Universidade do Minho
Escola de Engenharia

João Carlos Leitão Gomes

# Sistema Inteligente de Apoio à Decisão de Apostas em Jogos de Futebol

Outubro de 2015

Universidade do Minho
Escola de Engenharia

João Carlos Leitão Gomes

# Sistema Inteligente de Apoio à Decisão de Apostas em Jogos de Futebol

Dissertação de Mestrado Integrado em

Engenharia e Gestão de Sistemas de Informação

Trabalho efetuado sob a orientação de

Professor Doutor Carlos Filipe Portela

Professor Doutor Manuel Filipe dos Santos

Outubro de 2015

DECLARAÇÃO

Nome: João Carlos Leitão Gomes

Endereço eletrónico: joaogomes0991@gmail.com

Telefone: 916377302

Número do Bilhete de Identidade: 13973643

Título projeto de dissertação

Sistema Inteligente de Apoio à Decisão de Apostas em Jogos de Futebol

Orientador(es):

Professor Doutor Carlos Filipe da Silva Portela

Professor Doutor Manuel Filipe Vieira Torres dos Santos

Ano de conclusão: 2015

Designação do Mestrado: Mestrado Integrado em Engenharia e Gestão de Sistemas de Informação

DE ACORDO COM A LEGISLAÇÃO EM VIGOR, NÃO É PERMITIDA A REPRODUÇÃO DE QUALQUER PARTE DESTA DISSERTAÇÃO/TRABALHO

Universidade do Minho, __/__/____

Assinatura: _____

# AGRADECIMENTOS

O meu sincero agradecimento a todos aqueles que, direta ou indiretamente contribuíram para a finalização deste percurso académico.

Em primeiro lugar agradecer à minha família, em especial aos meus pais por todo o apoio e pelo investimento na minha formação académica.

Um agradecimento especial ao Professor Doutor Carlos Filipe Portela por toda a disponibilidade, paciência e partilha de conhecimento no desenrolar desta dissertação.

Aos amigos que fiz durante este percurso académico, é mais gratificante trabalhar e estudar quando estamos rodeados de boas pessoas.

Aos amigos de sempre por todo o apoio, amizade e troca de ideias.

À Sara Gomes por toda cumplicidade ao longo destes 4 magníficos anos, por todo o apoio incondicional, por todos os sorrisos que recompensaram todo o esforço durante o meu percurso académico.

# RESUMO

No último século o futebol é considerado o desporto com mais adeptos no mundo. No mundo das apostas, apostar em jogos de futebol, atingiu uma posição de destaque em relação a todos os outros mercados, chegando a movimentar milhões de euros num único jogo. O crescimento do número de casas de apostas nos últimos leva a concluir que este é um negócio rentável para as mesmas, em detrimento dos seus utilizadores, que têm dificuldades em obter lucros a médio/longo prazo. A probabilidade de acertar em uma aposta num jogo de futebol cresce consoante o conhecimento do apostador na temática em que efetua a sua aposta.

Casas de apostas como *Betfair*, *Bet365* e *Bwin* permitem efetuar apostas diferentes apostas, apostas no resultado final, no número de cantos, no número de golos entre outros.

Com o intuito de otimizar os lucros dos apostadores, bem como, diminuir os riscos que estão envolvidos em cada aposta, foi decidido criar um sistema inteligente que englobasse o maior conhecimento possível sobre cada jogo em que se pretende efetuar a aposta. Os dados estatísticos e alguns fatores extra futebol que são conhecidos antes de cada jogo podem servir de indicadores para diminuir os riscos em cada aposta. Através de técnicas de Data Mining (DM) é possível detetar padrões nesses dados de forma automática, sem ser necessário conhecimento algum por parte do apostador. Os modelos de DM induzidos neste projeto tinham como objetivo prever vários *targets*. Relativamente ao resultado final, os modelos foram induzidos para prever duas abordagens: as três saídas possíveis, vitória equipa visitada, empate ou vitória equipa visitante e as duas saídas, se o resultado é contra ou a favor da equipa visitada e também o contrário contra ou a favor a equipa visitante. Relativamente aos cantos foram induzidos modelos para prever o número de cantos, especificamente se existiram mais ou menos de '7,5', '8,5', '9,5' e '10,5' cantos. Por fim foram criados modelos para prever o número de golos, mais ou menos de '1,5', '2,5' e '3,5'. Os modelos estão preparados para serem induzidos em tempo real e todo processo ser executado automaticamente utilizando aprendizagem *on-line*.

Ao nível dos resultados obtidos sete das previsões efetuadas atingiram os valores dos parâmetros de qualidade definidos, o resultado com três saídas, o resultado contra ou a favor a equipa visitada, resultado contra ou a favor a equipa visitada, mais ou menos de 7,5 e 8,5 cantos e mais ou menos de 1,5 e 3,5 golos. Os modelos que cumprem todos os parâmetros de qualidade definidos foram implementados no protótipo.

Palavras-Chave: Sistemas de Apoio à Decisão em Jogos de Futebol, Previsão em Jogos de Futebol, Data Mining, Apostas em Jogos de Futebol

# ABSTRACT

In the last century football is considered the sport with the most fans in the world. In the betting world, betting on football games, reached a prominent position in relation to all other markets, moving millions of euros in a single game. The growth of the number of bookmakers leads to the conclusion that this is a profitable business for them, to the detriment of its users, who have difficulty to make a profit in the medium / long term. The probability of making the correct bet on a football game increases depending on the knowledge of the gambler on the market that makes his bet. Bookmakers like Betfair, Bet365 and Bwin allows you to make different bets, such as betting on the end result, the number of corners and goals, among many others in a single game. In order to optimize the profits of gamblers, as well as reducing the risks involved in each bet, it was decided to create an intelligent system that would include as much knowledge as possible about each game in which the user is trying to place the bet. Statistical data and some extra football factors, like precipitation, that are known before each game can serve as indicators to reduce the risk on each bet. Through techniques of Data Mining (DM) it is possible to detect patterns in these data automatically, without requiring the gambler to have any knowledge on the market. The DM models induced in this project aim to predict several targets. For the final result, the models were induced to predict two approaches: the three possible outcomes, home team win, draw or away team win and two outputs, if the result is in benefit or against the home team and also the other way against or in benefit of the visiting team. For corners were induced models to predict the number of corners, especially if there were more or less than '7,5', '8,5', '9,5' and '10,5' corners. Finally, models were created to predict the number of goals, more or less than '1,5', '2,5' and '3,5'. Models are prepared to be induced in real time and the entire process is performed automatically using online learning. In terms of results, seven of forecasts equal or exceed the values defined as quality parameters, these were: the result with three outputs, results in benefit or against the home team, result in benefit or against the home team, more or less than 7,5 and 8,5 corners and more or less than 1,5 to 3,5 goals. The models that comply with all the defined quality standards have been implemented in the prototype.

Keywords: Decision Support Systems in Football Games, Football Games Forecasts, Data Mining, Betting Football Games

# ÍNDICE

# LISTA DE FIGURAS

# LISTA DE TABELAS

# LISTA DE ABREVIATURAS, SIGLAS E ACRÓNIMOS

10FCV – *10-Folds Cross-Validation*

AD – Árvores de Decisão

BADC – *British Atmospheric Data Centre*

BI – *Business Intelligence*

C7,5 – 7,5 Cantos

C8,5 – 8,5 Cantos

C9,5 – 9,5 Cantos

C10,5 – 10,5 Cantos

CEDA – *Centre for Environmental Data Archival*

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

CV – *Cross Validation*

DCBC – Descoberta de Conhecimento de Bases de Dados

DM – *Data Mining*

DSR – *Design Science Research*

ETL – *Extract Trasformation and Load*

FN – Falsos Negativos

FP – Falsos Positivos

G1,5 – 1,5 Golos

G2,5 – 2,5 Golos

G3,5 – 3,5 Golos

HS – *Holdout Simple*

KNN – *K-Nearest Neighbours*

LL – *Lazy Learning*

MDM – Modelos de *Data Mining*

MIEGSI – Mestrado Integrado em Engenharia e Gestão de Sistemas de Informação

ML – *Machine Learning*

NB – *Naive Bayes*

R2SC – Resultado a favor ou contra a equipa visitada

R2SF – Resultado a favor ou contra a equipa visitante

R3S – Vitória equipa visitada, empare ou vitória equipa visitante

ROC – *Receiver Operating Curve*

ROCCH – *Receiver Operating Curve Convex Hull*

SAD – Sistemas de Apoio à Decisão

SIAD – Sistemas Inteligentes de Apoio à Decisão

SINO – Sistemas para a Inteligência do Negócio Organizacional

SVM – *Support Vector Machines*

UC – Unidade Curricular

VN – Verdadeiros Positivos

VP – Verdadeiros Negativos

# 1 INTRODUÇÃO

Neste primeiro capítulo será introduzida a temática e o enquadramento deste trabalho, e são ainda apresentados os objetivo, a abordagem metodológica seguida durante o trabalho e a estrutura do mesmo.

## 1.1. Enquadramento

A presente dissertação surge no âmbito da Unidade Curricular (UC) de Dissertação do plano de trabalho do curso Mestrado Integrado em Engenharia e Gestão de Sistemas de Informação (MIEGSI).

De forma a encontrar um tema a explorar nesta dissertação, foi decidido escolher um tema que englobasse o fascínio adquirido ao longo do MIEGSI, mais concretamente nas unidades curriculares de Sistemas para a Inteligência do Negócio Operacional (SINO) e Sistemas de Apoio à Decisão (SAD) por sistemas inteligentes, *Business Intelligence* (BI), Data Mining (DM) e SAD.

Surgiram várias ideias para explorar mas a mais motivadora foi, sem dúvida, o tema das apostas em jogos de futebol. Na altura da tomada desta decisão, em conjunto com um grupo de amigos, descobrimos que partilhávamos um interesse comum, fazer apostas em jogos de futebol. Estas eram capazes de tornar qualquer jogo de baixa intensidade num jogo interessante que nos levava a ficar focados no mesmo. Começamos por partilhar os resultados que íamos atingindo, o lucro que fazíamos e foi aí que todos nos deparamos com a mesma situação, praticamente todos não retiravam qualquer lucro destas apostas a médio/longo prazo. Decidimos então ajudar-nos uns aos outros. Começamos por criar um grupo no *Facebook* em que partilhávamos as nossas opiniões relativamente a cada equipa que participa num determinado jogo, estas informações complementavam-se e levaram a que os resultados que atingíamos melhorassem, permitindo obter algum lucro nestas apostas.

Foi decidido desenvolver um sistema que efetuasse este procedimento de partilha de informação de forma automática e inteligente, sem ser necessário estudar as informações relacionadas com os jogos de futebol e tentar descobrir através destas informações, se através da utilização de técnicas de DM é possível ajudar a escolher qual a melhor aposta que se deve efetuar num determinado jogo de futebol.

## 1.2. Objetivos

Esta dissertação de mestrado tem como principais objetivos:

- Obter modelos com boas capacidades de previsão, que visam o suporte à decisão para que os seus utilizadores realizem a melhor aposta num determinado jogo de futebol;
- Desenvolver um protótipo de um Sistema Inteligente de Apoio à Decisão (SIAD) que incorpore os modelos desenvolvidos.

O primeiro objetivo deste projeto desdobra-se nos seguintes:

- Recolher dados estatísticos sobre os jogos de futebol (número de golos, número de remates, etc..);
- Efetuar um tratamento dos dados;
- Criar modelos de previsão de Data Mining;
- Fazer uma avaliação das métricas dos modelos;
- Escolher o modelo que satisfaz os requisitos do projeto.

O segundo objetivo divide-se nos seguintes:

- Criar um protótipo Sistema Inteligente de Apoio à Decisão;
- Integrar os modelos previamente criados no protótipo.

Cientificamente, esta dissertação pretende responder à seguinte questão de investigação:

***Existe viabilidade na utilização de modelos de previsão como base fundamental de um sistema inteligente de apoio à decisão para apostas em jogos de futebol?***

## 1.3. Abordagem Metodológica

Este tópico da dissertação apresenta a metodologia que foi utilizada no desenrolar deste projeto, esta será uma combinação entre três metodologias, uma é a metodologia de investigação que se designa de *Design Science & Research* (DSR) e duas de desenvolvimento, o *Cross Industry Standard Process for* Data Mining (CRISP-DM) e as fases do processo de decisão. A metodologia de desenvolvimento adotada para desenvolver este projeto é uma mistura entre estas duas metodologias distintas porque o projeto consiste em construir um protótipo de um sistema de apoio à decisão que vai ser suportado pela utilização de técnicas de DM, daí a necessidade de se combinar ambas as metodologias.

### 1.3.1. Design Science and Research

A investigação é uma atividade complexa que requer o uso de diversas atividades cognitivas que são por vezes mal compreendidas como a criatividade e a intuição. É considerada uma atividade semiestruturada. Não está definida nenhuma forma "perfeita" de efetuar a investigação, existem sim, diretrizes que demonstram ter qualidade que podem ser seguidas.

Vaishnavi & Jr, 2007 definem a investigação como uma atividade que contribui para a compreensão de um determinado fenómeno.

A metodologia de investigação que serve como base de todo o trabalho de recolha de informação relacionada com este projeto denomina-se DSR.

Esta metodologia identifica três principais objetivos. O primeiro é a consistência da pesquisa prévia. O segundo é que o processo deve proporcionar um método de investigação para conduzir o processo de *Design Science*. O terceiro objetivo do modelo de processo DSR é fornecer um modelo mental para as características dos resultados da investigação.

Peffers et al., 2006 definem este como um processo que através de uma investigação prévia resulta num modelo que consiste em 6 atividades representadas na Figura 1. Apesar do processo estar bem definido por vezes os investigadores não o seguem da primeira à última atividade sequencialmente.

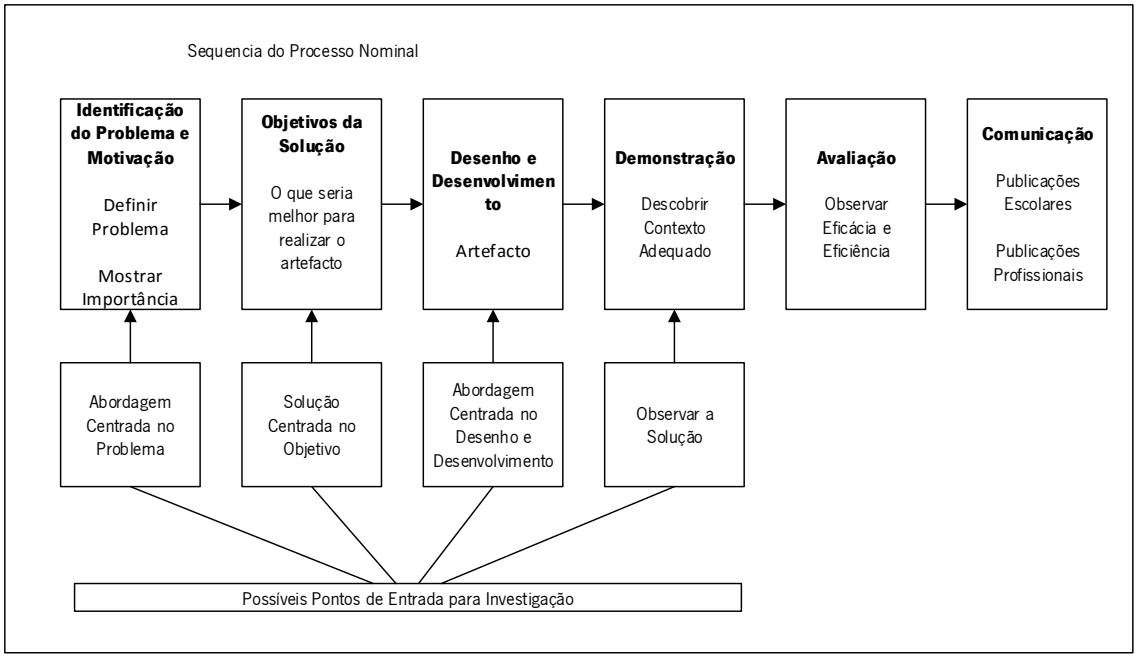


**Figura 1 - Fases Design Science Research adaptado de (Peffers et al., 2006)**

#### *1.3.1.1. Identificação e Motivação do Problema*

Esta primeira atividade consiste em definir a especificidade do problema de investigação e justificar o valor da mesma como solução. Desde a definição do problema, para desenvolver um artefacto como solução, pode ser útil focar o problema em pequenos detalhes para que a solução capte toda a complexidade do processo. Justificar o valor da solução motiva o investigador e o seu público-alvo a perseguir a solução, aceitar os resultados e ajuda a compreender o raciocínio associado à compreensão do problema do ponto de vista do investigador. Os recursos necessários para realizar esta atividade incluem o conhecimento do estado do problema e a importância que a solução do mesmo tem para o público.

#### *1.3.1.2. Objetivos da Solução*

Nesta atividade pressupor os objetivos da solução a partir da definição do problema é fundamental. Os objetivos podem ser quantitativos ou qualitativo e é espectável a criação de um artefacto para suportar a solução de problemas que até ao momento não tinham sido abordados. Os objetivos devem ser deduzidos racionalmente a partir da especificação do problema. Os recursos requeridos para esta atividade são o conhecimento do estado do problema e as suas soluções atuais e a respetiva eficácia.

#### *1.3.1.3. Desenho e Desenvolvimento*

Esta atividade foca-se na criação de uma solução em forma de artefacto. Estes artefactos são por norma construções, modelos, métodos ou instâncias (Hevner, March, Park, & Ram, 2004). As atividades incluem determinar a funcionalidade desejada do artefacto, a sua arquitetura e depois, por fim, a criação do artefacto em si. Os recursos utilizados para cumprir os objetivos e efetuar a conceção e desenvolvimento incluem o conhecimento da teoria que pode ser exercida como solução.

#### *1.3.1.4. Demonstração*

Esta atividade demonstra a eficácia do artefacto a resolver problemas. Pode envolver experimentação, simulação, um caso de estudo, provas ou outra atividade apropriada. Os recursos para esta demonstração incluem um conhecimento efetivo de como utilizar o artefacto para resolver o problema.

#### *1.3.1.5. Avaliação*

Nesta atividade o foco está em observar e medir corretamente como o artefacto suporta a solução do problema. Envolve comparar os objetivos da solução com os resultados observados depois da utilização do artefacto criado e demonstrado. É uma atividade que requer conhecimento em métricas relevantes e técnicas de análise. Dependendo da natureza do problema a avaliação do artefacto pode envolver a

comparação das suas funcionalidades com os objetivos definidos para a solução na atividade 2, comparar métricas de performance, verificar orçamentos, a satisfação na produção ou até verificar o feedback dos clientes. No final desta atividade os investigadores conseguem decidir se continuam para a atividade de comunicação, deixando trabalho a desenvolver futuramente ou se voltam para a atividade 3 para melhorar a eficiência do artefacto, sendo responsáveis por determinar a viabilidade da decisão.

#### *1.3.1.6. Comunicação*

A última atividade será comunicar o problema e a sua importância do artefacto, a sua utilidade e novidade, o rigor do seu *design* e a sua eficácia para os investigadores e outras entidades relevantes. A comunicação da investigação em trabalhos académicos podem utilizar o processo para estruturar todo o documento, bem como para efetuar todo o processo de investigação. Começando pela definição do problema, revisão da literatura, o desenvolvimento de hipóteses, recolha de dados, análise, resultados, discussão e conclusão. A divulgação dos resultados pode ser efetuada através da escrita de artigos científicos.

### 1.3.2. Cross Industry Standard Process for Data Mining

Esta metodologia é um processo que envolve seis fases sequenciais, começa com a compreensão do negócio e com a necessidade de utilização de DM no projeto e termina com um desenvolvimento de uma solução que satisfaça as mais específicas necessidades do negócio. Apesar de ser um processo sequencial é a maior parte das vezes favorável voltar atrás no processo visto que cada fase é consequência de trabalho anterior (E Turban, Sharda, & Aronson, 2008).

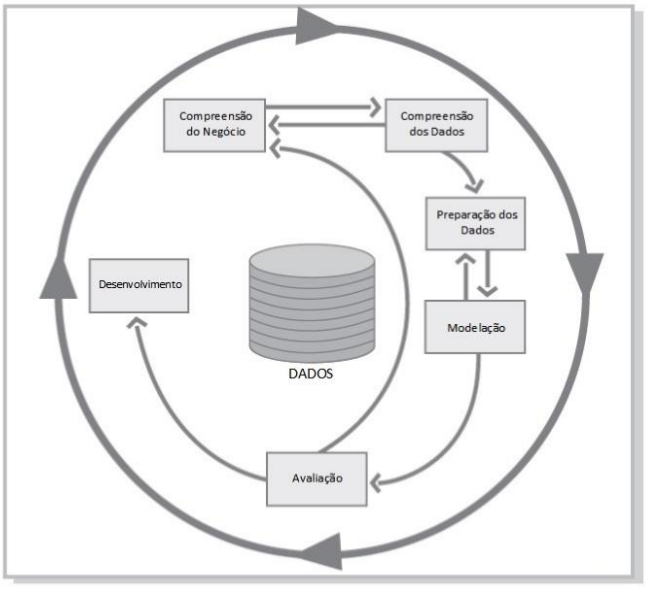

**Figura 2 - Fases do CRISP-DM retirada de (M. Y. Santos & Ramos, 2009)**

Em seguida encontra-se uma descrição de cada fase deste processo, que é apresentado na Figura 2, bem como as subtarefas que se encontram em cada uma delas.

#### *1.3.2.1. Compreensão do Negócio*

O elemento chave de cada estudo que envolva DM é entender qual o intuito do estudo. Focar essencialmente em entender os objetivos e requisitos a partir de uma perspetiva de negócio, convertendo o conhecimento na definição de um problema de DM. É também nesta fase que o orçamento do estudo deve ser estabelecido. Na Figura 3 estão apresentadas as subtarefas desta fase (Chapman et al., 2000; E Turban et al., 2008).

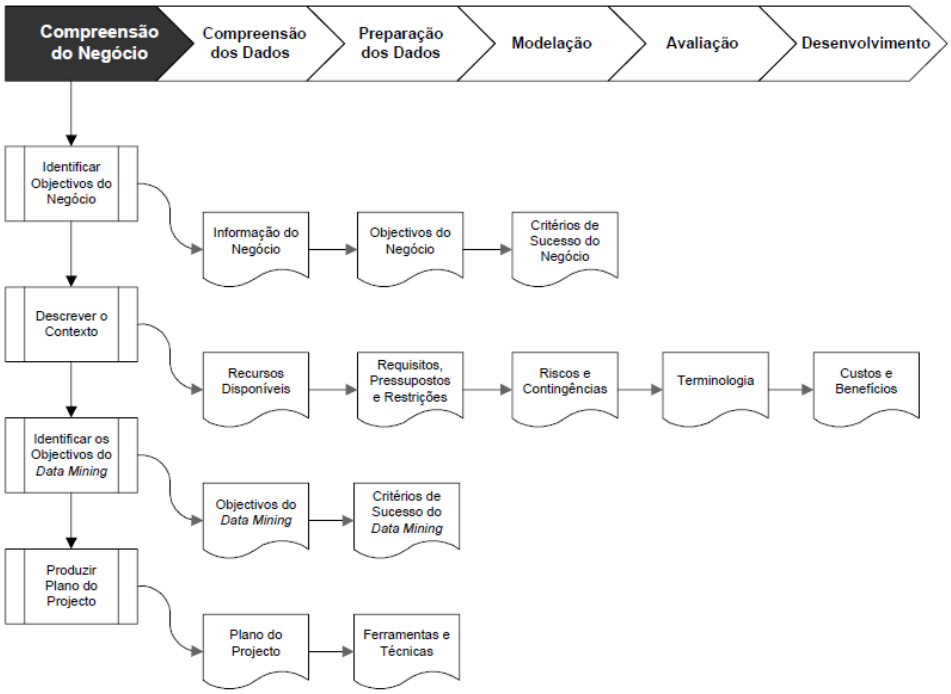


**Figura 3 - CRISP-DM Fase de Compreensão do Negócio retirada de (M. Y. Santos & Ramos, 2009)**

#### *1.3.2.2. Compreensão dos Dados*

Diferentes processos de negócio requerem diferentes tipos de dados. Nesta etapa a principal atividade será identificar quais os dados que são relevantes entre todos os existentes. Deve ser efetuada uma análise clara e concisa para que a informação relevante possa ser corretamente identificada. Na Figura 4 estão apresentadas as subtarefas desta etapa.

Para melhor entender os dados são feitos vários tipos de análises, usando variáveis estatísticas, técnicas gráficas, bem como uma simples sumarização estatística de cada variável, análise da correlação, caixas de bigodes, histogramas.

Um aspeto também bastante importante a ter em conta é uma seleção cuidadosa da fonte dos dados (E Turban et al., 2008).

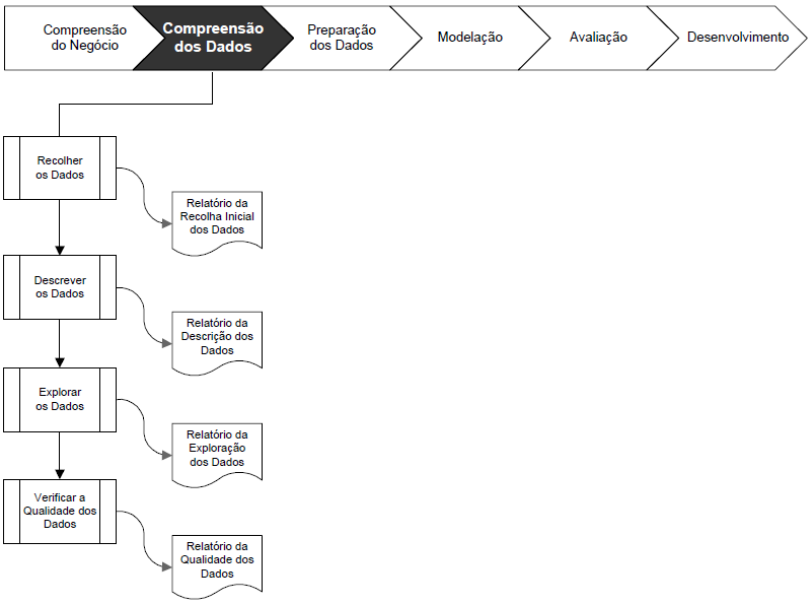


**Figura 4 - CRISP-DM Fase de Compreensão dos Dados retirada de(M. Y. Santos & Ramos, 2009)**

#### *1.3.2.3. Preparação dos Dados*

O propósito da preparação dos dados, como se verifica na Figura 5, é selecionar os dados identificados na fase anterior e preparar os mesmos para uma análise através de modelos de DM.

Esta fase comparada com as outras é a que ocupa maior parte do tempo e do esforço do trabalho. A razão do esforço ser enorme é porque é fundamental ter os dados estruturados de uma forma correta, os conjuntos de dados originais normalmente vêm com bastantes erros, sendo inconsistentes e incompletos. O processo inicia-se com a recolha e seleção dos dados, segue-se a limpeza dos mesmos, são tratados e por fim, são reduzidas as variáveis a utilizar para apenas as que são consideradas fundamentais para atingir o objetivo definido.

Dependendo da organização e dos seus objetivos, tipicamente esta fase segue segundo (Ibm, 2011; E Turban et al., 2008) as seguintes tarefas:

- Efetuar uma fusão dos conjuntos de dados;
- Selecionar uma subparte dos dados;

- Agregar atributos;
- Criar novos atributos;
- Ordenar os dados para modelar;
- Remover ou substituir os nulos ou valores inexistentes;
- Separar os dados, em data sets de treino e teste.

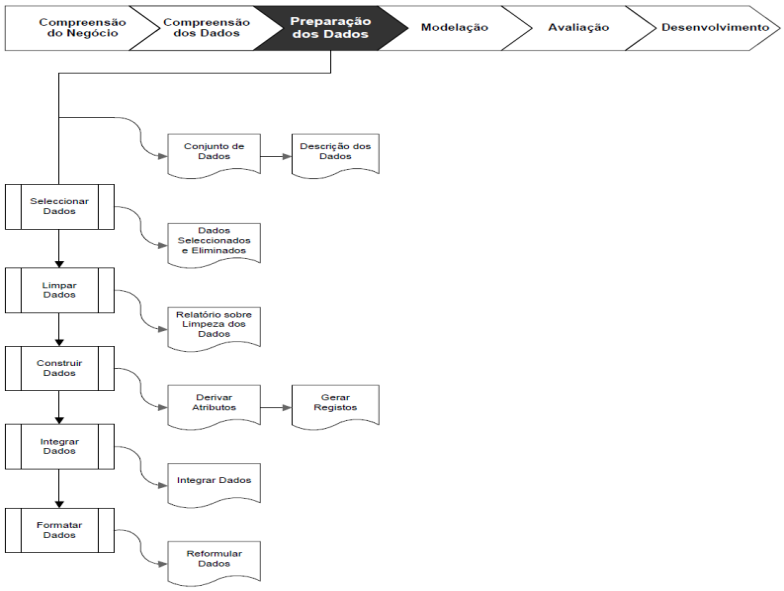


**Figura 5 - CRISP-DM Fase de Preparação dos Dados retirada de (M. Y. Santos & Ramos, 2009)**

#### *1.3.2.4. Modelação*

Nesta fase, como se pode verificar na Figura 6 são selecionados e aplicadas várias técnicas de modelação no conjunto de dados previamente preparados para a necessidade específica do negócio. A fase da construção de modelos também engloba a análise e comparação dos vários modelos construídos. Como não existe um algoritmo ou método considerado ideal a um nível universal para as tarefas de DM, devem ser usados vários tipos de modelos viáveis e ponderar uma estratégia para identificar o "melhor" método para um determinado propósito.

