dados não estão normalizados, como é possível verificar na Tabela VI. De referir que os mesmos algoritmos apresentam valores numa taxa equivalente, quer para dados não normalizados, quer para os dados normalizados (ver Figura 30). Isto é, o algoritmo com melhores resultados para os dados não normalizados (M5P com RMSE igual a 0.8832), é o mesmo para os dados normalizados (M5P com RMSE igual a 0.0162) e assim sucessivamente até ao que apresenta piores resultados (ZeroR).

Tabela VI Comparação da taxa de erro da raiz do valor quadrático médio (RMSE) para dados normalizados e não normalizados (ordenada ascendentemente pelo valor de RMSE)

Algoritmo	Dados Não Normalizados	Dados Normalizados
M5P	0.8832	0.0162
REPTree	0.93	0.0173
DecisionTable	2.0347	0.0361
MultilayerPerceptrum	2.6853	0.0493
LinearRegression	3.7223	0.0683
PaceRegression	3.7223	0.0683
SMOreg	4.0518	0.0743
SimpleLinearRegression	5.4779	0.1005
ZeroR	14.0609	0.258

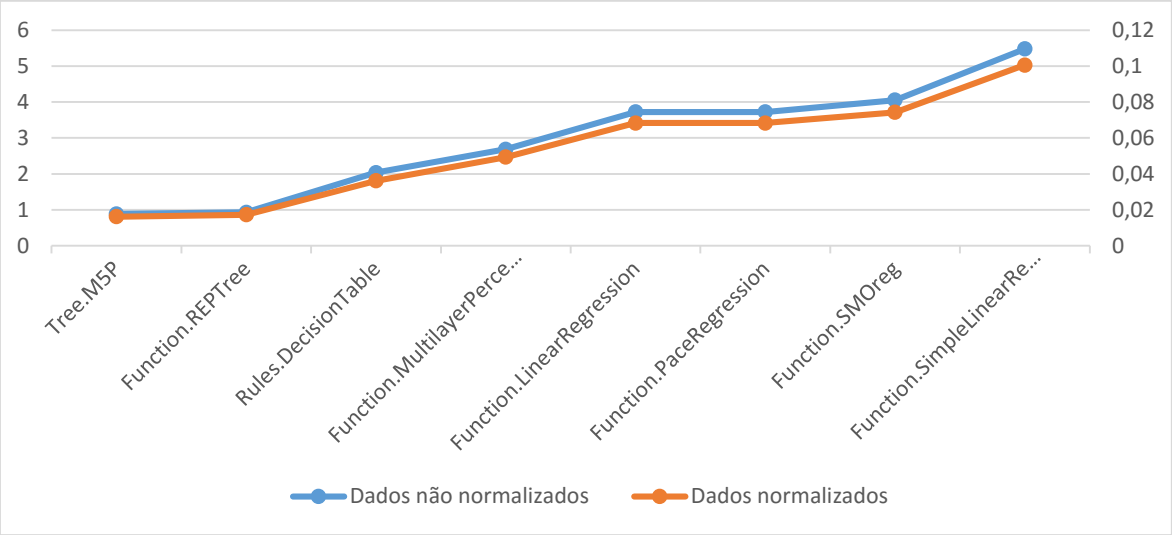


Figura 30 Exemplificação gráfica da comparação da taxa de erro da raiz do valor quadrático médio (RMSE) para dados normalizados e não normalizados em diferentes escalas

Recorrendo à Figura 30 é facilmente percetível a enorme diferença nos resultados para dados normalizados (linha cor de laranja) e dados não normalizados (linha azul). A normalização dos dados foi feita também para obter um maior coeficiente de correlação entre os atributos (*x*) e a variável alvo (*y*). A correlação pode ser positiva (1) quando os dados crescem juntos, ou negativa (-1) se *y* cresce quando *x* desce e vice-versa (Pierce, 2015), como está exemplificado na Figura 31. O valor de 0 significa que não existe correlação.

A normalização significa ajustar os valores medidos em diferentes escalas para uma escala subjetivamente comum (Dodge, 2006). Esta normalização foi considerada para o uso de mais do que um jogo para análise, uma vez que as leis do jogo permitem um intervalo de medidas para os terrenos de jogo (ver Figura 1). Estes dados necessitam então de ser normalizados para que o intervalo de valores seja sempre o mesmo independentemente de onde se realizou o jogo.

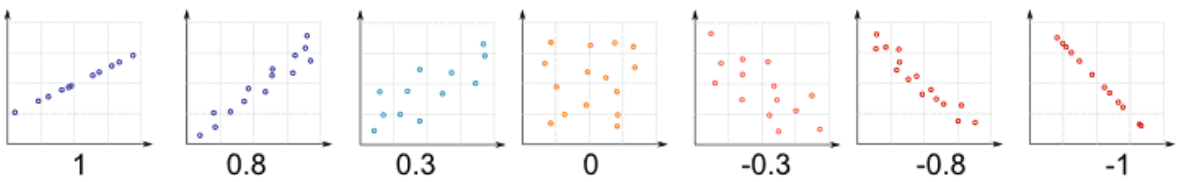


Figura 31 Representação gráfica do coeficiente de correlação entre os atributos e a variável alvo.

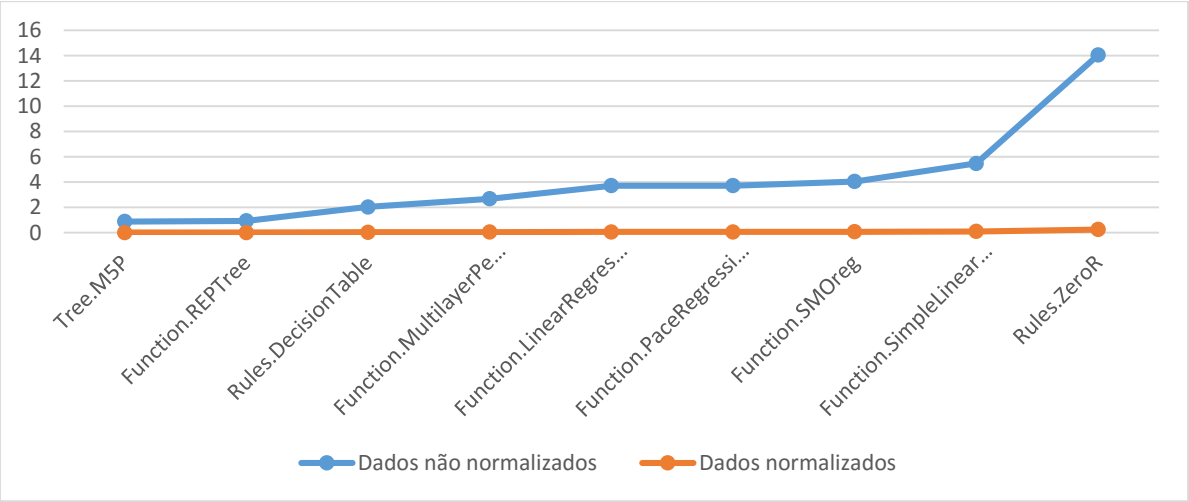


Figura 32 Exemplificação gráfica da comparação da taxa de erro da raiz do valor quadrático médio (RMSE) para dados normalizados e não normalizados na mesma escala

Tabela VII Taxa de erro RMSE para os diferentes modelos para a previsão dos AA

Algoritmos	Um Terço			Dois Terços			Três Terços		
	França	*Itália*	*90m*	*França*	*Itália*	*90m*	*França*	*Itália*	*90m*
Previsão da posição do AA Garcia – dados normalizados									
ZeroR	0.258	0.2627	0.26	0.2591	0.2622	0.2605	0.2591	0.2626	0.2601
M5P	0.0162	0.0189	0.0328	0.0131	0.0143	0.014	0.0118	0.0126	0.0124
LinearRegression	0.0683	0.0807	0.077	0.0686	0.0802	0.0777	0.0686	0.0804	0.0778
DecisionTable	0.0361	0.0572	0.0376	0.0206	0.0266	0.0268	0.0206	0.021	0.0218
REPTree	0.0173	0.0194	0.0196	0.0115	0.0133	0.0141	0.0097	0.0113	0.0115
SimpleLinearRegression	0.1005	0.0984	0.1013	0.1009	0.0973	0.1022	0.1008	0.0977	0.1022
PaceRegression	0.0683	0.0807	0.077	0.0686	0.0802	0.0777	0.0686	0.0804	0.0778
MultilayerPerceptron	0.0493	0.0572	0.0551	0.0481	0.0527	0.0535	0.0481	0.0549	0.0582

Algoritmos	Um Terço			Dois Terços			Três Terços		
	França	*Itália*	*90m*	*França*	*Itália*	*90m*	*França*	*Itália*	*90m*
Previsão da posição do AA Garcia – dados normalizados									
Previsão da posição do AA Otero – dados normalizados									
ZeroR	0.2084	0.1886	0.1986	0.2074	0.1886	0.199	0.2069	0.1895	0.1991
M5P	0.0388	0.028	0.0389	0.0279	0.0188	0.0274	0.0203	0.0155	0.022
LinearRegression	0.1507	0.1325	0.1487	0.1499	0.1314	0.1487	0.1495	0.1314	0.1485
DecisionTable	0.0381	0.0355	0.042	0.028	0.1314	0.0302	0.0218	0.0189	0.0247
REPTree	0.0393	0.0299	0.0432	0.0259	0.1433	0.0269	0.0201	0.0148	0.0209
SimpleLinearRegression	0.1625	0.1437	0.1535	0.1612	0.1433	0.1535	0.1609	0.1435	0.1533
PaceRegression	0.1507	0.1325	0.1487	0.1499	0.1314	0.1488	0.1495	0.1314	0.1485
MultilayerPerceptron	0.1412	0.0922	0.1309	0.1214	0.0964	0.1402	0.1272	0.09	0.1259

A avaliação dos diferentes algoritmos elaborada na Tabela VII foi feita com o propósito de perceber como estas classes se comportam para as diferentes experiências pensadas para este problema. Analisando a Figura 32 e a Figura 34 que mostram como os modelos se comportam conforme o volume de dados analisados para treino aumentam, podemos concluir que as árvores de decisão (M5P e REPTree) melhoram conforme a volumetria de dados é maior, sendo que essa diferença é ainda mais visível para o DecisionTable. Estes três algoritmos são os que menor RMSE apresentam.

Todos os algoritmos executados apresentam uma grande correlação entre os atributos e a variável alvo (ver anexo B) com valores perto de uma correlação perfeita positiva. Este valor é facilmente justificado pelo facto do AA ter que acompanhar sempre o penúltimo defensor. Se a equipa que o AA está responsável de acompanhar está a atacar, os seus jogadores vão todos subir no terreno de jogo,

ou seja, o valor de *x* vai subir e o AA naturalmente vai subir também. O oposto também acontece quando a respetiva equipa sofre um ataque.

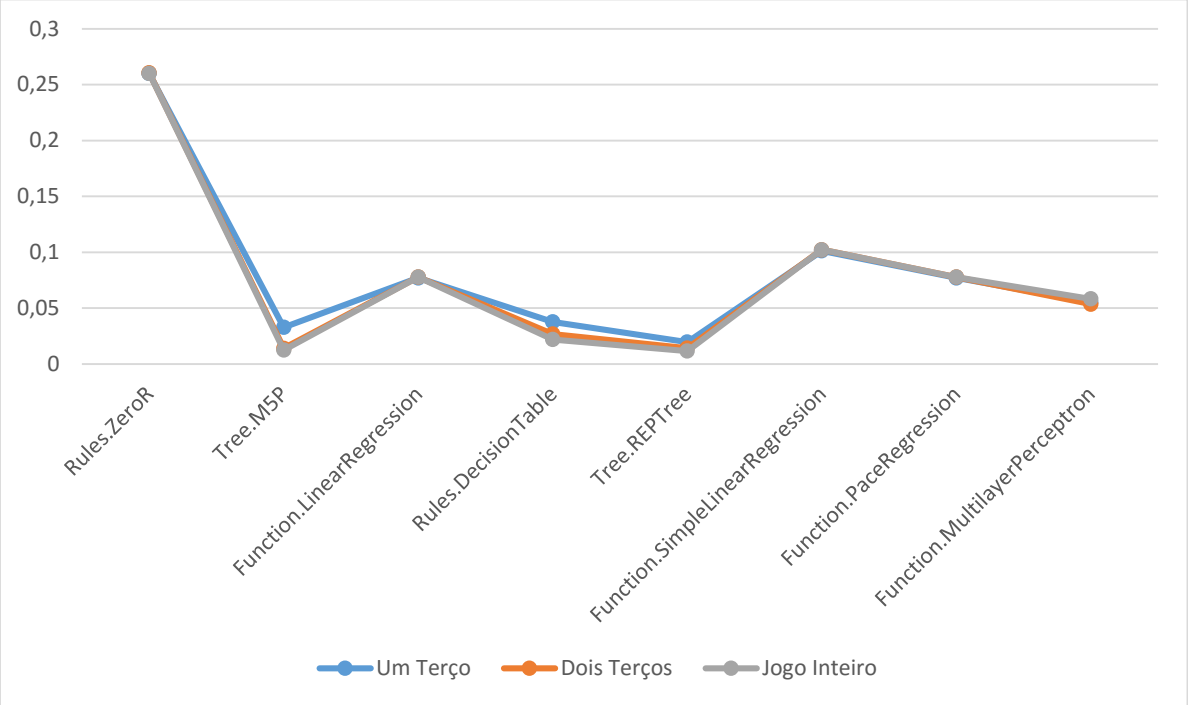


Figura 33 Diferença de valores para os vários algoritmos utilizados para os diferentes subconjuntos de dados dos 90 minutos da análise ao AA Garcia

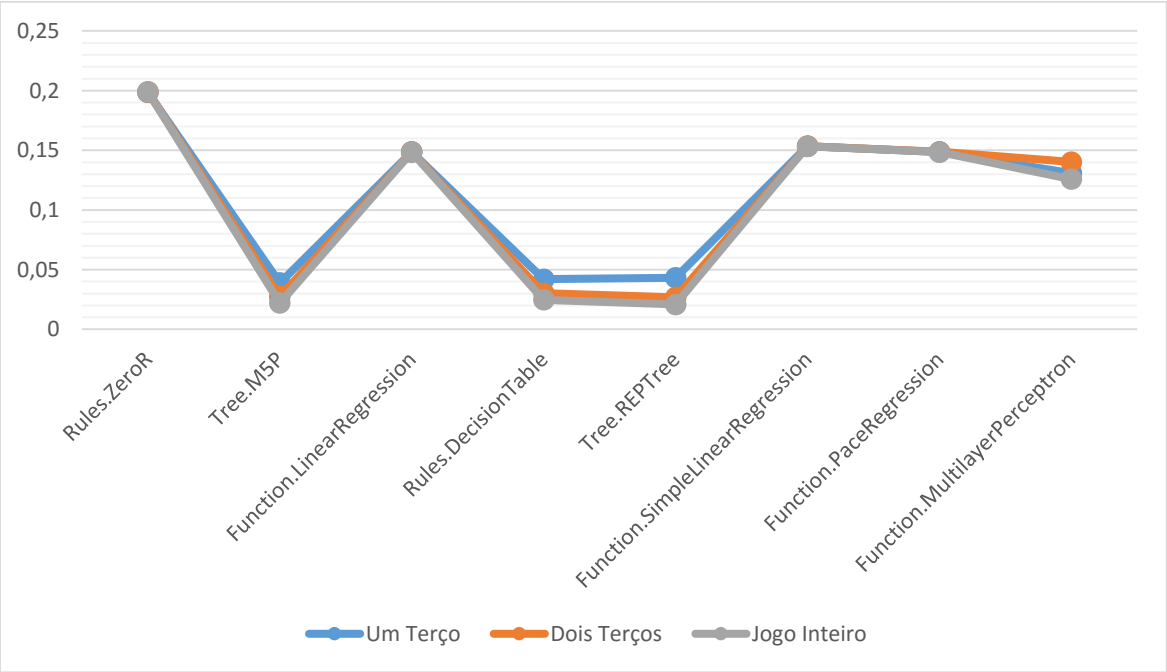


Figura 34 Diferença de valores para os vários algoritmos utilizados para os diferentes subconjuntos de dados dos 90 minutos da análise ao AA Otero

Como já referido, a Tabela VII considera apenas um indicador entre vários para avaliar os modelos matemáticos em questão. Esta ação não é suficiente e na próxima fase da metodologia é feita uma melhor avaliação dos modelos. A análise elaborada nesta tabela é, contudo, excelente para perceber como os modelos se comportam conforme o volume de dados é maior.

Estes registos foram obtidos executando o modelo de validação designado por *cross-validation*. Esta é uma técnica de validação do modelo para avaliar a forma como os resultados de uma análise estatística irá generalizar um conjunto de dados independentes (Varma & Simon, 2006). Foi considerado para este problema (este modelo de validação é utilizado principalmente em ambientes onde o objetivo é a previsão) para estimar com precisão como um modelo preditivo se irá comportar na prática (Bermingham et al., 2015).

Convém, no entanto, realçar que os valores apresentados na Tabela VII são todos muito otimistas, uma vez que os dados utilizados para treino são os mesmos que os utilizados para teste. O que justifica as pequenas taxas de erro para os vários modelos.

Para o árbitro, Elizondo, foram aplicadas várias técnicas de modelação e os parâmetros calibrados para a otimização. A Tabela VIII é indicativa dos modelos com melhores resultados, usando, uma vez mais o *cross-validation*. Ou seja, as taxas de erro apresentadas nesta fase serão muito otimistas.