Dependendo das necessidades do negócio, o DM pode ser do tipo previsão (ou de classificação ou regressão), associação, ou segmentação. Cada um destes pode usar uma variedade de métodos e algoritmos de DM (Chapman et al., 2000; E Turban et al., 2008).

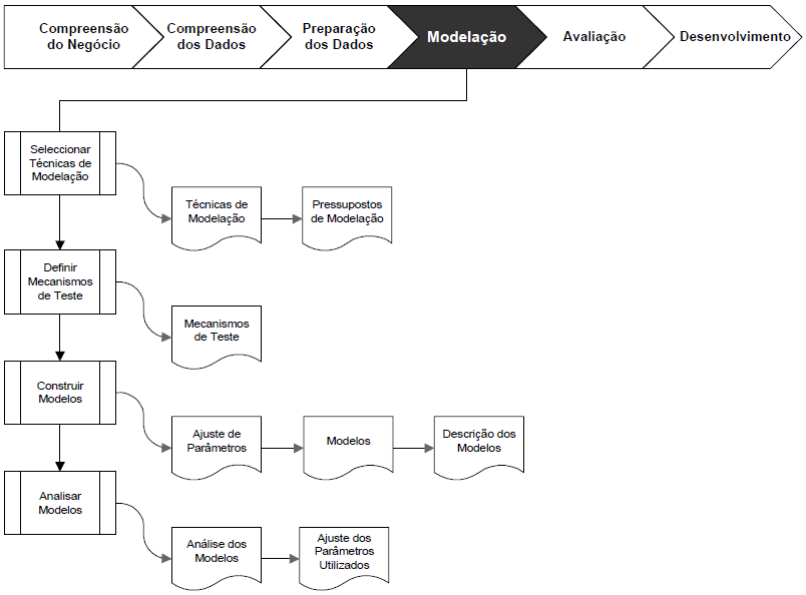


**Figura 6 - CRISP-DM Fase de Modelação retirada de (M. Y. Santos & Ramos, 2009)**

#### *1.3.2.5. Avaliação*

Nesta fase, como se pode verificar na Figura 7, os modelos previamente desenvolvidos são testados e avaliados tendo em conta a sua precisão. É necessário selecionar o modelo de acordo com os objetivos do negócio existentes, outra opção é testar os modelos desenvolvidos num cenário real se as restrições de tempo e orçamento assim o permitirem.

Esta é uma fase crítica e desafiadora do projeto, pois, o negócio não obtém valor nas fases de DM enquanto não forem descobertos padrões nos dados e consequentemente aplicá-los. O sucesso na identificação dos mesmos depende da interação entre analistas de dados, analistas de negócio e responsáveis por tomar a decisão. É necessária esta interação porque os primeiros não têm um total conhecimento dos objetivos do DM para o negócio e os restantes não têm o conhecimento técnico a interpretar os resultados através de sofisticadas soluções como tem o analista de dados (E Turban et al., 2008).

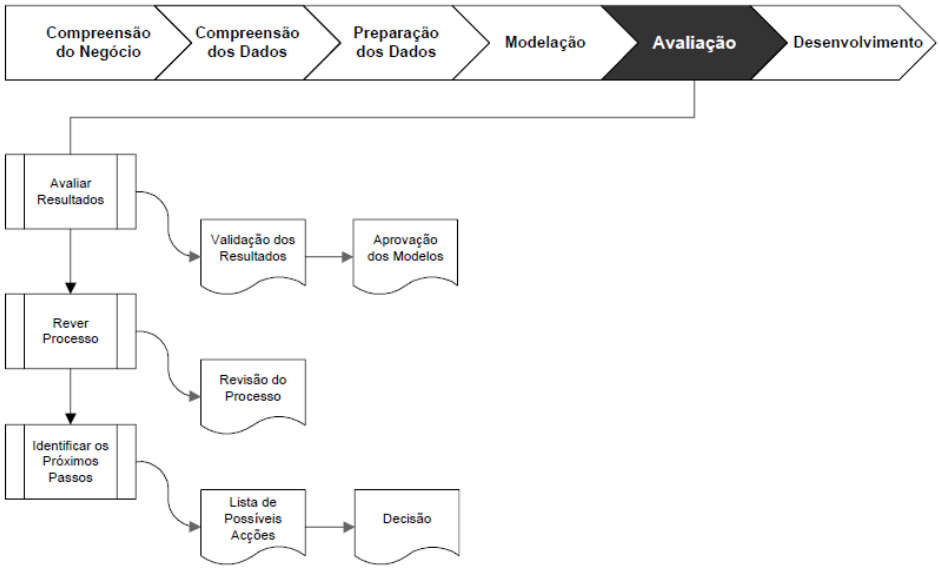


**Figura 7 - CRISP-DM Fase de Avaliação retirada de (M. Y. Santos & Ramos, 2009)**

### *1.3.2.6. Desenvolvimento*

Nesta fase, tal como se comprova na Figura 8, ocorre a geração de todos os documentos, bem como a implementação de plataformas e aplicações que tornem o processo de tomada de decisão mais simples para as organizações através de todo o processo de DM desenvolvido previamente (Chapman et al., 2000).

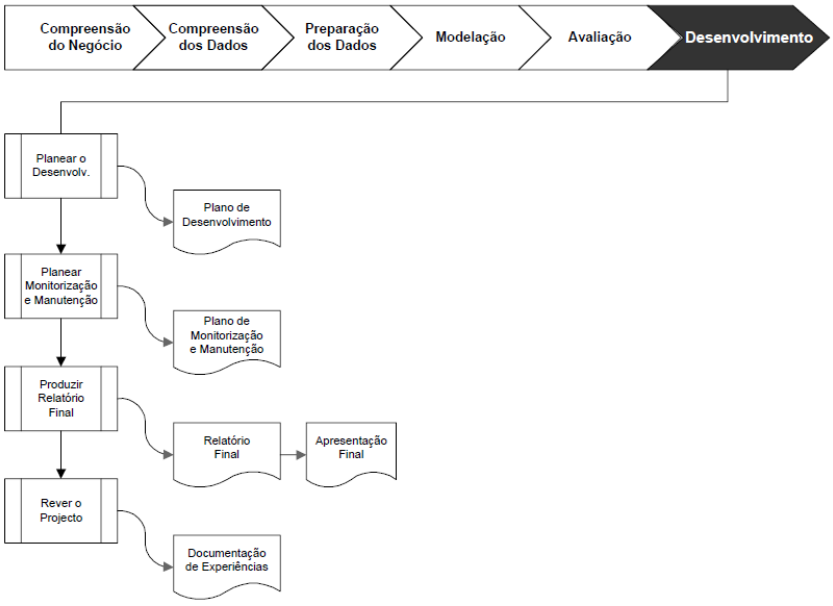


**Figura 8 - CRISP-DM Fase de Desenvolvimento retirada de (M. Y. Santos & Ramos, 2009)**

### 1.3.3. Fases da Tomada de Decisão

Os SADs seguem como base para o seu desenvolvimento as fases do processo de tomada de decisão. Simon (H. A. Simon, 1960) é o autor da metodologia que reúne um maior consenso entre a comunidade. Inicialmente este define o processo de tomada de decisão como tendo apenas três fases, a "Inteligência", a "Desenho" e a "Escolha". Anos mais tarde, Simon (H. a. Simon, 1977) e diversos outros autores defendem a divisão da fase "Escolha" definindo uma quarta fase a "Implementação", pois, defendiam que a implementação do que havia sido previamente decidido era bastante importante, a ponto de se criar uma fase individual para a mesma. Posteriormente (Efraim Turban, 2010) analisa a possibilidade de existência de uma outra fase, a "Monitorização" e explica-a como uma aplicação de uma fase Inteligência na fase Implementação, finalizando assim a metodologia que reúne consenso entre toda a comunidade científica. Na Figura 9 encontra-se a estrutura do processo de tomada de decisão.

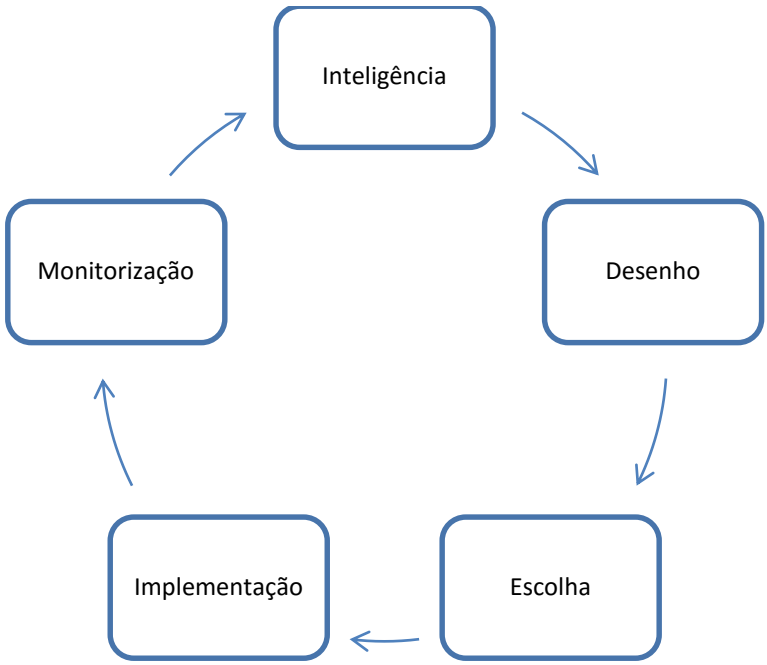


**Figura 9 - Fases do processo de tomada de decisão figura adaptada de (Efraim Turban, 2010)**

#### *1.3.3.1. Fase Inteligência*

"Inteligência" é a fase na qual o responsável por tomar decisões identifica, explica e define o problema que aparece no contexto de uma respetiva organização. É efetuada uma análise ao ambiente, recolhe-se os dados e avalia-se essa informação disponível de modo a permitir um rápido reconhecimento de sintomas e sinais que indiquem uma ação corretiva para melhorar o desemprenho de uma determinada tarefa. Nesta fase, faz-se uma comparação entre o corrente desempenho de um processo com aquilo

que foi previamente planeado. Faz-se uma categorização do problema, definindo se este é estruturado, semiestruturado ou não estruturado. Se este for um problema demasiado complexo, pode optar-se por subdividir o mesmo em pequenos problemas distintos. Atribui-se a responsabilidade a quem deverá resolver o problema. Sintetizando esta fase é responsável por formalizar o problema em si (Efraim Turban, 2010; Vercellis, 2009).

#### *1.3.3.2. Fase Desenho*

"Desenho" é a fase na qual as ações visam a resolução do problema definido na fase anterior, deve então desenvolver-se e planear. A este nível a experiência e criatividade dos responsáveis por tomar decisões torna-se critica, pois, estes são convidados a elaborar soluções viáveis de modo a que a finalidade pretendida seja alcançada. Quando o número de ações a tomar é pequeno, estes responsáveis, devem enumerar todas as possibilidades de forma a facilitar a identificação da melhor solução a escolher. Por outro lado se existirem inúmeras ações, a identificação da solução deve ser efetuada de forma implícita, usualmente através da descrição de regras que as ações viáveis devem satisfazer, estas regras podem identificar as limitações do modelo de otimização. De uma forma sumária esta fase envolve a descoberta de ações possíveis a desenvolver. Determina-se os critérios e objetivos da escolha, cria-se, testa-se e faz-se a validação do modelo de decisão, efetuando-se depois a modelação do mesmo (Efraim Turban, 2010; Vercellis, 2009).

#### *1.3.3.3. Fase Escolha*

Escolha é a fase na qual, depois de todas as ações alternativas serem identificadas, se faz uma avaliação às mesmas com base na performance dos critérios anteriormente definidos para definir a tomada de decisão a realizar. Esta fase procura e avalia os modelos criados na fase "Desenho" recomendando uma solução para o modelo.

Para escolher a melhor decisão a tomar pode-se fazer uma pesquisa de vários métodos como técnicas analíticas, algoritmos matemáticos, regras de procura que são também conhecidas como heurísticas e a procura cega que também é conhecida como procura completa. Depois desta pesquisa é necessário efetuar uma análise à robustez e para isso pode-se fazer uma análise de sensibilidade, ponderar cenários e descobrir que valores de entradas provocam determinados objetivos (Efraim Turban, 2010; Vercellis, 2009)

#### *1.3.3.4. Fase Implementação*

Nesta fase, depois de selecionada a melhor alternativa pelo responsável por tomar a decisão, é necessário transformar o plano em ações, basicamente é implementar o que está planeado. Esta fase envolve a atribuição de responsabilidades e regras a todos os envolvidos. Esta fase sumariamente é um processo que é responsável por pôr a solução a trabalhar. É um processo que envolve lidar com fatores como a resistência à mudança, treinar os utilizadores e dar suporte aos mesmos.

#### *1.3.3.5. Fase Monitorização*

"Monitorização" é a fase em que depois de a solução ter sido implementada, se verifica se as expectativas originais e os efeitos que se esperava que esta trouxesse estão ou não a corresponder. Em particular, as diferenças entre os valores dos indicadores de desempenho identificados na fase Escolha e os valores efetivamente observados no final do plano de execução devem ser medidos. Num SAD adequadamente planeado, os resultados dessas avaliações traduzem-se em experiências e informações, que são então transferidos para uma base de dados para serem utilizados durante os processos de tomadas de decisões posteriores (Vercellis, 2009).

### 1.3.4. Metodologia Adotada

As metodologias descritas previamente têm algumas características em comum e outras que se complementam, portanto para o sucesso do decorrer deste projeto foi considerado ideal efetuar uma combinação de todas elas. Esta combinação resulta na criação de uma metodologia que será mais completa e está mais preparada para responder a todos os eventuais problemas que possam vir a surgir. Esta metodologia é constituída por cinco fases, onde por exemplo na fase 1 combina a fase de compreensão do negócio do CRISP-DM com a fase "Inteligência" da tomada de decisão e com as fases "Identificação", "Motivação do Problema" e "Objetivos da solução" do DSR, na Tabela 1 encontra-se a combinação das fases das três metodologias utilizadas, essa combinação tem o seguinte resultado.

**Tabela 1 - Combinação das Metodologias**

| | | Metodologia Combinada | | | | |
|---|---|---|---|---|---|---|
| | | Fase 1 | Fase 2 | Fase 3 | Fase 4 | Fase 5 |
| **CRISP-DM** | Compreensão do Negócio | X | | | | |
| | Compreensão dos dados | | X | | | |
| | Preparação dos dados | | X | | | |
| | Modelação | | X | | | |
| | Avaliação | | | X | | |
| | Desenvolvimento | | | | X | |
| **Tomada de Decisão** | Inteligência | X | | | | |
| | Desenho | | X | | | |
| | Escolha | | | X | | |
| | Implementação | | | | X | |
| | Monitorização | | | | | X |
| **DSR** | Identificação e motivação do Problema | X | | | | |
| | Objetivos da Solução | X | | | | |
| | Design e Desenvolvimento | | X | | | |
| | Demonstração | | X | | | |
| | Avaliação | | | X | | |
| | Comunicação | | | | | X |

## 1.4. Estrutura do Documento

Este documento de dissertação está organizado com a seguinte estrutura:

- Introdução: é o capítulo da dissertação que apresenta uma introdução do tema que foi explorado, bem como uma descrição do enquadramento do projeto, quais os objetivos dos projetos e a metodologia utilizada no decorrer do mesmo.
- Estado da Arte e Enquadramento Conceptual: Neste tópico da dissertação está apresentada toda a revisão de literatura efetuada, os temas aprofundados foram a descoberta de conhecimento em bases de dados, o DM, os sistemas de apoio à decisão e foi também efetuada uma pesquisa com o objetivo de recolher informação sobre trabalhos já efetuados que tinham o mesmo objetivo deste projeto de dissertação.
- Trabalho Relacionado: Este capítulo contém o trabalho realizado previamente que deu origem a esta dissertação.
- Sistema Inteligente de Apoio à Decisão em Apostas de Jogos de Futebol: é a secção que apresenta o trabalho realizado neste projeto.
- Discussão de Resultados: Neste tópico estão discutidos os resultados obtidos com a criação dos modelos de DM e os testes realizados ao protótipo.
- Conclusão: No último capitulo é apresentada uma síntese de todo o trabalho e quais as contribuições cientificas do mesmo. Estão também identificadas algumas lacunas que podem ser exploradas futuramente e serem consideradas como orientações futuras.

# 2 ESTADO DA ARTE E ENQUADRAMENTO CONCEPTUAL

Para a elaboração do estado da arte e enquadramento conceptual foram consultados e utilizados diversos motores de pesquisa de publicações científicas. Entre os utilizados destacam-se:

- Web of Knowledge
- ScienceDirect
- Springer
- Scopus
- Google Scholar

A pesquisa efetuada teve como base as seguintes palavras-chaves/expressões:

- Data Mining
- Data Mining Methodologies
- Data Mining Classifiers
- Decision Support Systems
- Decision Support Systems for Football Betting
- Knowledge Discovery in Databases
- Soccer Predictions
- Football Predictions
- Soccer Game Result Prediction
- Football Game Result Prediction
- Soccer Bets
- Football Bets
- Systems for Sports Predictions

Para a revisão de literatura e artigos a utilizar foram considerados fatores como a relevância do autor na área, a reputação dos artigos publicados assim como o ano em que o artigo foi publicado, bem como a pesquisa dos mesmos termos em outras linguagens (ex. português).

## 2.1. Descoberta de Conhecimento em Bases de Dados

Segundo os autores Maimon, Oded & Rokach (Maimon, Oded; Rokach, 2010), Descoberta de Conhecimento em Bases de Dados (DCBD) é uma análise exploratória e modelação automática de grandes repositórios de dados. É um processo organizado que tem o intuito de identificar padrões úteis que se consigam entender em grandes e complexos conjuntos de dados.

É um processo interativo e iterativo no qual é exigida a interação, em diversas fases, de um responsável por tomar decisões (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

A DCBD é segundo (Ibm, 2011; E Turban et al., 2008) um processo complexo que tem como intuito identificar padrões válidos, originais, potencialmente utilizáveis e de fácil compreensão num conjunto de dados. A

A Figura 10 apresenta o processo de DCBD que é constituído por cinco distintas fases, este processo é iniciado pela recolha dos dados e termina com a obtenção de um novo conhecimento.

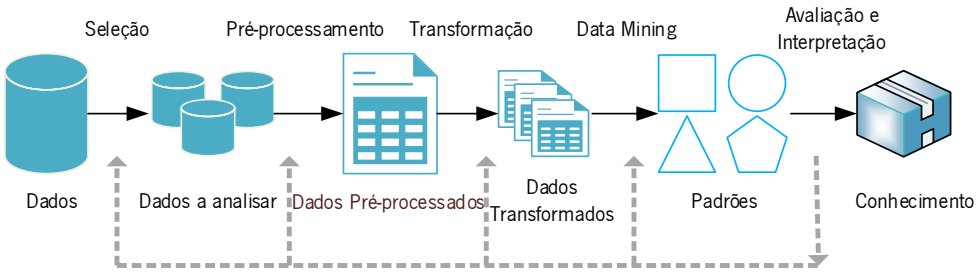


**Figura 10 - Processo de DCBD adaptado de (Vercellis, 2009)**

- Seleção: é a fase na qual é realizada a seleção ou, até mesmo, a criação de um conjunto de dados no qual a descoberta de conhecimento será realizada. Uma vez definidos quais os objetivos a realizar, os dados utilizados para a descoberta de conhecimento devem ser determinados. Isto inclui descobrir quais os dados disponíveis, obter dados adicionais, para uma posterior integração dos mesmos no processo DBCD, levando a que os atributos sejam considerados para o processo. Este processo é fundamental porque o Data Mining (DM) tem a capacidade de aprender e descobrir a partir dos dados disponíveis, sendo esta é a base de evidências que suporta a construção dos seus modelos. Se alguns atributos importantes não forem considerados no processo, poderá implicar que todo o estudo venha a ser um fracasso. Para que o processo seja bem-sucedido é necessário encontrar o número ideal de variáveis e considerar o maior número possível atributos nesta fase (Maimon, Oded; Rokach, 2010);

- Pré-processamento: esta fase inclui a realização de uma limpeza e um pré-processamento dos dados. São efetuadas operações básicas como remover o "ruído" existente nos dados, se for apropriado, decidir quais as estratégias a se devem utilizar para lidar com os registos que tenham campos em falta, representando informações e as suas respetivas mudanças em sequências temporais (Fayyad et al., 1996);
- Transformação: nesta fase é efetuada uma redução e projeção dos dados, tentando encontrar recursos para conseguir representar as dependências dos dados em relação aos objetivos da tarefa. É utilizada a redução de dimensionalidade ou métodos de transformação para reduzir o número efetivo de variáveis a ter em consideração (Fayyad et al., 1996);
- Data Mining: esta fase da DCBA irá ser profundamente analisada no ponto 2.2 desta dissertação;
- Avaliação: é a fase na qual se faz a avaliação e interpretação dos padrões encontrados que vão ao encontro dos objetivos definidos na primeira fase. As etapas de pré-processamento são consideradas em relação ao seu efeito sobre os resultados do algoritmo de DM. Esta etapa tem como foco a compreensão e utilidade do modelo a induzir. O conhecimento descoberto é documentado para uma posterior utilização (Maimon, Oded; Rokach, 2010).

## 2.2. Data Mining

Nas últimas décadas a tendência da maior parte das organizações tem sido o armazenamento de dados, informações relativas a clientes, a vendas, a compras, a cada área de negócio específica de cada uma delas. Com a evolução evidente do *hardware* é possível guardar cada vez mais informação (Han, Kamber, & Pei, 2012; J. Cios, Pedrycz, W. Swiniarski, & A. Kurgan, 2007). Com este crescimento de armazenamento de informação as técnicas tradicionais de análise e exploração dos dados deixou de ser adequada para tratar os enormes conjuntos de dados. Um problema associado a esta atividade é o facto de serem feitos grandes investimentos com o intuito de identificar e recolher informação útil nesses conjuntos de dados e por vezes nada ser encontrado. De forma a conseguir obter maior proveito de toda esta informação surgiu a técnica de DM (T. Larose & Larose, 2014).

Não existe uma definição consensual de DM, os autores Rajaraman & Ullman (Rajaraman & Ullman, 2011) sugerem como a mais comum definição de DM, a descoberta de modelos para os dados. Esses modelos podem, no entanto, corresponder a diferentes contextualizações como por exemplo, modelação

estatística, *Machine Learning* (ML), abordagens computacionais para a modelação, sumarização e extração de características. Os autores (E Turban et al., 2008) identificam, no início do século, que o DM a partir de bases de dados organizacionais será fundamental num futuro próximo. Segundo os autores, esta técnica se tornará tão importante que as organizações não poderão abdicar de quaisquer informações recolhidas sobre os seus clientes, arriscando-se a estar fora do negócio. Consideram-no, então, a próxima arma estratégica organizacional. Inicialmente definiam DM como um processo através do qual se detetavam padrões nos dados, mas a definição foi modificada aumentando assim, a sua abrangência a uma análise de dados que tem como objetivo aumentar a eficiência e eficácia das organizações. Sauter (Sauter, 2011) define DM como uma procura profunda num determinado conjunto de dados que tem como objetivo obter algo com valor. Considera-o um processo de extração de valiosos padrões a partir de uma grande quantidade de dados.

Uma definição mais concisa, mas também bastante abrangente de DM é que esta técnica consiste numa aplicação de técnicas e métodos em grandes bases de dados de modo a ser possível encontrar tendências ou padrões como forma de apoiar a descoberta de novo conhecimento (M. F. dos Santos & Azevedo, 2005). O termo de DM tem sido utilizado essencialmente pelas comunidades estatísticas, de analistas de dados e de gestão de sistemas de informação (Fayyad et al., 1996).

O DM está associado a diversas áreas podendo ser considerado uma combinação multidisciplinar (E Turban, Sharda, & Delen, 2011). A Figura 11 procura demonstrar essa multidisciplinariedade.

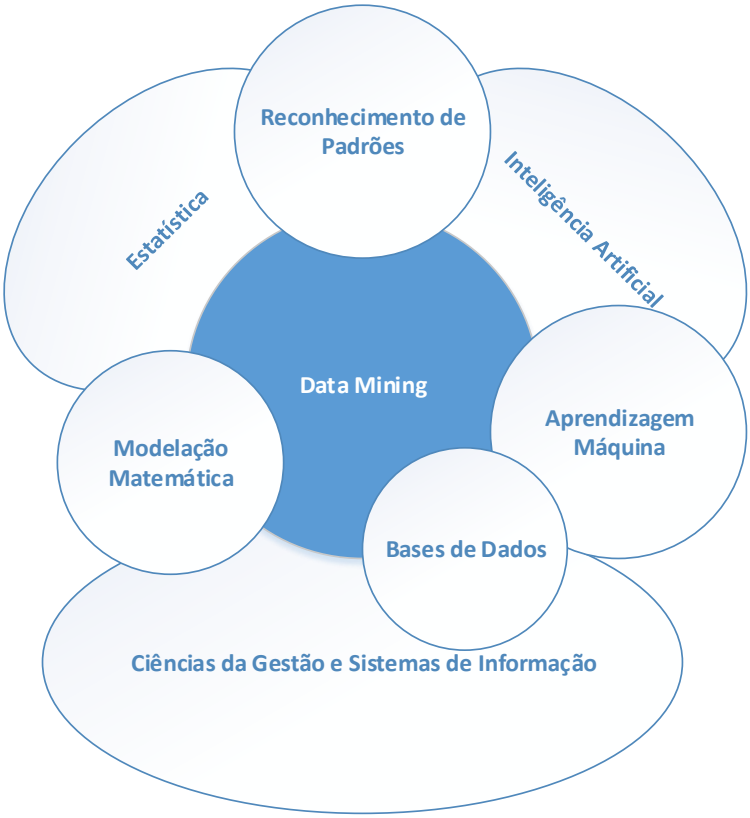


**Figura 11 - Áreas Associadas ao DM adaptado de (Efraim Turban, 2010)**

Os autores (Olson & Delen, 2008; Witten, Frank, & Hall, 2011) identificam as seguintes áreas como sendo as áreas onde o DM já foi aplicado com sucesso:

- Retenção de clientes: identificação de perfis para determinados produtos, venda cruzada;
- Bancos: identificar padrões para auxiliar na gestão de relacionamento com o cliente;
- Cartão de Crédito: identificar segmentos de mercado e identificar padrões de rotatividade;
- Cobrança: deteção de fraudes;
- Telemarketing: acesso facilitado aos dados do cliente;
- Eleitoral: identificação de um perfil para possíveis votantes;
- Medicina: indicação de diagnósticos mais precisos;
- Tomada de Decisões: filtrar as informações relevantes, fornecer indicadores de probabilidade.

O DM contém técnicas ou atividades que podem ser subdivididas em dois grandes focos de investigação, de acordo com a análise que se pretende realizar, podem ser análises interpretativas ou preditivas. Os tipos de padrões a detetar são diferentes em cada uma destas análises. Esses focos segundo (Vercellis, 2009) são:

- Previsão: estas tarefas têm como objetivo prever uma determinada classe;
- Interpretação: neste tipo de tarefas espera-se proporcionar ao utilizador final uma fácil análise aos padrões identificados, de modo a ir ao encontro do objetivo de cada tarefa de interpretação.

As tarefas de DM de previsão estão divididas em duas categorias:

- Classificação;
- Regressão.

Já as tarefas de interpretação são:

- Caracterização e Discriminação;
- Séries Temporais;
- Regras de Associação;
- Segmentação;
- Descrição e Visualização.

As tarefas de previsão mais comuns são a classificação, regressão, enquanto as de interpretação mais comuns são a segmentação e regras de associação (Gorunescu, 2011).

Na aprendizagem supervisionada, os dados incluem um atributo alvo, cujos valores podem ser estimados utilizando os atributos de entrada do registo. O objetivo de um algoritmo de DM utilizado nestas tarefas é aprender, a partir de um conjunto de dados, um modelo ou hipótese capaz de relacionar os valores dos atributos de entrada do registo com o valor do atributo alvo. As tarefas de descrição seguem o paradigma da aprendizagem não-supervisionada em que não existe um atributo alvo. De forma oposta às tarefas supervisionadas, as tarefas não-supervisionadas podem ser mais difíceis de avaliar, dado que não existe um atributo alvo com o qual se pode comparar e assim avaliar o desempenho do modelo (Gama, Carvalho, Faceli, Lorena, & Oliveira, 2012a).

### 2.2.1. Tarefas de Data Mining

As tarefas de DM tal como descrito anteriormente dividem-se em duas categorias distintas, a previsão e a descrição. Neste documento apenas vão estar descritas as tarefas de previsão.

Existem dois géneros de tarefas de previsão, a classificação e a regressão.

A classificação é a tarefa de DM utilizada mais frequentemente, esta tem como objetivo analisar dados históricos de um conjunto de dados e automaticamente gerar um modelo que consiga prever um comportamento futuro.

Os modelos de classificação têm como objetivo identificar relações recorrentes entre variáveis que são consideradas características de uma única classe. Estas relações estão traduzidas em regras de classificação que podem obter diferentes valores de acordo com o tipo de modelo utilizado (E Turban et al., 2008; Vercellis, 2009).

Num problema de classificação, existe um conjunto de dados que contém observações descritas em vários atributos explicativos e um atributo alvo que tem de ter formato categórico. Os atributos explicativos também conhecidos como variáveis de previsão podem ser categóricos e numéricos. O atributo alvo também é conhecido como classe e as observações podem ser chamadas de instâncias ou exemplos. A variável-alvo tem um número finito de valores. A classificação pode ser binária se o problema envolver apenas duas classes e multi-classe ou multi-categoria se este tiver mais que duas classes.

Dois casos práticos em que a tarefa de classificação poderia ser utilizada são, por exemplo, na tomada de decisão dos bancos, pois, quando estes tentam determinar se o empréstimo a um determinado cliente é considerado 'arriscado' ou então 'seguro'. Outro caso no qual pode ser aplicado é na criação de um modelo para um hospital que ajude um médico a determinar qual tratamento que um específico doente deve receber 'Tratamento A', 'Tratamento B' ou 'Tratamento C' (Han et al., 2012; E Turban et al., 2008).

As técnicas de DM mais comuns para elaborar modelos de classificação são, redes neuronais artificiais, árvores de decisão, classificadores de *Bayes* e algoritmos genéticos (Efraim Turban, 2010).

Nos últimos anos surgiram outras técnicas que têm tido excelentes resultados em problemas de classificação, como é o caso das *Support Vector Machine* (SVM).

As Redes neuronais são modelos de computação utilizados para o processamento de informações e são particularmente úteis para a identificação de uma relação fundamental entre um conjunto de variáveis ou até mesmo padrões. Elas cresceram a partir da pesquisa em inteligência artificial, mais especificamente tenta imitar a aprendizagem das redes neuronais biológicas, especialmente aquelas em que o cérebro humano pode conter mais neurónios interligados. Embora as redes neuronais artificiais sejam abstrações extremamente simples de sistemas biológicos são bastante limitadas em tamanho, capacidade e poder comparando com as redes neuronais biológicas. As redes artificiais e biológicas compartilham duas características muito importantes, o processamento paralelo de informações e a aprendizagem e generalização da experiência (Maimon, Oded; Rokach, 2010).

As árvores de decisão (AD) classificam os dados num número finito de classes que se baseiam no valor das variáveis de entrada. Este método é essencialmente uma hierarquia de instruções "$se \rightarrow então$" e são significativamente mais rápidos que as redes neuronais. São métodos mais apropriados para identificar dados categóricos e intervalos de dados (E Turban et al., 2008).

São os algoritmos de classificação mais conhecidos e mais utilizados em aplicações de DM. As razões para a sua popularidade estão na sua simplicidade conceitual, a sua facilidade de uso, a velocidade computacional, a robustez em relação à falta de dados e *outliers* e, acima de tudo, a sua capacidade de interpretar as regras que geram. Para separar as observações pertencentes a diferentes classes, os métodos baseados em árvores obtêm regras simples e explicativas para a relação existente entre a variável-alvo e variáveis de previsão (Vercellis, 2009).Na Figura 12 está um exemplo de uma árvore de decisão.


**Figura 12 - Exemplo Árvore de Decisão retirado de ("Genómica Funcional e Bioinformática," 2012)**

Classificação de *Bayes* é uma técnica estatística, relacionada com a probabilidade condicional, baseada no teorema de Thomas Bayes. É calculada a probabilidade posterior $P\ (y|x)$ que uma determinada observação pertence a uma classe específica de destino, uma vez que a probabilidade anterior $P\ (y)$ e as probabilidades condicionais classe $P\ (x|y)$ são conhecidas. A teoria de *Bayes* pode ser expressada pela seguinte expressão:

$$P\ (y|x) = \frac{P\ (x|y)P\ (y)}{P\ (x)}$$

Na qual $P\ (y|x)$representa a probabilidade *à posteriori*, $P\ (y)$ a probabilidade *à priori*, $P\ (x|y)$ a função densidade de probabilidade (a probabilidade da classe $x$) e $P\ (y|x)$ a função densidade de probabilidade incondicional (Langley, Iba, & Thompson, 1992)

*Support Vector Machine* (SVM) é um algoritmo de aprendizagem que visa resolver problemas de classificação de duas classes. A máquina conceitualmente coloca em prática a seguinte ideia. Os vetores de entrada são mapeados para um espaço de características de elevada dimensão de uma forma não linear e, neste espaço, é construída uma decisão, garantindo as características especiais deste espaço e tendo uma grande e generalizada capacidade de aprendizagem da máquina. Inicialmente este algoritmo foi desenvolvido especificamente para os casos onde os dados do conjunto de treino podiam ser separados sem erros mas, posteriormente, este objetivo foi alargado de modo a incluir dados dos conjuntos de treino que não estejam separados (Cortes & Vapnik, 1995). O SVM coloca todos os casos possíveis distribuídos no espaço tentando depois encontrar a separação ótima entre valores.

*Lazy Learners* (LL) ao contrário das técnicas de classificação descritas até agora utilizam um conjunto de dados de treino para aprender a classificar um novo registo. Assim, quando são submetidas a um novo registo elas já estão preparadas, ou seja, já aprenderam. Existe, no entanto, uma outra categoria de métodos, que somente realizam essa aprendizagem quando é solicitada a classificação de um novo registo. Neste caso, a aprendizagem é considerada tardia (*Lazy Learning*). Apesar de precisar de um tempo menor para efetuar a fase de treino, estes métodos são muito dispendiosos computacionalmente, pois necessitam de técnicas que armazenem e recuperem os dados de treino. Por outro lado, estes métodos permitem uma aprendizagem superior (W. Aha, 1997).

Os conjuntos *Fuzzy* foram propostos por Lotfi Zadeh (Zadeh, 1965), a ideia dos conjuntos *Fuzzy* é de em vez de se realizar um corte direto, as variáveis sejam caracterizadas e agrupadas em categorias e que a lógica *Fuzzy* seja aplicada para definição dos limites destas categorias. Com isso, ao contrário de se ter as categorias com limites de corte bem definidos, tem-se um certo grau de flexibilidade entre as categorias.

### 2.2.2. Avaliação na Classificação

O desenvolvimento de modelos de classificação Figura 13 consiste em três fases principais. A primeira é a fase de treino, na qual o algoritmo de classificação é aplicado em apenas uma parte dos dados, ao chamado conjunto de treino, a fim de obter regras de classificação que permitam a efetuar a correspondência da classe alvo a cada observação. A segunda é a fase de teste, na qual as regras geradas na fase de treino são utilizadas para classificar as observações que não foram introduzidas no conjunto de treino, isto para que a classe alvo seja então conhecida. Para ser avaliada a precisão do modelo é comparada a atual classe alvo de cada instância do conjunto de treino com a classe prevista

pelo modelo. A última fase é conhecida como fase de previsão, esta representa a utilização efetiva do modelo de classificação para atribuir a classe alvo para novas observações que serão gravadas no futuro. A previsão é obtido através da aplicação das regras geradas durante a fase de treino para as variáveis explicativas que descrevem a nova instância (Vercellis, 2009).

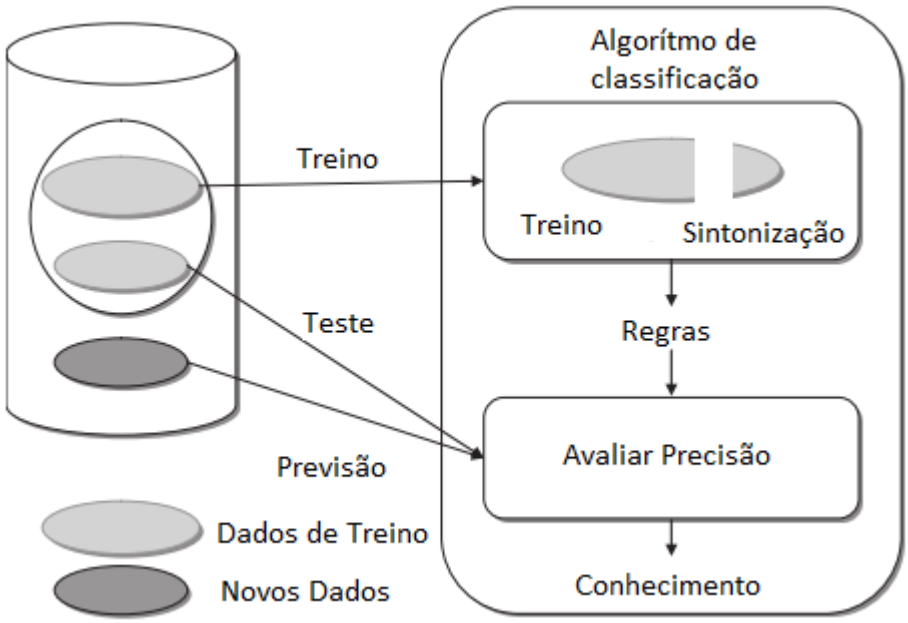


**Figura 13 - Fases dos modelos de classificação figura adaptada de (Vercellis, 2009)**

Para se poder avaliar o desempenho de um modelo é necessário definir o conjunto de dados de treino em que o valor da variável-alvo é conhecido, ficando os restantes dados para teste, garantindo sempre que os dados de treino são diferentes dos dados de teste. Essa divisão dos dados é necessária para que não se dê o fenómeno de *overfitting*. Este fenómeno dá-se quando o modelo fica dependente de um conjunto de dados específico e, ao ser submetido a outros conjuntos (com valores diferentes dos usados na construção e validação do modelo), apresenta resultados insatisfatórios. A divisão dos dados pode ser feita utilizando várias técnicas:

- *Holdout:* A partir de um conjunto de dados de tamanho N, divide-se numa proporção P*N para treino e (1-P)*N para teste. Esta abordagem é adequada quando há um grande volume de dados. A Figura 14 mostra a forma como é efetuada a divisão dos dados através da técnica *holdout* (Gama, Carvalho, Faceli, Lorena, & Oliveira, 2012).

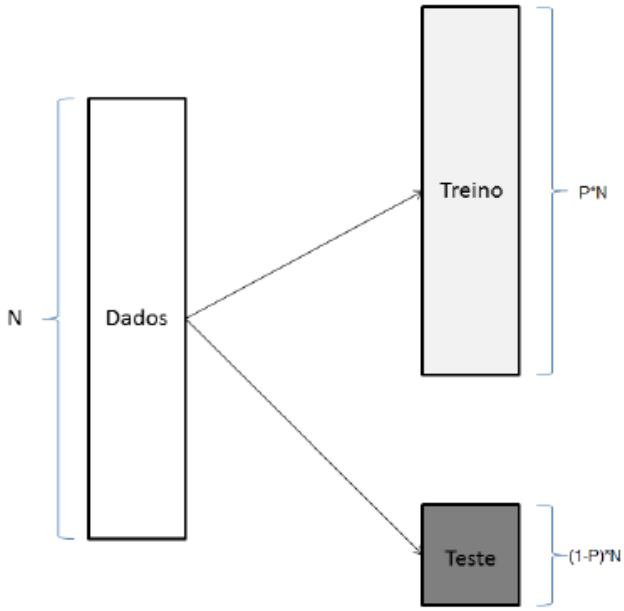


**Figura 14 - Divisão do Conjunto de Dados Holdout retiradode (Gama et al., 2012a)**

- Amostragem Aleatória: A amostragem aleatória contraria a dependência existente no *holdout* executando o método *holdout* diversas vezes com partições de teste aleatórias como se pode ver na Figura 15. As proporções P para treino e (1-P) para teste mantêm-se em todas as iterações. Os resultados deste método são dados pela média dos diferentes testes (Gama et al., 2012).

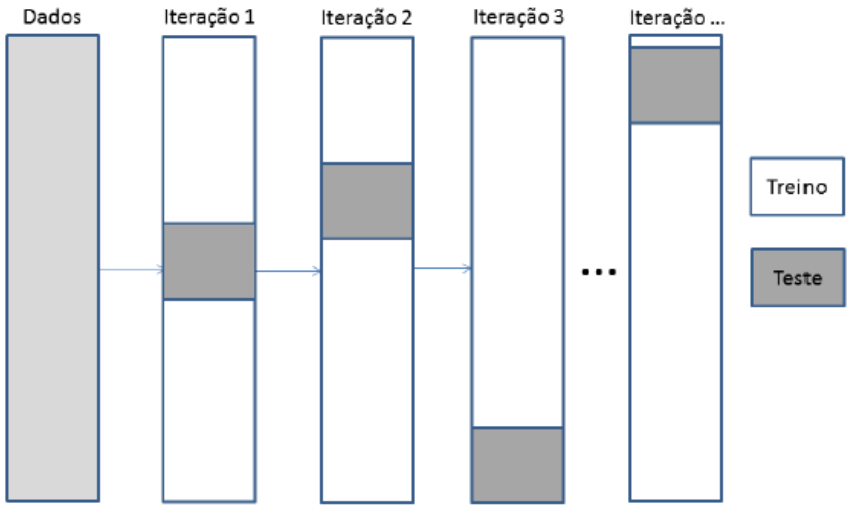


**Figura 15 - Divisão do Conjunto de Dados Amostragem Aleatória retirado de (Gama et al., 2012)**

- *K-Fold Cross-Validation:* O conjunto de dados é dividido em K subconjuntos de tamanho aproximadamente ou até mesmo igual. Uma das partições é usada para teste, enquanto as

restantes são utilizadas no treino do método. Este processo é realizado K vezes, utilizando em cada ciclo uma partição diferente para teste. O desempenho final é dado pela média dos desempenhos observados sobre cada subconjunto de teste (Gama et al., 2012). A Figura 16representa o funcionamento do método *K-fold cross-validation*.

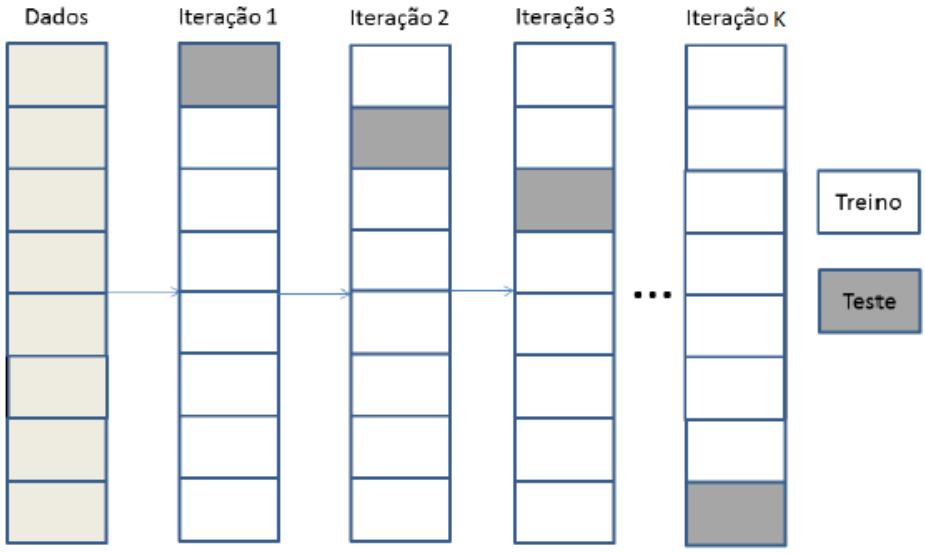


**Figura 16 - Divisão do Conjunto de Dados *K-Folds Cross-Validation* retirado de (Gama et al., 2012)**

- Bootstrap: No *Bootstrap* são gerados x subconjuntos de treino a partir do conjunto de exemplos original. Os exemplos são amostrados aleatoriamente desse conjunto, com reposição. O resultado é dado pela média do desempenho em cada subconjunto de teste (Gama et al., 2012).

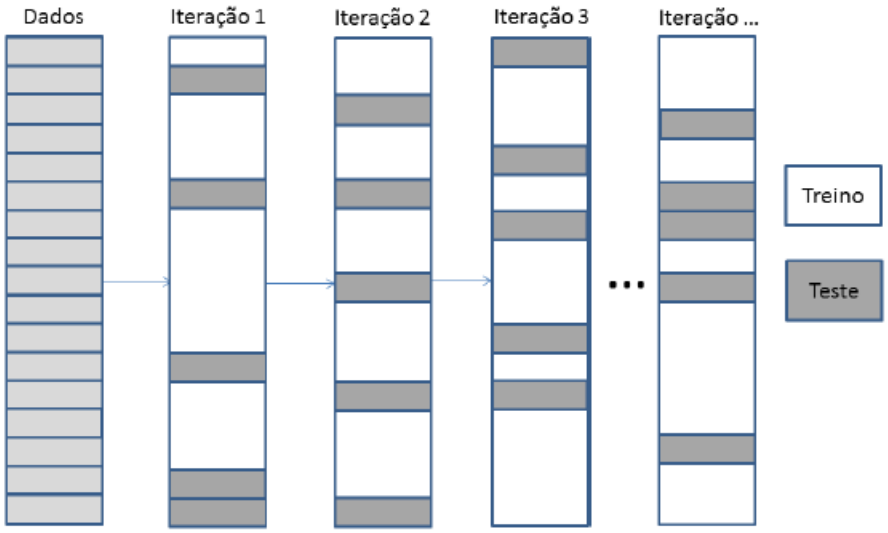


**Figura 17 - Divisão do Conjunto de Dados Bootstrap retirado de (Gama et al., 2012)**

Em problemas de classificação a principal fonte de consideração da precisão é conhecida como a matriz de confusão (Tabela 2), esta mostra a tabulação dos resultados de classificação de duas diferentes classes.

**Tabela 2 - Matriz de confusão adaptado de (Efraim Turban, 2010)**

| **Matriz Confusão** | | **Classe Realidade** | |
|---|---|---|---|
| | | Positivo | Negativo |
| **Classe de Previsão** | Positivo | VP | FP |
| | Negativo | FN | VN |

A matriz de confusão (tabela 2) permite obter:

- Verdadeiros Positivos (VP): correspondem ao número de exemplos positivos classificados como tal (corretamente);
- Falsos Positivos (FP): correspondem ao número de exemplo positivos classificados como negativos (incorretamente);
- Falsos Negativos (FN): correspondem ao número de exemplo negativos classificados como positivos (incorretamente).
- Verdadeiros Negativos (VN): correspondem ao número de exemplo negativos classificados como tal (corretamente).

A partir da matriz de confusão é possível retirar as seguintes métricas que servem para avaliar os modelos de classificação criados (E Turban et al., 2008).

- Acuidade – calcula a proporção de casos classificados corretamente;

$$Acuidade = \frac{VP + VN}{VP + VN + FP + FN} \times 100\%$$

- Sensibilidade – é a proporção de verdadeiros positivos que são corretamente identificados como positivos pelo classificador;

$$Sensibilidade = \frac{VP}{VP + FN} \times 100\%$$

- Especificidade – é a proporção dos verdadeiros negativos e está relacionada com a capacidade do classificador identificar resultados negativos.

$$Especificidade = \frac{VN}{VN + FP} \times 100\%$$

Outra medida de avaliação dos modelos de classificação são as curvas denominadas de *Receiver Operating Characteristics* (ROC) que permitem uma avaliação de desempenho de um classificador, a sua utilização é possível quando existem duas classes de previsão. Permite visualizar a relação entre a sensibilidade e a especificidade do modelo. Numa situação ideal o modelo deveria possuir indicadores máximos de sensibilidade e especificidade, ambos iguais a um. A partir da curva ROC é possível utilizar duas técnicas, a *Area Under Curve* (AUC) e a análise ROC *Convex Hull* (ROCCH). A primeira consiste numa métrica de desempenho do classificador que é obtida através do cálculo da área que se encontra por baixo da curva ROC, esta assume valores entre zero e um. A segunda permite declarar um subconjunto de classificadores como potencialmente ótimos. Incluídos todos os pontos que constituem as curvas ROC de todos os diferentes classificadores e formada a *convex hull* que lhe corresponde, é realizada uma análise dos pontos que se encontram acima da linha. Se um dos pontos está acima da linha, existe então uma linha tangente ao mesmo que tenha uma sensibilidade superior, sendo o classificador representado por esse ponto considerado ótimo sob a distribuição assumida correspondente a essa inclinação (M. F. dos Santos & Azevedo, 2005). Na Figura 18 está apresentado um exemplo duma curva ROC.

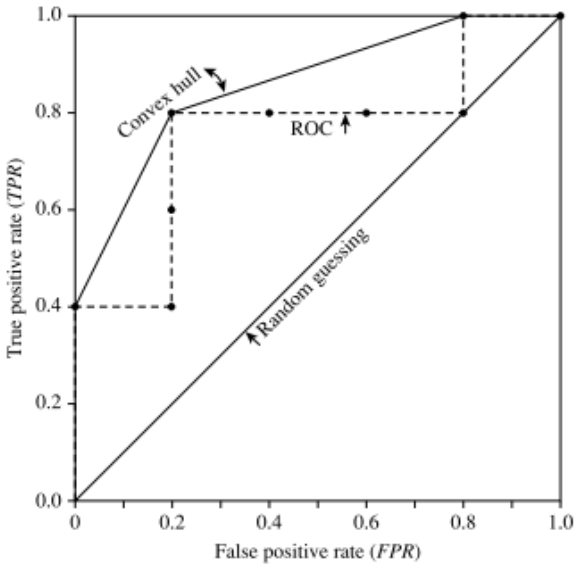


**Figura 18 - Curva ROC figura retirada de (Han et al., 2012)**

### 2.2.3. Algoritmos de classificação

Neste tópico são apresentados os algoritmos de classificação utilizados na dissertação, um algoritmo relacionado com os classificadores de *bayes,* um com os SVM, outro com as AD e por último um de LL. Todos os algoritmos apresentados são disponibilizados pela ferramenta WEKA e têm como objetivo induzir modelos de DM.

#### *2.2.3.1. NaiveBayes*

O funcionamento deste algoritmo baseia-se no teorema de *Thomas Bayes* e está explicado no tópico 2.2.1**Erro! A origem da referência não foi encontrada.** deste documento.

#### *2.2.3.2. LibSVM*

É um algoritmo de aprendizagem que visa resolver problemas de classificação de duas classes. A máquina conceitualmente coloca em prática a seguinte ideia. Os vetores de entrada são mapeados para um espaço de características de elevada dimensão de uma forma não linear e, neste espaço, é construída uma decisão, garantindo as características especiais deste espaço uma grande e generalizada capacidade de aprendizagem da máquina. Inicialmente este algoritmo foi desenvolvido especificamente para os casos onde os dados do conjunto de treino podiam ser separados sem erros mas, posteriormente, este objetivo foi alargado de modo a incluir dados dos conjuntos de treino que não estejam separados (Cortes & Vapnik, 1995).

#### *2.2.3.3. J48*

O algoritmo J48 é a implementação na ferramenta WEKA do algoritmo conhecido como c4.5. Este é o algoritmo de uma árvore de decisão e usa a estratégia de dividir o problema para resolver uma determinada questão, um problema que implique uma tomada de decisão. Um problema complexo é, então, dividido em problemas mais simples aos quais é aplicada recursivamente a mesma estratégia (Saravanan & Ramachandran, 2009).

#### *2.2.3.4. Kstar*

O algoritmo *Kstar* é um algoritmo de LL baseado em instâncias que procuram identificar métricas de valores idênticos para encontrar novas instâncias o mais idênticas possíveis, conseguindo assim, efetuar a classificação dessa mesma instância (Garner, 1995).

## 2.3. Sistemas de Apoio à Decisão

Este processo, basicamente, é um ciclo que se inicia pela fase Inteligência, seguem-se as fases Desenho, Escolha, Implementação e por fim a Monitorização, apesar de ser um processo sequencial existe sempre a possibilidade de retroceder algumas fases. A última fase pode originar a repetição de todo o processo. Os Sistemas de Apoio à Decisão (SAD) são definidos por diversos autores de uma forma relativamente uniforme, estes consideram os SADs como um sistema de computador interativo que ajuda os responsáveis a tomar decisões baseadas em atributos, metas e objetivos, a fim de resolver dois tipos de problemas: semiestruturados e não estruturados. O primeiro tipo diz respeito a são problemas que envolvem dados bem estruturados e dados não estruturados, já os segundos apenas contêm dados não-estruturados. São, portanto, sistemas que apoiam os gestores na tomada de decisão, fornecendo e analisando diversas alternativas, para tal pesquisam o histórico de decisões tomadas e qual a influência que estas tiveram no contexto organizacional permitindo assim dar auxílio à resolução de problemas (Nemati, Steiger, Iyer, & Herschel, 2002; Sauter, 2011; Shim et al., 2002; Efraim Turban, 2010; Vercellis, 2009).

De forma simples, um SAD relaciona dados com modelos matemáticos e mostra a informação obtida através deles num interface gráfico que facilita a interação de um responsável por tomar a decisão com o sistema como demonstrado na Figura 19 (Vercellis, 2009).

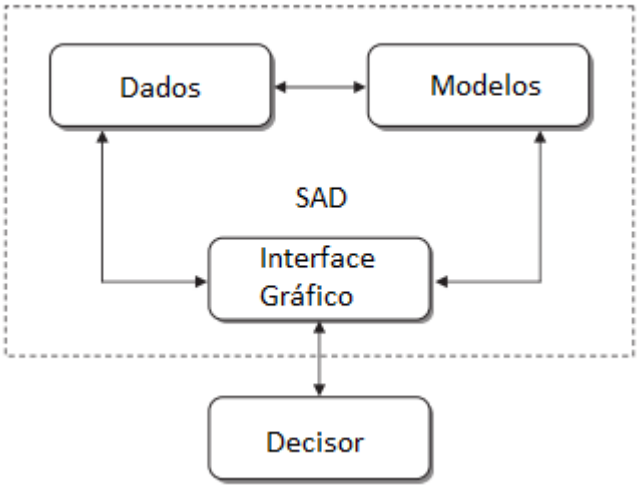


**Figura 19 - Estrutura de um SAD figura adaptada de (Vercellis, 2009)**

O processo de tomada de decisão e as suas respetivas fases estão apresentadas no tópico 1.3.3 deste documento.

## 2.4. Sistemas de Suporte em Apostas de Futebol

O jogo das apostas é uma grande indústria, com muitas empresas a operar em diferentes estruturas, como, as empresas de apostas tradicionais, as empresas *online* e bolsas de apostas. Os casos mais famosos são empresas *online* como a *Betfair*, a *Bwin* e a *Bet365*. Cada empresa de apostas *online* tem o seu próprio mercado de probabilidades e bolsas de apostas. O desporto, mais particularmente o futebol é um dos mercados mais comuns apresentando várias possibilidades de apostas como o número de cantos, o número de golos, o resultado final entre outras. O resultado final além do clássico 1, X ou 2, permite ainda apostar num mercado conhecido como *double chance* que permite duas apostas distintas, vitória da equipa visitada contra o empate e vitória da equipa visitante, e a segunda aposta a ser vitória da equipa visitante e empate ou então empate.

O processo de apostas em jogos de futebol é um processo simples. Os utilizadores podem fazer apostas em tempo real ou antes do jogo começar. Quanto mais o apostador perceber sobre futebol melhor ele pode controlar as variáveis do jogo aumentando a probabilidade de ganhar. As casas de apostas oferecem várias opções para apostar. Primeiro, o apostador tem de escolher um mercado e uma aposta para fazer. Em seguida, o jogador escolhe a quantidade de dinheiro que vai investir na aposta. Depois de apostar, o dinheiro é retirado da conta e em seguida, o jogador precisa esperar até o resultado da

aposta ser conhecido. Se ele perdeu (caso a aposta efetuada seja diferente do resultado real) nada acontece (porque o dinheiro já foi retirado da conta). Caso contrário, se ele vencer, ele ganha o dinheiro apostado multiplicado pelo valor da odd (se ele apostou 10 euros em uma aposta em que a *odd* seria de 2.00 ele ganha 20 euros obtendo um lucro de 10 euros).

### 2.4.1. Trabalho Científico Existente

Nos dias de hoje, a utilização de técnicas de DM na previsão de acontecimentos relacionados com o futebol é uma realidade. De modo a se ter uma visão mais alargada dos trabalhos já existentes foi efetuada uma análise a alguns trabalhos de caracter científico.

Todos os trabalhos analisados têm o objetivo de efetuar a previsão em jogos de futebol, mas estas previsões podem ser inseridos em três tipos de categorias:

- Efetuar previsão de resultados de uma equipa específica, por exemplo os resultados do FC Barcelona;
- Efetuar a previsão de um determinado campeonato, no qual existem jogos semanalmente e as equipas se defrontam todas entre si e no fim a equipa que obtiver mais pontos é a vencedora, o campeonato que recolhe mais interesse nestes estudos é a liga inglesa;
- Fazer a previsão de um determinado torneio com uma fase de grupos inicial e uma posterior fase eliminatória, a Liga dos Campeões é um exemplo que se adequa nesta categoria.

Da primeira categoria previamente identificada, foram selecionados dois trabalhos:

- Owramipur *et a*l (Owramipur, Eskandarian, & Mozneb, 2013) realizam um estudo que pretende efetuar a previsão de resultados do FC Barcelona. O período de estudo é a época 2008/2009 da primeira Liga Espanhola na qual os autores propõem a utilização de redes de *bayes* para efetuar a previsão do resultado dos jogos. Os fatores que entram como variáveis são algumas características do próprio jogo, um histórico de resultados e também de golos, a forma recente de cada equipa e algumas características físicas e também mentais. De forma a validar o trabalho compara os resultados obtidos com os resultados que aconteceram na realidade e a taxa de acerto obtida está na ordem do 92%.
- Joseph *et al* (Joseph, Fenton, & Neil, 2006) realizam um estudo semelhante ao anterior, mas com uma equipa distinta numa liga diferente, o *Tottenham Hotspur FC*. Foram utilizadas quatro técnicas de DM, as redes de *Bayes*, as AD, o NB e o *K-Nearest Neighbors* (KNN). O objetivo do estudo era obter a maior taxa de acerto possível, para tal foram comparadas as métricas relativas

à acuidade que foram geradas para cada modelo criado com estas diferentes técnicas. As variáveis que os modelos tiveram em consideração foram o ranking de cada equipa, características físicas e o fator casa. A taxa de acuidade que obteve o melhor resultado foi de 59,21%.

A segunda categoria foi a categoria da qual resultou uma maior recolha de informação, a maioria dos estudos foca-se na previsão de resultados de um determinado campeonato e também no respetivo vencedor.