Tabela VIII Taxa de erro RMSE para os diferentes modelos para a previsão de Elizondo

Algoritmo	Primeira Parte	Segunda Parte
Todos as coordenadas (*x,y*) para todos os jogadores e bola, exceto os AA e Elizondo (só *x*)		
REPTree	0.007	0.0077
RandomSubSpace	0.0031	0.0033
Bagging	0.0035	0.004
IBK	0.0009	0.0009
Somente coordenadas x		
REPTree	0.0075	0.0087

Algoritmo	Primeira Parte	Segunda Parte
RandomSubSpace	0.0035	0.0038
Bagging	0.0039	0.0046
IBK	0.0009	0.0009
	Somente coordenadas y	
REPTree	0.0128	0.0117
RandomSubSpace	0.0057	0.0046
Bagging	0.0061	0.0054
IBK	0.0012	0.001
	Coordenadas x,y juntas	
REPTree	0.0147	0.016
RandomSubSpace	0.0098	0.0069
Bagging	0.0101	0.0089
IBK	0.0071	0.002

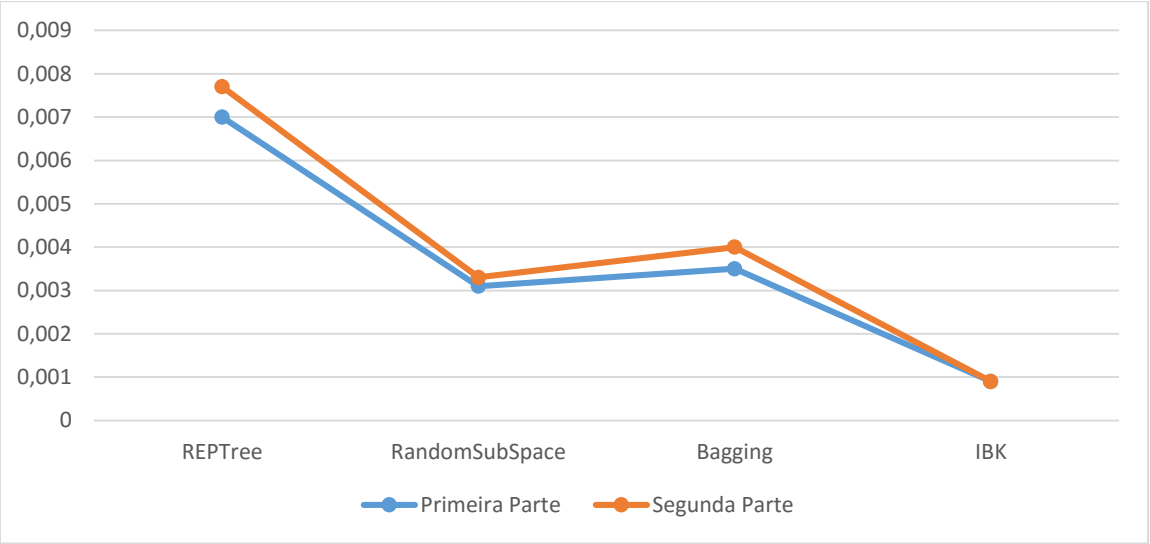


Figura 35 Comparação do valor de RMSE para os modelos matemáticos com melhores resultados na primeira e segunda parte do tempo regulamentar para a experiência de cada jogador com duas variáveis, x e y

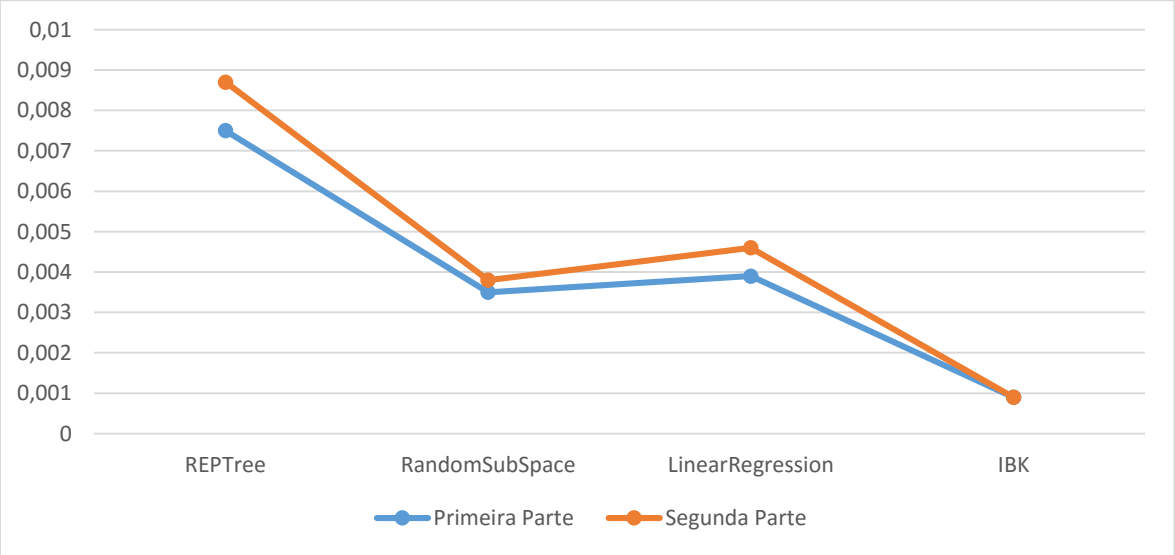


Figura 36 Comparação do valor de RMSE para os modelos matemáticos com melhores resultados na primeira e segunda parte do tempo regulamentar para a experiência limitada à posição x

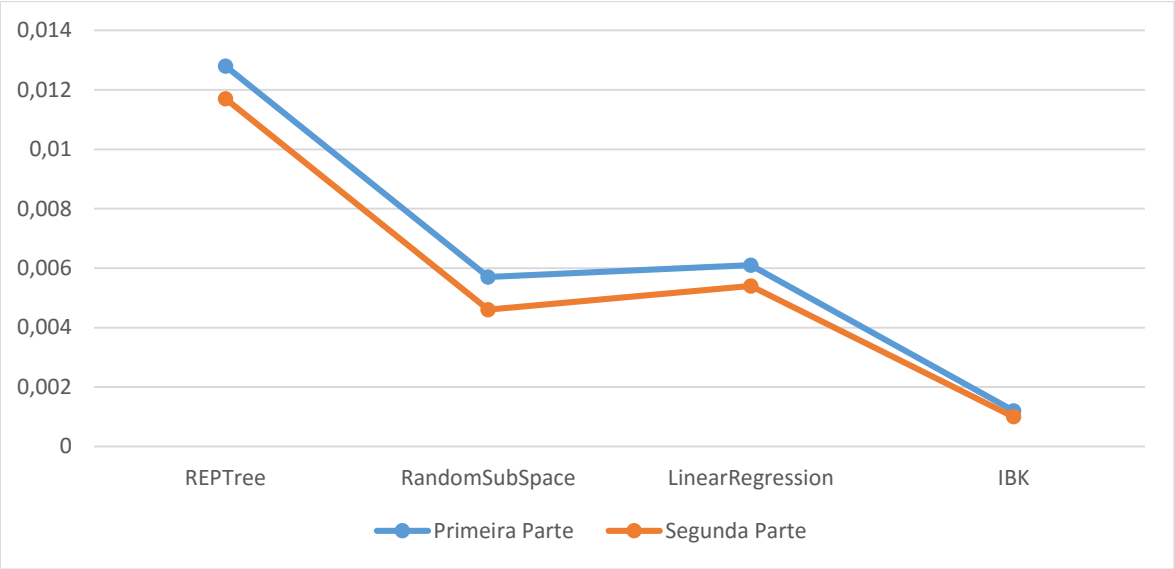


Figura 37 Comparação do valor de RMSE para os modelos matemáticos com melhores resultados na primeira e segunda parte do tempo regulamentar para a experiência limitada à posição y

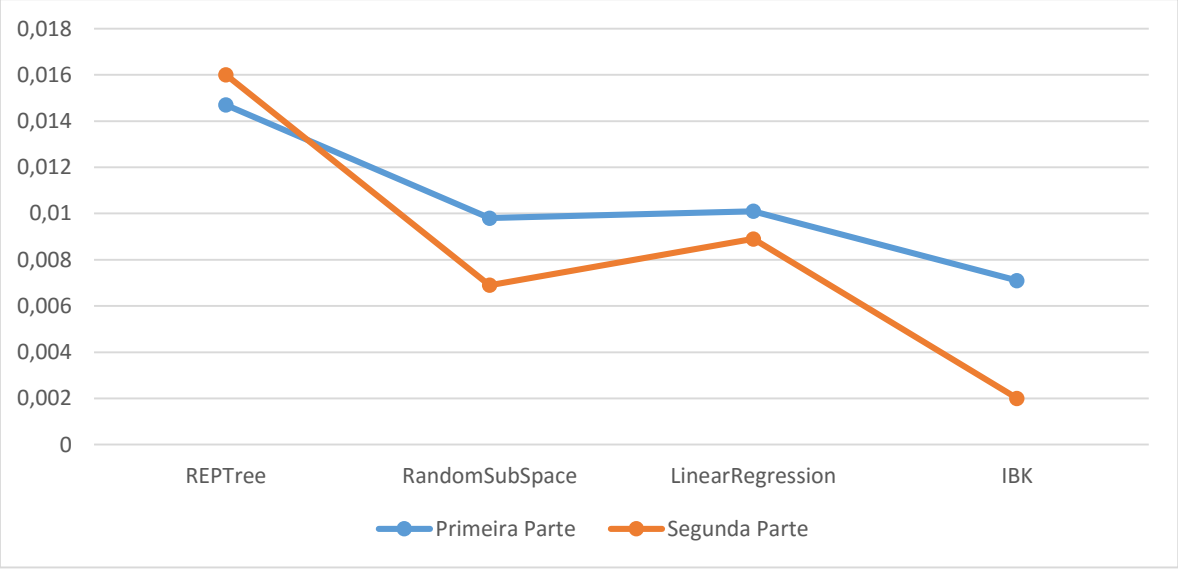


Figura 38 Comparação do valor de RMSE para os modelos matemáticos com melhores resultados na primeira e segunda parte do tempo regulamentar para a experiência com cada jogador a ter a posição x,y numa só variável

Após análise dos gráficos (34 a 37) obtidos a partir da informação recolhida na Tabela VIII confirma-se que em todos os casos o modelo matemático REPTree é o que apresenta piores resultados para o indicador de RMSE. Por outro lado o modelo matemático IBK é o que apresenta melhores resultados em todas as experiências realizadas.

Quanto à avaliação da primeira/segunda parte, os resultados destas comparativamente com a homóloga varia de experiência para experiência e como tal não é possível concluir qual o conjunto de dados serviria melhor para treino do modelo.

É sempre importante lembrar que os valores representados na Tabela VIII são o resultado de uma avaliação *cross-validation* e portanto muito otimistas. Não é expectável que se consigam resultados semelhantes quando os dados de treino e de teste sejam distintos.

6.5. Avaliação

A avaliação é o ponto fulcral para existir progresso efetivo na mineração dos dados. Existem vários métodos distintos para tirar conclusões de dados, mas nem todos são indicados. Segundo Witten & Frank, (2005) a questão do desempenho de previsão baseada num conjunto limitado de dados é interessante e controversa. Apesar da existência de várias técnicas, o *cross-validation* destaca-se das restantes e é, neste momento, o método de avaliação de eleição em situações com poucos dados. Por outro lado, quando os dados disponíveis são muitos, o melhor método será dividir um grande conjunto de dados para treino e outro grande conjunto de dados para teste. No entanto, apesar da grande quantidade de dados, a qualidade destes é, em alguns casos, escassa. Por exemplo, usar um conjunto de dados para treino no qual apenas existe uma situação de pontapé de canto e, de seguida, utilizar para teste um conjunto de dados no qual existem dez situações de jogo deste tipo. Os valores apresentados pelo modelo não serão indicadores de nada, não servirão de referência.

O método de *cross-validation* reserva um certo número de registos para teste e usa o restante para treino. Como este processo é feito de forma aleatória, a amostra para treino pode não ser representativa – uma situação extrema seria nenhuma das instâncias no teste ter sido utilizado para treino. Como tal, é necessário garantir que a amostra aleatória é feita para que todas as situações são propriamente representadas tanto no treino como no teste. Este procedimento é designado de estratificação (Witten & Frank, 2005).

Uma maneira de apaziguar qualquer viés causado por uma amostra em particular para validação é repetir o processo todo com diferentes amostras aleatórias. Em cada iteração, por exemplo, dois terços de dados aleatórios são selecionados para treino e o restante para teste. As taxas de erro para cada iteração são calculados para produzir uma taxa de erro global.

Durante este trabalho serão utilizadas dez iterações. Witten & Frank, (2005) afirma que testes extensivos a inúmeros conjuntos de dados, com diferentes técnicas de aprendizagem, provam que dez é o número ideal de iterações para obter a melhor estimativa de erro.

Obviamente o interesse é o desempenho para prever novas situações e não de dados antigos. Essa informação já é conhecida e por isso mesmo serve como treino para o modelo (Aggarwal, 2015). Logo surge a questão: será a taxa de erro de dados históricos um bom indicador da taxa de erro de novos dados? Numa palavra, não. Não se os dados antigos foram utilizados durante o processo de treino da classe, pois a classe aprendeu a partir dos dados de treino e qualquer estimativa de desempenho baseada nesses dados será extremamente otimista (Witten & Frank, 2005).

Para prever o desempenho de um algoritmo em novos dados, é necessário avaliar a sua taxa de erro de um conjunto de dados que não desempenhou qualquer papel no treino deste. Este conjunto de dados é designado de teste. Evidentemente ambos os conjuntos de dados (treino e teste) são amostras representativas do problema subjacente. Naturalmente a taxa de erro do conjunto de teste apresenta um indicador real de previsão. Quanto maior a amostra de dados de treino e de teste, mais assertivo será esta estimativa.

Em previsões numéricas a forma de avaliar um modelo difere um pouco. O cálculo dos erros é distinto e não é só um indicador, mas sim vários meios de avaliar o sucesso de uma previsão numérica. A forma de cálculo também difere, pois são utilizados dois tipos de variáveis. Os valores de previsão são $p_1,p_2,\ldots,p_n$ e os valores reais são $a_1,a_2,\ldots,a_n$. O valor de p significa o valor numérico da previsão para a n instância de teste (Witten & Frank, 2005).

Mean-squared error (MSE), e *Root Mean-squared Error* (RMSE) são os principais indicadores e os mais utilizados. Muitos modelos matemáticos, tal como a regressão linear, usam este indicador porque tende a ser a medida mais fácil de manipular matematicamente – é matematicamente "bem comportada".

$$\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{n}$$

Equação 1 MSE: Mean-squared error

$$\sqrt{\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{n}}$$

Equação 2 RMSE: Root Mean-squared error

Mean absolute error (MAE) é uma alternativa. Faz a média da magnitude dos erros individuais sem ter em conta os seus operadores. Ao contrário do MSE, que tende a exagerar o efeito dos *outliers* – instâncias no qual a previsão de erro é maior – todos os erros são tratados de forma uniforme com a sua magnitude.

$$\frac{|p_1-a_1|+\ldots+|p_n-a_n|}{n}$$

Equação 3 MAE: Mean absolute error

É possível o erro relativo, em certas situações, ter mais importância do que o absoluto. Por exemplo, se um erro de 10% é igualmente importante num erro de 50 numa previsão de 500 ou um erro de 0.2 numa previsão de 2, então as médias de erro absoluto não fazem sentido – os erros relativos são apropriados.

Relative squared error (RSE) refere-se, no entanto, a uma situação diferente. O erro é calculado em relação ao que seria se um simples modelo matemático tivesse sido usado. O algoritmo em questão é apenas a média dos valores reais dos dados de treino. Assim, o RSE assume o erro e normaliza-o dividindo pelo total do *squared error* do modelo matemático padrão.

$$\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{(a_1-\bar{a})^2+\ldots+(a_n-\bar{a})^2}, \text{ where } \bar{a}=\frac{1}{n}\sum_i a_i$$

Equação 4 RSE: Relative squared error

$$\sqrt{\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{(a_1-\bar{a})^2+\ldots+(a_n-\bar{a})^2}}$$

Equação 5 RRSE: Root relative squared error

Relative absolute error (RAE) é o total de erro absoluto com o mesmo tipo de normalização que o indicador anterior.

$$\frac{|p_1 - a_1| + \ldots + |p_n - a_n|}{|a_1 - \bar{a}| + \ldots + |a_n - \bar{a}|}$$

Equação 6 RAE: Relative absolute error

Somente após o estudo da aplicação é que pode ser determinado qual dos indicadores é o mais apropriado para uma situação definida. Os indicadores de *squared error* e *root squared error* atribuem uma maior importância a grandes discrepâncias de erro do que a pequenas, enquanto os indicadores de *absolute error* não. O *square error* reduz os números para ter a mesma dimensão que a quantidade a ser prevista. Os valores dos *relative errors* tentam compensar pela previsibilidade básica ou imprevisibilidade da variável alvo. Isto é, se esta tende a estar próxima do seu valor médio, então é esperado que a previsão seja boa e o valor relativo seja compensado. Caso contrário se o valor de erro numa situação é muito maior que numa outra situação, pode ser que a quantidade numa primeira situação é inerentemente mais variável e, portanto, mais difícil de prever, e não porque o modelo matemático é pior.