- Rotshtein *et al* (Rotshtein, Posner, Rakityanskaya, Lev, & National, 2005) efetuaram um estudo que pretende prever os resultados da liga finlandesa através de modelos *Fuzzy*, As variáveis que entram neste trabalho são os resultados dos últimos cinco jogos de cada equipa e os dois últimos confrontos diretos das mesmas. Os autores definem cinco previsões distintas, se uma equipa vence por mais de 3, se ganha por 1 ou 2, se empata, se perde por 1 ou 2 e se perde por mais de 3. A previsão que obtém uma taxa de acuidade superior foi obtida quando pretendiam prever se determinada equipa perde por mais de 3 golos, a taxa obtida foi de 87,5%.
- Os autores Tsakonas & Dounias (Tsakonas & Dounias, 2002) através do seu estudo tinham como objetivo prever os resultados da liga ucraniana e qual seria o vencedor do campeonato. Utilizaram três técnicas distintas para criar modelos, *Fuzzy sets*, redes neuronais e algoritmos genéticos. As variáveis todas consideração foram características do próprio jogo, o histórico de golos, a posição em que cada equipa se encontra e a forma atual de cada uma delas. Utilizam como medidas de avaliação a taxa de acuidade e o erro quadrático médio. Os melhores valores foram adquiridos através de redes neuronais, obtendo uma taxa de acuidade de 64%.
- Os autores Nunes & Sousa, (Nunes & Sousa, 2006) no seu estudo tiveram como objetivo utilizar técnicas de DM em dados de futebol em alguns campeonatos europeus, e por fim criaram um modelo que prevê o resultado para a liga portuguesa. Os campeonatos dos quais são recolhidos dados são o português, o inglês, espanhol, italiano, francês e alemão. O trabalho efetuado com estes dados focou-se na visualização dos mesmos, os dados recolhidos são principalmente sobre o jogo em si e registos de golos e histórico de resultados. É sugerido um modelo para efetuar previsão na liga portuguesa que segue apenas duas regras, quando a equipa visitante é o "FC Porto", "SL Benfica" e "Sporting CP" sugere a derrota da equipa visitada em todos os outros casos sugere sempre vitória da equipa visitada. Este modelo não sugere qualquer empate. A sua taxa de acuidade é de 59,1%.

- Para a liga inglesa os autores Ulmer & Fernandez (Ulmer & Fernandez, 2013) tiveram como objetivo fazer a previsão dos resultados, estes utilizaram como dados de treino dez épocas, entre 2002/2003 e 2011/2012 e como dados de teste as épocas 2012/20013 e 2013/2014. As técnicas de DM utilizadas foram o NB, o SVM Gaussiano, o SVM Linear e o *Random Forest*. Como medida de avaliação utilizam a taxa de erro. Os dados que tiveram em conta são as características do jogo, o desempenho recente das equipas e a posição em que se encontram na tabela classificativa antes do respetivo jogo. O modelo em que a taxa de erro foi inferior continha um classificador linear e obteve uma taxa de erro 48%.

Da terceira categoria identificada foram recolhidos dois trabalhos um acerca da principal competição europeia a nível de clubes, a Liga dos Campeões, e o Mundial de Futebol de 2006:

- Os autores Hucaljuk & Rakipovic (Hucaljuk & Rakipovic, 2011) fizeram o estudo com o objetivo de prever os resultados dos jogos na liga dos campeões. Quando estavam a efetuar a recolha dos dados depararam-se com um problema, equipas diferentes participam nesta competição todos os anos, o que levava a que não possuíssem dados de qualidade que lhes permitisse fazer a previsão. Decidiram então recolher os dados do próprio ano em que o estudo se desenvolveu. Escolheram três estratégias, na primeira a fase de treino continha as três primeiras rondas da fase de grupos e como teste era utilizadas as outras três rondas, a segunda estratégia utilizava 4 rondas como treino e duas como teste e por fim, a terceira estratégia utilizava as 5 primeiras rondas como treino e apenas a última como teste. Foram utilizados os seguintes algoritmos, o NB, as redes de *Bayes*, LogitBoost, KNN, *Random Forest* e redes neuronais artificiais. O método de amostragem utilizado foi o *10-Folds Cross-Validation* (10FCV). As variáveis que foram recolhidas foram, a forma recente das equipas nos últimos seis jogos, o histórico de confrontos entre as equipas, a posição em que se encontram, o número de jogadores lesionados da primeira equipa e o número de golos marcados e sofridos. O modelo que obteve melhores resultados foi gerado através de uma rede neuronal artificial que teve uma taxa de acuidade de 68%.
- Suzuki *et al* (Suzuki, Salasar, Leite, & Louzada-Neto, 2010) efetuaram um trabalho que tinha como objetivo prever os resultados dos jogos do mundial de 2006 utilizando uma metodologia de *Bayes*. O modelo teve em conta a opinião de especialistas na área, o ranking das seleções e o fator casa. O modelo proposto era um modelo estatístico que utiliza o teorema de De Finetti, através do mesmo foi possível classificar corretamente a previsão de 57,81% dos resultados.

### 2.4.2. Sistemas Semelhantes

As seguintes plataformas são *websites* que contêm as probabilidades de ocorrer um de três resultados (vitória da equipa da casa, empata ou vitória da equipa visitante), bem como outras estatísticas que consideram importantes, sugerindo depois qual a aposta que consideram ideal. Todas as seguintes plataformas indicadas são idênticas à apresentada na Figura 20, na qual se pode verificar qual o jogo em que se está a efetuar a previsão, a probabilidade de ocorrer a vitória da equipa visitada, o empate ou a vitória da equipa visitante, em seguida a sugestão de aposta a efetuar e a respetiva *odd* a que essa se encontra. Estes são alguns dos sistemas mais comuns no suporte em apostas de jogos de futebol. Estes sistemas têm essencialmente como base da sua construção a utilização de cálculos matemáticos. O protótipo desenvolvido neste trabalho distingue-se das soluções existentes devido à utilização de técnicas de DM.

- http://soccervista.com/
- http://vitibet.com/
- http://pt.zulubet.com/
- http://www.footwin.net/
- http://www.predictz.com/
- http://www.forebet.com/
- https://www.statarea.com/predictions
- http://www.windrawwin.com/predictions/

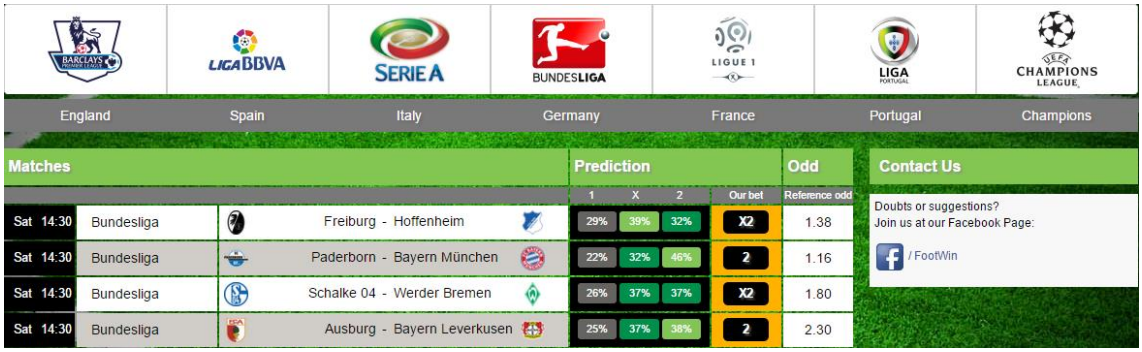


**Figura 20 - Interface do *website* footwin.net**

Além dos *websites* mencionados existe um outro *website* que contém um sistema que se baseia em probabilidades estatísticas, no qual, é necessário introduzir as estatísticas de um determinado jogo que se pretenda obter uma previsão. O sistema é bastante simples, é necessário preencher os campos com o número de golos marcados durante toda a época das equipas intervenientes no jogo, o número total

de vitórias, empates e derrotas de ambas as equipas em toda a época e o número de vitórias, empates e derrotas de cada equipa em todos os jogos realizados entre as duas equipas. Depois destes valores inseridos no sistema, este calcula a probabilidade de ocorrer vitória da equipa visitada, empate ou vitória da equipa visitante, sugerindo também um provável resultado final:

- http://spotwin.net/football-betting-system-7.html

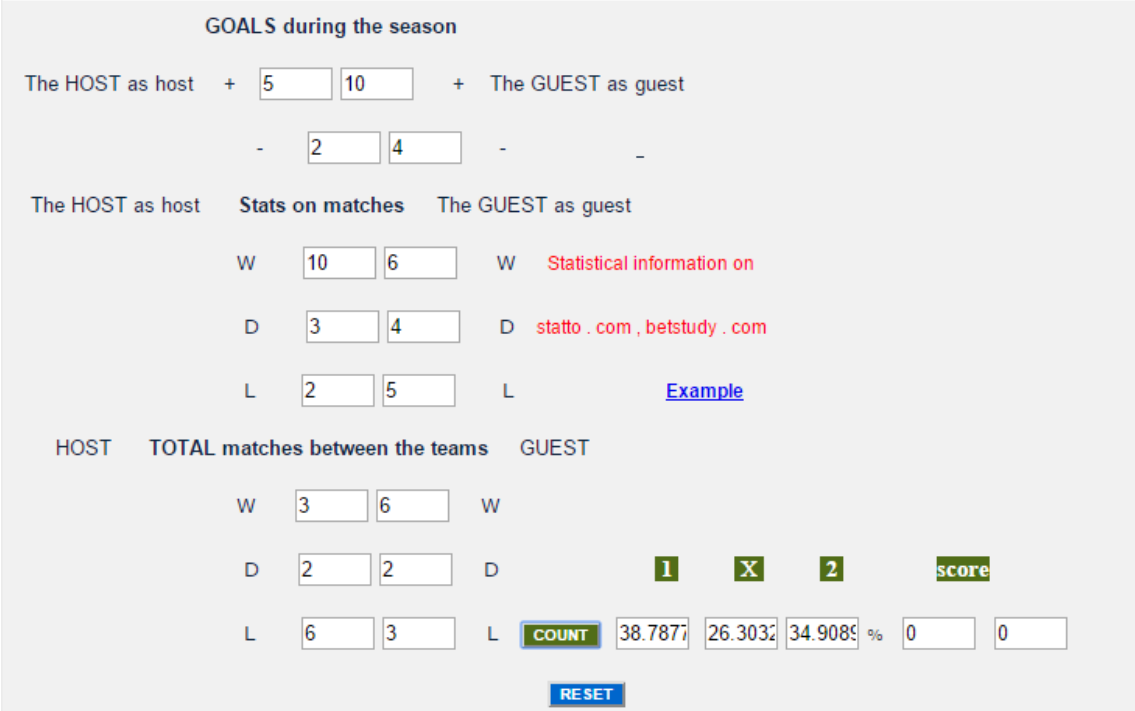


**Figura 21 - Interface do sistema *spotwin***

Existem também aplicações para os *smartphones*, estas encontram-se disponíveis na *App Store* ou na *Play Store* que têm o mesmo objetivo dos *websites* descritos previamente. Na Figura 22 pode-se ver o interface da aplicação *KickOff*

- *KickOff – Smart Betting Made Simple*
- *Smart BET Prediction*
- *FootWin – Sports Prediction*


**Figura 22 - Interface da aplicação *Kickoff***

E por fim foi encontrado um *website* que tem como base de cálculo as redes neuronais que apresenta também a probabilidade de ocorrer a vitória da equipa visitada, do empate e da vitória da equipa visitante e ainda é sugerida a aposta a ser concretizada:

- http://www.prosoccer.gr/

# 3 TRABALHO RELACIONADO

Este projeto de dissertação teve como base um trabalho desenvolvido na Unidade Curricular (UC) de Sistemas de Apoio à Decisão (SAD) do quarto ano de Mestrado Integrado em Engenharia e Gestão de Sistemas de Informação (MIEGSI). Nesta UC foi desenvolvido um SAD que suportava a decisão dos utilizadores, num determinado jogo de futebol, sobre qual a aposta que deveria realizar, se seria a vitória da equipa visitada, o empate ou vitória da equipa visitante. Esta sugestão era efetuada através de uma análise à estatística que está presente em cada jogo, como por exemplo o número de golos, número de remates, número de faltas.

O trabalho iniciou-se então com a identificação do problema, verificou-se que os utilizadores das casas de apostas tinham mais vezes prejuízo que lucro, o que tem levado a um aumento do número de casas de apostas, visto este ser um negócio rentável para as mesmas.

Foi efetuada uma recolha de dados estatísticos. Estes dados eram relativos a catorze épocas da primeira liga inglesa que estavam armazenados em ficheiros **csv* com os dados relativos a 380 jogos em cada época, o que corresponde a 5320 jogos no total.

O conjunto de dados originais está apresentado na Tabela 3.

**Tabela 3 - Atributos originais trabalho anterior**

| Atributos Originais | Descrição |
|---|---|
| Date | Data do Jogo (dd/mm/yy) |
| HomeTeam | Equipa Visitada |
| AwayTeam | Equipa Visitante |
| FTHG | Golos Equipa Visitada Tempo Final |
| FTAG | Golos Equipa Visitante Tempo Final |
| FTR | Resultado Tempo Final (H=Vitória casa, D=Empate, A=Vitória Fora) |
| HTHG | Golos Equipa Visitada ao Intervalo |
| HTAG | Golos Equipa Visitante ao Intervalo |
| HTR | Resultado ao Intervalo (H=Vitória casa, D=Empate, A=Vitória Fora) |
| Attendance | Número de Espectadores |

| Atributos Originais | Descrição |
|---|---|
| Referee | Árbitro |
| HS | Remates Equipa Visitada |
| AS | Remates Equipa Visitante |
| HST | Remates à Baliza da Equipa Visitada |
| AST | Remates à Baliza da Equipa Visitante |
| HC | Cantos Equipa Visitada |
| AC | Cantos Equipa Visitante |
| HF | Faltas Cometidas Equipa Visitada |
| AF | Faltas Cometidas Equipa Visitante |
| HO | Foras-de-Jogo Equipa Visitada |
| AO | Foras-de-Jogo Equipa Visitante |
| HY | Cartões Amarelos Equipa Visitada |
| AY | Cartões Amarelos Equipa Visitante |
| HR | Cartões Vermelhos Equipa Visitada |
| AR | Cartões Vermelhos Equipa Visitante |

Depois de recolhidos e analisados os dados foi necessário fazer um tratamento dos mesmos. Para isso foi efetuado um processo *Extract Transform Load* (ETL),representado na Figura 23.

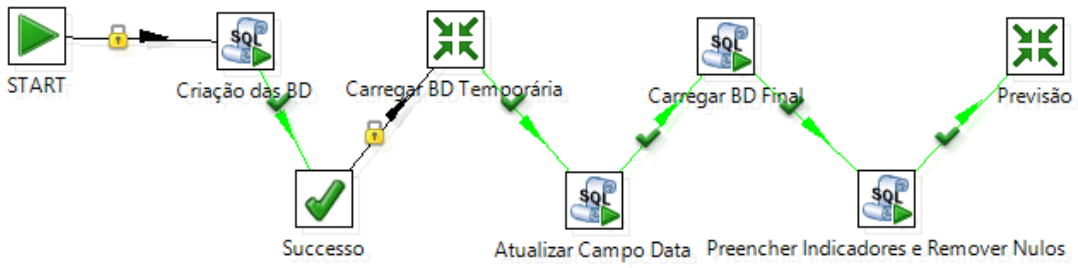


**Figura 23 - Processo ETL Anterior**

Este é um processo de ETL bastante simples, começa pela criação de duas bases de dados, uma temporária e outra que será preenchida mais tarde. Os dados que estão contidos no conjunto original

são carregados para a tabela temporária. O campo data é separado em três, dia, mês e ano. É então carregada a outra base de dados com os dados contidos na temporária. De seguida é efetuado um tratamento aos dados onde foram analisadas as anomalias existentes nos mesmos e tratados os nulos. Por fim, a base de dados final é exportada para um ficheiro de texto de forma a ser utilizada na ferramenta Exsys Corvid.

Como para o resultado final de um jogo de futebol é possível obter três distintas previsões somente se pode efetuar uma análise da precisão / acuidade dos modelos como apresentado na Tabela 4. Os modelos foram induzidos através de um único cenário que continha todas as variáveis presentes no conjunto de dados. Foram utilizados dois métodos de amostragem, o *Holdout Simple* (HS) e o 10-*Folds Cross-Validation* (10FCV). As técnicas utilizadas foram o *Naive Bayes*, as *Support Vector Machines* e as árvores de decisão.

**Tabela 4 - Precisão dos Modelos de DM trabalho anterior**

| Modelos | Treino | Técnica | Precisão |
|---|---|---|---|
| **Modelo 1** | HS | Naïve Bayes | 0,487 |
| **Modelo 2** | HS | J48 | 0,472 |
| **Modelo 3** | HS | LibSVM | 0,508 |
| **Modelo 4** | 10FCV | NaiveBayes | 0,492 |
| **Modelo 5** | 10FCV | J48 | 0,476 |
| **Modelo 6** | 10FCV | LibSVM | 0,492 |

A precisão obtida foi idêntica em todos os modelos, mas o modelo 3 demonstrou ter uma ligeira superioridade em relação aos outros e por isso foi o modelo escolhido. Esta precisão apesar de não atingir os valores desejados atingiu um valor superior aos 33% se a decisão fosse tomada aleatoriamente. De seguida foi criado um primeiro protótipo que funciona como um sistema de scores, este sistema além da variável precisão tem em conta outras variáveis. Para descobrir essas novas variáveis o sistema faz as seguintes perguntas ao utilizador:

- "Qual a classificação da equipa visitada?";
- "Qual a classificação da equipa visitante?";
- "Quantos titulares, na opinião do utilizador, não estão disponíveis na equipa visitada?";
- "Quantos titulares, na opinião do utilizador, não estão disponíveis na equipa visitante?".

Cada uma destas perguntas gera uma nova variável e através de um sistema de scores foi efetuado um cálculo com o objetivo de obter um score para cada equipa onde este pode obter valores até 1. Com ambos os scores definidos o sistema vai sugerir um dos 3 possíveis resultados:

- Vitória da equipa visitada, se a diferença dos scores entre as duas equipas foi superior a 0,17 e o score da equipa visitada for superior ao da equipa visitante;
- Vitória da equipa visitante, se a diferença dos scores entre as duas equipas foi superior a 0,17 e o score da equipa visitante for superior ao da equipa visitada;
- Empate, se a diferença entre os dois scores for inferior a 0,17.

Depois de tudo definido e implementado, começaram os primeiros testes e apesar dos valores obtidos nos modelos não terem valores muito elevados de precisão, o lucro que o sistema geraria aos utilizadores é considerável, como se pode verificar pelos testes realizados ao sistema que simulam algumas jornadas da época 2013/2014 da primeira liga inglesa, tal como se pode verificar na Tabela 5.

**Tabela 5 - Resultados dos Testes ao Protótipo**

| Jornada | % de apostas acertadas | Retorno (Apostas de 100€) |
|---|---|---|
| **Jornada 5** | 80% | 689 € |
| **Jornada 10** | 30% | -418 € |
| **Jornada 15** | 40% | 11 € |
| **Jornada 20** | 70% | 713 € |
| **Jornada 25** | 60% | 480 € |
| **Jornada 30** | 40% | -245 € |
| **Jornada 35** | 60% | 179 € |
| **Total** | Média = 54,29% | 1409 € |

Como verificado acima, a efetuar apostas de 100 € em cada jogo das sete jornadas em que os testes foram realizados o lucro obtido seria de 1409 €, tendo em conta que seriam apostados ao todo 7000 € o lucro obtido andaria a rondar 20,13%.

# 4 Sistema Inteligente de Apoio à decisão em Apostas de Jogos de Futebol

A criação do protótipo do sistema envolve, como está previamente descrito no tópico 1.3 deste documento, a utilização de três diferentes metodologias no desenvolvimento do projeto. Uma metodologia de investigação denominada por *Design Science and Research (DSR)* e duas metodologias de desenvolvimento, o *Cross Industry Standard Process for Data Mining (CRISP-DM)* e as fases do processo de tomada de decisão. Estas metodologias complementam-se e da união das três, tal como está demonstrado na Tabela 1, resultou em 5 fases distintas. Essas são as fases seguidas para a elaboração da parte prática do projeto.

## 4.1. Fase 1

A primeira fase resulta da junção da fase "Compreensão do Negócio" do CRISP-DM, da fase "Inteligência" do processo de tomada de decisão e das fases "Identificação e Motivação do Problema" e "Objetivos da Solução".

Nesta fase foi identificado um problema ou oportunidade: um grupo de amigos tinha como *hobby* apostar e depois de diversas conversas sobre os resultados que iam obtendo chegaram à conclusão que poucos eram os que conseguiam a médio/longo prazo obter lucro com as apostas que iam efetuando. Decidiram ajudar-se uns aos outros, faziam uma análise específica de cada jogo em que pretendiam apostar e por fim discutiam a informação que tinham recolhido e repararam que na maioria das situações os dados que recolhiam se complementavam, isso permitiu-lhes começar a retirar maior lucro das suas apostas. Com isto verificou-se que a análise prévia a cada jogo estava a tornar este *hobby* rentável.

Surgiu então a ideia de tornar este processo automático, algo que diminuísse o trabalho que eles tinham em analisar aprofundadamente as equipas que iam participar num determinado jogo, simplificar de tal modo o processo que mesmo quem não perceba nada de futebol consiga obter uma informação fidedigna que permita obter lucros em apostas nos jogos de futebol.

A criação de um sistema inteligente de apoio à decisão foi, portanto, a solução encontrada para este problema. O sistema seria inteligente pois irá utilizar técnicas de Data Mining (DM) para efetuar as previsões, complementando o sistema com a inteligência humana:

Foi também efetuada uma recolha de informação ao ambiente de tudo o que estaria relacionado com esta área, identificando o que ocorre antes do início deste projeto e o que se espera que ocorra depois do final do mesmo, para existir a possibilidade de efetuar comparação do que acontece antes e o que é esperado que aconteça.

Este projeto tem como objetivo de negócio suportar os apostadores e ajudá-los a obter lucro nas apostas a médio/longo prazo diminuindo o risco que correm quando efetuam as apostas.

O objetivo de Data Mining (DM) é elaborar modelos de DM viáveis e capazes de efetuar previsões que levem ao cumprimento do objetivo de negócio, capazes de obter métricas que cumpram os parâmetros de qualidade definidos.

## 4.2. Fase 2

A segunda fase do projeto resulta da união de seis fases, "Compreensão dos dados", "Preparação dos Dados" e "Modelação" do CRISP-DM, da fase "Desenho" do processo de tomada de decisão e as fases "Desenho e Desenvolvimento" e "Demonstração" do DSR.

Nesta fase foram identificadas as possíveis restrições existentes no projeto e foi profundamente analisados os dados que podiam ser recolhidos e úteis ao projeto. Foi também nesta fase que ficou definitivamente decidido que o protótipo a desenvolver seria um sistema inteligente de apoio à decisão, este é um sistema que tem por base técnicas de DM mas também necessita da perceção humana para ser mais fiável, pois, a maioria dos dados estatísticos recolhidos são conhecidos previamente ao início do jogo (ex. equipas, golos marcados, golos sofridos, outras), no entanto existem outros dados que só são identificados à hora do jogo, como por exemplo as condições climatéricas.

Após uma análise à informação existente relativamente a dados estatísticos relacionados com os jogos de futebol foi efetuada uma recolha dos mesmos a partir de um conjunto de dados encontrado no *website* "football-data-co.uk", a informação encontrada neste site era bastante completa e continha todas as variáveis necessárias relativas ao jogo de futebol. Foram recolhidos dados estatísticos de cada jogo de 14 épocas da Primeira Liga Inglesa, também conhecida como *Barclays Premier League*, desde a época 2000/2001 até à época 2013/2014. Foram então recolhidos dados de 5320 diferentes jogos.

### 4.2.1. Recolha dos Dados

A Tabela 6 contém apenas as variáveis originais existentes no conjunto de dados recolhido que continham registos contínuos entre o ano 2000 e 2014 de jogos de futebol que envolveram 41 distintas equipas, as variáveis relacionadas com o intervalo do jogo não foram consideradas.

**Tabela 6 - Dados Recolhidos**

| Variáveis Originais | Descrição | Mínimo | Máximo | Média |
|---|---|---|---|---|
| **Date** | Data do Jogo (dd/mm/AA) | 2000/08/19 | 2014/05/11 | - |
| **HomeTeam** | Equipa Visitada | - | - | - |
| **AwayTeam** | Equipa Visitante | - | - | - |
| **FTHG** | Golos Marcados Equipa Visitada | 0 | 9 | 1,53 |
| **FTAG** | Golos Marcados Equipa Visitante | 0 | 0 | 1,118 |
| **FTR** | Resultado Final (H=Vitória Equipa Visitada, D=Empate, A=Vitória Equipa Visitante) | - | - | - |
| **HTHG** | Golos Marcados Equipa Visitada ao Intervalo | - | - | - |
| **HTAG** | Golos Marcados Equipa Visitante ao Intervalo | - | - | - |
| **HTR** | Resultado ao Intervalo (H=Vitória Equipa Visitada, D=Empate, A=Vitória Equipa Visitante) | - | - | - |
| **Referee** | Nomo do Árbtiro do jogo | - | - | - |
| **HS** | Remates Equipa Visitada | 0 | 39 | 13,342 |
| **AS** | Remates Equipa Visitante | 0 | 30 | 10,282 |
| **HST** | Remates à Baliza Equipa Visitada | 0 | 24 | 6,908 |
| **AST** | Remates à Baliza Equipa Visitante | 0 | 20 | 5,236 |
| **HC** | Cantos Equipa Visitada | 0 | 20 | 6,264 |
| **AC** | Cantos Equipa Visitante | 0 | 19 | 4,813 |
| **HF** | Faltas Cometidas Equipa Vitsitada | 0 | 33 | 11,771 |

| Variáveis Originais | Descrição | Mínimo | Máximo | Média |
|---|---|---|---|---|
| **AF** | Faltas Cometidas Equipa Vitsitante | 1 | 29 | 12,342 |
| **HO** | Foras-de-jogo Equipa Visitada | - | - | - |
| **AO** | Foras-de-jogo Equipa Visitante | - | - | - |
| **HY** | Amarelos Equipa Visitada | 0 | 7 | 1,338 |
| **AY** | Amarelos Equipa Visitante | 0 | 8 | 1,778 |
| **HR** | Vermelhos Equipa Visitada | 0 | 3 | 0,068 |
| **AR** | Vermelhos Equipa Visitante | 0 | 2 | 0,098 |
| **ODDS** | Odds relativas a diversas casas de apostas | - | - | - |

Além destas variáveis estatísticas, uma variável que foi considerada fundamental de se introduzir no modelo está relacionada com as condições climatéricas. A precipitação, é um fator que a "olho nu" qualquer pessoa consegue identificar que influência a forma de jogar das equipas. Os dados relativos à meteorologia do Reino Unido não estão disponíveis a qualquer utilizador, pelo que foi necessário contactar o *British Atmospheric Data Centre* (BADC), mais especificamente o departamento *Centre for Environmental Data Archival* (CEDA), que depois de um longo questionário para verificar quem eram as pessoas que pretendiam utilizar os dados e qual seria o objetivo da utilização dos mesmos cederam um acesso a alguns conjuntos de dados por tempo limitado. Foi necessário garantir a confidencialidade dos dados. O conjunto de dados recolhido é aquele que permite obter informação sobre a precipitação.

O CEDA possui uma grande quantidade de conjuntos de dados, para conseguir recolher a precipitação foi necessário focar no conjunto de tabelas denominado por "*UK Daily Rainfall Data*", sendo este constituído pelas variáveis apresentadas na Tabela 7.

**Tabela 7 - Dados Recolhidos BADC**

| Variáveis | Descrição |
|---|---|
| **Id** | Número Identificador do Pluviómetro |

| Variáveis | Descrição |
|---|---|
| **id_type** | Tipo do Identificador |
| **ob_date** | Data da Observação |
| **version_num** | Número da Versão da Observação |
| **met_domain_name** | Tipo de Mensagem |
| **ob_end_ctime** | Hora no Final da Observação |
| **ob_day_cnt** | Contagem de Dias de Observação |
| **src_id** | Número Identificador da Estação |
| **rec_st_ind** | Indicador do Estado para Registo |
| **prcp_amt** | Quantidade de Precipitação em Milímetros |

Os dados recolhidos estavam divididos em diferentes ficheiros **.txt*, tendo sido necessário efetuar a recolha de 15 ficheiros correspondentes a cada ano desde 2000 até 2014.

O processo seguinte foi efetuar a seleção das estações das quais se iria efetuar a recolha da variável "prcp_amt". O primeiro passo deste processo foi identificar em que condado cada clube, dos que existem registos no conjunto de dados referentes às estatísticas de cada jogo de futebol, tem o seu estádio construído como se verifica na Tabela 8.