Felizmente, na grande maioria das situações práticas, o melhor método de previsão numérica é o melhor em todos os indicadores de erro (Witten & Frank, 2005).

Existem ainda outras formas de avaliar o desempenho de um modelo, tais como o tempo de construção do modelo de treino, o tempo necessário para usar o modelo, robustez para lidar com ruído nos dados e valores omissos, interoperabilidade pela compreensão e conhecimento munido pelo modelo (Faria, 2013).

Nesta fase do projeto já foram construídos um ou mais modelos que parecem ter grande qualidade numa perspetiva de análise de dados. Antes de avançar para a fase de implementação do modelo é importante avaliar cuidadosamente este e rever os passos executados para a construção do modelo para ter a certeza que alcance os objetivos do negócio. Um objetivo chave é determinar se há algum problema importante que não tenha sido considerado o suficiente. No final desta fase deve ser obtida uma decisão sobre a utilização dos resultados de *data mining* (Wirth & Hipp, 2000).

Foram avaliados inúmeros modelos para os vários cenários pensados. Destes foram sempre selecionados os quatro melhores modelos para cada caso.

Tabela IX Avaliação dos modelos matemáticos da previsão do AA Garcia usando um conjunto de dados para treino correspondente à primeira parte do tempo regulamentar e outro para teste correspondente à primeira parte do prolongamento

	AdditiveRegression	RandomSubSpace	LeastMedSq	IBK
Correlation Coe.	0.9585	0.9815	0.9587	0.9398
MAE	0.0678	0.0541	0.0595	0.0676
RMSE	0.0798	0.0638	0.0775	0.088
RAE	34.2013%	27.3105%	30.0052%	34.1154%
RRSE	33.9036 %	27.0972%	32.9405%	37.3836%

Tabela X Avaliação dos modelos matemáticos da previsão do AA Garcia usando *cross-validation* correspondente à primeira parte do tempo regulamentar e à primeira parte do prolongamento

	REPTree	M5P	IBK	IBK*
Correlation Coe.	0.9992	0.999	1	1
MAE	0.0057	0.0079	0.0007	0.0007
RMSE	0.0102	0.0114	0.0011	0.0011
RAE	2.7171%	3.7691%	0.3332%	0.3336 %
RRSE	4.0222%	4.513 %	0.4236%	0.4243 %

*com dados ordenados aleatoriamente

Através da análise das tabelas 9 e 10 conseguimos perceber, com resultados práticos, a diferença da utilização de avaliação dos modelos matemáticos recorrendo a um conjunto de dados para treino (primeira parte do tempo regulamentar – 45 minutos de jogo) e um conjunto de dados para teste

(primeira parte do prolongamento – 15 minutos de jogo) sem qualquer participação na fase de aprendizagem do modelo, para uma avaliação *cross-validation*.

Para esta análise e com o auxílio da Figura 39 é constatável que o algoritmo com melhores resultados é o *RandomSubSpace* em todos os indicadores.

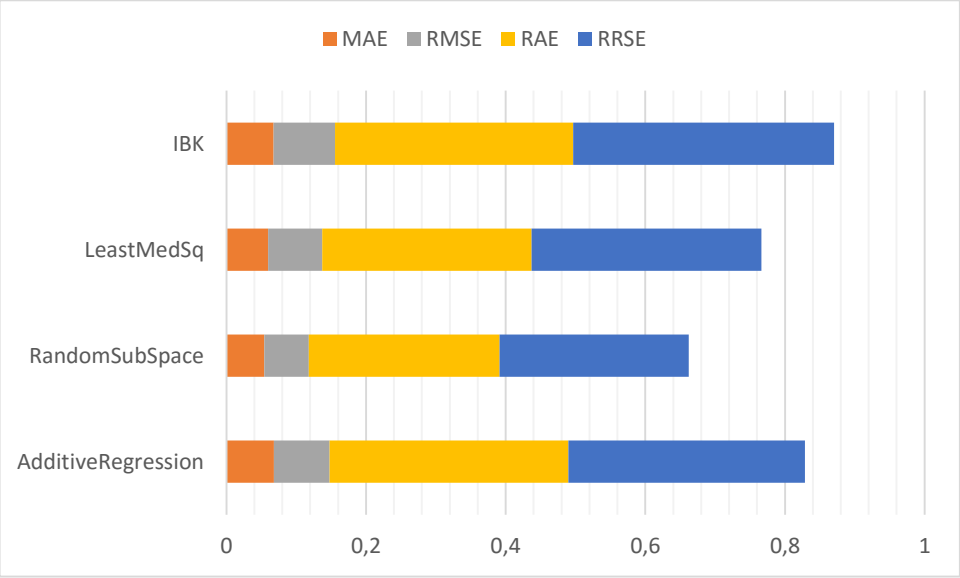


Figura 39 Avaliação dos modelos matemáticos da previsão do AA Garcia. Gráfico representativo da Tabela IX

O gráfico da Figura 39, que representa os valores da Tabela IX, não considera o coeficiente de correlação, uma vez que quanto mais próximo de um este indicador se encontrar, melhor. Ao contrário dos restantes indicadores, que se pretendem que sejam o mais baixo possíveis. O algoritmo *RandomSubSpace* apresenta uma menor taxa de erro absoluto e relativo surgindo naturalmente com a barra menor.

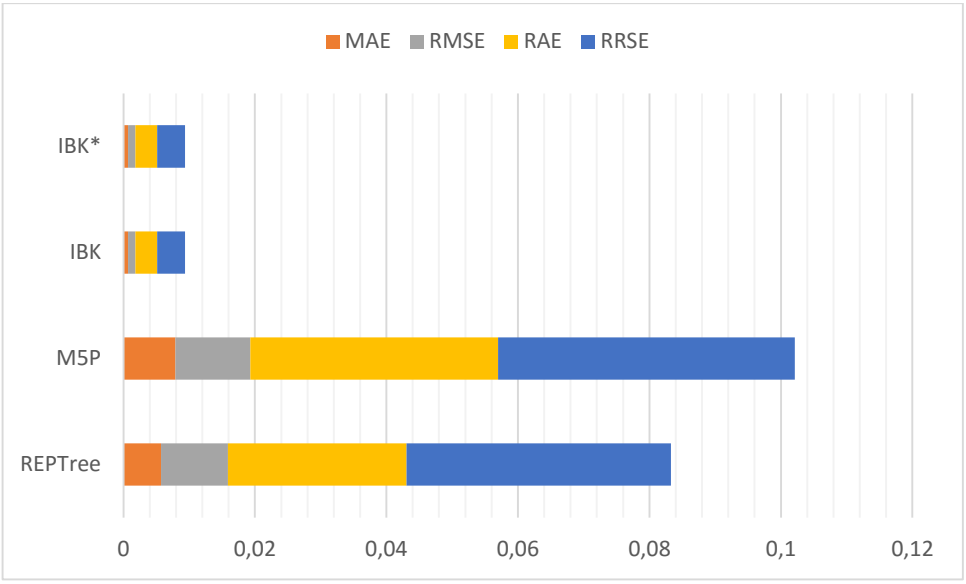


Figura 40 Avaliação dos modelos matemáticos da previsão do AA Garcia. Gráfico representativo da Tabela X

Analisando a Tabela X e o gráfico da Figura 40 resultante da mesma avaliação feita na Tabela IX e gráfico 38 com a diferença do uso do *cross-validation* como método de avaliação e não dos dados de treino e teste, percebe-se que os resultados dos modelos são imensamente melhores.

Esta grande diferença (o mesmo algoritmo, *IBK*, teve uma melhoria de 33.7822% de erro absoluto) deve-se, sobretudo, à falta de exemplos de treino para os dados de teste. Isto é, apesar de existirem dados correspondentes a 45 minutos de jogo para treino e os dados de teste incidirem em 15 minutos de jogo, um jogo de futebol tem muitas fases e essas alterações fazem-se sentir ao longo do jogo. Uma mudança óbvia do período de tempo correspondente ao treino para o período de tempo correspondente ao teste é o cansaço físico e psicológico dos atletas (nesta fase muitas equipas já só pensam nos pontapés da marca de grande penalidade). As alterações táticas que os treinadores foram ajustando para melhor anular a equipa adversária e substituições alteram a forma da equipa de jogar. Por último, e após uma análise ao jogo, a primeira parte do tempo regulamentar é equilibrada, com ambas as equipas a disputar o jogo por igual enquanto na primeira parte do prolongamento "só" a França atacou. É mais ou menos natural, então, que os dados de treino não prevejam todas as situações existentes nos dados de teste.

Por outro lado, o modelo matemático *IBK - Nearest Neighbour* – apresenta para o mesmo conjunto de dados uma taxa de erro mínima, a roçar o modelo perfeito, pois utiliza os mesmos dados de teste para treino – como explicado no início deste capítulo. Para o algoritmo *IBK* também foi feita a

experiência de ordenar de forma aleatória todos os dados para perceber de que forma este modelo matemático melhorava/piorava a sua avaliação e a diferença foi mínima.

Tabela XI Avaliação dos modelos matemáticos da previsão do AA Garcia usando um conjunto de dados para treino correspondente à primeira parte do tempo regulamentar e outro para teste correspondente à segunda parte do tempo regulamentar

	IsotonicRegression	Bagging	AdditiveRegression	LeastMedSq
Correlation Coe.	0.9523	0.9449	0.9451	0.9415
MAE	0.0544	0.0667	0.0662	0.0694
RMSE	0.0802	0.0881	0.0861	0.0911
RAE	23.9738 %	29.4011 %	29.2079%	30.6107%
RRSE	29.3223 %	32.1991 %	31.469 %	33.3284%

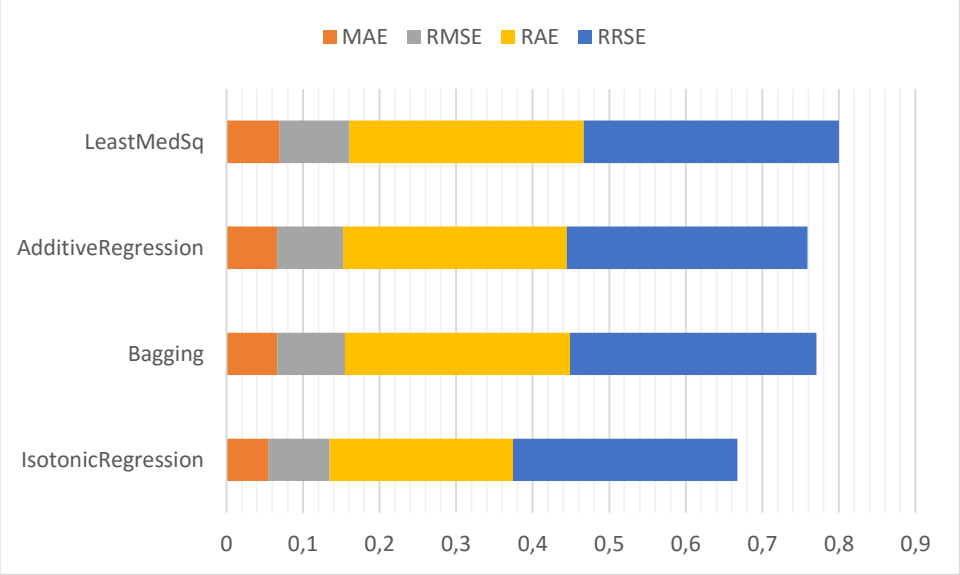


Figura 41 Avaliação dos modelos matemáticos da previsão do AA Garcia. Gráfico representativo da Tabela XI

Tabela XII Avaliação dos modelos matemáticos da previsão do AA Garcia usando *cross-validation* correspondente à primeira parte do tempo regulamentar e à segunda parte do tempo regulamentar

	REPTree	M5P	RandomSubSpace	IBK
Correlation Coe.	0.999	0.9989	0.9997	1
MAE	0.0058	0.0084	0.0037	0.0007
RMSE	0.0115	0.0124	0.0063	0.001
RAE	2.6551 %	3.8154 %	1.6738 %	0.3014%
RRSE	4.4116 %	4.7587 %	2.4211 %	0.3989%

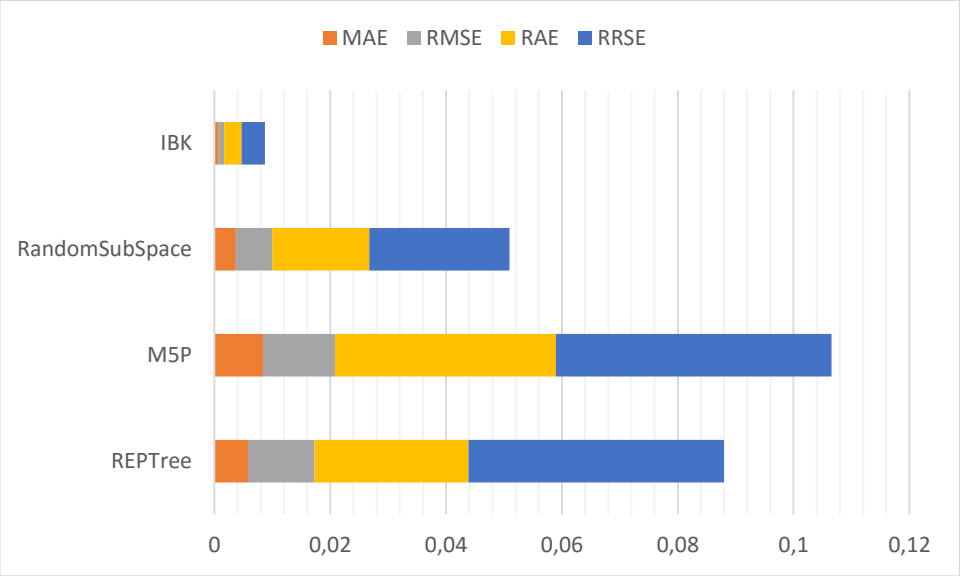


Figura 42 Avaliação dos modelos matemáticos da previsão do AA Garcia. Gráfico representativo da Tabela XII

Esta segunda experiência serve para perceber como o modelo se comporta ao treinar o AA com uma equipa (França) e depois fazer o teste noutra equipa (Itália). De acordo com o gráfico, o melhor modelo matemático foi o *IsotonicRegression*.

Estas taxas de erro são justificadas pela razão óbvia de nem todas as equipas se movimentarem da mesma forma. Considerando que a taxa de erro diminuiria à medida que existissem mais dados para treino, seria de esperar que estes valores fossem mais baixos, uma vez que o AA, independentemente

de como uma equipa ataca ou defende, deve-se colocar sempre em linha com o penúltimo defensor. Como tal, a aprendizagem de uma equipa de futebol deveria ser suficiente para todas as outras.

Tal como verificado no caso anterior, o método *cross-validation* consegue obter, uma vez mais, resultados extraordinariamente positivos. Apesar de os resultados serem otimistas, não deixam de evidenciar que com um maior conjunto de dados de treino este modelo conseguirá obter excelentes resultados.

Segue-se a avaliação referente ao modelo do posicionamento do árbitro. Para esta análise foram consideradas várias experiências para tentar perceber qual seria a melhor forma de obter um modelo para o posicionamento do árbitro respeitando as posições da bola, jogadores e assistentes.

Tabela XIII Avaliação dos modelos matemáticos da previsão do árbitro Elizondo usando um conjunto de dados para treino correspondente à primeira parte do tempo regulamentar e outro para teste correspondente à segunda parte do tempo regulamentar.

	RandomSubSpace	Bagging	IBK
Correlation Coe.	0.923	0.8926	0.8429
MAE	0.0966	0.0995	0.117
RMSE	0.1221	0.127	0.1479
RAE	50.874 %	52.3976%	61.609 %
RRSE	52.7546 %	54.8904%	63.9193 %

Esta experiência foi pensada com o intuito de perceber como o posicionamento do árbitro no terreno de jogo era influenciada pelas posições dos jogadores com um valor para a variável x e outro para a variável y. Ou seja, cada jogador é representado por exemplo como: *Zidane_X* e *Zidane_Y*. A exceção foi feita para os AA nos quais apenas foi considerado o posicionamento de *x*, pois o valor de *y* é praticamente fixo.

Foi utilizada a 1ª parte para treino e a 2ª parte para teste.

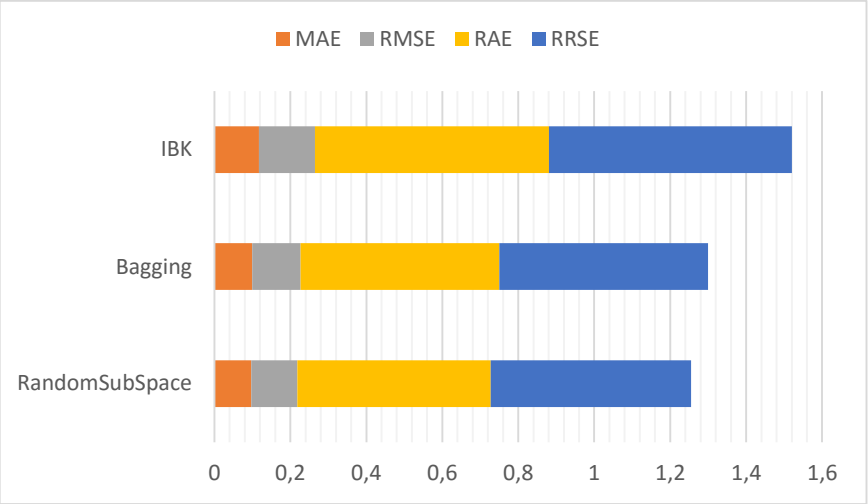


Figura 43 Avaliação dos modelos matemáticos da previsão do árbitro Elizondo. Gráfico representativo da Tabela XIII

De acordo com a Figura 43 o melhor resultado foi conseguido pelo modelo matemático RandomSubSpace. Contudo, os valores apresentados não são satisfatórios. Estes, no entanto, podem ser justificados pelo insuficiente número de casos para treino.