**Tabela 8 - Clube e Condado**

| Clube | Condado | Clube | Condado | Clube | Condado |
|---|---|---|---|---|---|
| **Reading** | Berkshire | Middlesbrough | Cleveland | Aston Villa | West Midlands |
| **Chelsea** | Greater London | Derby | Derbyshire | Hull | Humberside |
| **Coventry** | Warwickshire | Stoke | Staffordshire | Arsenal | Greater London |
| **Sheffield United** | South Yorkshire | Birmingham | West Midlands | Charlton | Greater London |
| **Fulham** | Greater London | Swansea | West Glamorgan | Middlesbrough | Cleveland |
| **Leicester** | Leicestershire | Blackpool | Lancashire | Ipswich | Suffolk |
| **Liverpool** | Merseyside | Wolves | Staffordshire | QPR | Greater London |
| **Bradford** | West Yorkshire | Watford | Hertforshire | Bolton | Greater Manchester |

| Clube | Condado | Clube | Condado | Clube | Condado |
|---|---|---|---|---|---|
| **Man City** | Greater Manchester | Portsmouth | Hamshire | Newcastle | Tyne & Wear |
| **Man United** | Greater Manchester | Norwich | Norfolk | Crystal Palace | Greater London |
| **Sunderland** | Durham | West Ham | Greater London | Leeds | West Yorkshire |
| **Southampton** | Hampshire | West Brom | West Midlands | Cardiff | South Glamorgan |
| **Burnley** | Lancashire | Tottenham | Greater London | Everton | Merseyside |
| **Wigan** | Lancashire | Blackburn | Lancashire | | |

De seguida, utilizando o exemplo do *Reading*, foi necessário descobrir qual a estação mais próxima do estádio existente no condado de *Berkshire*. Para isso foram selecionadas todas as estações existentes nesse condado que continham informação entre os anos 2000 e 2014, e foram comparadas as suas localizações com a do estádio e a que tivesse uma distância inferior seria a escolhida, pois, o valor registado pela estação mais próxima é o mais fiável. Neste caso, a estação mais próxima é a "*Reading University: Whiteknites*" que tem como "src_id" o número 830. Depois de descoberta a estação, foi necessário verificar se esta tinha efetuado registos diários da precipitação ao longo dos anos e foi aí que surgiu um problema, poucas eram as estações que efetuaram o registo de forma continua durante os 14 anos, verificou-se então que esta estação não efetuou o registo da precipitação diariamente nos anos 2006, no qual tem registos em apenas 358 dias, em 2012 só regista 233, no ano de 2013 não foi efetuado qualquer registo e por fim, no ano de 2014 regista 94 dias. Foi então necessário recorrer à segunda estação mais próxima do estádio, a "*Englefield Estate*", para complementar os dados em falta. Nesta estação apenas se verificou os anos em falta na estação anterior, no ano de 2006 e 2013 tem os registos completos mas nos anos 2012 e 2014 regista apenas 335 dias, o que leva a que seja necessário repetir o processo até se obter registos de todos os anos sem falhar qualquer dia. No caso do Reading foi necessário fazer um levantamento das 4 estações mais próximas, como indicado na Tabela 9. Este foi um processo bastante exaustivo, e apesar de com o *Reading* ser "apenas" necessário explorar até à quarta estação mais próxima existem situações em que foi necessário recorrer à sexta estação mais próxima como foi o caso do *Queens Park Rangers*. De salientar que este procedimento foi executado para todas as 41 equipas consideradas (Tabela 8).

**Tabela 9 - Preenchimento Campo "prcp_amt"**

| Clube | Condado | Estação mais próxima | 2ª Estação mais próxima | 3ª Estação mais próxima | 4ª Estação mais próxima |
|---|---|---|---|---|---|
| **Reading** | Berkshire | Reading University: Whiteknites – src_id=830 | Englefield Estate – src_id=5973 | Bracknell S WKS – src_id=6165 | Bucklebury – src_id=5963 |
| **Anos** | 2000 | X | | | |
| | 2001 | X | | | |
| | 2002 | X | | | |
| | 2003 | X | | | |
| | 2004 | X | | | |
| | 2005 | X | | | |
| | 2006 | | X | | |
| | 2007 | X | | | |
| | 2008 | X | | | |
| | 2009 | X | | | |
| | 2010 | X | | | |
| | 2011 | X | | | |
| | 2012 | | | X | |
| | 2013 | | X | | |
| | 2014 | | | | X |

### 4.2.2. Processo Extract Transform and LoadL

Depois de recolhida a informação da qual existem registos contínuos foi criado o processo *Extract, Transform and Load* (ETL) apresentado na Figura 24.

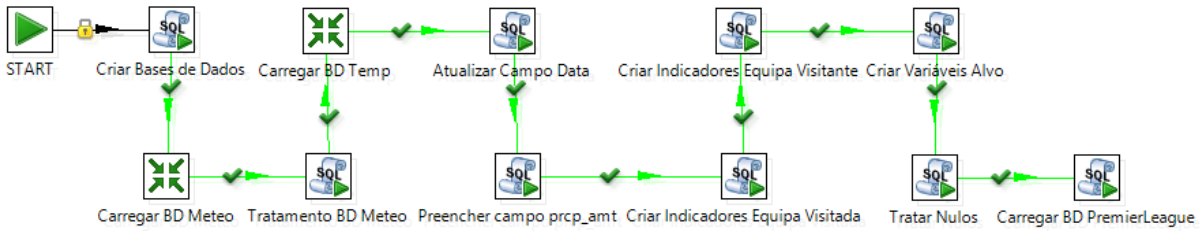


**Figura 24 - Processo ETL**

#### *4.2.2.1. Criar Tabelas*

Este processo inicia-se com a criação de três tabelas, a tabela "Meteo" que contém os dados relativos à precipitação, a tabela "Temp" de temporária que vai conter os dados relacionados com as estatísticas e que vai servir de estágio para uma tabela final denominada de "PremierLeague" que terá os dados das duas tabelas precedentes.

#### *4.2.2.2. Carregar Meteo*

Os dados recolhidos relativos à precipitação, encontravam-se divididos em 14 ficheiros *.*csv* foi então necessário carrega-los todos para uma única tabela como se verifica na Figura 25.

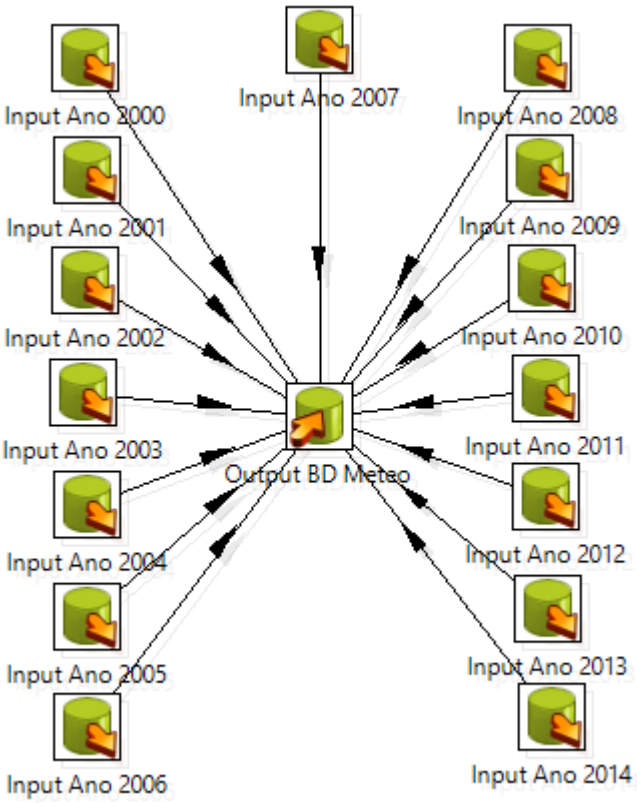


**Figura 25 - Carregar Tabela "Meteo"**

#### *4.2.2.3. Tratamento Meteo*

De seguida foram tratados os dados que se encontravam nulos e também, os registos que se encontravam duplicados, caso o valor fosse diferente foi efetuada uma média dos valores registados.

#### *4.2.2.4. Carregar Tabela "Temp"*

Tal como a tabela "Meteo", na tabela "Temp" também foi necessário efetuar o carregamento dos ficheiros para uma única tabela como se pode comprovar na Figura 26.

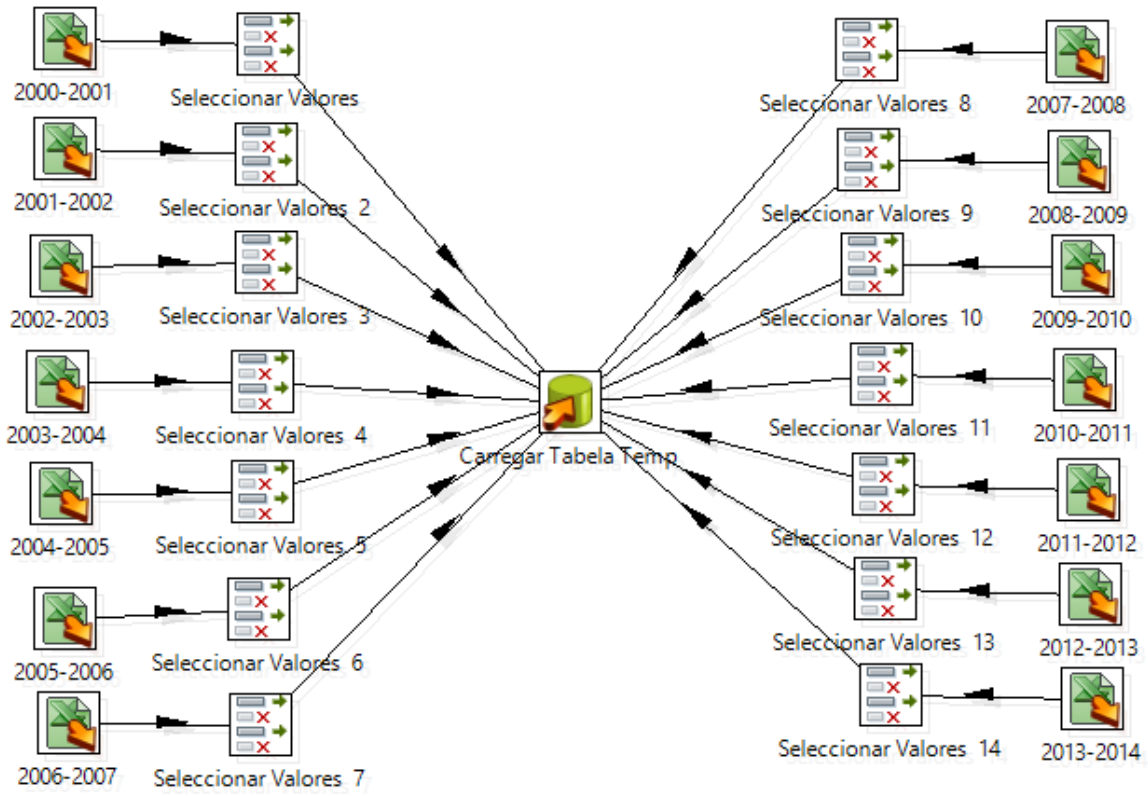


**Figura 26 - Carregar Tabela "Temp"**

#### *4.2.2.5. Atualizar Data*

Para facilitar o manuseamento dos dados foi efetuada uma separação do campo "Date" que é do tipo Date (YYYY/MM/DD) em três campos distintos, Dia, Mês e Ano.

#### *4.2.2.6. Preencher Campo "prcp_amt"*

Este campo é preenchido através duma *query sql,* esta *query* faz o processo explicado anteriormente para o exemplo do Reading, comparando o id de cada estação com cada ano que esta deve colocar o registo para uma respetiva equipa.

#### *4.2.2.7. Criar Indicadores Equipa*

Como a informação que o conjunto de dados recolhidos contém é histórica, são dados que não são conhecidos antes do jogo se iniciar, por exemplo, o número de remates de cada jogo obviamente só é conhecido no final de cada encontro, foram então criados indicadores que posteriormente serão aplicados ao modelo. Esses indicadores são conhecidos antes do encontro se realizar, como por exemplo a média de remates de cada equipa, foram criados indicadores tantos para a equipa visitada como para a equipa visitante.

#### *4.2.2.8. Atualizar Variáveis Alvo*

Nesta fase do ETL foram definidos os campos target que o modelo de DM vai ter, neste caso, são criadas as variáveis alvo através das que já existem. São variáveis relacionadas com o resultado final do jogo, com o número de golos e com o número de cantos existentes em cada jogo. Para isso foi necessário categorizar as variáveis em estudo e agrupá-las de modo a que fosse possível prever os seguintes resultados:

- Vitória equipa visitada, empate ou vitória equipa visitante (R3S);
- A favor ou contra a equipa visitada (R2SC);
- A favor ou contra a equipa visitante (R2SF);
- Mais ou menos de 7,5 cantos (C7,5);
- Mais ou menos de 8,5 cantos (C8,5);
- Mais ou menos de 9,5 cantos (C9,5);
- Mais ou menos de 10,5 cantos (C10,5);
- Mais ou menos de 1,5 golos (G1,5);
- Mais ou menos de 2,5 golos (G2,5);
- Mais ou menos de 3,5 cantos (G3,5).

#### *4.2.2.9. Tratar Nulos*

Tal como o nome indica, foi efetuado um tratamento dos dados que estão na tabela temporária de modo a eliminar todos os registos que continham campos com o valor nulo.

#### *4.2.2.10. Carregar Tabela "PremierLeague"*

Nesta última fase do ETL foi realizado o carregamento da tabela final "PremierLeague", uma tabela completamente tratada em que apenas se encontram os dados que serão necessários para realizar a induzir os modelos de DM, as variáveis presentes nesta tabela final, incluindo os novos indicadores, estão apresentados na Tabela 10.

**Tabela 10 - Variáveis Tabela "PremierLeague"**

| Variável | Tipo | Mínimo | Máximo | Média |
|---|---|---|---|---|
| **Época** | String(50) | 2000/2001 | 2013/2014 | - |
| **Dia** | Int(10) | 1 | 31 | - |
| **Mês** | Int(10) | 1 | 12 | - |
| **Ano** | Int(10) | 2000 | 2014 | - |
| **Equipa Visitada** | String(50) | - | - | - |
| **Equipa Visitante** | String(50) | - | - | - |
| **Árbitro** | String(50) | - | - | - |
| **Média Golos Equipa Visitada** | Decimal(4,2) | 0,53 | 3,58 | 1,53 |
| **Média Golos Concedidos Equipa Visitada** | Decimal(4,2) | 0,32 | 2,26 | 1,118 |
| **Média Remates Equipa Visitada** | Decimal(4,2) | 7,79 | 20,74 | 13,342 |
| **Média Remates Concedidos Equipa Visitada** | Decimal(4,2) | 5,47 | 16,74 | 10,282 |
| **Média Remates à Baliza Equipa Visitada** | Decimal(4,2) | 3,42 | 12,11 | 6,908 |
| **Média Remates Concedidos à Baliza Equipa Visitada** | Decimal(4,2) | 2,68 | 10,11 | 5,236 |
| **Média Cantos Equipa Visitada** | Decimal(4,2) | 3,68 | 9,32 | 6,264 |
| **Média Cantos Concedidos Equipa Visitada** | Decimal(4,2) | 2,21 | 7,79 | 4,813 |
| **Média de Faltas Equipa Visitada** | Decimal(4,2) | 8,26 | 16,32 | 11,771 |
| **Média de Faltas Contra Equipa Visitada** | Decimal(4,2) | 7,74 | 17,74 | 12,342 |

| Variável | Tipo | Mínimo | Máximo | Média |
|---|---|---|---|---|
| **Média Amarelos Equipa Visitada** | Decimal(4,2) | 0,53 | 2,42 | 1,338 |
| **Média Vermelhos Equipa Visitada** | Decimal(4,2) | 0 | 0,32 | 0,069 |
| **Vitórias Últimos 5 Jogos Equipa Visitada** | Int(10) | 0 | 5 | 2,258 |
| **Vitórias Últimos 5 Confrontos Diretos Equipa Visitada** | Int(10) | 0 | 5 | 1,729 |
| **Média Golos Equipa Visitante** | Decimal(4,2) | 0,42 | 2,53 | 1,118 |
| **Média Golos Concedidos Equipa Visitante** | Decimal(4,2) | 0,47 | 2,89 | 1,53 |
| **Média Remates Equipa Visitante** | Decimal(4,2) | 5,21 | 16,79 | 10,282 |
| **Média Remates Concedidos Equipa Visitante** | Decimal(4,2) | 7,42 | 20,37 | 13,342 |
| **Média Remates à Baliza Equipa Visitante** | Decimal(4,2) | 2,47 | 9,68 | 5,236 |
| **Média Remates Concedidos à Baliza Equipa Visitante** | Decimal(4,2) | 3,16 | 11,16 | 6,908 |
| **Média Cantos Equipa Visitante** | Decimal(4,2) | 2,95 | 7,53 | 4,813 |
| **Média Cantos Concedidos Equipa Visitante** | Decimal(4,2) | 3,26 | 9,11 | 6,264 |
| **Média de Faltas Equipa Visitante** | Decimal(4,2) | 8 | 17,53 | 12,342 |
| **Média de Faltas Contra Equipa Visitante** | Decimal(4,2) | 8,26 | 16,95 | 11,771 |
| **Média Amarelos Equipa Visitante** | Decimal(4,2) | 0,89 | 2,79 | 1,778 |

| Variável | Tipo | Mínimo | Máximo | Média |
|---|---|---|---|---|
| **Média Vermelhos Equipa Visitante** | Decimal(4,2) | 0 | 0,37 | 0,099 |
| **Vitórias Últimos 5 Jogos Equipa Visitante** | Int(10) | 0 | 5 | 2,138 |
| **Vitórias Últimos 5 Confrontos Diretos Equipa Visitante** | Int(10) | 0 | 5 | 1,124 |
| **Precipitação** | Int(10) | 0 | 4 | |

O número de ocorrências de cada classe de cada variável-alvo está apresentado nas figuras seguintes, a Figura 27 contém a percentagem de ocorrências de cada classe a prever.

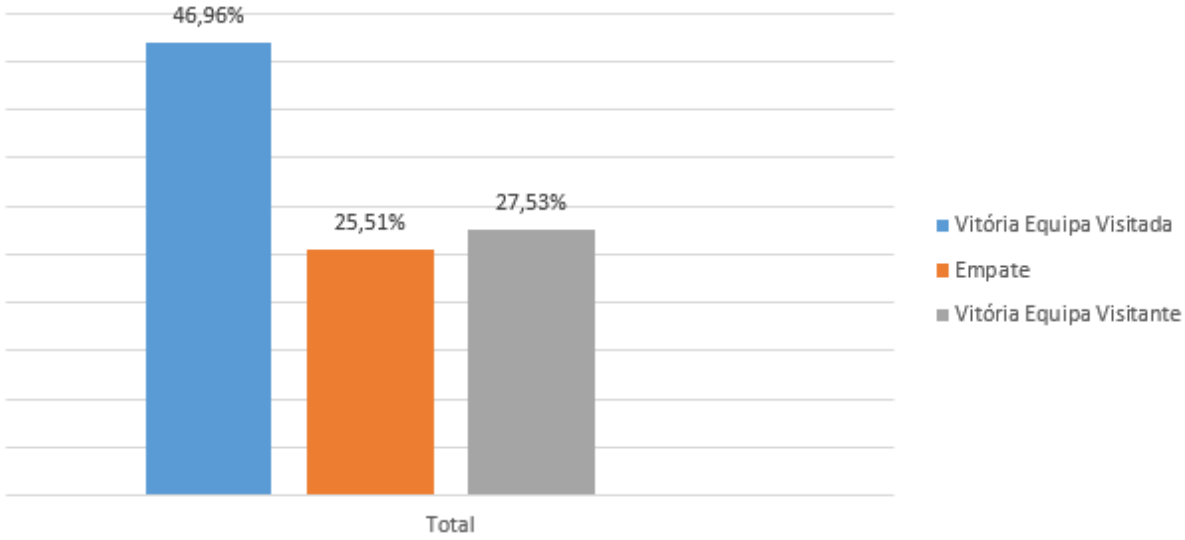


**Figura 27 - Número de Ocorrências R3S**

A Figura 28 apresenta a percentagem do número de ocorrências de cada classe na variável-alvo a favor ou contra a equipa visitada.

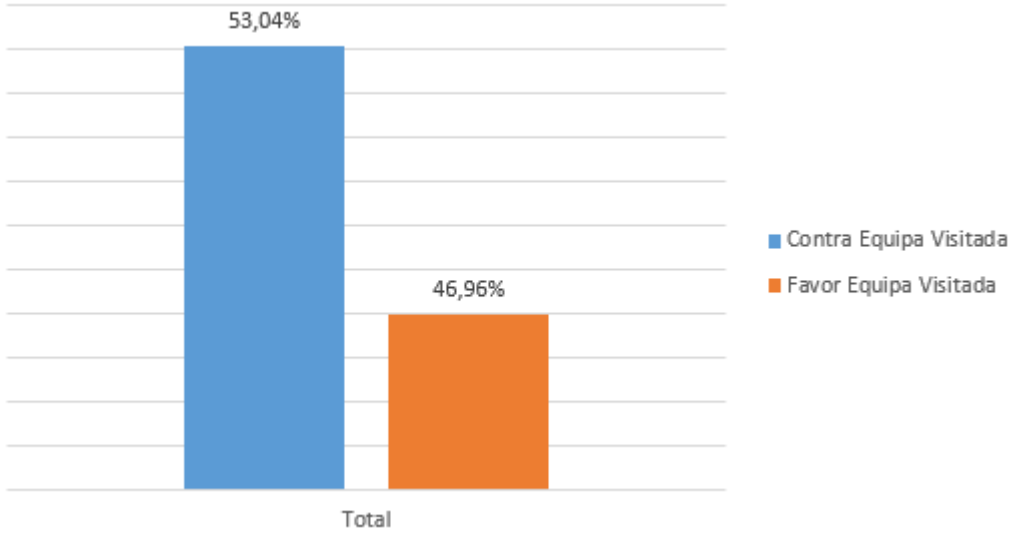


**Figura 28 - Número de Ocorrências R2SC**

A Figura 29 estão apresentas as percentagens das ocorrências de cada classe da variável-alvo a favor ou contra a equipa visitante.

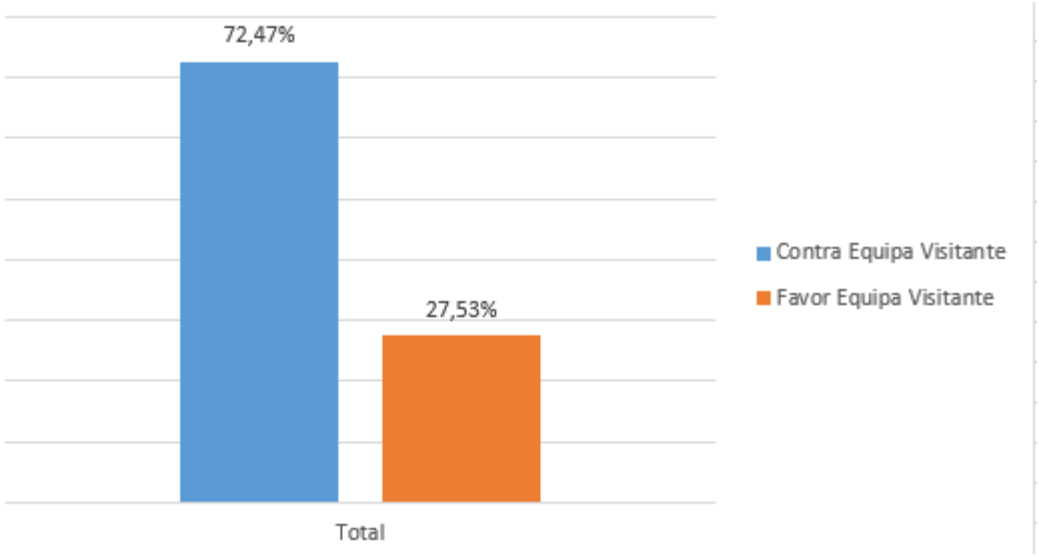


**Figura 29 - Número de Ocorrências R2SF**

Na Figura 30 estão apresentadas as percentagens do número de ocorrências de ambas as classes da variável-alvo mais ou menos de 7,5 cantos.

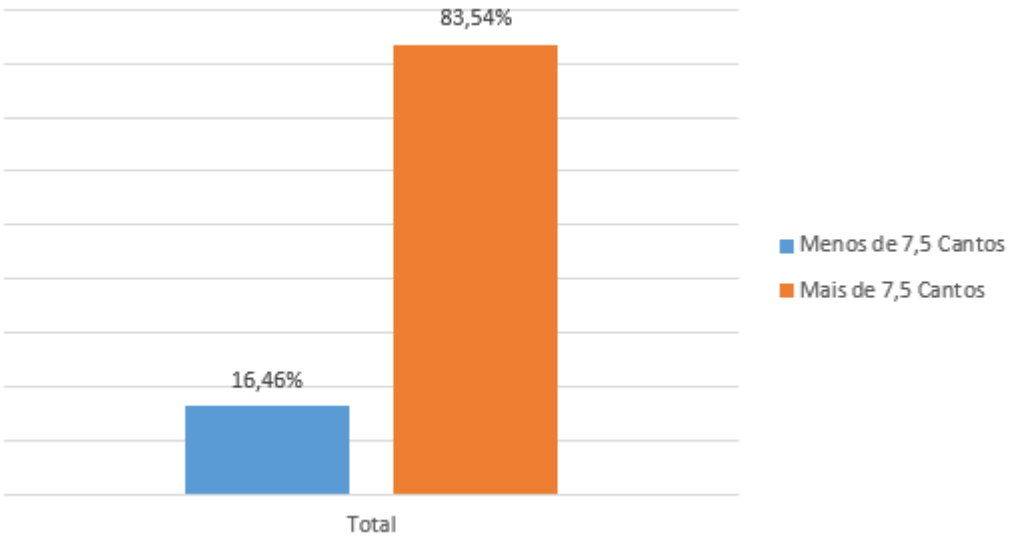


**Figura 30 - Número de Ocorrências C7,5**

Na Figura 31 estão apresentadas as ocorrências da variável-alvo mais ou menos de 8,5 cantos.

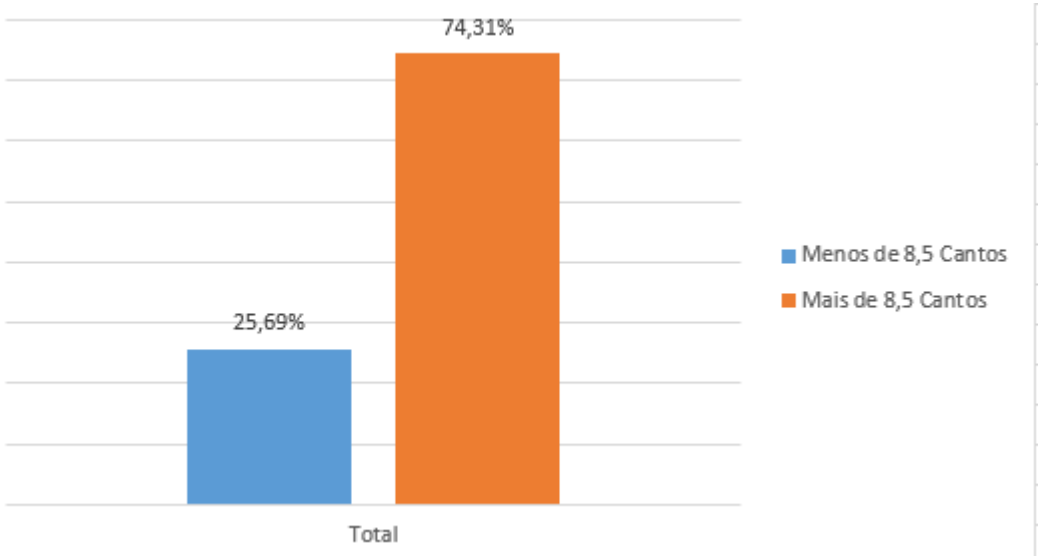


**Figura 31 - Número de Ocorrências C8,5**

Na Figura 32 estão apresentadas as ocorrências de cada classe da variável-alvo mais ou menos de 9,5 cantos.

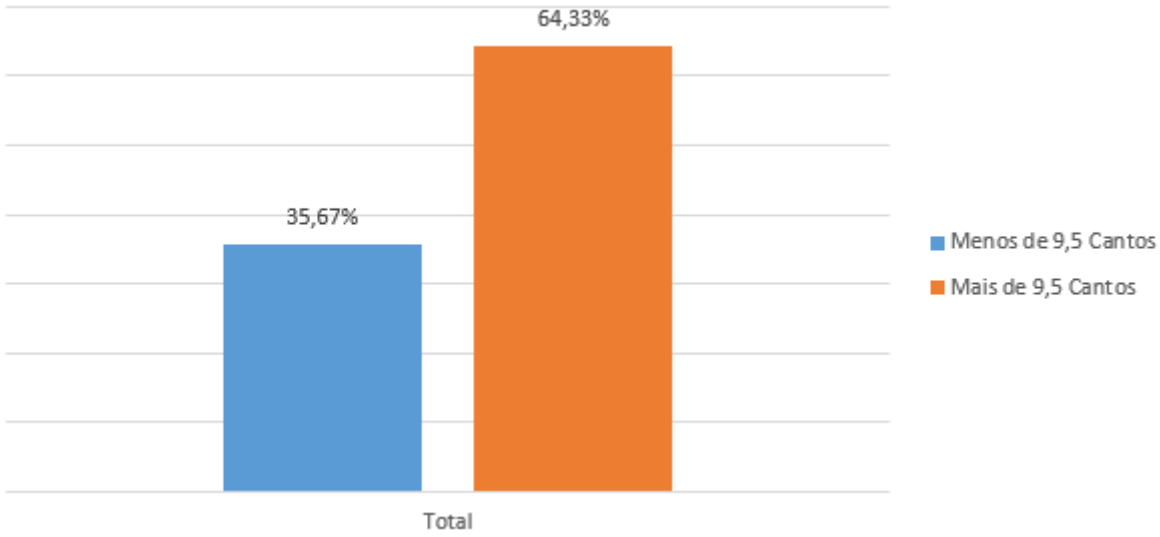


**Figura 32 - Número de Ocorrências C9,5**

Na Figura 33 é apresentado o número de ocorrências das classes mais ou menos de 10,5 cantos.