Tabela XIV Avaliação dos modelos matemáticos da previsão do árbitro Elizondo usando um conjunto de dados para treino correspondente à primeira parte do tempo regulamentar e outro para teste correspondente à segunda parte do tempo regulamentar. Somente para a coordenada x

	LeastMedSq	RandomSubSpace	Bagging
Correlation Coe.	0.9449	0.9239	0.837
MAE	0.0893	0.1066	0.1137
RMSE	0.1122	0.136	0.1492
RAE	47.0419%	56.1366%	59.9029%
RRSE	49.0194%	59.4113%	65.217%

Esta experiência elaborada consistiu em restringir os valores de todos os intervenientes à variável *x*. Ou seja, um jogador é representado apenas por *jogador_X*. O objetivo deste teste é perceber de que

forma, a movimentação dos atletas em terreno de jogo no plano horizontal do terreno afeta o posicionamento do árbitro e como este pode ser previsto em situações semelhantes.

Esta experiência resultou em valores algo semelhantes à realizada para as variáveis de *x* e *y*.

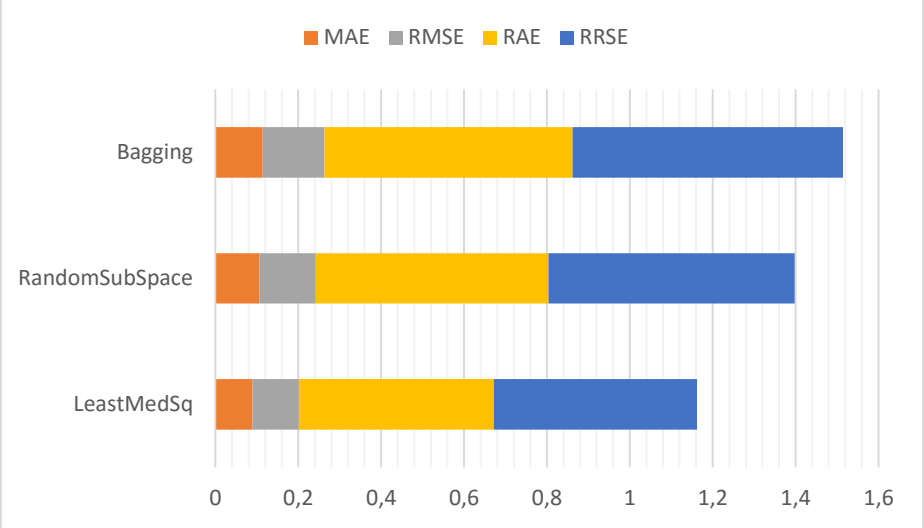


Figura 44 Avaliação dos modelos matemáticos da previsão do árbitro Elizondo. Gráfico referente à Tabela XIV

O modelo que conseguiu melhores resultados foi o *LeastMedSq* para todos os indicadores. O coeficiente de correlação entre os atributos é quase perfeitamente positiva. Este valor é compreensível uma vez que os jogadores tendem a deslocar na mesma direção. Ou seja, a equipa que ataca sobe as linhas e a equipa que defende desce as linhas. O árbitro, naturalmente, acompanha este movimento, assim como os assistentes.

O motivo apontado para as taxas de erro calculadas é a falta de dados de treino.

Tabela XV Avaliação dos modelos matemáticos da previsão do árbitro Elizondo usando um conjunto de dados para treino correspondente à primeira parte do tempo regulamentar e outro para teste correspondente à segunda parte do tempo regulamentar. Somente para a coordenada y

	LeastMedSq	Bagging	RandomSubSpace
Correlation Coe.	0.6361	0.3958	0.5073
MAE	0.2393	0.233	0.2104
RMSE	0.277	0.2834	0.2568

RAE	114.4756 %	111.4654 %	100.6455 %
RRSE	111.4282 %	114.0168 %	103.327 %

Esta experiência é semelhante à organizada anterior correspondente à Tabela XV. A diferença ocorreu na restrição da variável. Ou seja, a restrição foi feita ao valor de *y*. Por exemplo, *jogador_Y*. Os resultados foram muito fracos e não houve nenhum modelo com valores aceitáveis.

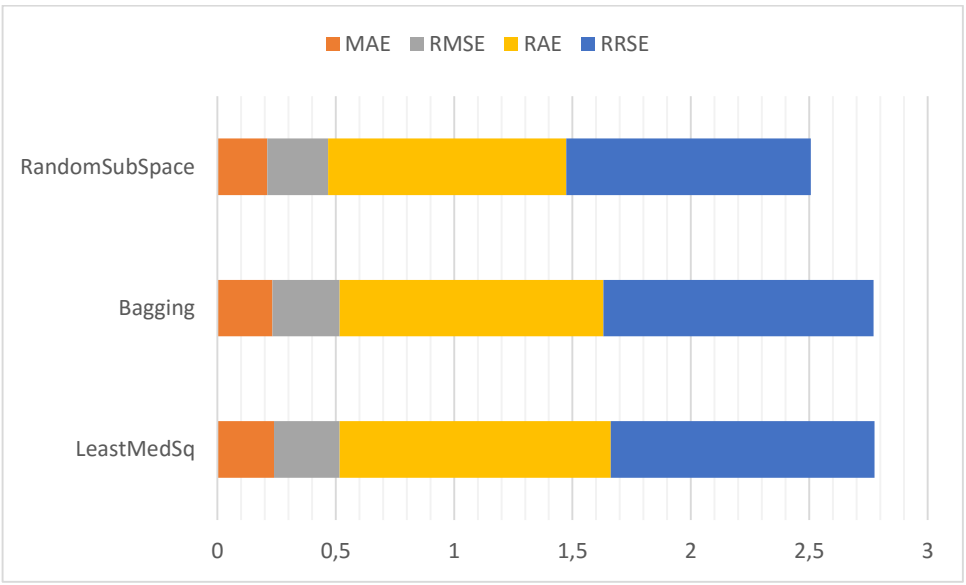


Figura 45 Avaliação dos modelos matemáticos da previsão do árbitro Elizondo. Gráfico referente à Tabela XV

De todos os modelos matemáticos o que obteve resultados menos negativos foi o *RandomSubSpace*. Verifica-se que o posicionamento dos jogadores no plano vertical (variável *y*) não é suficiente para determinar o posicionamento do árbitro em terreno de jogo.

Tabela XVI Avaliação dos modelos matemáticos da previsão do árbitro Elizondo usando um conjunto de dados para treino correspondente à primeira parte do tempo regulamentar e outro para teste correspondente à segunda parte do tempo regulamentar. Valores de x e y juntos numa variável

	LeastMedSq	**RandomSubSpace**	**IBK**
Correlation Coe.	0.9392	0.9242	0.9075
MAE	0.0765	0.0836	0.0979

RMSE	0.1039	0.116	0.1301
RAE	29.2208%	31.9444%	37.4031%
RRSE	34.5298%	38.5545%	43.2347%

Para esta experiência foram utilizados os valores referentes às posições de *x* e *y* numa só variável. Uma vez mais foi utilizado a primeira parte para treino e a segunda parte para teste.

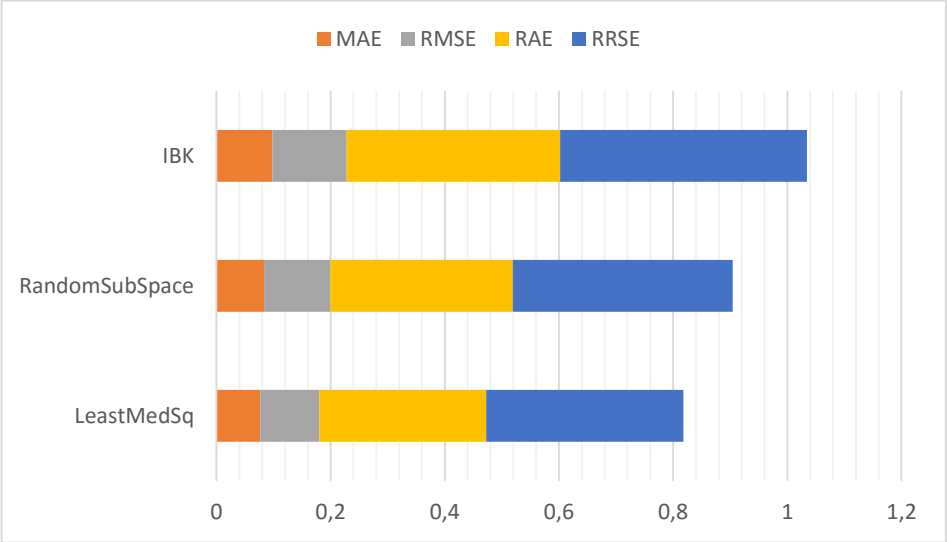


Figura 46 Avaliação dos modelos matemáticos da previsão do árbitro Elizondo. Gráfico referente à Tabela XVI

O algoritmo com melhores resultados é o *LeastMedSq*. Este modelo matemático, para além de detetar um grande coeficiente de correlação nos dados é capaz de prever o posicionamento do árbitro em terreno de jogo com uma taxa de erro absoluto de 29%. Este foi, inclusive, o melhor modelo para a análise feita às várias experiências.

Tabela XVII Avaliação dos modelos matemáticos da previsão do árbitro Elizondo numa situação de pontapé de canto. Usado para treino duas situações semelhantes.

	IBK	**KStar**
Correlation Coe.	0.9883	0.9727
MAE	0.0662	0.0898

RMSE	0.0778	0.1133
RAE	18.2653%	24.7615%
RRSE	19.0358%	27.706 %

Esta análise foi elaborada com o intuito de avaliar o modelo para um caso de jogo específico. O lance de jogo selecionado foi o pontapé de canto por ser um momento do jogo que é facilmente detetado nos dados. Foram usados para treino quatro situações de pontapé de canto e uma situação para teste do mesmo tipo de lance. Esta é a avaliação mais esclarecedora do modelo, uma vez que é uma situação controlada e os dados de treino são úteis para a análise pretendida.

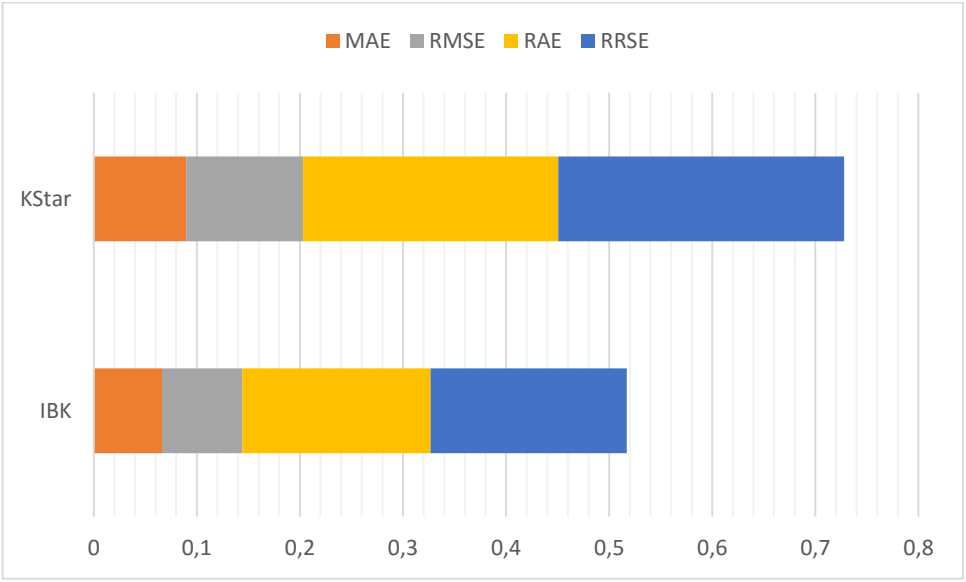


Figura 47 Avaliação dos modelos matemáticos da previsão do árbitro Elizondo. Gráfico representativo da Tabela XVII

Após a análise ao gráfico é possível concluir que este é um modelo com uma taxa de sucesso de previsão muito positiva. É possível concluir, com este resultado, que o algoritmo *K Nearest Neighbour* consegue com mais dados de treino (mais situações de pontapés de canto) conseguiria prever com uma percentagem ainda mais alta de sucesso a localização ideal do árbitro.

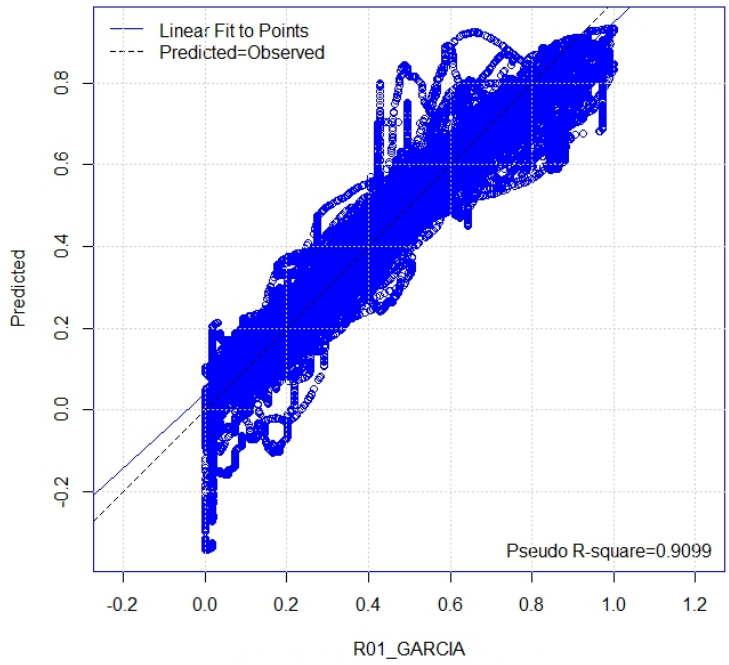


Figura 48 Regression Scatter Plot (RSP) comparando os valores previstos com os observados para o AA Garcia.

O gráfico *Regression Scatter plots*, usado para fazer uma comparação entre o previsto e o observado, ou vice-versa, é uma das principais alternativas para avaliar os modelos de previsão (Piñeiro, Perelman, Guerschman, & Paruelo, 2008). De acordo com Witten & Frank, (2005) o uso de um conjunto de dados externo garante que o modelo apenas treina com um conjunto de dados para depois avaliar um outro conjunto de dados. Esta justificação explica a razão de os valores representativos do RMSE na Tabela XVIII serem consideravelmente maiores para o caso supracitado.

O gráfico RSP dispõe os valores previstos comparativamente com os valores reais. Para além destes valores representados pelos círculos azuis, são representadas duas linhas. Uma é o ajuste linear para os pontos reais e a outra corresponde ao ajuste perfeito se os valores previstos forem os mesmos que as observações. O *Pseudo R-square* é uma medida que tenta imitar o *R-squared*. É calculado como o quadrado da correlação entre os valores previstos e observados. Quanto mais perto for este valor, melhor. Devido à grande quantidade de dados avaliados, o gráfico tem uma aparência confusa, uma vez que cada ponto valor corresponde a um círculo azul.

Através da análise da Figura 47 é percetível uma grande correlação entre a posição prevista e a posição real. Este é um modelo eficiente e consegue prever com precisão elevada, qual a posição ideal para o AA em terreno de jogo de forma a reduzir o número de erros.

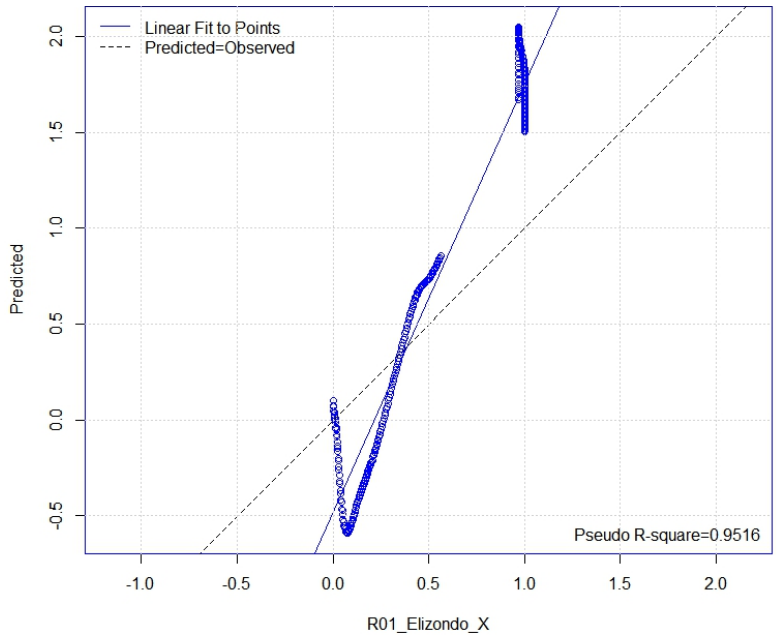


Figura 49 RSP comparando os valores previstos com os observados para o árbitro num pontapé de canto

Através da informação disponibilizada pela ferramenta Rattle aquando da execução do gráfico RSP para comparar os valores previstos com os valores observados conseguimos concluir que este modelo consegue prever com um bom grau de certeza (*pseudo R-square* muito próximo de 1*)* a posição do árbitro num pontapé de canto.