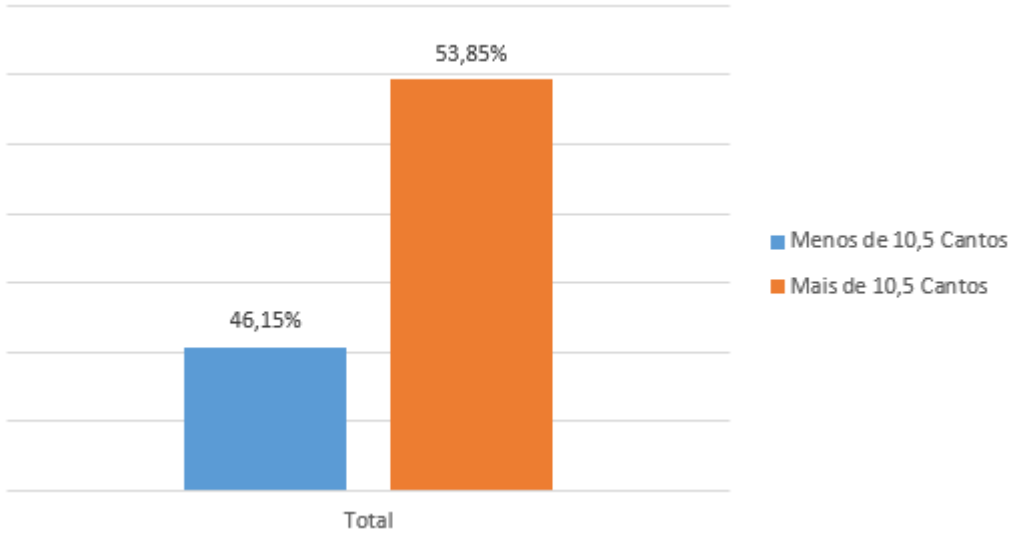


**Figura 33 - Número de Ocorrências C10,5**

Na Figura 34 está apresentado a percentagem do número de ocorrências de cada classe da variável-alvo mais ou menos de 1,5 golos.

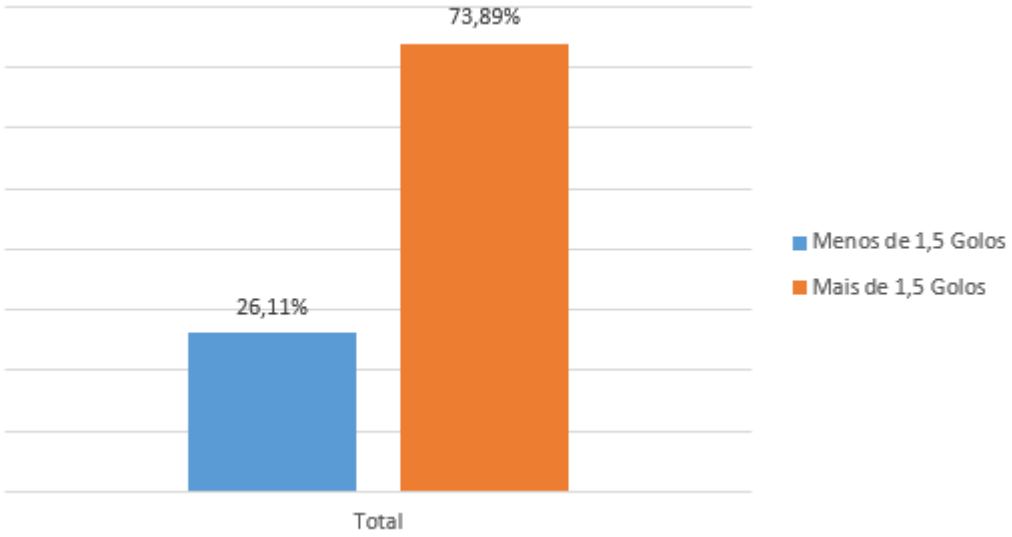


**Figura 34 - Número de Ocorrências G1,5**

Na Figura 35 são apresentadas as percentagens do número de ocorrências de cada classe da variável-alvo mais ou menos de 2,5 golos.

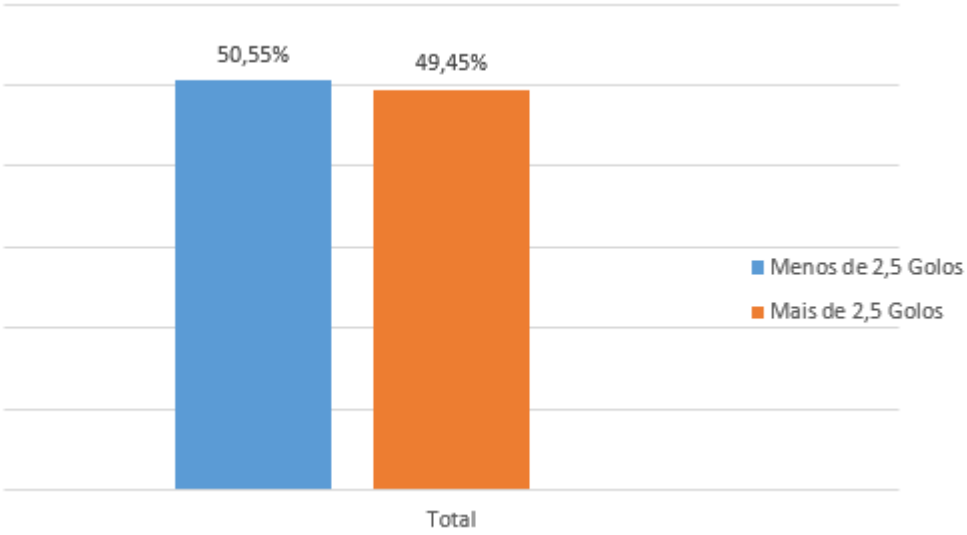


**Figura 35 - Número de Ocorrências G2,5**

Na Figura 36 estão apresentadas as percentagens das ocorrências de mais ou menos de 3,5 golos.

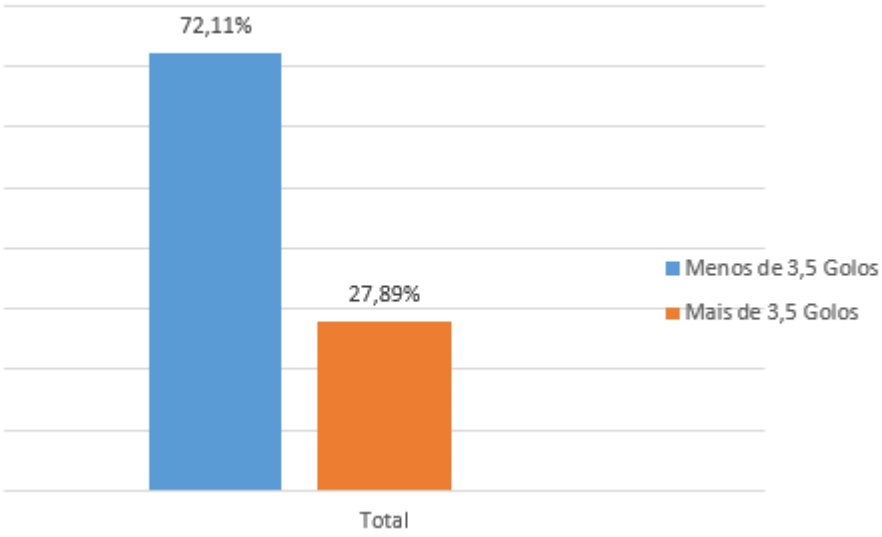


**Figura 36 - Número de Ocorrências G3,5**

### 4.2.3. Criação Modelos de Data Mining

Nesta fase do projeto foram também criados os modelos de DM, estes foram induzidos a partir dos dados tratados e processados anteriormente, que se encontram na tabela "PremierLeague" através de quatro técnicas de DM distintas: o *Naive Bayes* (NB), os *Support Vector Machines* (SVM), as Árvores de Decisão (AD) e *Lazy Learning* (LL). Para aplicar estas técnicas na criação dos modelos foi utilizada a ferramenta Weka que através dos algoritmos NaiveBayes, LibSVM, J48 e Kstar permite aplicar as técnicas supracitadas respetivamente.

Para o desenvolvimento dos modelos foram utilizados dois diferentes métodos de amostragem, o método "*Holdout Simple*" (HS) que utiliza 66% dos dados para treino e 34% para teste e o método de amostragem "*10-Folds Cross-Validation*" (10FCV).

As variáveis carregadas na tabela "PremierLeague" foram agrupadas em diferentes grupos de modo a que fosse possível definir diferentes cenários na indução dos modelos, para tal foi necessário focar nas características e processos existentes em cada jogo de futebol. Os grupos aqui mencionados foram criados após a execução de algumas tarefas de ETL extras. Cada um desses grupos exceto o "Estado Climatérico" e as "*Odds*" é constituído por dois conjuntos, um relacionado com as variáveis associadas à equipa visitada e outro pelas variáveis da equipa visitante. Os grupos e as respetivas variáveis são os seguintes:

- Agressividade:
    - Nome do Árbitro;
    - Média de Faltas;

    - Média de Faltas Sofridas;
    - Média de Cartões Amarelos;
    - Média de Cartões Vermelhos.
- Ataque:
    - Média de Golos;
    - Média de Remates;
    - Média de Remates à Baliza;
    - Média de Cantos.
- Defesa:
    - Média de Golos Sofridos;
    - Média de Remates Concedidos;
    - Média de Remates à Baliza Concedidos;
    - Média de Cantos Concedidos.
- Forma da equipa:
    - Número de Vitórias nos últimos 5 jogos.
- Confronto Direto:
    - Número de Vitórias nos últimos 5 confrontos diretos.
- Estado climatérico:
    - Precipitação.
- Odds:
    - Odd da Vitória da Equipa Visitada;
    - Odd do Empate;
    - Odd da Vitória da Equipa Visitante.

De seguida foi necessário definir os cenários através dos quais os modelos DM seriam induzidos. Foram definidos 19 cenários:

- CA – Todas as variáveis;
- CB – Agressividade Equipa Visitada VS Agressividade Equipa Visitante;
- CC – Ataque Equipa Visitada VS Ataque Equipa Visitante;
- CD – Defesa Equipa Visitada VS Defesa Equipa Visitante;
- CE – Ataque Equipa Visitada VS Defesa Equipa Visitante;
- CF – Defesa Equipa Visitada VS Ataque Equipa Visitante;

- CG – Agressividade Equipa Visitada VS Defesa Equipa Visitante;
- CH – Defesa Equipa Visitada VS Agressividade Equipa Visitante;
- CI – Odds;
- CJ – CB,CC;
- CK – CF,CG;
- CL – CD,CE;
- CM – Odds, Ataque Equipa Visitada, Ataque Equipa Visitante, Confronto Direto, Forma da Equipa Visitada, Forma da Equipa Visitante, Estado Climatérico, Cartões Vermelhos;
- CN – Odds, Confronto Direto, Forma da Equipa Visitada, Forma da Equipa Visitante, Estado Climatérico;
- CO – Média de Cantos Equipa Visitada VS Média de Cantos Equipa Visitante; Média de Cantos Sofridos Equipa Visitada VS Média de Cantos Sofridos Equipa Visitante;
- CP – Média de Cantos Equipa Visitada VS Média de Cantos Equipa Visitante;
- CQ – CC,CD;
- CR – Média de Golos Marcados e Sofridos Equipa Visitada VS Média de Golos Marcados e Sofridos Equipa Visitante;
- CS – Média de Golos Marcados Equipa Visitada VS Média de Golos Marcados Equipa Visitante.

Então, a criação dos Modelos de DM (MDM) é constituída por:

- Dezanove cenários:
    - CA, CB, ... CS.
- Dois métodos de amostragem:
    - 10FCV;
    - HS.
- Quatro Técnicas de DM:
    - NB;
    - SVM;
    - AD;
    - LL
- Dez variáveis alvo (*target*):
    - Vitória equipa visitada, empate ou vitória da equipa visitante (R3S);
    - Resultado a favor ou contra a equipa visitante (R2SF);

- Resultado a favor ou contra a equipa visitada (R2SC);
- Mais ou menos de 7,5 cantos (C7,5);
- Mais ou menos de 8,5 cantos (C8,5);
- Mais ou menos de 9,5 cantos (C9,5);
- Mais ou menos de 10,5 cantos (C10,5);
- Mais ou menos de 1,5 golos (G1,5);
- Mais ou menos de 2,5 golos (G2,5);
- Mais ou menos de 3,5 golos (G3,5).

No total, inicialmente, foram criados 752 modelos, 336 estão relacionados com as variáveis-alvo associadas ao resultado final de cada jogo, a R3S, a R2SF e a R2SC, 192 com as variáveis alvo relacionadas com o número de golos, a variável G1,5, a G2,5 e a G3,5 e os restantes 224 modelos dizem respeito às variáveis relacionadas com o número de cantos, a variável-alvo C7,5, a C8,5, a C9,5 e a C10,5. Depois para as variáveis-alvo R3S, R2SF, 'C7,5', 'C8,5', 'C9,5', 'G1,5' e 'G3,5' (que apresentavam valores desequilibrados no número de casos de cada classe) foram induzidos mais 24 modelos, recorrendo a técnica de *oversampling*, para cada atributo alvo, um total de 168. No total foram induzidos 920 modelos de DM. Para utilizar esta técnica foi executada uma função existente no WEKA, o SMOTE. Esta função replica em 100% o número de ocorrências da classe que contém menos ocorrências e pode ser utilizada as vezes necessárias até as classes ficarem com um número de ocorrências semelhante.

Os modelos de Data Mining (MDM) podem, portanto, ser representado pela seguinte expressão:

$$MDM = < \Delta, \alpha, MADM, VADM, CENVAR >$$

Onde,

- $\Delta$, representa as regras de DM;
- $\alpha$, representa a configuração do modelo de DM;
- MADM é o respetivo método de amostragem;
- VADM, corresponde à variável-alvo;
- CENVAR são as variáveis que podem ser utilizadas por cada cenário (CA - CS).

Por exemplo, se o modelo escolhido for composto pelo CI, utilizando como método de amostragem o 10FCV e como técnica de DM as AD e a variável-alvo for o RF3, o mesmo pode ser representado por:

$$MDM = < \Delta, \alpha, AD, 10FCV, RF3, EquipaVisitada, EquipaVisitante, OddVitóriaEquipaVisitada, OddEmpate, OddVitóriaEquipaVisitante >$$

## 4.3. Fase 3

A terceira fase deste projeto é a reunião de três fases distintas das metodologias utilizadas para o seu desenvolvimento. A fase "Avaliação" do CRISP-DM e do DSR que têm exatamente a mesma denominação e a fase "Escolha" do processo de tomada de decisão.

Nesta fase do projeto é efetuada uma análise detalhada aos resultados obtidos com a criação do modelo e é feita uma avaliação das métricas que foram tomadas em consideração.

Por último é efetuada a escolha de qual o melhor modelo a utilizar para efetuar a previsão de cada uma das variáveis-alvo. Os modelos induzidos são avaliados através da matriz de confusão, como explicado no tópico 2.2.2. deste documento, esta matriz é composta por verdadeiros positivos (VP) – número de exemplos positivos classificados corretamente, falsos positivos (FP) – número de exemplos positivos classificados como negativos, falsos negativos (FN) – número de exemplos negativos classificados como positivos e por último os verdadeiros negativos (VN) – que corresponde ao número de exemplos negativos classificados corretamente. As quatro métricas definidas para fazer a avaliação utilizam os resultados da matriz de confusão e as fórmulas matemáticas também descritas no tópico 2.2.2. Assim foi possível calcular a precisão, a especificidade, a sensibilidade e ainda apresentar a *Area Under Curve* (AUC). Para cada atributo alvo foram definidos parâmetros de qualidade para garantir a qualidade dos modelos e ao mesmo tempo facilitar a escolha do melhor modelo em cada uma das variáveis-alvo. Se os modelos não atingirem os parâmetros definidos é possível concluir que os modelos não têm a qualidade necessária para suportar os apostadores nessa determinada aposta. Com base na revisão de literatura efetuada e tendo e conta que não foi possível contactar um *expert* em apostas em jogos de futebol de modo a perceber quais seriam os *tresholds* que os modelos deveriam atingir, foram definidos como parâmetros de qualidade os valores mínimos de 65% nas métricas acuidade, especificidade, sensibilidade e AUC, a acuidade foi a métrica mais relevante para efetuar a avaliação dos modelos induzidos. Posteriormente

com a utilização do sistema através de técnicas de otimização e adaptabilidade será possível refinar os valores definir métricas mais específicas para cada um dos *targets.*

### 4.3.1. Vitória Equipa Visitada, Empate ou Vitória Equipa Visitante

Os melhores modelos gerados para esta variável-alvo estão presentes na Tabela 11 e na Tabela 12, a primeira contem os três melhores modelos obtidos pelo método de amostragem 10FCV e os restantes três pelo HS.

**Tabela 11 - Melhores Modelos R3S 10FCV (%)**

| Modelo | Cenário | Técnica | Acuidade | AUC |
|---|---|---|---|---|
| **1** | CA | SVM | 57,29 | 0,64 |
| **2** | CM | SVM | 57,08 | 0,64 |
| **3** | CN | SVM | 55,89 | 0,63 |

**Tabela 12 - Melhores Modelos R3S HS (%)**

| Modelo | Cenário | Técnica | Acuidade | AUC |
|---|---|---|---|---|
| **4** | CA | SVM | 57,50 | 0,64 |
| **5** | CL | SVM | 55,95 | 0,61 |
| **6** | CM | SVM | 56,36 | 0,64 |

Para este *target* foram também induzidos modelos depois de aplicado um *oversampling* ao conjunto de dados, devido ao desequilíbrio do número de exemplos existentes de cada classe que se pretendia prever, os melhores modelos e respetivas métricas obtidas estão apresentados na Tabela 13.

**Tabela 13 - Melhores Modelos R3S Oversampling (%)**

| Modelo | Cenário | M. Amostragem | Técnica | Acuidade | AUC |
|---|---|---|---|---|---|
| **7** | CA | 10FCV | LL | 76,08 | 0,92 |
| **8** | CA | 10FCV | AD | 70,31 | 0,89 |
| **9** | CM | 10FCV | LL | 74,49 | 0,90 |

Os modelos 7, 8 e 9, que foram induzidos depois de aplicada a técnica de *oversampling* ao conjunto de dados inicial foram os únicos que cumpriram os parâmetros de qualidade. A acuidade foi a métrica utilizada para distinguir o melhor modelo entre os três que atingiram os parâmetros de qualidade e nesta situação o modelo 7 foi o modelo que obteve uma maior percentagem de acuidade.

O modelo escolhido para esta previsão é, então, o modelo 7 que pode ser representado pela seguinte expressão:

$$M7 = < \Delta, \alpha, SVM, 10FCV, R3S, EquipaVisitada, EquipaVisitante, OddEquipaVisitada, OddEmpate, OddEquipaVisitante, Arbitro, MediaGolosEquipaVisitada, MediaGolosConcedidosEquipaVisitada, MediaRematesEquipaVisitada, MediaRematesConcedidosEquipaVisitada, MediaRematesBalizaEquipaVisitada, MediaRemateBalizaconcedidosEquipaVisitada, MediaCantosEquipaVisitada, MediaCantosConcedidosEquipaVisitada, MediaFaltasEquipaVisitada, MediaFaltasSofridasEquipaVisitada, MediaAmarelosEquipaVisitada, MediaVermelhosEquipaVisitada, VitoriasUltimos5JogosEquipaVisitada, Ultimos5ConfrontosDiretosEquipaVisitada, MediaGolosEquipaVisitante, MediaGolosConcedidosEquipaVisitante, MediaRematesEquipaVisitante, MediaRematesConcedidosEquipaVisitante, MediaRematesBalizaEquipaVisitante, MediaRemateBalizaconcedidosEquipaVisitante, MediaCantosEquipaVisitante, MediaCantosConcedidosEquipaVisitante, MediaFaltasEquipaVisitante, MediaFaltasSofridasEquipaVisitante, MediaAmarelosEquipaVisitante, MediaVermelhosEquipaVisitante, VitoriasUltimos5JogosEquipaVisitante, Ultimos5ConfrontosDiretosEquipaVisitante, Precipitacao >$$

Na Figura 37, Figura 38 e Figura 39 são apresentadas as curvas ROC das três classes do modelo 7.

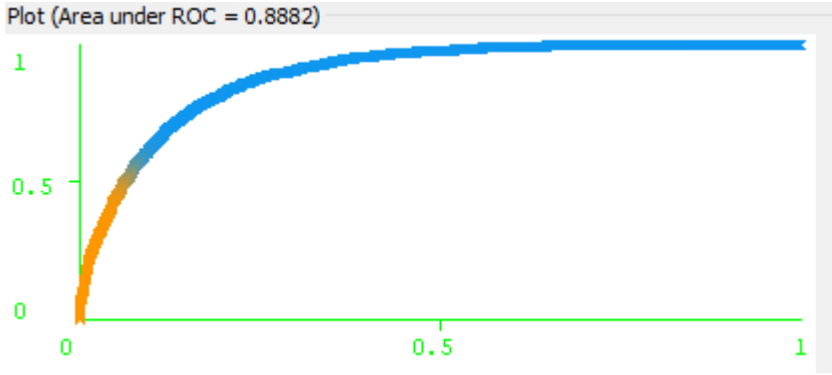


**Figura 37 - Classe Vitória Equipa Visitante**

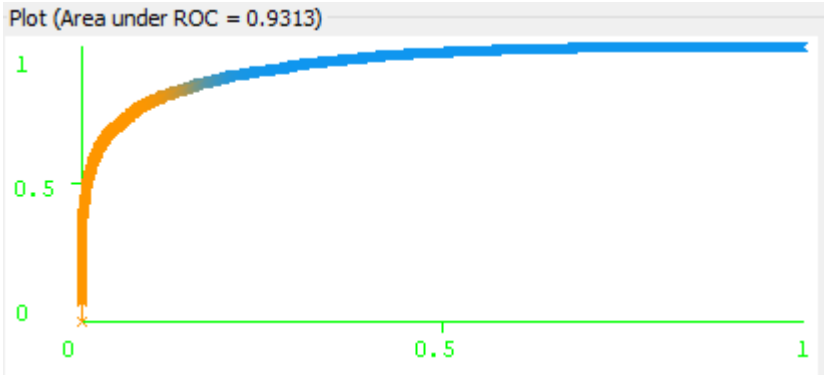


**Figura 38 - Classe Empate**

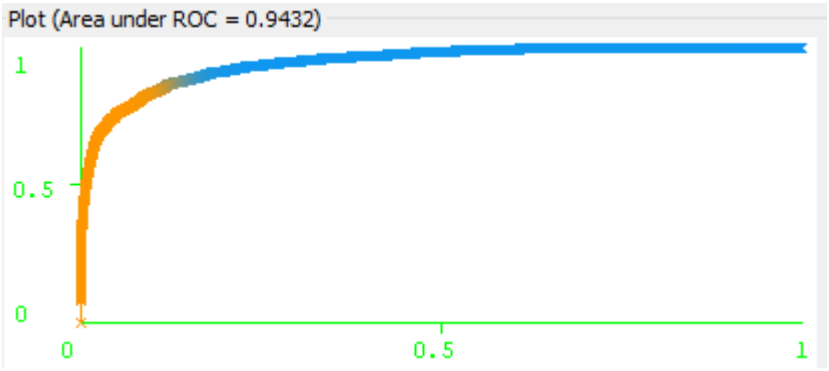


**Figura 39 - Classe Vitória Equipa Visitada**

### 4.3.2. Resultado a favor ou contra a equipa visitada

Na Tabela 14 e na Tabela 15 estão identificados os três melhores modelos induzidos pelos dois métodos de amostragem utilizados na criação dos mesmos.

**Tabela 14 - R2SC Melhores Modelos 10FCV (%)**

| Modelo | Cenário | Técnica | Especificidade | Sensibilidade | Acuidade | AUC |
|---|---|---|---|---|---|---|
| **1** | CA | NB | 61,38 | 75,88 | 69,07 | 0,75 |
| **2** | CA | SVM | 67,77 | 72,38 | 70,20 | 0,70 |
| **3** | CM | SVM | 69,01 | 71,45 | 70,30 | 0,70 |

**Tabela 15 - R2SC Melhores Modelos HS (%)**

| Modelo | Cenário | Técnica | Especificidade | Sensibilidade | Acuidade | AUC |
|---|---|---|---|---|---|---|
| **4** | CA | NB | 59,95 | 75,22 | 68,10 | 0,74 |
| **5** | CA | SVM | 67,35 | 72,21 | 69,94 | 0,70 |
| **6** | CM | SVM | 66,33 | 71,25 | 69,17 | 0,69 |

Os modelos que cumprem todos os parâmetros de qualidade são o modelo 2, o 3, o 5 e o 6, optou-se por escolher o modelo 3, pois, a métrica que é considerada mais relevante é a acuidade e o modelo 3 atinge um valor superior aos restantes, ficando então tomada a decisão de introduzir o modelo 3 posteriormente no funcionamento do protótipo.

Então, o modelo 3 pode ser representado pela expressão:

$$M3 = < \Delta, \alpha, SVM, 10FCV, R2SC, EquipaVisitada, EquipaVisitante, OddEquipaVisitada, OddEquipaVisitante, MediaGolosEquipaVisitada, MediaRematesEquipaVisitada, MediaVermelhosEquipaVisitada, Ultimos5ResultadosEquipaVisitada, ConfrontoDiretoEquipaVisitada, MediaGolosEquipaVisitante, MediaRematesEquipaVisitante, Ultimos5ResultadosRquipaVisitante, ConfrontoDiretoEquipaVisitante, Precipitacao >$$

A Figura 40 apresenta a curva ROC do modelo 3.

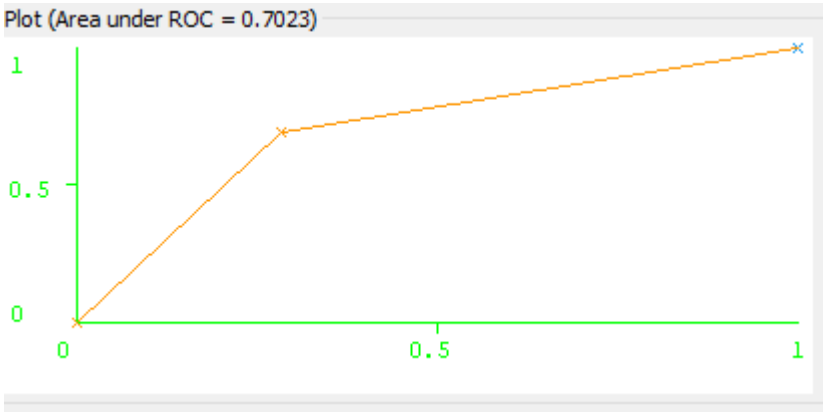


**Figura 40 - Cruva ROC Melhor Modelo R2SC**

### 4.3.3. Resultado a favor ou contra a equipa visitante

Na Tabela 16 e na Tabela 17 estão apresentados os melhores modelos de cada um dos métodos de amostragem utilizados no desenvolvimento dos mesmos. E na Tabela 18 os melhores modelos depois de apicado o *oversampling* ao conjunto de dados.

**Tabela 16 - R2SF Melhores Modelos 10FCV (%)**

| Modelo | Cenário | Técnica | Especificidade | Sensibilidade | Acuidade | AUC |
|---|---|---|---|---|---|---|
| **1** | CA | SVM | 95,61 | 22,87 | 75,59 | 0,59 |
| **2** | CA | AD | 93,91 | 26,99 | 75,49 | 0,64 |
| **3** | CL | AD | 93,21 | 27,65 | 75,16 | 0,68 |

**Tabela 17 - R2SF Melhores Modelos HS (%)**

| Modelo | Cenário | Técnica | Especificidade | Sensibilidade | Acuidade | AUC |
|---|---|---|---|---|---|---|
| **4** | CC | AD | 94,60 | 26,26 | 76,01 | 0,63 |
| **5** | CF | AD | 94,03 | 27,35 | 75,89 | 0,64 |
| **6** | CM | SVM | 95,83 | 22,32 | 75,83 | 0,59 |

**Tabela 18 - R2SF Melhores Modelos *Oversampling* (%)**

| Modelo | M. Amostragem | Cenário | Técnica | Especificidade | Sensibilidade | Acuidade | AUC |
|---|---|---|---|---|---|---|---|
| **7** | 10FCV | CI | LL | 53,27 | 92,83 | 77,13 | 0,87 |
| **8** | 10FCV | CM | LL | 73,41 | 88,58 | 81,49 | 0,91 |
| **9** | 10FCV | CM | AD | 80,47 | 72,77 | 77,14 | 0,85 |

Os modelos que cumprem os parâmetros de qualidade previamente definidos, foram induzidos depois de aplicada a técnica de *oversampling* no conjunto de dados, são os modelos 8 e 9, destes dois, o modelo escolhido foi o 8, pois, foi o modelos que apresentou maior acuidade. Este modelo pode ser representado pela seguinte expressão:

$$M8 = < \Delta, \alpha, LL, 10FCV, R2SF, EquipaVisitada, EquipaVisitante, OddEquipaVisitada, OddEquipaVisitante, Arbitro, MediaGolosEquipaVisitada, MediaRematesEquipaVisitada, MediaVermelhosEquipaVisitada, VitoriasUltimos5JogosEquipaVisitada, Ultimos5ConfrontosDiretosEquipaVisitada, MediaGolosEquipaVisitante, MediaRematesEquipaVisitante, MediaVermelhosEquipaVisitante, VitoriasUltimos5JogosEquipaVisitante, Ultimos5ConfrontosDiretosEquipaVisitante, Precipitacao >$$

Na Figura 41 é apresentada a curva ROC do modelo 8.

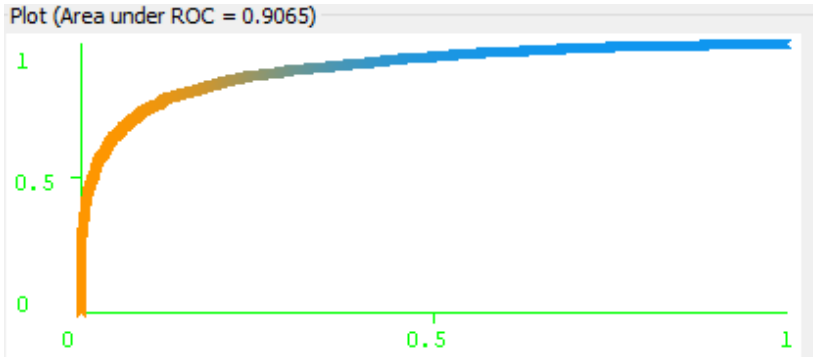

**Figura 41 - Curva ROC Melhor Modelo R2SF**

### 4.3.4. Mais ou menos de 7,5 cantos

No caso da variável-alvo C7,5 foram também recolhidos os seis melhores modelos, três provenientes dos modelos criados através do método de amostragem 10FCV e os restantes três através do HS. A Tabela 19 e a Tabela 20 apresentam as métricas que resultaram desses modelos. Na Tabela 21 estão apresentados os três melhores modelos obtidos depois de aplicado o *oversampling* ao conjunto de dados.

**Tabela 19 - C7,5 Melhores Modelos CV (%)**

| Modelo | Cenário | Técnica | Especificidade | Sensibilidade | Acuidade | AUC |
|---|---|---|---|---|---|---|
| **1** | CF | NB | 0,74 | 99,83 | 83,52 | 0,56 |
| **2** | CO | NB | 0,62 | 99,88 | 83,54 | 0,61 |
| **3** | CP | LL | 0,37 | 99,39 | 83,10 | 0,54 |

**Tabela 20 - C7,5 Melhores Modelos %Split (%)**

| Modelo | Cenário | Técnica | Especificidade | Sensibilidade | Acuidade | AUC |
|---|---|---|---|---|---|---|
| **4** | CC | NB | 0,67 | 99,93 | 82,32 | 0,55 |
| **5** | CF | NB | 1,34 | 99,86 | 82,38 | 0,55 |
| **6** | CO | NB | 1,01 | 100 | 81,86 | 0,6 |

**Tabela 21 - Melhores Modelos C7,5 Oversampling**

| Modelo | M. Amostragem | Cenário | Técnica | Especificidade | Sensibilidade | Acuidade | AUC |
|---|---|---|---|---|---|---|---|
| **7** | HS | CO | LL | 89,16 | 71,34 | 80,99 | 0,9 |
| **8** | 10FCV | CO | AD | 83,29 | 72,67 | 78,42 | 0,81 |
| **9** | 10FCV | CO | LL | 95,63 | 72,28 | 84,93 | 0,90 |

Os modelos 7, 8 e 9 cumprem todos os parâmetros mínimos de qualidade definidos, destes três o que tem uma acuidade superior é o modelo 7. Este modelo pode ser representado pela expressão:

$$M7 = \\ < \Delta, \alpha, LL, 10FCV, 'C7{,}5', EquipaVisitada, EquipaVisitante, CantosEquipaVisitada, \\ CantosConcedidosEquipaVisitada, CantosEquipaVisitante,$$

$CantosConcedidosEquipaVisitante >$

Na Figura 42 está apresentada a curva ROC do modelo:

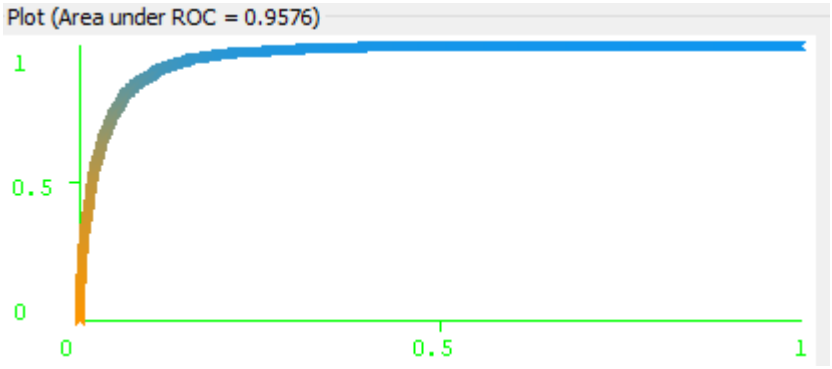


**Figura 42 - Curva ROC Melhor Modelo C7,5**

### 4.3.5. Mais ou menos de 8,5 cantos

Também para esta variável-alvo foram destacados os seis melhores modelos. Estes estão apresentados na Tabela 22 e na Tabela 23. Na Tabela 24 são apresentados os três melhores modelos obtidos depois de aplicada a técnica de *oversampling* ao conjunto de dados.

**Tabela 22 - C8,5 Melhores Modelos 10FCV (%)**

| Modelo | Cenário | Técnica | Especificidade | Sensibilidade | Acuidade | AUC |
|---|---|---|---|---|---|---|
| **1** | CF | NB | 2,29 | 98,69 | 73,93 | 0,57 |
| **2** | CO | NB | 4,33 | 98,17 | 74,07 | 0,62 |
| **3** | CP | NB | 1,18 | 99,24 | 74,05 | 0,58 |

**Tabela 23 - C8,5 Melhores Modelos HS (%)**

| Modelo | Cenário | Técnica | Especificidade | Sensibilidade | Acuidade | AUC |
|---|---|---|---|---|---|---|
| **1** | CF | NB | 1,77 | 98,61 | 72,50 | 0,56 |
| **2** | CO | NB | 3,53 | 98,45 | 72,86 | 0,61 |
| **3** | CP | NB | 1,99 | 98,61 | 72,56 | 0,55 |

**Tabela 24 - Melhores Modelos C,85 *Oversampling***

| Modelo | M. Amostragem | Cenário | Técnica | Especificidade | Sensibilidade | Acuidade | AUC |
|---|---|---|---|---|---|---|---|
| **7** | HS | CO | LL | 79,25 | 69,9 | 74,54 | 0,84 |
| **8** | 10FCV | CO | LL | 87,89 | 68,40 | 78,32 | 0,89 |
| **9** | 10FCV | CO | J48 | 78,00 | 67,79 | 72,76 | 0,84 |

Os três modelos induzidos depois de aplicada a técnica de *oversampling* obtiveram os valores mínimos dos parâmetros de qualidade definidos, sendo então aprovados. Destes modelos, o modelo 8 foi o que obteve uma acuidade superior e foi então o escolhido. O modelo 8 pode ser representado pela expressão:

$$M8 = \\ < \Delta, \alpha, LL, 10FCV, '8{,}5', EquipaVisitada, EquipaVisitante, CantosEquipaVisitada, \\ CantosConcedidosEquipaVisitada, CantosEquipaVisitante, \\ CantosConcedidosEquipaVisitante >$$

Na Figura 43 está apresentada a curva ROC deste modelo.

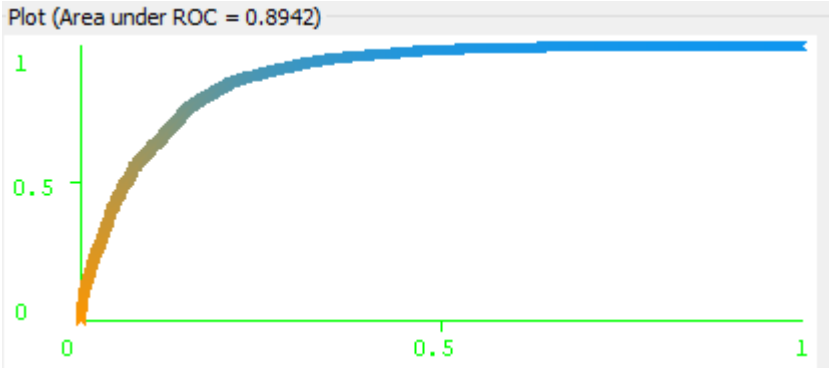


**Figura 43 - Curva ROC Melhor Modelo C8,5**

### 4.3.6. Mais ou menos de 9,5 cantos

Também para esta variável-alvo são escolhidos os três melhores modelos de cada método de amostragem abordado. Estão apresentados na Tabela 25 e na Tabela 26.

**Tabela 25 - C9,5 Melhores Modelos 10FCV (%)**

| Modelo | Cenário | Técnica | Especificidade | Sensibilidade | Acuidade | AUC |
|---|---|---|---|---|---|---|
| **1** | CF | NB | 10,67 | 93,52 | 63,97 | 0,57 |
| **2** | CO | NB | 19,92 | 89,18 | 64,47 | 0,61 |
| **3** | CP | NB | 12,03 | 91,66 | 63,26 | 0,58 |

**Tabela 26 -C9,5 Melhores Modelos HS (%)**

| Modelo | Cenário | Técnica | Especificidade | Sensibilidade | Acuidade | AUC |
|---|---|---|---|---|---|---|
| **4** | CF | NB | 14,61 | 90,60 | 62,74 | 0,55 |
| **5** | CO | NB | 21,10 | 89,28 | 64,29 | 0,62 |
| **6** | CP | NB | 14,29 | 89,66 | 62,02 | 0,56 |

**Tabela 27 - Melhores Modelos C9,5 *Oversampling* (%)**

| Modelo | M. Amostragem | Cenário | Técnica | Especificidade | Sensibilidade | Acuidade | AUC |
|---|---|---|---|---|---|---|---|
| **7** | 10FCV | CO | LL | 81,70 | 60,07 | 71,44 | 0,81 |
| **8** | HS | CO | LL | 78,24 | 55,19 | 66,97 | 0,81 |
| **9** | 10FCV | CP | LL | 79,54 | 57,33 | 68,90 | 0,81 |

Para este atributo-alvo, nenhum dos modelos induzidos atingiu todos os parâmetros de qualidade mínimos aceitáveis. Nestes casos todos os modelos foram descartados.

### 4.3.7. Mais ou menos de 10,5 cantos

São escolhidos também para esta variável-alvo os seis melhores modelos, os três melhores criados através do método de amostragem 10FCV e os restantes através do HS como está apresentado na Tabela 28 e na Tabela 29.

**Tabela 28 - C10,5 Melhores Modelos 10FCV (%)**

| Modelo | Cenário | Técnica | Especificidade | Sensibilidade | Acuidade | AUC |
|---|---|---|---|---|---|---|
| **1** | CA | SVM | 47,19 | 72,22 | 60,67 | 0,60 |
| **2** | CO | NB | 47,63 | 68,20 | 58,70 | 0,62 |
| **3** | CO | SVM | 49,47 | 72,67 | 61,96 | 0,61 |

**Tabela 29 - C10,5 Melhores Modelos HS (%)**

| Modelo | Cenário | Técnica | Especificidade | Sensibilidade | Acuidade | AUC |
|---|---|---|---|---|---|---|
| **4** | CA | SVM | 46,14 | 72,84 | 60,48 | 0,60 |
| **5** | CO | NB | 47,69 | 68,29 | 58,75 | 0,62 |
| **6** | CO | SVM | 48,07 | 74,17 | 62,08 | 0,61 |

Para este atributo alvo nenhum dos modelos criados cumpre os requisitos mínimos de qualidade definidos, portanto não será sugerida qualquer previsão.

### 4.3.8. Mais ou menos de 1,5 golos

Foram selecionados os 3 melhores modelos originados por cada um dos dois métodos de amostragem utilizados para o desenvolvimento do projeto. Os valores obtidos através da matriz de confusão estão apresentados na Tabela 30 e na Tabela 31. Na Tabela 32 estão apresentados os três melhores modelos obtidos depois de aplicar a técnica de *oversampling* no conjunto de dados.

**Tabela 30 - G1,5 Melhores Modelos 10FCV (%)**

| Modelo | Cenário | Técnica | Especificidade | Sensibilidade | Acuidade | AUC |
|---|---|---|---|---|---|---|
| **1** | CQ | NB | 11,71 | 93,78 | 72,35 | 0,61 |
| **2** | CF | NB | 4,11 | 96,47 | 72,35 | 0,57 |
| **3** | CR | LL | 11,40 | 93,97 | 72,41 | 0,60 |

**Tabela 31 - G1,5 Melhores Modelos HS (%)**

| Modelo | Cenário | Técnica | Especificidade | Sensibilidade | Acuidade | AUC |
|---|---|---|---|---|---|---|
| **4** | CQ | NB | 12,99 | 94,01 | 71,73 | 0,62 |
| **5** | CQ | AD | 11,47 | 93,68 | 71,07 | 0,57 |
| **6** | CR | LL | 14,07 | 93,27 | 71,49 | 0,59 |

**Tabela 32 - Melhores Modelos G1,5 *Oversampling***

| Modelo | M. Amostragem | Cenário | Técnica | Especificidade | Sensibilidade | Acuidade | AUC |
|---|---|---|---|---|---|---|---|
| **7** | 10FCV | CQ | J48 | 89,87 | 70,63 | 76,54 | 0,91 |
| **8** | HS | CQ | LL | 89,45 | 72,39 | 81,46 | 0,91 |
| **9** | 10FCV | CQ | LL | 90,08 | 71,87 | 81,65 | 0,91 |

Dos modelos induzidos, apenas os induzidos depois da aplicação da técnica de *oversampling* no conjunto de dados cumprem os parâmetros de qualidade, destes o modelo que apresentou maior acuidade foi o modelo 9. Este foi o modelo escolhido e pode ser representado pela expressão:

$$M9 = < \Delta, \alpha, LL, 10FCV,' G1{,}5', EquipaVisitada, EquipaVisitante, MediaGolosEquipaVisitada, MediaGolosConcedidosEquipaVisitada, MediaRematesEquipaVisitada, MediaRematesConcedidosEquipaVisitada, MediaRematesBalizaEquipaVisitada, MediaRemateBalizaconcedidosEquipaVisitada, MediaCantosEquipaVisitada, MediaCantosConcedidosEquipaVisitada, MediaGolosEquipaVisitante, MediaGolosConcedidosEquipaVisitante, MediaRematesEquipaVisitante, MediaRematesConcedidosEquipaVisitante, MediaRematesBalizaEquipaVisitante, MediaRemateBalizaconcedidosEquipaVisitante, MediaCantosEquipaVisitante, MediaCantosConcedidosEquipaVisitante >$$

Na Figura 44 está apresentada a curva ROC do modelo 9.

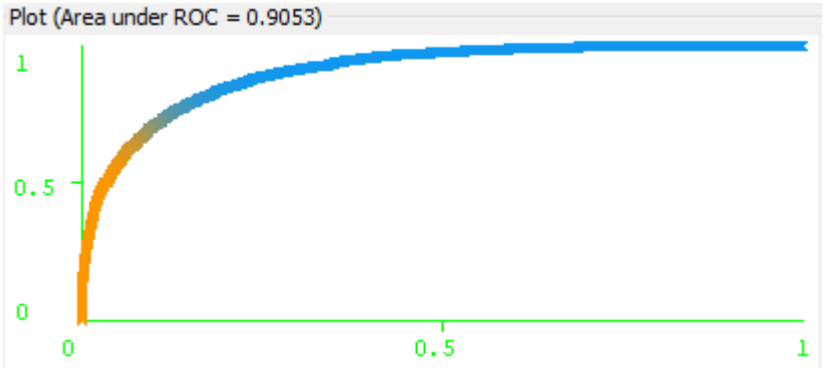


**Figura 44 - Curva ROC Melhor Modelo G1,5**

### 4.3.9. Mais ou menos de 2,5 golos

Foram selecionados os seis melhores modelos, os três melhores resultantes de cada um dos métodos de amostragem utilizados no desenvolvimento do projeto. Os valores obtidos estão apresentados na Tabela 33 e Tabela 34.

**Tabela 33 - G2,5 Melhores Modelos 10FCV (%)**

| Modelo | Cenário | Técnica | Especificidade | Sensibilidade | Acuidade | AUC |
|---|---|---|---|---|---|---|
| **1** | CQ | SVM | 65,00 | 58,82 | 60,95 | 0,61 |
| **2** | CR | NB | 66,12 | 52,07 | 59,17 | 0,63 |
| **3** | CR | SVM | 67,48 | 57,43 | 62,51 | 0,63 |

**Tabela 34 - G,2,5 Melhores Modelos HS (%)**

| Modelo | Cenário | Técnica | Especificidade | Sensibilidade | Acuidade | AUC |
|---|---|---|---|---|---|---|
| **4** | CQ | SVM | 63,80 | 56,15 | 60,06 | 0,60 |
| **5** | CR | NB | 66,82 | 52,01 | 59,58 | 0,63 |
| **6** | CR | SVM | 68,80 | 57,25 | 63,15 | 0,63 |

Nenhum dos modelos induzidos cumpre os parâmetros mínimos de qualidade definidos, portanto, todos estes modelos foram descartados.

### 4.3.10. Mais ou menos de 3,5 golos

Para a última variável-alvo que se pretendia efetuar a previsão foram também definidos como nos casos anteriores os seis melhores modelos. Os valores obtidos pelas suas métricas estão apresentados na Tabela 35 e Tabela 36. Na Tabela 37 estão apresentados os valores obtidos nas métricas, depois de efetuada a técnica de *oversampling* ao conjunto de dados.

**Tabela 35 - G3,5 Melhores Modelos 10FCV (%)**

| Modelo | Cenário | Técnica | Especificidade | Sensibilidade | Acuidade | AUC |
|---|---|---|---|---|---|---|
| **1** | CQ | NB | 25,40 | 85,26 | 68,56 | 0,60 |
| **2** | CR | NB | 15,02 | 93,29 | 71,46 | 0,62 |
| **3** | CS | NB | 13,43 | 93,07 | 70,85 | 0,58 |

**Tabela 36 - G3,5 Melhores Modelos HS (%)**

| Modelo | Cenário | Técnica | Especificidade | Sensibilidade | Acuidade | AUC |
|---|---|---|---|---|---|---|
| **4** | CQ | NB | 24,79 | 88,37 | 70,36 | 0,61 |
| **5** | CD | NB | 9,24 | 94,35 | 70,24 | 0,58 |
| **6** | CR | NB | 14,71 | 94,10 | 71,61 | 0,54 |

**Tabela 37 - Melhores Modelos G3,5 *Oversampling***

| Modelo | M. Amostragem | Cenário | Técnica | Especificidade | Sensibilidade | Acuidade | AUC |
|---|---|---|---|---|---|---|---|
| **7** | 10FCV | CQ | LL | 90,08 | 71,87 | 81,65 | 0,91 |
| **8** | HS | CQ | LL | 90,1 | 66,52 | 79,05 | 0,91 |
| **9** | 10FCV | CQ | AD | 89,84 | 72,64 | 80,35 | 0,90 |

Dos modelos induzidos apresentados, os modelos 7, 8 e 9 cumprem os parâmetros de qualidade definidos, destes, o modelo que obteve uma acuidade superior foi o modelo 7, portanto, este foi o modelo escolhido para ser implementado no protótipo. O modelo 7 pode ser representado pela expressão:

$$M7 = < \Delta, \alpha, LL, 10FCV,' G3{,}5', EquipaVisitada, EquipaVisitante, GolosMarcadosEquipaVisitada, GolosSofridosEquipaVisitada, RematesEquipaVisitada, RematesConcedidosEquipaVisitada, MediaRematesBalizaEquipaVisitada, MediaRematesBaliza\ ConcedidiosEquipaVisitada, MediaCantosEquipaVisitada, MediaCantosConcedidosEquipaVisitada, GolosMarcadosEquipaVisitante, GolosSofridosEquipaVisitante, RematesEquipaVisitante, RematesConcedidosEquipaVisitante, MediaRematesBalizaEquipaVisitante, MediaRematesBaliza\ ConcedidiosEquipaVisitante, MediaCantosEquipaVisitante, MediaCantosConcedidosEquipaVisitante >$$

A curva ROC deste modelo está apresentada na Figura 45.

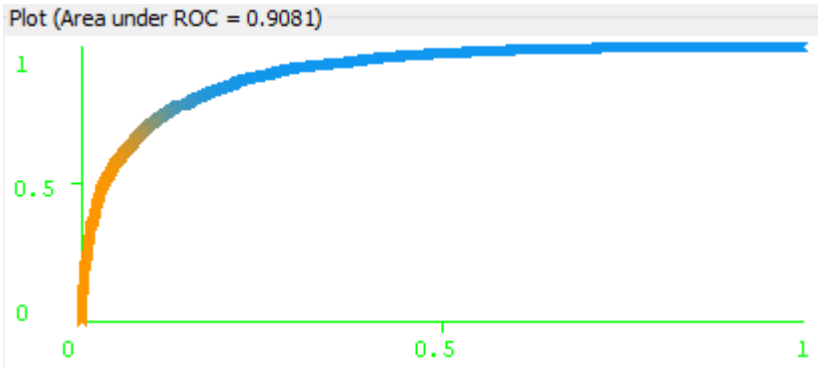


**Figura 45 - Curva ROC Melhor Modelo G3,5**

## 4.4. Fase 4

A fase quatro é composta pela combinação de duas fases, a fase "Desenvolvimento" do CRISP-DM e a fase "Implementação" do processo de tomada de decisão.

Foi nesta fase que se iniciou a construção do protótipo do sistema que permite efetuar previsões inteligentes de vários eventos, em tempo real (Filipe Portela et al., 2013; Filipe Portela et al., 2011), em jogos de futebol.

Foi decidido criar uma plataforma *web*, pois, esta permite um fácil acesso a partir de qualquer local num maior número de dispositivos.

Na Figura 46 encontra-se a arquitetura pela qual o protótipo pode ser representado.

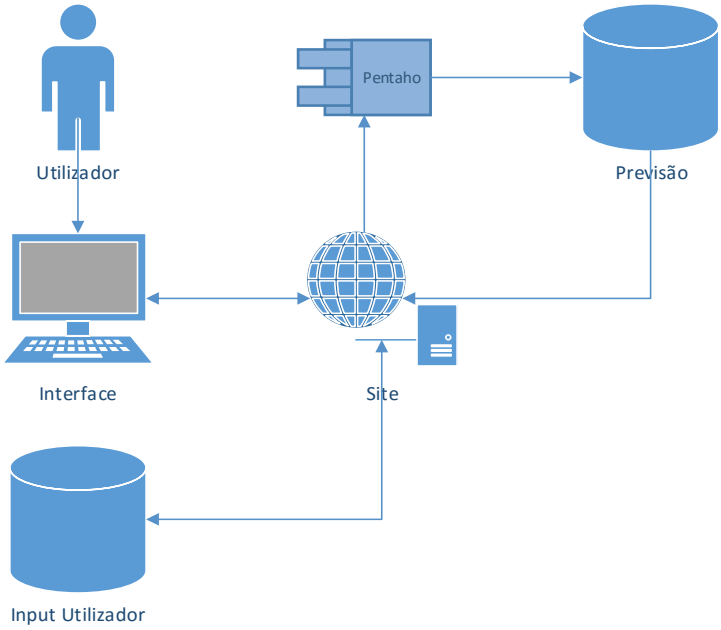


**Figura 46 - Arquitetura Protótipo**

Para a utilização do protótipo o utilizador começa por inserir informações necessárias para o sistema saber de que jogo se quer efetuar uma das possíveis previsões. Esta informação é armazenada numa base de dados que está representada na Figura 46 pelo "Input Utilizador". A plataforma vai depois utilizar essas informações e efetuar um pedido, através de um ficheiro *.bat* que inicia automaticamente o processo desenhado na ferramenta Pentaho (Figura 47).

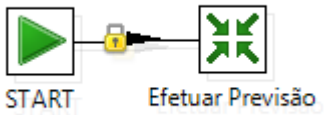


**Figura 47 - Pentaho Previsão *Job***

Depois de iniciado o *job* no Pentaho, a informação inserida previamente pelo utilizador é utilizada pelo modelo de DM criado anteriormente, representado na figura por "Weka Scoring", para gerar uma previsão que é guardada na base de dados "Previsão" (Figura 48).

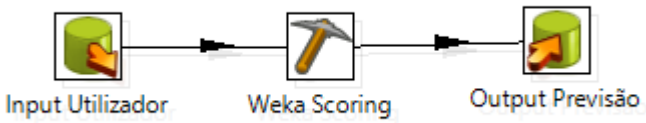

**Figura 48 - *Transformation* Pentaho**

A previsão gerada é depois enviada para a plataforma, ficando depois disponível para o utilizador do protótipo apresentado na Figura 49.


**Figura 49 - Página Inicial Protótipo**

Neste protótipo o utilizador começa por selecionar uma de três grupos de previsões, um grupo que engloba as previsões relacionadas com o resultado final, outro com o número de cantos e ainda outro com o número de golos. Clicando, por exemplo, no botão "número de golos" surgem dois novos botões, o "Mais ou Menos de 1,5 Golos e Mais ou Menos de 3,5 Golos. Se se clicar num deles surge o formulário que o utilizador necessita de preencher para passar a informação para o modelo de DM e submetendo esse formulário atualiza automaticamente a *div* seguinte com a previsão efetuada pelo protótipo como se pode ver na Figura 50

**Figura 50 - Preencher Formulário do Protótipo**

Se o utilizador quiser efetuar outra previsão apenas tem de clicar no botão "Carregar Aqui!" que o levará novamente para a página inicial.

De salientar que o protótipo permite efetuar previsão do R3S, R2SC, R2SF, 'C7,5', 'C8,5', 'G1,5', 'G3,5'.

## 4.5. Fase 5

A fase 5 resulta da combinação da fase "Monitorização" do processo de tomada de decisão e "Comunicação" do DSR.

A fase "Monitorização" não foi abordada no desenvolvimento desta dissertação. No entanto foram definidos alguns indicadores que facilitem a monitorização e avaliação do desemprenho do protótipo. Alguns indicadores que poderiam facilitar a monitorização e avaliação da performance do protótipo seriam:

- A Disponibilidade;
- A Velocidade de carregamento das páginas;
- A Velocidade de resposta das previsões;
- O Número de visitas;
- A Facilidade de utilização;
- Percentagem de previsões efetuadas corretamente;
- Lucro obtido;
- Investimento efetuado com base nas previsões.

Relativamente à comunicação foram publicados alguns artigos científicos (em anexo)

- "***Decision Support System for predicting Football Game result***"

Os outros dois artigos estão já aprovados e estão a aguardar a publicação, os seus títulos são:

- ***"Predicting 2-Way Football Results by Means of Data Mining"***
- ***"Real-Time Data Mining Models to Predict Football 2-Way Result"***

# 5 DISCUSSÃO DE RESULTADOS

A discussão de resultados pode ser dividida em dois subconjuntos de avaliação, os resultados obtidos através das métricas geradas pelos modelos de Data Mining (DM) e os resultados originados pelos testes feitos ao protótipo.

Para induzir os modelos, em todas as variáveis alvo foram utilizados dois métodos de amostragem, o *10-Folds Cross-Validation* (10FCV) e o *Holdout Simple* (HS). Foram utilizadas quatro técnicas de DM, o *Naive Bayes* (NB), o Support Vector Machine (SVM), as Árvores de Decisão (AD) e o *Lazy Learning* (LL).

As variáveis-alvo que tinham classes muito desequilibradas, por exemplo, a variável-alvo mais ou menos de 1,5 golos (G1,5) possui 74% de exemplos de mais de 1,5 golos, o que leva a que exista um desequilíbrio no modelo. Nestas situações foi utilizada a técnica de *oversampling* para equilibrar o conjunto de dados e os modelos que tinham os melhores valores nas métricas voltaram a ser induzidos.

## 5.1. Vitória Equipa Visitada, Empate ou Vitória Equipa Visitante

Inicialmente para esta variável-alvo foram induzidos 112 modelos de DM.

Esta variável-alvo tem três possibilidades de previsão, portanto, apenas as métricas de acuidade e a *Area Under Curve* (AUC) são consideradas.

A acuidade obtida nos modelos de DM não varia significativamente quando são comparadas as métricas geradas através dos dois métodos de amostragem utilizados, a acuidade varia cerca de 1% a 2%, sendo por vezes a acuidade superior quando o método de amostragem é o *10-Folds Cross-Validation* (10FCV) e noutras quando os modelos são gerados através do método do *Holdout Simple* (HS).

Em relação às técnicas de DM foram utlizadas os classificadores *Bayes*, que são aplicados através do algoritmo *NaiveBayes* (NB), a técnica *Suppor Vector Machine* (SVM), que é aplicada através do algoritmo *LibSVM*, a árvore de decisão aplicada pelo algoritmo J48 e por fim a técnica *Lazy Learning* (LL) que é aplicada pelo algoritmo *Kstar*. A acuidade obtida nos modelos induzidos pela técnica LL destacam-se pela negativa, obtendo como valor mais elevado apenas 50,18%, as técnicas NB e J48 que obtêm valores idênticos sendo o valor de acuidade mais elevado em ambos de 54%. Por último, a técnica que obtém melhores valores de acuidade é o SVM que obtém valores na ordem dos 57%.

Como este é um atributo-alvo no qual as classes a prever têm um número de exemplos desequilibrado no conjunto de dados, decidiu-se aplicar a técnica de *oversampling* no mesmo. Foram induzidos novamente os seis melhores modelos através das quatro técnicas de DM e dos dois métodos de amostragem, ou seja, mais 24 modelos. Agora, os modelos induzidos através das técnicas de DM SVM e NB não demonstraram ser úteis e os valores obtidos nas métricas foram idênticos, as duas restantes técnicas obtiveram valores relevantes, cumprindo todos os parâmetros de qualidade definidos. Destes, o LL destaca-se e obtém uma acuidade 76% e uma AUC de 93%, sendo então superior em 4% em relação às AD. Em relação aos métodos de amostragem o 10FCV obtém valores superiores nas métricas na ordem dos 3% em relação ao HS.

Estes valores são superiores aos casos identificados no tópico 2.4.1 deste documento, nos quais o valor de acuidade mais elevado era de 59% nas previsões do resultado final em que existe três possibilidades a prever.

Por fim, tendo em conta os parâmetros de qualidade definidos esta é uma variável-alvo que tem capacidade para ser implementada no protótipo.

## 5.2. Resultado a favor ou contra a equipa visitada

Foram induzidos para esta variável-alvo 112 modelos de DM. Existindo várias métricas que podem ser tidas em consideração, pois, ao contrário da variável-alvo discutida no tópico 5.1 esta tem apenas duas classes possíveis.