Foi feita uma análise aos diferentes tipos de avaliação disponibilizados pelo *software* Weka e analisado o comportamento destes, para o mesmo conjunto de dados de treino e avaliação. A Tabela XVIII mostra essa mesma discrepância de valores para o indicador RMSE. Como é possível confirmar, a avaliação *cross-validation* é a que apresenta melhores resultados pois, uma vez mais, garante que todos os dados de teste estão também nos dados de treino. Desta forma o modelo matemático consegue fazer previsões com baixa percentagem de erro.

O uso de dados externos para teste (garante que estes dados não têm qualquer interação durante a fase de aprendizagem), é o que tipo de avaliação que apresenta taxas de erro mais altas. Estas taxas de erro são justificadas por situações que decorrem durante o período de tempo de jogo utilizado para teste que não aconteceram no treino. Um exemplo disso é a expulsão do Zidane já perto do final do jogo (foi a única expulsão do jogo e como tal não existe uma situação semelhante nos dados de treino). Durante este período de tempo todos os intervenientes têm um comportamento totalmente

distinto do normal. Por exemplo, ninguém presta atenção à bola, estão todos no mesmo local e, naturalmente, as taxas de erro aumentam. A divisão controlada de 70% dos dados para treino e os restantes para teste apresenta uma taxa de erro ligeiramente maior que o *cross-validation* mas muito menor que os dados externos pois os dados para teste, aleatoriamente selecionados pela ferramenta, não coincidiram com situações que não foram aprendidas previamente pelo modelo.

Tabela XVIII Comparação dos valores de RMSE da avaliação elaborada

	Dados de treino	Dados externos	Cross-Validation (10)	Divisão controlada (70%)
		AA - Garcia		
M5P	0.0105	0.0971	0.0124	0.0133
REPTree	0.0082	0.0943	0.0107	0.013
		AA - Otero		
M5P	0.0161	0.238	0.0216	0.0251
REPTree	0.0133	0.2493	0.0215	0.024
		Elizondo		
IBK	0	0.0778	0.0069	0.0015
KStar	0.0001	0.1133	0.0068	0.0009

Infelizmente devido ao ruído dos dados do AA, Otero, não será possível fazer a análise deste. Logo ficamos limitados ao outro AA, Garcia, para a construção do modelo para a posição ideal de AA.

6.6. Implementação

Geralmente a criação de um modelo não é o final do projeto. Normalmente, o conhecimento adquirido necessita de ser organizado e apresentado numa forma que o cliente consiga utilizar. Dependendo dos requisitos, a fase de implementação pode ser tão simples como gerar um relatório ou tão

complexa como implementar um processo de *data mining* repetitivo. Em muitas situações será o utilizador e não o analista de dados que realiza os passos de implementação. Em qualquer caso, é importante perceber desde o início que serão necessárias ações para realmente fazer uso dos modelos criados (Wirth & Hipp, 2000).

Para o desafio enfrentado e após a consecução de um modelo capaz de determinar o posicionamento correto da equipa de arbitragem é necessário implementá-lo. Esta fase, para este problema, consiste no cumprimento de um processo de *data mining* repetitivo.

Para a realização deste processo é necessário a obtenção de dados relativos a várias competições internacionais arbitrados por árbitros experientes. Após este passo é necessário uma análise de especialistas que observem o posicionamento do árbitro e o corrijam se necessário. Desta forma é criado um largo conjunto de dados de qualidade para servirem de treino para os modelos em questão. Assim, o modelo tornar-se cada vez mais eficiente na previsão da localização ideal do árbitro de futebol para situações de jogos futuros.

Este é um modelo com grande valor para, por exemplo, academias de arbitragem.

A próxima figura pretende ilustrar o processo necessário para a obtenção deste modelo.

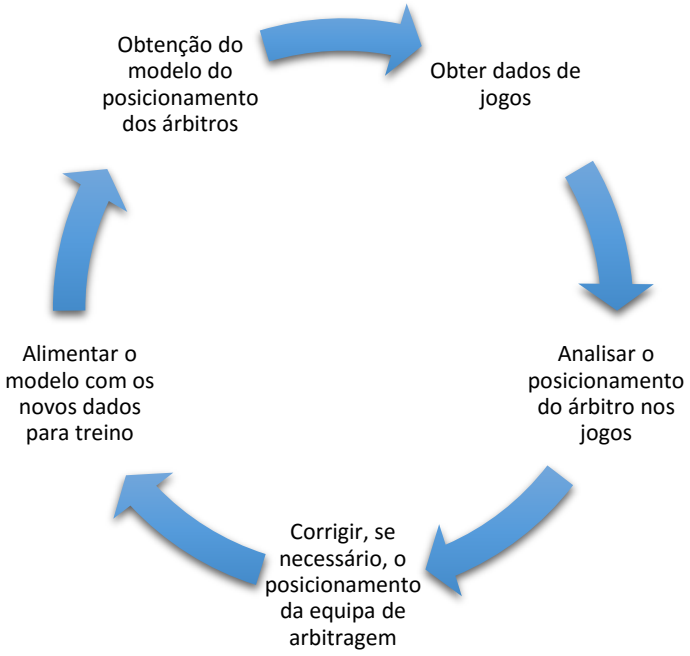


Figura 50 Workflow para a implementação do modelo – processo repetitivo de *data mining*

6.7. Conclusões

Durante o desenvolvimento do modelo, baseado na metodologia CRISP-DM, foram executadas várias tarefas numa ordem lógica que permitiu chegar a uma conclusão.

Numa primeira instância foi feito um estudo do negócio, isto é, qual o desafio proposto, os objetivos e por último a ação para se conseguir alcançar tais metas.

O segundo passo para a construção deste modelo coincidiu com o estudo dos dados. Esta foi uma tarefa muito importante uma vez que consistiu em compreender como os dados estavam construídos, o formato e a qualidade destes dados. Isto é, uma vez que não houve participação na extração destes dados, era necessário perceber se estes tinham muitos erros, como era processado quando um jogador, árbitro ou bola saiam do campo de visão das câmaras, o processo de substituições e a margem de erro. Ou seja, comparando as posições dos vários elementos em campo com imagens reais do jogo e perceber qual era a diferença entre a posição nos dados e a posição efetiva.

A preparação dos dados consistiu na adaptação destes ao desafio anunciado no estudo do negócio e comportou algumas alterações aos dados, nomeadamente o tratamento de valores omissos, *outliers* e eliminação de atributos em conformidade com as experiências em mão. Outra alteração aos dados foi a correção da posição do árbitro em lances duvidosos e nos quais o árbitro tomou a decisão errada. Estes dados foram alterados para a posição que deveria ser mais correta e que o teria auxiliado a tomar a decisão correta. Este passo foi feito com a ajuda do árbitro da primeira liga (Liga NOS) e serve para melhorar o modelo de aprendizagem ao máximo.

A fase da modelação serviu, para além da aprendizagem dos softwares Weka e Rattle, para perceber que tipo de modelos matemáticos melhor se ajustavam para a solução do problema, quais as variáveis a considerar e quais deveriam ser removidas da análise. É nesta fase que são enumeradas, também, as várias experiências a realizar na fase seguinte da metodologia.

A quinta fase, a avaliação, como o próprio nome indica, serve para fazer a avaliação da eficácia dos modelos de previsão. Esta fase foi também aproveitada para compreender como as taxas de erro dos vários modelos para problemas de previsão numérica são calculadas e portanto, qual o relevo que estas têm no resultado final do modelo. É, ainda, feita uma análise aos vários métodos de avaliação existentes, com o *cross-validation* e o conjunto de dados de treino e teste.

Por fim, na implementação, o conhecimento adquirido pelos modelos é organizado e apresentado.

7. Resultados e Discussão

Este capítulo serve como apresentação dos resultados. São primeiramente indicados os resultados esperados, seguidos dos resultados obtidos. Assim estamos em condições para de seguida comparar os resultados esperados dos obtidos.

7.1. Resultados esperados

Durante este trabalho foram desenvolvidos dois modelos para o posicionamento ideal do árbitro em terreno de jogo. Um modelo para o posicionamento do AA e outro modelo para o posicionamento do árbitro.

Os resultados esperados para cada um destes modelos são distintos devido à complexidade de um modelo comparativamente com o outro.

O modelo referente ao AA é considerado mais simples devido às instruções da FIFA, que refere que estes se devem colocar sempre em linha com o penúltimo defensor ou com a bola se esta estiver mais próxima da linha de baliza. Obviamente que existe uma margem de tolerância aceitável (de acordo com o capítulo 3.4 o AA encontra-se em condições de decidir corretamente se estiver apenas a 0.81m atrasado e 0.77m adiantado) e esta figura de AA é um ser humano que tem limitações físicas e acompanhar a bola é, por exemplo, impossível em algumas situações. Todavia, é expectável que o modelo apresente taxas de erro pequenas.

Relativamente ao modelo referente ao posicionamento do árbitro é considerado um exercício com um grau de complexidade muito maior. Isto porque ao contrário do AA no qual a sua posição depende apenas de uma equipa e/ou da bola e apenas se movimenta na coordenada x, o árbitro tem uma liberdade muito maior e pode chegar a todos os pontos do terreno durante um jogo. A sua posição é também dependente da posição e movimentação da bola, de ambas as equipas e inclusive da posição dos AA. São, portanto, muitas variáveis com influência. Existe contudo, um padrão na movimentação deste, tal como a diagonal que é uma recomendação da FIFA.

É esperado que se chegue a um modelo satisfatório e que se comprove que com dados de mais jogos para treino o modelo ficaria ainda mais robusto. Apesar de não ser esperado taxas de erro tão baixas como para os modelos dos AA, é, contudo, previsto obter um modelo muito positivo para situações

específicas do jogo que serão usadas para treino e posteriormente para avaliação de uma situação semelhante.

7.2. Resultados obtidos

Após proceder à modelação e avaliação dos vários modelos matemáticos é possível observar os resultados obtidos dos vários exercícios elaborados ao longo do trabalho.

Para os modelos referentes aos AA foram conseguidos alguns modelos com pequenas taxas de erro e resultados muito satisfatórios para a quantidade de dados em mão.

Foi conseguido um modelo com 73% de sucesso. Este valor foi conseguido usando a primeira parte para treino e a primeira parte do prolongamento para teste. Ou seja, o AA em questão estava a acompanhar a mesma equipa tanto para um conjunto de dados como para outro.

O resultado de 77% de sucesso para a experiência de usar a primeira parte para treino e a segunda para teste valida o objetivo deste trabalho. Isto é, usando um conjunto de dados para treino do AA a acompanhar uma equipa (França), e o conjunto de dados de teste quando este acompanhava outra equipa (Itália), confirma que este modelo pode ser aplicado a qualquer situação. Ou seja, com uma volumetria de dados de treino suficiente, é possível prever o posicionamento do AA em qualquer jogo, para qualquer equipa.

Tal como esperado, os modelos conseguidos para o árbitro apresentam taxas de erro muito mais altas do que o conseguido para o AA. Isto é, se for feita uma análise semelhante à elaborada ao AA – uso de um conjunto de dados abrangente para treino como os 45 minutos iniciais e outro amplo conjunto de dados para teste, por exemplo os segundos 45 minutos - foram alcançados resultados nos quais a percentagem de sucesso era inferior a 50%. Todavia, quando foi feita uma análise a situações específicas do jogo, como um pontapé de canto, os resultados alcançados apresentam uma percentagem de sucesso de 82%.

7.3. Discussão dos resultados

Após a elaboração de todos os modelos e reflexão sobre os resultados obtidos é tempo de discutir estes e, acima de tudo, compará-los com os resultados esperados.

O modelo obtido para o posicionamento do AA em terreno de jogo foi de encontro com o esperado e conclusivo de que este pode ser aplicado a diferentes jogos. Isto é, o modelo consegue aprender a partir de um conjunto de dados e aplicar a previsão com sucesso a outros jogos, independentemente da equipa e do AA. O uso do método de avaliação *cross-validation* no qual são obtidos valores muito altos de sucesso pode ser indicativo de que, com um grande volume de dados de treino é possível prever com eficácia a posição do AA.

Por sua vez o modelo de posicionamento para o árbitro apresenta resultados menos positivos. Estes resultados, contudo, vão de acordo com o esperado pois a complexidade do modelo para o árbitro é muito maior.

No entanto, se for utilizado um conjunto de dados relevantes para o exercício em questão, entenda-se reduzir o ruído dos dados, e treinar o modelo com estes mesmos modelos podem-se obter resultados muito interessantes.

Devido à falta de estudos relacionados com esta temática, e sendo este um modelo completamente recente é capaz de prever, com sucesso, onde é que o árbitro se deve colocar no terreno de jogo para, numa fase inicial, lances de bola parada – pontapés de canto, pontapés de baliza, pontapés de grande penalidade, livres diretos e indiretos em locais frontais à baliza, nos cantos da área de grande penalidade ou faltas cometidas perto do meio campo. Numa versão futura, e com dados de treino suficientes, este modelo deverá ser capaz de prever a localização do árbitro e o movimento deste para situações mais complexas, tais como contra-ataques com superioridade numérica de uma equipa, reagir a diversas situações do jogo dependendo do comportamento dos jogadores.

Com estes resultados obtidos é possível confirmar que é possível fazer a previsão com sucesso da posição ideal do árbitro para situações específicas do jogo. Assim, com este modelo é possível determinar qual o posicionamento ideal em lances, tais como, pontapés de canto, lançamentos laterais, pontapés de baliza. Para isso é necessário que os dados de treino tenham bons exemplos das situações descritas.

8. Conclusões e Trabalho Futuro

O último capítulo desta dissertação tem como objetivo fazer o balanço final deste trabalho.

Dessa forma o capítulo centra-se nos contributos do trabalho realizado, assim como nas limitações que foram sentidas para o desenvolvimento deste. Para finalizar são anunciadas todas as conclusões alcançadas e os trabalhos futuros.

8.1. Contributos do trabalho realizado

Os principais contributos deste trabalho são referentes a aspetos teóricos, mas com o objetivo de serem aplicados a situações práticas.

O estudo realizado contribuiu para a construção de um modelo para o posicionamento da equipa de arbitragem em terreno de jogo. Para isso são usados dados de equipas de arbitragem experientes com jogos importantes e com as devidas correções no posicionamento destes quando erravam. Estas correções foram elaboradas, após várias visualizações atentas do jogo de futebol em questão. Os lances erradamente avaliados foram corrigidos por um especialista na área, o AA da primeira liga, Jorge Oliveira.

Este modelo vem também justificar a importância de um bom posicionamento em terreno de jogo para a tomada da decisão correta.

8.2. Limitações do trabalho

Para o desenvolvimento deste trabalho foram enfrentadas algumas limitações que impediram a realização de novas experiências e aprendizagens mais completas para a construção do modelo.

Infelizmente só foi conseguida informação das posições dos jogadores e árbitros de um jogo de futebol de grande nível (para os dados de treino serem o melhor possível). Outra limitação sentida foi a impossibilidade de obter dados em campeonatos de futebol amadores por falta de meios e da proibição do uso de dispositivos nos atletas.

8.3. Conclusões

Este trabalho de investigação tinha como objetivo a construção de um modelo formal que determinasse o melhor posicionamento da equipa de arbitragem no terreno de jogo em conformidade com a posição da bola, dos jogadores e inclusive, dos restantes membros da equipa.

Para alcançar este objetivo foram estudados vários algoritmos de *data mining*. Para este modelo foram utilizados dados referentes a um jogo de futebol profissional, a final do Mundial de 2006, que continha informação das posições cartesianas (*x,y*) de todos os intervenientes.

Os objetivos foram atingidos e foram construídos dois modelos matemáticos que determinam qual o posicionamento ideal do árbitro e dos seus assistentes em terreno de jogo. Concluiu-se para o AA que o modelo elaborado consegue determinar com uma precisão de 77% a posição deste independentemente da equipa que este acompanha ou do jogo. Para este modelo foram utilizados os 90 minutos de jogo para treino e os 30 minutos do prolongamento para teste numa situação e os primeiros 45 minutos para treino numa e os segundos 45 minutos para teste noutra situação. Para a determinação do modelo do árbitro foram utilizadas situações concretas de jogo, como pontapés de canto, para prever a posição correta deste com um sucesso de 82%.

Este estudo vem acrescentar um novo conhecimento numa área na qual não existia nada semelhante. Os resultados obtidos para os modelos de previsão da equipa de arbitragem podem acrescentar um enorme saber a novos, e mesmo experientes, árbitros de futebol. Estes modelos podem, e devem, ser utilizados em academias de arbitragem de forma a melhor formar os seus formandos.

Convém, contudo, realçar que os resultados obtidos para estes modelos foram alcançados a partir um conjunto de dados referente a um único jogo. Esta limitação condicionou, em parte, os resultados atingidos. Um maior conjunto de dados contribuiria para resultados ainda mais positivos.

Os resultados conseguidos abrem portas a novas pesquisas nesta área, ainda inexplorada, e incentiva também a um maior investimento na área da deteção dos jogadores por métodos não intrusivos.

Concluindo, os resultados obtidos nestes modelos poderão contribuir para uma redução dos erros em jogos de futebol resultantes do mau posicionamento da equipa de arbitragem em terreno de jogo.

8.4. Trabalhos Futuros

As propostas de trabalho futuro resultam das necessidades identificadas, e que, por motivos diversos, não foram contempladas nesta dissertação. Contudo, devido à relevância destas, devem ser referidas com a intenção de abrir novas perspetivas para a realização de trabalhos que possam dar continuidade ao que foi conseguido nesta dissertação.

A primeira proposta é, recorrendo a um trabalho semelhante ao realizado aqui, o uso de mais do que um jogo de futebol de competições de alto nível. Dessa forma é possível usar vários jogos de treino para o árbitro e no final fazer o teste a um jogo completo de equipas totalmente distintas. Este trabalho viria a provar quão completo este modelo poderá ser para prever a posição ideal de árbitros de futebol no terreno de jogo.

Uma segunda proposta é o uso de uma ferramenta de simulação de futebol, semelhante às utilizadas atualmente em videojogos como o *Football Manager,* e no futebol robótico. Através dos registos das posições dos jogadores, e o acréscimo das figuras dos árbitros à simulação, seria possível criar um modelo interessante que conseguisse indicar, com uma interface gráfica, a posição ideal do árbitro.

Uma terceira proposta seria aplicar testes estatísticos para verificar a significância estatística de Wilcoxon de dados não normalizados e normalizados. De forma a perceber a grande diferença entre estas duas formas de abordagem ao tema.

9. Referências Bibliográficas

Abreu, P. (2010). *Artificial Intelligence Methodologies Applied in the Analysis and Optimization of Soccer Teams Performance*. Universidade do Porto.

Aggarwal, C. C. (2015a). *Data Mining*. Cham: Springer International Publishing. http://doi.org/10.1007/978-3-319-14142-8

Aggarwal, C. C. (2015b). *Data Mining: The Textbook*. Springer. Retrieved from https://books.google.com/books?id=cfNICAAAQBAJ&pgis=1

Almeida, R. M. F. de. (2009, May 12). *Análise e Previsão das Formações das Equipas no Domínio do Futebol Robótico.* FEP. Retrieved from http://repositorio-aberto.up.pt/handle/10216/20583

Alves, B. (2011). *Sistema de Observação e Registo do Desempenho Táctico-Técnico em Jogos Desportivos Colectivos*. Universidade do Porto.

Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. In *IADIS European Conference on Data Mining 2008, Amsterdam, The Netherlands, July 24-26, 2008. Proceedings* (pp. 182–185). Retrieved from http://www.researchgate.net/publication/220969845_KDD_SEMMA_and_CRISP-DM_a_parallel_overview

Beetz, M., Gedikli, S., Bandouch, J., Kirchlechner, B., Hoyningen-Huene, N. V., & Perzylo, A. (2007). Visually tracking football games based on TV broadcasts, 2066–2071. Retrieved from http://dl.acm.org/citation.cfm?id=1625275.1625609

Bermingham, M. L., Pong-Wong, R., Spiliopoulou, A., Hayward, C., Rudan, I., Campbell, H., ... Haley, C. S. (2015). Application of high-dimensional feature selection: evaluation for genomic prediction in man. *Scientific Reports*, *5*, 10312. http://doi.org/10.1038/srep10312

Catteeuw, P., Gilis, B., García-Aranda, J.-M., Tresaco, F., Wagemans, J., & Helsen, W. (2010). Offside decision making in the 2002 and 2006 FIFA World Cups. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/02640414.2010.491084#.VLANbSusWuk

Decisive Facts. (2010). Metodologia SEMMA. Retrieved September 25, 2015, from https://decisionstats.files.wordpress.com/2011/10/metodo-semma.jpg

Demšar, U. (2006). Data mining of geospatial data: combining visual and automatic methods. KTH. Retrieved from http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A9900&dswid=9622

Dodge, Y. (2006). *The Oxford Dictionary of Statistical Terms*. Oxford University Press. Retrieved from https://books.google.com/books?id=_OnjBgpuhWcC&pgis=1

Elsworthya, N., Burkeb, D., & Ben J. Dascombea, C. (2014). Factors relating to the decision-making

performance of Australian football officials. *International Journal of Performance Analysis in Sport*.

Faria, B. M. T. de. (2013). Patient classification for intelligent wheelchair adaptation. Universidade de Aveiro. Retrieved from http://ria.ua.pt/handle/10773/11507

Flinders, K. (2002). Football injuries are rocket science. Retrieved September 30, 2015, from http://www.v3.co.uk/v3-uk/news/1950164/football-injuries-rocket-science

Forbes. (2014, July 16). The World's 50 Most Valuable Sports Teams 2014 - Forbes. Retrieved January 16, 2015, from http://www.forbes.com/sites/kurtbadenhausen/2014/07/16/the-worlds-50-most-valuable-sports-teams-2014/

Giannotti, F., Nanni, M., & Pedreschi, D. (2006). Efficient Mining of Temporally Annotated Sequences. *Proceedings*, 12. http://doi.org/10.1137/1.9781611972764.31

Gomes, J. B., Phua, C., & Krishnaswamy, S. (2013). *Data Warehousing and Knowledge Discovery*. (L. Bellatreche & M. K. Mohania, Eds.) (Vol. 8057). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-40131-2

Helsen, W., Gilis, B., & Weston, M. (2006). Errors in judging "offside" in association football: test of the optical error versus the perceptual flash-lag hypothesis. *Journal of Sports Sciences*, *24*(5), 521–8. http://doi.org/10.1080/02640410500298065

Hevner, A., & Chatterjee, S. (2010). *Design Research in Information Systems* (Vol. 22). Boston, MA: Springer US. http://doi.org/10.1007/978-1-4419-5653-8

IASI. (2013). IASI : About Us. Retrieved September 30, 2015, from http://www.iasi.org/aboutus.html

INMOTIO. (2012). Local Position Measurement (LPM) technology is the world's most precise sports tracking system. Retrieved February 7, 2015, from http://www.inmotio.eu/en-GB/20/lpm-technology.html

Internation Football Association Board. (2013). *Leis do Jogo 2013/2014*.

International Association of Computer Science in Sport. (n.d.). Objectives: IACSS - International Association of Computer Science in Sport. Retrieved September 30, 2015, from http://www.iacss.org/index.php?id=31

Keles, I., Ozer, M., Toroslu, I. H., & Karagoz, P. (2015). *New Frontiers in Mining Complex Patterns*. (A. Appice, M. Ceci, C. Loglisci, G. Manco, E. Masciari, & Z. W. Ras, Eds.) (Vol. 8983). Cham: Springer International Publishing. http://doi.org/10.1007/978-3-319-17876-9

Klein, H. K. (2003). Crisis in the IS Field? A Critical Reflection on the State of the Discipline. *Journal of the Association for Information Systems*. Retrieved from http://aisel.aisnet.org/jais/vol4/iss1/10

Krustrup, P., Helsen, W., Randers, M. B., Christensen, J. F., MacDonald, C., Rebelo, A. N., & Bangsbo, J. (2009). Activity profile and physical demands of football referees and assistant referees in international games. *Journal of Sports Sciences*, *27*(11), 1167–76. http://doi.org/10.1080/02640410903220310

Liu, J., Tong, X., Li, W., Wang, T., Zhang, Y., & Wang, H. (2009). Automatic player detection, labeling and tracking in broadcast soccer video. *Pattern Recognition Letters*, *30*(2), 103–113. http://doi.org/10.1016/j.patrec.2008.02.011

Machine Learning Group at the University of Waikato. (2015). Weka 3 - Data Mining with Open Source Machine Learning Software in Java. Retrieved October 10, 2015, from http://www.cs.waikato.ac.nz/ml/weka/index.html

Mallo, J., Frutos, P. G., Juárez, D., & Navarro, E. (2012). Effect of positioning on the accuracy of decision making of association football top-class referees and assistant referees during competitive matches, 9.

Mallo, J., Veiga, S., López de Subijana, C., & Navarro, E. (2010). Activity profile of top-class female soccer refereeing in relation to the position of the ball. *Journal of Science and Medicine in Sport / Sports Medicine Australia*, *13*(1), 129–32. http://doi.org/10.1016/j.jsams.2008.09.006

Mambo Studio. (2015). Match Analysis - What We Do. Retrieved October 22, 2015, from http://matchanalysis.com/process.htm

Marques, F. T. (2010). *Generic Coordination Methodologies Applied to the RoboCup Simulation Leagues*. Faculdade de Engenharia da Universidade do Porto.

Mascarenhas, D. R. D., Dicks, M., O'Hare, D., & Button, C. (2009). Physical Performance and Decision Making in Association Football Referees: A Naturalistic Study. *The Open Sports Sciences Journal*, *2*(1), 1–9. http://doi.org/10.2174/1875399X00902010001

Mingkhwan, A. (2006). WI-FI Tracker: an Organization WI-FI Tracking System. In *2006 Canadian Conference on Electrical and Computer Engineering* (pp. 231–234). IEEE. http://doi.org/10.1109/CCECE.2006.277387

Monreale, A., Pinelli, F., Trasarti, R., & Giannotti, F. (2009). WhereNext: a location predictor on trajectory pattern mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09* (p. 637). New York, New York, USA: ACM Press. http://doi.org/10.1145/1557019.1557091

Mughal, K. U. (2014). Top 10 Most Popular Sports in The World. Retrieved December 19, 2014, from http://sporteology.com/top-10-popular-sports-world/

Naidoo, W. C., & Tapamo, J. R. (2006). Soccer video analysis by ball, player and referee tracking. In *Proceedings of the 2006 annual research conference of the South African institute of computer*

scientists and information technologists on IT research in developing couuntries - SAICSIT '06 (pp. 51–60). New York, New York, USA: ACM Press. http://doi.org/10.1145/1216262.1216268

National Research Council (U.S.). Committee on the Future of the Global Positioning System, Administration, N. A. of P., & System, N. R. C. (U. S.). C. on the F. of the G. P. (1995). *The global positioning system: a shared national asset: recommendations for technical improvements and enhancements*. National Academies Press. Retrieved from http://books.google.com/books?id=FAHk65slfY4C

Needham, C. J. (2003, March 1). *Tracking and modelling of team game interactions*. University of Leeds. Retrieved from http://etheses.whiterose.ac.uk/1320/1/needham.pdf

Nillius, P., Sullivan, J., & Carlsson, S. (2006). Multi-Target Tracking - Linking Identities using Bayesian Network Inference. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06)* (Vol. 2, pp. 2187–2194). IEEE. http://doi.org/10.1109/CVPR.2006.198

Oliveira, M. C. de, Orbetelli, R., & Neto, T. L. de B. (2011). Call Accuracy and Distance from the Play: A Study with Brazilian Soccer Referees.

Orwell, J., & Jones, G. A. (2004). A general framework for 3d soccer ball estimation and tracking. In *2004 International Conference on Image Processing, 2004. ICIP '04.* (Vol. 3, pp. 1935–1938). IEEE. http://doi.org/10.1109/ICIP.2004.1421458

Oudejans, R. R. D., Bakker, F. C., Verheijen, R., Gerrits, J. C., Steinbrückner, M., & Beek, P. J. (2005). How position and motion of expert assistant referees in soccer relate to the quality or their offside judgements during actual match play. *International Journal of Sport Psychology*, *36*.

PCWorld. (2010). Ascensio Match Expert specs. Retrieved February 24, 2015, from http://www.pcworld.com/product/971999/ascensio-match-expert.html

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77. http://doi.org/10.2753/MIS0742-1222240302

Pierce, R. (2015). Correlation. Retrieved October 19, 2015, from http://www.mathsisfun.com/data/correlation.html

Piñeiro, G., Perelman, S., Guerschman, J. P., & Paruelo, J. M. (2008). How to evaluate models: Observed vs. predicted or predicted vs. observed? *Ecological Modelling*, *216*(3-4), 316–322. http://doi.org/10.1016/j.ecolmodel.2008.05.006

Prozone. (2014). About Prozone Sports. Retrieved December 11, 2014, from http://www.prozonesports.com/about/

Reutemann, P., Pfahringer, B., & Frank, E. (2004). Proper: A Toolbox for Learning from Relational Data with Propositional and Multi-Instance Learners. In *17th Australian Joint Conference on Artificial Intelligence (AI2004)*. Springer-Verlag.

Rouse, M. (2008, June). Defenition ultra wideband. Retrieved February 6, 2015, from http://whatis.techtarget.com/definition/ultra-wideband

Santiago, C. (2011). *Vision and Knowledge Representation Methodologies for Game Analysis*. Universidade do Porto.

Santiago, C. B., Sousa, A., & Reis, L. P. (2012). Vision system for tracking handball players using fuzzy color processing. *Machine Vision and Applications*, *24*(5), 1055–1074. http://doi.org/10.1007/s00138-012-0471-z

SAS. (2015). Business Intelligence & Analytics Software | SAS. Retrieved September 25, 2015, from http://www.sas.com/en_gb/software/business-intelligence.html

Sato, K., & Aggarwal, J. K. (2005). Tracking soccer players using broadcast TV images. In *Proceedings. IEEE Conference on Advanced Video and Signal Based Surveillance, 2005.* (pp. 546–551). IEEE. http://doi.org/10.1109/AVSS.2005.1577327

Schumaker, R. P., Solieman, O. K., & Chen, H. (2010a). *Sports Data Mining*. Springer Science & Business Media. Retrieved from https://books.google.com/books?id=r0h2nVGb3qIC&pgis=1

Schumaker, R. P., Solieman, O. K., & Chen, H. (2010b). *Sports Data Mining*. Springer US. http://doi.org/10.1007/978-1-4419-6730-5

Seo, Y., Choi, S., Kim, H., & Hong, K.-S. (2005). Where are the ball and players? Soccer game analysis with color-based tracking and image mosaick, *1311*. http://doi.org/10.1007/3-540-63508-4

SKY MONDIALE 1. (2014). *Italy vs France Full Match 1 1 5 3 HD World Cup 2006 Final English Commentary YouTube*. Retrieved from https://www.youtube.com/watch?v=rcLiQReHuw4

Sullivan, J., & Carlsson, S. (2006). Tracking and Labelling of Interacting Multiple Targets. In *Proc. European Conf. on Computer Vision (ECCV)*.

Togaware Pty Ltd. (2015). Togaware: Rattle: A Graphical User Interface for Data Mining using R. Retrieved October 10, 2015, from http://rattle.togaware.com/

USC - Viberti School of Engeneering. (2006). USC - Viterbi School of Engineering - USC Electrical Engineering: Innovation and Excellence. Retrieved February 6, 2015, from http://viterbi.usc.edu/news/news/2006/usc-electrical-engineering.htm

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, *7*, 91. http://doi.org/10.1186/1471-2105-7-91

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In

Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001 (Vol. 1, pp. I–511–I–518). IEEE Comput. Soc. http://doi.org/10.1109/CVPR.2001.990517

Wang, L., Zeng, B., Lin, S., Xu, G., & Shum, H.-Y. (2004). Automatic extraction of semantic colors in sports video. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 3, pp. iii–617–20). IEEE. http://doi.org/10.1109/ICASSP.2004.1326620

Weisstein, E. W. (2015). Outlier. Wolfram Research, Inc. Retrieved from http://mathworld.wolfram.com/Outlier.html

Weka Documentation. (2015). All Classes. Retrieved October 16, 2015, from http://weka.sourceforge.net/doc.dev/allclasses-noframe.html

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. In *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining* (pp. 29–39).

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann. Retrieved from https://books.google.com/books?id=QTnOcZJzlUoC&pgis=1

Wolfram. (2014). Partitioning Data into Clusters—Wolfram Language Documentation. Retrieved September 29, 2015, from https://reference.wolfram.com/language/tutorial/PartitioningDataIntoClusters.html

Xu, M., Orwell, J., & Jones, G. (2004). Tracking football players with multiple cameras. In *2004 International Conference on Image Processing, 2004. ICIP '04.* (Vol. 5, pp. 2909–2912). IEEE. http://doi.org/10.1109/ICIP.2004.1421721

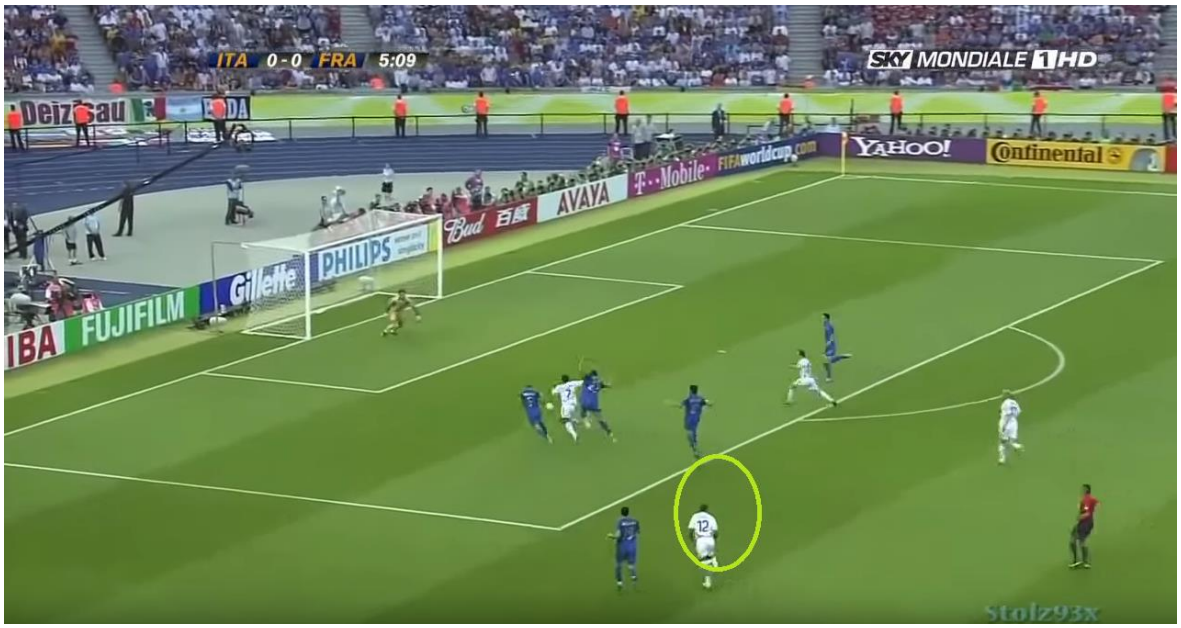
Ye, N. (2014). *Data Mining - Theories, Algorithms, and Examples*. New York, New York, USA: CRC Press.

ZEROZERO. (2015). Hertha-Berliner Sport-Club von 1892. Retrieved October 4, 2015, from http://www.zerozero.pt/equipa.php?id=101&epoca_id=145

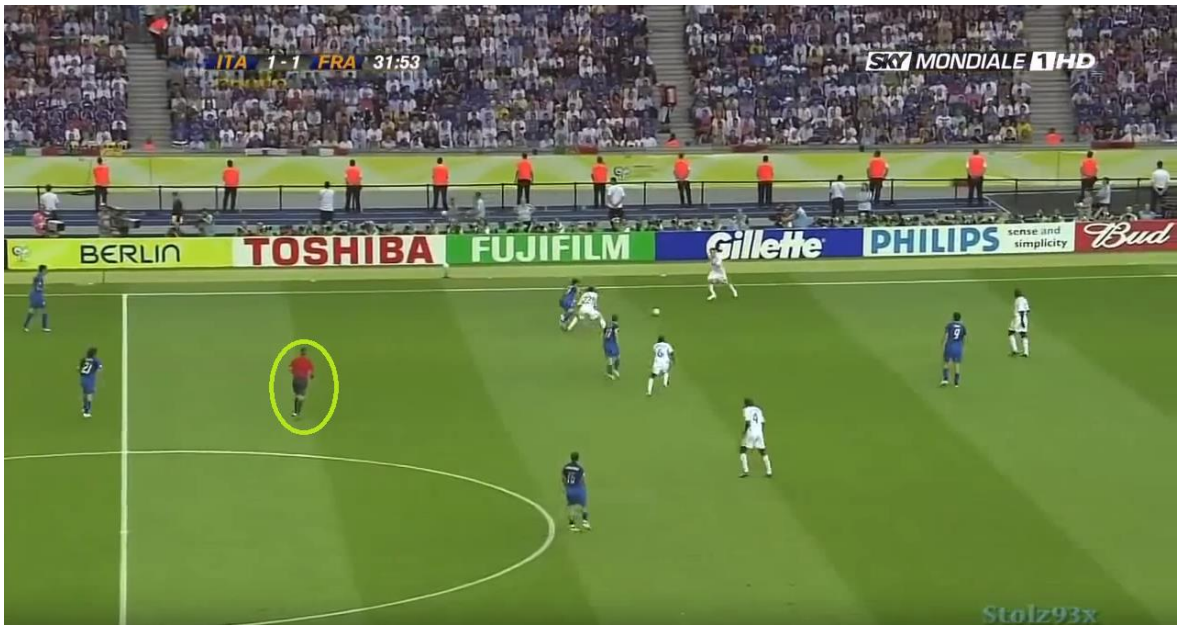
Zhao, Y. (2012). *R and Data Mining: Examples and Case Studies - RDataMining.com: R and Data Mining*. Elsevier. Retrieved from http://www.rdatamining.com/docs/r-and-data-mining-examples-and-case-studies

Apêndice A – Posicionamento do árbitro real vs ideal

ID: (1)



ID: (2)



ID: (3)


ITA 1 - 1 FRA 52:40
SKY MONDIALE 1 HD
Mobile
FIFAworldcup.com
YAHOO!
MasterCard
Stolz93x

ID: (4)


ITA 1 - 1 FRA 65:55
SKY MONDIALE 1 HD
YAHOO!
Stolz93x

ID: (5)


ITA 1 - 1 FRA 71:37
SKY MONDIALE 1 HD
TOSHIBA
FUJIFILM
Gillette
PHILIPS
Bud
Stolz93x

ID: (6)


ITA 1 - 1 FRA 79:14
SKY MONDIALE 1 HD
Gillette
PHILIPS
Bud
AVAYA
T-Mobile
FIFAworldcup.com
YAHOO!
Stolz93x

Apêndice B – Exemplos de dois modelos e respetiva previsão

LeastMedSq

=== Run information ===

Scheme:weka.classifiers.functions.LeastMedSq -S 4 -G 0
Relation: primeiraparte_todosjuntos_normalizado
Instances: 68971
Attributes: 26
R01_TNM_Ball
R01_TNM_Buffon
R01_TNM_Grosso
R01_TNM_Materazzi
R01_TNM_Cannavaro
R01_TNM_Zambrotta
R01_TNM_Camoranesi_Piero
R01_TNM_Perrotta_Iaquinta
R01_TNM_Pirlo
R01_TNM_Gattuso
R01_TNM_Totti_Rossi
R01_TNM_Toni
R01_TNM_Barthez
R01_TNM_Thuram
R01_TNM_Abidal
R01_TNM_Sagnol
R01_TNM_Gallas
R01_TNM_Vieira_Diarra
R01_TNM_Ribery
R01_TNM_Makelele
R01_TNM_Malouda
R01_TNM_Zidane
R01_TNM_Henry
R01_TNM_Garcia
R01_TNM_Otero
R01_TNM_Elizondo
Test mode:user supplied test set: 45506instances

=== Classifier model (full training set) ===

Linear Regression Model

R01_TNM_Elizondo =

0.0471 * R01_TNM_Ball +
-0.0137 * R01_TNM_Buffon +
0.1041 * R01_TNM_Grosso +
-0.0215 * R01_TNM_Materazzi +
-0.043 * R01_TNM_Cannavaro +
-0.022 * R01_TNM_Zambrotta +
0.001 * R01_TNM_Camoranesi_Piero +
0.1505 * R01_TNM_Perrotta_Iaquinta +
0.016 * R01_TNM_Pirlo +
0.0501 * R01_TNM_Gattuso +
0.0878 * R01_TNM_Totti_Rossi +
0.0337 * R01_TNM_Toni +
0.0284 * R01_TNM_Barthez +
-0.0206 * R01_TNM_Thuram +
0.1382 * R01_TNM_Abidal +
-0.0106 * R01_TNM_Sagnol +
0.0511 * R01_TNM_Gallas +
0.2234 * R01_TNM_Vieira_Diarra +
0.0082 * R01_TNM_Ribery +
0.0833 * R01_TNM_Makelele +
0.1495 * R01_TNM_Malouda +
0.0166 * R01_TNM_Zidane +
0.0718 * R01_TNM_Henry +
-0.0891 * R01_TNM_Garcia +
0.0531 * R01_TNM_Otero +
-0.0577

Time taken to build model: 38.01 seconds

=== Evaluation on test set ===
=== Summary ===

Correlation coefficient	**0.9392**
Mean absolute error	**0.0765**
Root mean squared error	**0.1039**
Relative absolute error %	**29.2208 %**
Root relative squared error %	**34.5298 %**
Total Number of Instances	**45506**

Exemplo de previsão da posição do árbitro num pontapé de canto: Em primeiro lugar é dado um ID à instância, em segundo o valor que realmente o árbitro se encontra, em terceiro a previsão do modelo e em último a diferença entre o real e o previsto.

=== Classifier model (full training set) ===

IB1 instance-based classifier

using 1 nearest neighbour(s) for classification

Time taken to build model: 0 seconds

=== Predictions on test split ===

inst#,	actual,	predicted,	error
1	1	0.942	-0.058
230	0.968	0.917	-0.051
231	0.968	0.917	-0.051
232	0.968	0.917	-0.051
250	0.509	0.373	-0.136
251	0.505	0.373	-0.132
254	0.495	0.354	-0.141
255	0.492	0.546	0.055
256	0.488	0.546	0.058
257	0.485	0.546	0.062
258	0.481	0.546	0.065
273	0.427	0.546	0.119
274	0.423	0.546	0.123
275	0.419	0.546	0.127
276	0.415	0.546	0.131
277	0.412	0.546	0.134
278	0.407	0.546	0.139
279	0.404	0.546	0.142
280	0.4	0.546	0.146
281	0.396	0.546	0.15
282	0.393	0.546	0.153
283	0.388	0.546	0.158
295	0.341	0.254	-0.088
296	0.337	0.254	-0.083
297	0.334	0.254	-0.08
213	0.968	0.922	-0.046
214	0.968	0.922	-0.046
215	0.968	0.922	-0.046
216	0.968	0.926	-0.042
406	0.022	0.148	0.126
407	0.02	0.148	0.128
408	0.02	0.148	0.128
409	0.018	0.148	0.13
410	0.018	0.148	0.13
411	0.017	0.148	0.131
412	0.016	0.148	0.132
413	0.015	0.148	0.133
414	0.014	0.148	0.134
415	0.013	0.148	0.135
416	0.012	0.148	0.136
417	0.011	0.148	0.137
418	0.01	0.148	0.138

=== Summary ===

Correlation coefficient	0.9883
Mean absolute error	0.0662
Root mean squared error	0.0778
Relative absolute error	18.2653 %

Root relative squared error	19.0358 %
Total Number of Instances	426

Apêndice C – Artigo Submetido ao WorldCist'16

Intelligent System for Soccer Referee's Position Analysis

Rego, Carlos Moreira[1], Reis, Luís Paulo[1], Meneses, Filipe[1],

[1] Information Systems Department, University of Minho, Campus de Azurém, 4800-058 Guimarães, Portugal
pg25200@alunos.uminho, {lpreis, meneses}@dsi.uminho.pt

Abstract. The role of referees and the decisions they take have, in some cases a direct influence on the course of the game. Unfortunately these decisions aren't always correct. In a large amount of times, these mistakes are due to referee's bad position. So it's fundamental to develop a model that may determine the correct position for a soccer referee. For this research work, data with all the game participants' positions in the field during the match has been used as its basis. The data included, players, ball and referees positions gathered by an automatic system. Referees with large experience in international competitions were used to obtain correct position knowledge. The achieved results are very positive, both for predicting the assistant referee position, regardless of the team he's accompanying or the match, as to predict the referee position, in controlled game situations. In short, despite the limitations, the models are conclusive and can efficiently determine the correct positioning of the referee.

Keywords: data mining, position, modelling, data preparation, refereeing, soccer

1 Introduction

This research work it's an innovative approach to determine the correct position of the referee's team in a game in every game situation. There're some papers that focus on the assistant referees position to have the best condition to analyse the off-side, and some work that evaluate the referee ideal distance to the ball. Although, there isn't any research that points the position, nor the angle, that best suits the referee to call the right decision.

In this research, all conclusions from previous works are taken into account. Still, as pointed out, this work is innovative and there isn't any research in the field with the same approach to achieve the goal of predicting the optimal position of the referee team on the pitch.

Some limitations were faced though the development of this model. The limitations of data mining itself, since it needs historical data to learn, to obtain knowledge. So this data needs to be a good example of how a referee should move on the pitch according to the game situation (direction of the ball, number of players and how they move). And consequently the lack of data, there is only one game (Italy vs France in the 2006 World Cup Final).

During the development of this research work, there've been taken some assumptions. With the exception of the plays in which the referee made the wrong decision, this plays were analysed by an expert and hence corrected for the position that would have made the right call, it was assumed that the referee's position was correct.

The aim of this research work is to build a model that can, using data mining to obtain knowledge from past games with quality referee's teams, determine the correct position of the referee's team in any game situation, reducing the amount of judging errors during soccer games.

2 Related Work

In this section it'll be presented all previous papers that had some influence in this research work.

2.1. Referees

The soccer referee's main duty to supervise the application of the Laws of the Game. Thus, being in the right place at the right time is important in order to get a better view of the play and a correct evaluation of it (Mallo, Frutos, Juárez, & Navarro, 2012).

According to Internation Football Association Board (2015) the position of the referee on the pitch it's recommended as follows:

- The play should be between the referee and the lead assistant referee
- The lead assistant referee should be within the referee's field of vision. The referee should use a wide diagonal system
- Staying towards the outside of the play makes it easier to keep play and the lead assistant referee within the referee's field of vision
- The referee should be close enough to see play without interfering with play

2.2. Assistant Referees

As the Internation Football Association Board (2015) points out, the assistant referee's help the referee to enforce the Laws of the Game. Two assistant referees may be appointed whose duties, subject to the decision of the referee, are to indicate:

- When the whole of the ball leaves the field of play
- Which team is entitled to a corner kick, goal kick or throw-in

- When a player may be penalised for being in an offside position
- When a substitution is requested
- When misconduct or any other incident occurs out of the view of the referee
- When offences have been committed whenever the assistant referees have a better view that the referee
- Whether, at penalty kicks, the goalkeeper moves off the goal line before the ball is kicked and if the ball crosses the line

Of all this duties, the most important of all for an assistant referee is the offside. With that in mind, they are instructed to constantly be in line with the second last defender of a team, or the last two, or the ball if it's nearer the goal line in order to be the best conditions to analyse.

The **Fig. 1** illustrates the zones of action of each member of the referee's team. The referee must move in a diagonal system due to the assistant referee's zones.

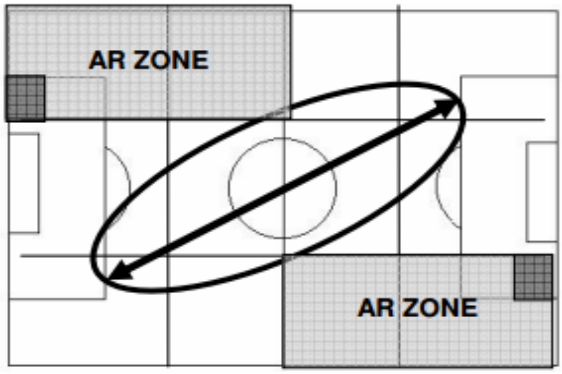


Fig. 1. Referee and assistant referee's zones of action

2.3. Data Mining

Data Mining is the study of collecting, cleaning, processing, analysing and gaining useful insights from data. A wide variation exists in terms of the problem domain, applications, formulations, and data representations that are encountered in real applications. Therefore, data mining is a broad umbrella term that is used to describe these different aspects of data processing (Aggarwal, 2015).

Data immensity is the direct result of advantages in technology and computerization of all aspects of modern life. It's therefore natural to analyse whether it's possible to extract concise ideas and possibly knowledge of the available data for specific purposes applications. It's in these situations that data mining is required. The original data can be arbitrary, unstructured or until a format that's not suitable for automatic processing. For example, manual data collection may originate from multiple sources and in different formats. To solve this problem it's necessary to develop a process in which

data are collected, processed and cleaned in a single format. This process causes the majority of work is related to the preparation of data (Ye, 2014).

The data mining process can be described by **Fig. 2** as Aggarwal (2015) explains. This process meets the various stages of CRISP-DM.

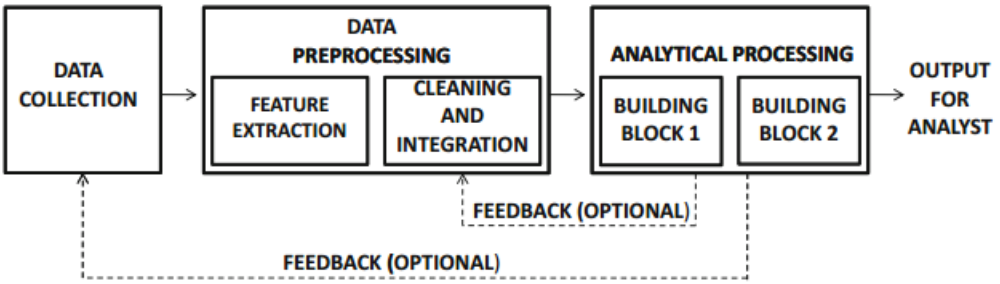


Fig. 2 Data Mining process (Source: Aggarwal, 2015)

Data collection – this phase is very specific to the application used and extremely important as correct choices at this stage can have a significant impact on the data mining process. After collection, the data is usually stored in databases or data warehouses.

Feature extraction and cleaning – the data collected is often at little proper format for processing. For example, the data can be encoded in complex logs or free writing papers (a text box is a good example because people can write anything and various type of data can be arbitrarily mixed). To make the data ready for processing is essential to turn these in a friendly format for data mining algorithms. It's crucial to extract relevant characteristics for the data mining process. The characteristics extraction stage is usually done in parallel with the data cleaning because the missing or incorrect values are corrected or eliminated.

Analytical processing – the final part of the data mining process is the design of effective analytical methods based on the processed data. In many cases it may not be possible to directly use a standard data mining problem, such as patterns of association, clustering, classification and outlier detection. However, these problems have such a wide coverage that many applications can be broken up into components that use these different building blocks.

The most typical problems of data mining can be divided into two categories: prediction and description. The prediction is characterized by problems with specific goals, with datasets of the past as a basis for what you want to predict. The description on the other hand, has the purpose of detecting information on a complex database to increase the knowledge to be extracted (Almeida, 2009).

Data mining has been use in sports since a long time now, and an important contribution is the ability to predict when a player may be having physical breaks

through injuries prognosis. AC Milan[5] monitors the training of their athletes (Flinders, 2002). The software compares the performance of an athlete's performance training and any sub-yield sign may indicate that the player is injured.

According to the work of (Almeida, 2009) using log records of various *RoboCup Soccer* games it can identify what's the best tactical formation that a team should adopt when facing another team. To get this knowledge, he used several tests and classification algorithms to these records games.

In short, the data mining in sports is used in order to create a competitive advantage over the opponent to facilitate the challenge.

Currently in the scientific community there are no studies on the prediction of the location of athletes in the pitch. However, it's possible to draw some conclusions from existing work in location prediction area, such as: people use common paths and they follow the crowd. So it's easy to see that a player in a certain position of the respective tactical formation will use repeated movements throughout the game so it's possible to determine what will be their behaviour to a given situation. The same principle can be applied to the referee's teams. That's the work it'll be developed in this research. Thus, in the existence of sufficient data to model behaviours of top referee teams, you can determine the ideal position of the referee team in a given moment.

3 Methodology

The Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology is a model of hierarchical processes, comprising four levels of abstraction (in descending order): phases, generic tasks, specific tasks and process instances (see **Fig. 3** left side).

According to the described by Wirth & Hipp (2000) at the top level, the data mining process is organized into number of stages, each of which consists of several generic tasks (so called because they cover all possible situations of data mining). The third level, the level of specific tasks, is the place to describe how actions in the generic tasks must be performed in specific situations. Finally, the level of process instances is a registration actions, decisions, and results of a real data mining application. A process instance is organized according to the tasks set at the previous levels but represents what actually happened in a specific job, rather than what happens in general.

The description of phases and tasks as discrete steps performed in a specific order represents an idealized sequence of events (see **Fig. 3** right side). In practice, many of the tasks can be performed in a different order and will often be necessary to go back to previous tasks and repeat certain actions. The CRISP-DM process model doesn't attempt to capture all of these possible routes through the data mining process, because

[5] One of the greatest Italian and European soccer clubs

it requires an overly complex process model and the expected benefits would be greatly reduced.

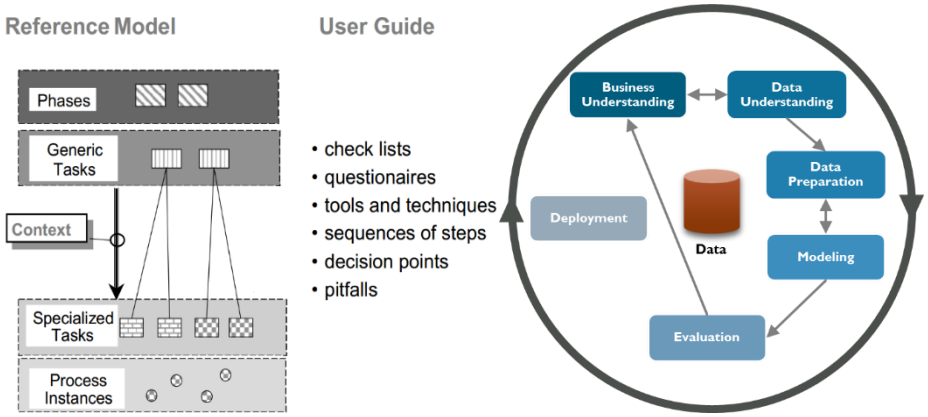


Fig. 3 On the left the four levels of CRISP-DM for Data Mining (source: Wirth & Hipp, 2000) and the phases of CRISP-DM Process Model for Data mining on the right (source: Decisive Facts, 2015)

4 Data understanding and preparation

In order to be able to predict the ideal position of all the referee's in the pitch during a game, using data mining techniques in a dataset with the Cartesian positions of all intervenient of the game (ball, players and referees), it's needed, in a first approach to the problem, understand the data available.

The data understanding phase begins with the familiarization of data to identify data quality problems, discover the first perceptions of the data or detect interesting subsets to form hypotheses for hidden information.

The data represents all the x and y positions of all the players and referees and the x, y and z of the ball. 24 records for second of all the 120 minutes of game denotes all the data available. The game was played in the Olympic Stadium "*Olympiastadion*", in Berlin. The field has 105x68m of dimensions.

The data quality is very good with the biggest margin of error less than 1 meter. This conclusion was get after comparing the dataset records with a video image.

The next phase of this data mining was the data preparation. According to Wirth & Hipp, (2000) this stage encompasses all activities that lead to the construction of the final dataset (the data that will feed the modelling tools from the initial data).

The data preparation was made according to the knowledge that is sought to obtain at each time. Thus, the data were prepared in two main points. First, to prepare the prediction model of the ideal location of the assistant referees and, second, to prepare the prediction model of the ideal position of the referee.

For the first problem have been removed all the values for the y-coordinate as this has no influence on the assistant referee position, since his position is influenced by the second last position (or second biggest according to which side he is). For the second problem, the ideal location of the prediction of the referee, changes to data were bigger. After analysing the game, it were pointed out all the bad decisions of the referee during the game due to bad position on the pitch. The position of the referee was edited in the data according to what the asked expert said - Assistant Referee of Portuguese First Soccer League – Liga NOS (see **Table 1**).

Table 1 Edited data according to the expert feedback in plays that the referee missed

ID	Time	Play description	Real Location	Correct Location
1	05:09	Malouda suffers a slight hint. The referee, although poorly positioned, made the right call	(27.5,13.84)	(17.5,12)
2	31:53	Perrota is stepped on. The referee is well positioned. This bad decision isn't due to his position	(58,22.85)	(58,22.85)
3	52:41	Penalty to France. Referee poorly positioned	Out of scene	(15.5,14.84)
4	65:55	Iaquinta receives the ball with his hand. The referee is far and with athletes in front of him	(43,18.7)	(37.5,13.84)
5	71:37	The misplaced referee is in a pass line of Italy. He's surprised with the ball and touch's it.	(66.25,21.85)	(71.75,20.85)
6	79:14	Cannavaro jumps negligently on Zidane. Nothing is indicated	(72,54.16)	(74.75,54.16)

B y default the value of "-65000" was set when some "object" (referee, players or ball) was out of view from all the cameras and for instance substitutes. Whatever one of this appeared, it were deleted. The substitutions were made manually, replacing the "-65000" values when the player left the game for the position of the new player.

Because not all soccer fields have the same dimensions, the data was normalized. The normalization mean adjusting the values measured on different scales for a subjectively common scale (Dodge, 2006). This way, the range of values is always the same regardless of where the game was held.

At the stage of data modelling, various modelling techniques are selected and applied and their parameters calibrated to optimal values. Typically there are several techniques for the same type of data mining problem and some of these techniques require specific data formats.

To obtain the most accurate possible model, various experiments were made. These experiences consisted of:

- For assistant referee
    - Use of subsets of data from one third, two thirds and all data
    - Restriction to the x coordinate
    - Restriction to the team that he was accompanying
    - With the data normalized and not normalized
    - Use of 90 minutes for training and 30 minutes for test
    - Change the order of the data (random)
- For the referee
    - Use of subsets of data from one third, two thirds and all data
    - Restriction on the coordinated x, y and both
    - With the data normalized and not normalized
    - Use of 90 minutes for training and 30 minutes for test
    - Data from just one team and both
    - Use of the first 15 minutes of the first half for training and the first 15 minutes of the second half for test
    - Use of controlled game situations – just corner kicks
    - Change the order of the data (random)

These experiments were designed in order to select the most suitable regression algorithm to predict the ideal position of the referee's team relative to the position of the ball and players.

Evaluation is the key point to have effective progress in data mining. There are several different methods to get data knowledge, but not all are specified. According to Witten & Frank (2005) the issue of predicting performance based on limited dataset is interesting and controversial. Despite the existence of various techniques, the cross-validation distinguished from the others and is, at the moment, the evaluation method of choice in situations with little data.

On the other hand, when the dataset is large, the best method is to divide a large dataset for training and other large dataset for testing. However, despite the large amount of data, quality thereof is, in some cases, scarce. For example, using a dataset for training in which there is only one corner kick situation and then, is used for testing a dataset in which there are ten play situations of a corner kick. The values shown by the model will not indicate nothing.

The cross-validation method reserves a certain number of records for testing and uses the remaining for training. As this process is done randomly, the sample for training may not be representative – an extreme situation would be none of the instances in the test has been used for training. As such, it must ensure that the random sample is made so that all situations are properly represented in both, the training and the test. This procedure is called stratification (Witten & Frank, 2005).

For the first problem, the best mathematic model to predict the ideal position of the referee on the pitch got a success rate of 76.03% using for this the first 45 minutes of the game for training and the second 45 minutes for testing. Although, using the cross-validation method for the same quantity of data, the results are much better with the algorithm K-nearest neighbour with more than 99% of success. The cross-validation

result can confirm, that, using enough data from a very large amount of games, it's possible to predict effectively the ideal position of an assistant referee on the pitch.

For the second problem, the position of the referee, the analysis was performed in order to evaluate the model for a case of specific game situations. The selected game play was a corner kick (because it's simple to detect this situation on the data). Four situations of corner kick were used for training and one for test. This is the most clear model evaluation since it's a controlled condition and the training data is very useful for analysis.

Table 2 Model evaluation to predict the referee position in a corner kick situation

	IBK	**KStar**
Correlation Coe.	0.9883	0.9727
MAE	0.0662	0.0898
RMSE	0.0778	0.1133
RAE	18.2653%	24.7615%
RRSE	19.0358%	27.706 %

In **Table 2** are represented the two best mathematical models for the described scenario. The most close to 1 the correlation coefficient is, the better. So, both algorithms can detect an almost perfect positive correlation between the target variable and the attributes (this means that as the target variable increases, the attributes increase as well, and vice-versa). The rest of the indicators show the errors, and this ones the lowest they are, the better. The **mean absolute error (MAE)** is a quantity used to measure how close forecasts or predictions are to the eventual outcomes. The **mean squared error** (**MSE**) and the **root mean squared error (RMSE)** of an estimator measures the average of the squares of the "errors", that is, the difference between the estimator and what is estimated. The **relative absolute error (RAE)** is the total of the absolute error with the same type of normalization as the RSE. The **root relative squared error (RRSE)** is calculated in relation to what would be a simple mathematical model would have been used. The algorithm in question is just the average of the actual amounts of training data. Thus, RSE assumes the error and normalize it by dividing by the total squared error of the standard mathematical error.

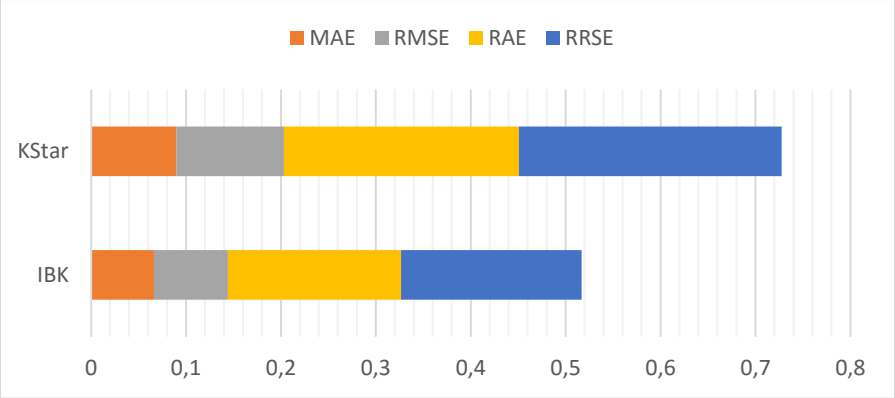


Fig. 4 Model evaluation to predict the referee position in a corner kick situation – graphic representation of Table 2

After analysing the chart it can be concluded that IBK is a model with a very positive success rate (as lower error rates for every indicator). With this result, it's clear that K-Nearest Neighbour algorithm, with more training data (more situations of corner kicks), can predict whit a even higher percentage of success the ideal position of the referee.

Some other experiences resulted in worst results because it exists lots of data but this doesn't fit to any situation. For instance, we can use a uncontrolled game situation, for example, all of the first half of the regular time for training and test that data on the second half of the regular time. The mathematical model will not perform well because there are lots of situations that the model is trying to predict based on the training data that doesn't occur in the test data. That is, if in the first half there aren't any corner kick situation, or the game is full of counter-attacks and in the second half there are ten corner kicks and it's everytime the same team that has the possession of the ball, the mathematical model will perform poorly.

5 Results and Discussion

After proceed to the modelling and evaluation of various mathematical models, it's possible to see the results of the various exercises developed throughout the work.

The models related to the assistant referee's problem were achieved with a small error rates and very satisfactory results for the available data.

The result of 77% of success for the experience of using the first part for training and the second for test, validates the objective of this research work. That is, using a dataset of training with the assistant referee follow one team (France), and the dataset for testing when he's accompanying another team (Italy) confirms that this model can be applied to any situation. That way, with a sufficient volumetric data for training, it's possible to predict the position of the assistant referee in any game with any team.

The model obtained for the assistant referee position on the pitch showed results has expected and it's conclusive that this mathematic model can be applied to different games. That is, the model can learn from a dataset and can be successfully applied to predict the position in other games, regardless of the team. The use of cross-validation evaluation method with such high success rates can indicate that with a large amount of data for training, it's possible to predict very effectively the position of the referees on the pitch.

As expected, the models obtained for the referee, present a higher error rates that that achieved for the assistant referee. If a similar analysis to the one developed for the assistant referee is made – use of a large dataset for training such as the first 45 minutes and another large dataset for test, for example the last 45 minutes of the regular time – the results achieved have been of less than 50% of success percentage. However, when an analysis was made to a specific game situation, such as a corner kick, the results achieved have a percentage of success of 82%.

Due to the lack of studies related to this theme, and being this topic completely new, with no previous work, it can predict, with success, where does the referee should be placed in the pitch for some soccer situations such as, corner kicks, goal kicks, direct and indirect free kicks in front of the goal or near the edge of the penalty box and penalty kicks. In a future release, with sufficient data for training, this model should be able to predict the location of the referee and the movement of him, in some more complex situations, such as counter-attacks with numerical superiority of one team and react to various situations of the game depending the player' behaviour.

With these results it's possible to predict the ideal position successfully for a specific situation of the game.

6 Conclusions and Future Work

This research was aimed at building a formal model to determine the best placement of the referee team on the pitch in accordance to the position of the ball, the players and even the remaining team members.

To accomplish this, various data mining algorithms were studied. For this model was used data from professional soccer game, the final of 2006 World Cup, which contained information of the Cartesian positions (x,y) of players, referees and ball.

The objectives were achieved and were built two mathematical models that determine what the ideal position of the referees is. It was concluded that for the assistant referee model developed can be determined with an accuracy of 77% regardless of the team. For this model it was used the 90 minutes for training and the 30 minutes of extra time for test in a situation, and the first 45 minutes for training and the second 45 minutes of regular time for test in other situation. To determine the referee's model, specific game situations were used as corner kicks, to predict the correct position of this with a success of 82%.

This study adds a further knowledge in the area in which there was nothing similar. The results obtained of the prediction models for the referee team can add tremendous

knowledge to new and even experienced soccer referees. These models can, and should, be used in refereeing academies so that it can improve their students.

It should, however, be noted that the results obtained for these models have been achieved from a dataset related to a single game. This limitation has conditioned, in part, the obtained results. A larger dataset would had help to achieve better results.

The obtained results open the doors to new research in this area, still unexplored, and also encourages greater investment in the area of detection of the players for not intrusive systems.

In conclusion, the results from these models will contribute to a reduction of errors in soccer games due to the poor position of the referee team on the pitch.

References

Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer. Retrieved from https://books.google.com/books?id=cfNICAAAQBAJ&pgis=1

Almeida, R. M. F. de. (2009, May 12). *Análise e Previsão das Formações das Equipas no Domínio do Futebol Robótico.* FEP. Retrieved from http://repositorio-aberto.up.pt/handle/10216/20583

Decisive Facts. (2015). Phases of the Current CRISP-DM Process Model for Data Mining. Retrieved from http://www.decisivefacts.nl/wp-content/uploads/2015/01/crisp_dm.png

Dodge, Y. (2006). *The Oxford Dictionary of Statistical Terms*. Oxford University Press. Retrieved from https://books.google.com/books?id=_OnjBgpuhWcC&pgis=1

Flinders, K. (2002). Football injuries are rocket science. Retrieved September 30, 2015, from http://www.v3.co.uk/v3-uk/news/1950164/football-injuries-rocket-science

Internation Football Association Board. (2015). *FIFA - Laws of the Game 2015/2016*.

Mallo, J., Frutos, P. G., Juárez, D., & Navarro, E. (2012). Effect of positioning on the accuracy of decision making of association football top-class referees and assistant referees during competitive matches, 9.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. In *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining* (pp. 29–39).

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann. Retrieved from https://books.google.com/books?id=QTnOcZJzlUoC&pgis=1

Ye, N. (2014). *Data Mining - Theories, Algorithms, and Examples*. New York, New York, USA: CRC Press.