O método de amostragem 10FCV destaca-se em relação ao HS, obtendo valores superiores praticamente em todas as métricas de todos os modelos induzidos.

Quanto às técnicas de DM, o LL volta a destacar-se pela negativa obtendo valores inferiores em todas as métricas, em relação à acuidade, os modelos induzidos através da técnica SVM obtêm valores superiores, cerca de 70% de acuidade, os modelos criados por NB e AD obtêm valores idênticos mas quase sempre inferiores aos obtidos por SVM. Quanto à sensibilidade, o modelo que atinge uma sensibilidade maior foi induzido através da técnica de NB e o seu valor é cerca de 84%, mas é um modelo um pouco desequilibrado, pois, o valor da acuidade e da especificidade é inferior comparado com outros modelos, o mesmo acontece com a especificidade, o modelo que tem valores superiores é igualmente criado através de NB e é também este um modelo desequilibrado. Os modelos induzidos através da técnica SVM são os que têm valores mais equilibrados e ainda uma acuidade superior, e é a métrica considerada

mais importante para a avaliação destes modelos. A *Area Under Curve* (AUC) tem valores superiores quando induzida por NB mas não se destaca muito em relação aos que são induzidos por SVM, já nas outras duas técnicas obtém valores inferiores.

Os modelos induzidos para esta variável-alvo cumprem também os parâmetros de qualidade mínima definidos e, portanto, foram implementados no protótipo.

## 5.3. Resultado a favor ou contra a equipa visitante

Foram, inicialmente, induzidos para esta variável-alvo 122 modelos de DM. Esta variável-alvo tem duas possibilidades de previsão, sendo igualmente avaliados como a variável-alvo apresentada no tópico anterior desta dissertação.

Os métodos de amostragem utilizados não se destacam um do outro, sendo por vezes os valores das métricas superiores uns valores no método de amostragem 10FCV e noutros no HS, variando cerca de 2% no máximo.

Uma técnica de DM que se comprovou não ser eficaz a efetuar a previsão desta variável-alvo foi o SVM, foram obtidos valores de sensibilidade demasiado reduzidos, o que indica que o algoritmo não encontrou grandes padrões nos dados e sugere como previsão a classe da qual tem mais exemplos. A técnica LL tem os valores mais baixos obtidos em todas as métricas avaliadas, sendo também descartada. As técnicas NB e AD são as que têm os valores mais equilibrados, o modelo em que a acuidade tem um valor superior é induzido por uma AD e tem o valor de 76%. A principal diferença entre os resultados obtidos nas duas métricas é o valor da sensibilidade e da especificidade, nos modelos induzidos por AD os valores de especificidade obtêm valores de cerca de 90% enquanto no NB os valores encontram-se na ordem dos 70%, e a sensibilidade enquanto as AD obtém os melhores valores na ordem dos 35% os NB estão entre 40 a 50%, sendo então os modelos induzidos por AD os que têm os modelos mais equilibrados e ainda com uma acuidade superior.

Como o conjunto de dados apresentava 72% de exemplos em que o resultado é contra a equipa visitante, foi aplicada a técnica de *oversampling* no conjunto de dados para equilibrar as duas classes e foram novamente induzidos os seis melhores modelos através das quatro técnicas de DM, ou seja, mais 48 modelos no total. Os valores obtidos pelas técnicas SVM, NB e AD não sofreram alterações relevantes, já na situação do LL, que curiosamente era a técnica que tinhas os valores mais baixos obtidos anteriormente, a acuidade sobe 6% em relação aos modelos induzidos anteriormente. A sensibilidade

sobe de valores inferiores a 30% para valores na ordem dos 90%, tornando este um modelo mais equilibrado na relação sensibilidade/especificidade.

Os últimos modelos induzidos cumprem os requisitos de qualidade mínima e o modelo escolhido no tópico 4.3.3 deste documento foi implementado no protótipo.

## 5.4. Mais ou menos de 7,5 cantos

Foram induzidos 56 modelos de DM para prever se ocorrerão mais ou menos de 7,5 cantos em cada encontro.

Os modelos induzidos através da técnica de amostragem 10FCV obtêm valores superiores aos induzidos pelo HS.

As técnicas SVM e AD não encontram qualquer padrão nos dados, as técnicas NB e LL detetam alguns padrões mas os valores de especificidade e sensibilidade encontram-se muito desequilibrados, sendo a sensibilidade superior a 90% na maior parte das previsões e a especificidade muito reduzida. Por conseguinte, a AUC obtém valores muito próximos de 50%.

Com base nos primeiros resultado foi então decidido aplicar a técnica de *oversampling* no conjunto de dados. Os valores obtidos não variaram consideravelmente com os métodos de amostragem utilizados, já nas técnicas de DM utilizadas, o LL obtém melhores valores nas métricas sendo capaz de cumprir com todos os parâmetros de qualidade, a sensibilidade obtida no melhor modelo foi de 72%, a especificidade 95% a acuidade 85% e a AUC 0,9.

O modelo obtido é bastante equilibrado e cumpre todos os requisitos, portanto, foi implementado no protótipo.

## 5.5. Mais ou menos de 8,5 cantos

Para esta variável-alvo foram induzidos 56 modelos de DM.

A análise aos resultados obtidos através das métricas é semelhante à primeira análise efetuada no tópico 5.4 sendo esta uma previsão com pouco valor pois o modelo indica a maior parte das situações, como classe de saída, a classe da qual existem mais exemplos.

Foi então decidido utilizar, novamente, a técnica de *oversampling* nesta variável-alvo e tal como no tópico 5.4, os valores obtidos pela técnica LL são superiores relativamente a todas as outras técnicas e o método

de amostragem 10FCV obtém valores superiores relativamente ao HS. Os valores obtidos pelo melhor modelo foram 74% de acuidade, 70% de sensibilidade, 79% de especificidade e 0,84 de AUC.

Este modelo cumpre todos os parâmetros de qualidade definidos e é o que tem uma acuidade superior, por isso foi também implementado no protótipo.

## 5.6. Mais ou menos de 9,5 cantos

Foram induzidos 56 modelos de DM também para esta variável-alvo relacionada com o número de cantos que ocorre num determinado jogo de futebol.

Os valores obtidos nas métricas quando o modelo é induzido através do método de amostragem 10FCV são idênticos aos obtidos através do HS, não se destacando muito.

As técnicas SVM e AD não identificam qualquer padrão nos dados e são, portanto, descartados. O valor mais elevado de acuidade é obtido através da técnica NB e tem o valor de 64%, a sensibilidade é igualmente superior nestes modelos e obtém valores na ordem dos 90%. Já os modelos induzidos pela técnica LL têm uma especificidade maior mas uma sensibilidade e acuidade inferiores. Os valores da AUC são idênticos entre estas duas técnicas.

Como esta é uma variável-alvo que ainda tem um desequilíbrio no número de exemplos de cada classe existentes no conjunto, foi utilizada a técnica de *oversampling* para o equilibrar. Os métodos de amostragem não evidenciam diferenças relevantes nos valores das métricas. As técnicas de DM SVM e NB não sofreram alterações em relação aos modelos induzidos sem o *oversampling*. As AD e o LL obtêm valores melhores e o modelo que tem uma maior acuidade é de 71%.

Nenhum destes modelos obtém uma sensibilidade superior ao parâmetro mínimo definido, 65%, portanto, os modelos são descartados e não foram inseridos no protótipo.

## 5.7. Mais ou menos de 10,5 cantos

Para esta variável foram também criados 56 modelos de DM.

Não é possível identificar um método de amostragem que obtenha melhores valores, os resultados obtidos nas métricas têm valores idênticos em cada um dos métodos de amostragem utilizados, sendo por vezes superior num e inferior noutro, dependendo do cenário definido.

Quanto às técnicas de DM utilizadas o LL é a que obtém modelos com piores valores nas suas métricas. A acuidade e sensibilidade mais elevadas são de 62% e 81% respetivamente e são obtidas através da técnica SVM, apenas a especificidade é superior noutra técnica, a NB, e a apenas 3% superior, tendo como valor 52%. A AUC é equilibrada em todas as técnicas mas mais uma vez o SVM obtém os melhores valores 49%.

Nenhum dos modelos induzidos cumpre todos os requisitos mínimos de qualidade definidos, logo, esta variável-alvo foi descartada e consequentemente não foi introduzida no protótipo.

## 5.8. Mais ou menos de 1,5 golos

Para a variável-alvo mais ou menos de 1,5 golos foram, inicialmente, induzidos 64 modelos de DM.

Os métodos de amostragem não permitem fazer uma avaliação dos resultados, pois, cada método obtém melhores métricas num ou outro parâmetro consoante o cenário que esteja em questão.

As técnicas de DM SVM e LL não detetaram quaisquer padrão nos dados, portanto os modelos sugerem sempre a mesma classe de saída. Os modelos induzidos através das outras duas técnicas obtêm uma especificidade demasiado baixa sendo o valor mais elevado de 11% e apesar da acuidade ser até bastante razoável, cerca de 73%, o desequilíbrio existente entre a sensibilidade e especificidade leva a descartar todos os modelos até aqui induzidos para esta variável-alvo.

Foi então decidido utilizar a técnica de *oversampling* no conjunto de dados. As métricas obtidas nos modelos melhoram consideravelmente. O método de amostragem 10FCV destaca-se em relação ao HS. Em relação às técnicas de DM utilizadas, o SVM e o NB melhoram relativamente aos modelos anteriormente induzidos, mas não se aproximam dos valores obtidos através das técnicas AD e LL, sendo entre estas duas o LL que obtém valores superiores. O modelo induzido com uma maior taxa de acuidade obtém 80% nessa métrica, 89% de especificidade, 70% de sensibilidade e 0,91 de AUC.

Este modelo foi aprovado, pois, cumpre todos os requisitos de qualidade definidos e foi, portanto, implementado no protótipo.

## 5.9. Mais ou menos de 2,5 golos

Foram induzidos 64 modelos desta variável-alvo.

Os modelos obtêm resultados idênticos independentemente do método de amostragem utilizado na indução dos modelos.

O valor de acuidade mais elevado é obtido a partir de um modelo induzido com a técnica SVM que obtém valores de 63% e a AUC que tem o mesmo valor, a especificidade é também superior nesta técnica e obtém como valor mais elevado 69%. Apenas a sensibilidade obtém valores superiores quando utilizadas

AD e tem como valor máximo 69%. Portanto o SVM nesta previsão é uma técnica que obtém valores de melhor qualidade em relação a todas as outras técnicas.

Os modelos porém, não atingiram os parâmetros de qualidade definidos e foram portanto descartados, não sendo inseridos no protótipo.

## 5.10. Mais ou menos de 3,5 golos

Para esta variável-alvo foram também induzidos 64 modelos de DM.

Os modelos criados através dos dois métodos de amostragem apresentam valores semelhantes nas métricas avaliadas.

As técnicas de SVM e AD foram descartadas, pois, não encontram qualquer padrão nos dados. Os classificadores criados pela técnica LL obtiveram piores valores em todas as métricas em relação aos criados pelos NB. A acuidade registada de maior valor é de 71,61%, a sensibilidade é de 97,51%, de especificidade de 30,25% e a AUC 62%. A técnica NB é portanto aquela com a qual se encontram melhores modelos para esta variável-alvo.

Não foram induzidos modelos que cumprissem todos os requisitos mínimos de qualidade. Como esta variável-alvo contem um número de exemplo de cada classe no conjunto de dados desequilibrados, em 72% das situações o número de golos existente em cada jogo é inferior a 3,5 golos e apenas em 28% dos exemplos é superior decidiu-se utilizar a técnica de *oversampling* no conjunto de dados.

Os modelos induzidos depois de aplicada esta técnica obtiveram valores idênticos quando aplicados os dois métodos de amostragem, mas no caso das técnicas de DM utilizadas, apesar de em todas os valores obtidos serem mais equilibrados dos que existiam previamente, a técnica LL destaca-se das outras e o modelo escolhido, obteve 81% de acuidade, nas restantes métricas obteve 71% na sensibilidade, 90 na especificidade e 0,91 de AUC.

As métricas obtidas no modelo cumpre todos os requisitos mínimos de qualidade definidos para os modelos e foi então aprovado e implementado no protótipo.

## 5.11. Testes Protótipo

Foram efetuados testes ao protótipo através da simulação de 5 jornadas da época 2014/15 da Primeira liga Inglesa. As variáveis-alvo utilizadas são aquelas que, no tópico 4.3 desta dissertação, tinham um

modelo que cumpriu todos os parâmetros de qualidade definidos para cada atributo-alvo. Como existe registo do valor das odds relativas ao resultado final de cada jogo é possível determinar os lucros/prejuízos que seriam obtidos nas variáveis-alvo R3S, R2SC e R2SF. Na tabela 38 estão apresentados os lucros/prejuízos obtidos em cada jornada simulada e a respetiva taxa de acerto do protótipo e Anexo II – Testes Protótipo estão apresentados os resultados obtidos em cada jogo mais detalhadamente.

**Tabela 38 - Testes Protótipo I**

| Jornada | Medidas Performance | R3S | R2SC | R2SF |
|---|---|---|---|---|
| **1** | Lucro Jornada | 318 | 108 | -211 |
| | Taxa Acerto | 60% | 80% | 60% |
| **6** | Lucro Jornada | 266 | 77 | 95 |
| | Taxa Acerto | 70% | 70% | 90% |
| **11** | Lucro Jornada | 1034 | 295 | 455 |
| | Taxa Acerto | 70% | 70% | 80% |
| **16** | Lucro Jornada | 116 | 147 | 197 |
| | Taxa Acerto | 60% | 80% | 90% |
| **21** | Lucro Jornada | 180 | -114 | 90 |
| | Taxa Acerto | 50% | 60% | 80% |
| **Média Lucro Jornada** | | 1914 | 513 | 626 |
| **Taxa de Lucro** | | 38,28% | 10,26% | 12,52% |
| **Média Taxa Acerto** | | 62% | 72% | 80% |

Comparativamente ao sistema desenvolvido na Unidade Curricular (UC) Sistemas de Apoio à Decisão, apresentado no tópico 3 deste documento, a taxa de lucro obtida pelo novo protótipo atinge uma taxa de lucro superior na variável-alvo que têm em comum, sobe de 20,13% para 38,28% o que leva a obter um lucro superior.

Para as restantes variáveis-alvo implementadas no protótipo não existe registo das *odds* o que impossibilita o cálculo do lucro/prejuízos que seriam obtidos na simulação destas jornadas, portanto na Tabela 39 é apenas apresentada a taxa de acerto de cada uma delas.

**Tabela 39 - Teste Protótipo II**

| Jornada | C7,5 | C8,5 | G1,5 | G3,5 |
|---|---|---|---|---|
| **1** | 80% | 60% | 80% | 70% |
| **6** | 70% | 50% | 70% | 80% |
| **11** | 80% | 60% | 60% | 60% |
| **16** | 80% | 70% | 70% | 70% |
| **21** | 70% | 80% | 60% | 60% |
| **Média Taxa Acerto** | 76,00% | 64,00% | 68,00% | 68,00% |

# 6 CONCLUSÃO

Neste último capítulo da dissertação são apresentadas as conclusões gerais desta dissertação. Este capítulo encontra-se dividido em dois diferentes subtópicos. No primeiro é apresentada uma síntese do trabalho prático efetuado, todo o trabalho que levou até à construção do protótipo que tinha como objetivo indicar aos apostadores qual seria a melhor aposta a realizar para cada uma das variáveis-alvo analisadas. Neste mesmo subtópico são ainda apresentadas as contribuições científicas que este projeto originou, bem como uma breve análise comparativa aos objetivos propostos com os objetivos alcançados. No segundo subtópico são apresentadas as sugestões de um possível trabalho futuro, com vista a melhorar a performance do protótipo.

## 6.1. Síntese e Contribuições Cientificas

Como referido ao longo deste projeto, obter lucros a médio/longo prazo em apostas de jogos de futebol não se tem verificado uma tarefa fácil, sendo por norma prejudicial para o utilizador e benéfico para as casas de apostas, daí se verificar um aumento destas nos últimos anos, pois, este tem-se revelado um negócio rentável para as casas de apostas. O desafio proposto nesta dissertação foi o de contrariar esta tendência e levar a que os apostadores obtenham lucros, correndo ainda menos riscos.

O trabalho realizado levou, então, à criação de um protótipo de um sistema inteligente de apoio à decisão em apostas de jogos de futebol. Inicialmente foi necessário efetuar uma recolha de dados, dados estes relacionados com um determinado jogo de futebol e ainda de eventos dos quais os intervenientes do jogo não têm qualquer influência, como a precipitação. Foi necessário efetuar o tratamento dos dados, de modo a garantir a sua qualidade. Através desses dados foram induzidos modelos de Data Mining (DM) utilizando quatro técnicas distintas e analisados os seus resultados. Foi então criado o protótipo que aplicava estes modelos de DM e foram guardados os resultados obtidos nos testes efetuados nele próprio. Este trabalho levou ao estudo de dez distintos atributos-alvo. O objetivo era prever, com a maior fiabilidade possível, a ocorrência de determinados eventos durante um jogo de futebol e ainda qual seria o resultado final do mesmo. Foram definidos parâmetros de qualidade para cada uma destas previsões e para 5 delas este processo revelou valores interessantes.

O primeiro estudo envolvia o estudo da previsão do resultado em jogos de futebol, tendo três classes como possíveis saídas, a vitória da equipa visitada, o empate e a vitória da equipa visitante. Este foi o seguimento do trabalho iniciado na Unidade Curricular Sistemas de Apoio à Decisão e foi possível verificar uma melhoria dos modelos em relação ao mesmo neste projeto de dissertação. As taxas de acuidade subiram de 28%, para 76%, o que permitirá apresentar mais lucros do que os apresentados no trabalho anterior. Estes valores são também superiores aos registados no tópico 2.4.1, no qual os trabalhos que tinham o mesmo objetivo de previsão que esta variável-alvo obtiveram valores máximos de acuidade de 64%.

O segundo estudo focou-se também na previsão do resultado do jogo de futebol, mas para esta variável-alvo apenas foram consideradas duas saídas, a vitória da equipa visitada ou o conjunto das duas outras duas possibilidades, o empate e a vitória da equipa visitante. Os modelos induzidos obtêm valores de acuidade de 70%, para esta variável foram identificados modelos bastante equilibrados que apresentam valores interessantes. A um nível de contribuição cientifica estes modelos suportam os utilizadores de casas de apostas em escolher qual a aposta que devem realizar de modo a correr menos riscos.

O terceiro estudo é idêntico ao segundo, invertendo apenas as classes alvo, a primeira é vitória da equipa visitante e a segunda empate ou vitória da equipa visitada. Os modelos induzidos para esta variável obtiveram valores superiores aos do segundo estudo, atingindo níveis de acuidade de 81%. São também modelos de qualidade para serem utilizados na previsão do resultado e que ajudarão os apostadores a aumentar os seus lucros.

O quarto estudo engloba a previsão do número de cantos, mais especificamente se existem mais ou menos de 7,5 cantos em cada jogo. Esta previsão inicialmente não tinha grande qualidade devido ao desequilíbrio existente no conjunto de dados dos exemplos de cada classe, a classe de saída menos de 7,5 cantos contem mais exemplos em relação à saída mais de 7,5 cantos o que leva a um desequilíbrio nas previsões. Depois da aplicação da técnica *oversampling* os modelos passaram a ser mais equilibrados e obtiveram valores de acuidade de 81%.

O quinto estudo pretendia também efetuar a previsão do número de cantos, neste caso a variável-alvo é mais ou menos de 8,5 cantos e tal como apresentado no estudo anterior a previsão inicial obtida nestes modelos não atingia os valores dos parâmetros de qualidade definidos, portanto, a previsão que poderia efetuar podia não ser a mais adequada. Foi então utilizada a técnica de *oversampling* no conjunto de dados e foram obtidos modelos com uma taxa de acuidade de 74%, as outras métricas atingiram também

os valores mínimos dos parâmetros de qualidade definidos sendo então aprovada a previsão para esta variável-alvo.

O sexto estudo continua focado no número de cantos existente em cada jogo, nesta variável-alvo o objetivo é prever se num determinado jogo existem mais ou menos de 9,5 cantos. Foi um estudo de previsão que cumpriu todos os requisitos definidos, obtendo valores de acuidade de cerca de 64% e valores de sensibilidade de 90% e especificidade superior a 20%. Decidiu-se também para esta variável-alvo efetuar um *oversampling* ao conjunto de dados, os valores das métricas obtidos por estes modelos subiram consideravelmente mas a sensibilidade não atinge o valor mínimo dos parâmetro de qualidade definidos, por isso, esta previsão não foi implementada no protótipo.

O estudo que se seguiu foi também relacionado com o número de cantos em cada jogo, neste caso a previsão tinha como objetivo prever se existiram mais ou menos de 10,5 cantos no jogo e tal como no estudo anterior, não foram cumpridos os requisitos de qualidade existente.

O oitavo estudo efetuado está relacionado com o número de golos, a previsão aqui realizada diz respeito ao número de golos existente num determinado jogo de futebol, especificamente o objetivo era prever se existem mais ou menos de 1,5 golos. Os modelos de DM criados, inicialmente, não cumpriram os parâmetros de qualidade pois, tinham uma relação entre a sensibilidade e especificidade demasiado desproporcional, o que levava a que o modelo sugerisse quase sempre a mesma classe de saída, acima de 1,5 golos por jogo. Portanto, foi efetuado um *oversampling* ao conjunto de dados para equilibrar o numero de exemplos existente em cada uma das classes a ser previstas. Depois de aplicado o modelo com maior acuidade, 80%, cumpriu todos os parâmetros de qualidade definidos. Sendo esta previsão implementada no protótipo.

O nono conjunto de modelos induzidos tinha como objetivo prever se existiam mais ou menos de 2,5 golos em cada jogo. Os modelos de DM induzidos atingem taxas de acuidade um pouco reduzidas para este género de previsão, o modelo com melhor valor atinge 63%, não cumprindo assim os parâmetros de qualidade definidos, sendo, então, descartados.

O último estudo efetuado neste projeto tinha também como intuito prever o número de golos existentes num jogo de futebol, se existiram mais ou menos de 3,5 golos. Os modelos inicialmente induzidos obtiveram uma taxa de acuidade interessante, 71%, mas o desequilíbrio existente entre a sensibilidade e especificidade levavam a que a classe prevista fosse na maioria das situações menos de 3,5 golos. Neste caso decidiu-se aplicar um *oversampling* ao conjunto de dados para equilibrar o número de exemplos existentes de cada classe, a acuidade obtida depois de aplicada esta técnica subiu para 81% e todas as

métricas cumprem os requisitos de qualidade definidos, portanto, esta previsão foi implementada no protótipo.

Por fim, foi criado um protótipo de um sistema inteligente de apoio à decisão que tem como intuito suportar os apostadores de jogos de futebol a definir qual é a aposta que devem realizar, neste protótipo foram implementados os modelos de DM que cumpriram os parâmetros de qualidade definidos, para receber uma sugestão o utilizador apenas tem de preencher um simples formulário que, depois de comunicar com os modelos imprime a resposta.

A utilização deste protótipo pode, como os testes indicam, suportar os apostadores de uma forma sustentada a obter lucros em médio longo/prazo. As três variáveis-alvo das quais existe um registo das *odds*, R3S, R2SC e R2SF permitiram a obtenção de uma taxa de lucro de 32,28%, 10,26% e 12,56%, respetivamente. Caso o apostador investisse 100 € em cada uma dessas três apostas, nas 5 jornadas simuladas, investia um total de 15000 € e obteria um lucro de 3053 €, uma taxa de lucro de cerca de 21%. Este é um valor que levaria a um retorno considerável comparativamente com os instrumentos de investimento existentes no mercado financeiro.

## 6.2. Trabalho Futuro

Para trabalho futuro, de modo a continuar o trabalho de investigação iniciado e no âmbito de suportar os apostadores a aumentar os seus lucros nas apostas em jogos de futebol são sugeridas as seguintes orientações:

- Utilizar diferentes técnicas de DM, outros artigos, efetuar testes mais exaustivos para identificar quais os algoritmos que permitem retirar melhores resultados dos modelos, bem como definir novos cenários para a criação dos mesmos;
- Estudar fatores extra futebol a serem introduzidos nos dados:
    - No momento da realização dos testes ao protótipo foi possível verificar que existem duas equipas que a previsão efetuada não fazia muito sentido, equipas estas que receberam injeção de capital apenas em anos mais recentes, o que lhes permitiu atingir níveis desportivos contrários aos que os modelos esperariam, o principal caso é do *Manchester City*, mas também nos jogos em que o Leicester participa isso se nota. Estas eram equipas que tinham resultados medíocres e que nos últimos anos o investimento realizado nestas equipas permitiu subir os índices de qualidade da equipa;

    - Verificar em que fase da época determinado jogo ocorre, isto porque por vezes existem equipas que já têm os seus objetivos assegurados e fazem uma gestão alargada do plantel;
    - Verificar se os jogadores chave das equipas estão disponíveis para cada jogo.
    - Verificar se existiu uma troca de treinadores na semana que antecede o jogo, pois, este é por norma um fator motivacional para todas as equipas que sofre estas "chicotadas psicológicas";
- Uma visão mais alargada do que poderia levar o sistema a obter melhores resultados seria estudar cada elemento de cada equipa, criando um índice de performance para cada jogador e estando a atualizar semanalmente esses dados de modo a terem influência na previsão do resultado;
- Seria também interessante tentar que os modelos se otimizassem automaticamente, que fossem aprendendo com os *inputs* que os utilizadores lhes fossem introduzindo.

Como nota final, algumas das variáveis acima mencionadas como passíveis de serem incluídas em futuros modelos não foram incorporados nos modelos atuais devido à falta de informação das mesmas. A solução passa por uma exploração exaustiva de todas as fontes de informação disponíveis relacionadas com os jogos a fim de criar uma base de dados com informações extras.

# BIBLIOGRAFIA


Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Rudiger, W. (2000). Crisp-Dm 1.0. *CRISP-DM Consortium*, 76.

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. doi:10.1111/j.1747-0285.2009.00840.x

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). Knowledge Discovery and Data Mining : Towards a Unifying Framework. *Kdd*. doi:10.1.1.27.363

Gama, J., Carvalho, A., Faceli, K., Lorena, A. C., & Oliveira, M. (2012). *Extração de Conhecimento de Dados - Data Mining* (1st ed.). Edições Sílabo.

Garner, S. R. (1995). WEKA: The Waikato Environment for Knowledge Analysis. *Proceedings of the New Zealand Computer Science*, 57–64.

Genómica Funcional e Bioinformática. (2012). Retrieved October 10, 2015, from http://web.tecnico.ulisboa.pt/ana.freitas/bioinformatics.ath.cx/bioinformatics.ath.cx/indexf23d.html?id

Gorunescu, F. (2011). *Data mining concepts, models and techniques*. Springer - Verlag Berlin Heidelberg.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. *MIS Quarterly*, *28*(1), 75–105. doi:10.2307/25148625

Hucaljuk, J., & Rakipovic, A. (2011). Predicting football scores using machine learning techniques. *2011 Proceedings of the 34th International Convention MIPRO*, *48*, 1623–1627.

Ibm. (2011). IBM SPSS modeler CRISP-DM guide.

J. Cios, K., Pedrycz, W., W. Swiniarski, R., & A. Kurgan, L. (2007). *Data Mining - A Knowledge Discovery Approach* (1st ed.). Springer US. doi:10.1007/978-0-387-36795-8

Joseph, a., Fenton, N. E., & Neil, M. (2006). Predicting football results using Bayesian nets and other machine learning techniques. *Knowledge-Based Systems*, *19*(7), 544–553. doi:10.1016/j.knosys.2006.04.011

Langley, P., Iba, W., & Thompson, K. (1992). An Analysis of Bayesian Classifiers. *Research Gate*, 15.

Maimon, Oded; Rokach, L. (2010). *Data Mining and Knowledge Discovery Handbook* (2nd ed.). doi:10.1007/978-0-387-09823-4

Nemati, H. R., Steiger, D. M., Iyer, L. S., & Herschel, R. T. (2002). Knowledge warehouse: An architectural integration of knowledge management, decision support, artificial intelligence and data warehousing. *Decision Support Systems*, *33*, 143–161. doi:10.1016/S0167-9236(01)00141-5

Nunes, S., & Sousa, M. (2006). Applying data mining techniques to football data from European championships. *Actas Da 1ª Conferência de Metodologias de Investigação Científica (CoMIC06)*, (December 2005). Retrieved from http://repositorio-aberto.up.pt/handle/10216/282

Olson, D. L., & Delen, D. (2008). *Advanced Data Mining Techniques*. Springer - Verlag Berlin Heidelberg.

Owramipur, F., Eskandarian, P., & Mozneb, F. S. (2013). Football Result Prediction with Bayesian

Network in Spanish League-Barcelona Team. *International Journal of Computer Theory and Engineering*, *5*(5), 812–815. doi:10.7763/IJCTE.2013.V5.802

Peffers, K., Tuunanen, T., Gengler, C. E., Rossi, M., Hui, W., Virtanen, V., & Bragge, J. (2006). The Design Science Research Process: A Model for Producing and Presenting Information Systems Research. *The Proceedings of Design Research in Information Systems and Technology DESRIST'06*, *24*, 83–106. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:The+Design+Science+Research+Process:+A+Model+for+Producing+and+presenting+Information+Systems+Research#0

Portela, F., Gago, P., Santos, M. F., Machado, J., Abelha, A., Silva, Á., & Rua, F. (2013). Implementing a pervasive real-time intelligent system for tracking critical events with intensive care patients. *International Journal of Healthcare Information Systems and Informatics*, *8*(4), 1–16. doi:10.4018/ijhisi.2013100101

Portela, F., Santos, M. F., Gago, P., Silva, Á., Rua, F., Abelha, A., ... Neves, J. (2011). *Enabling real-time intelligent decision support in intensive care*. *1. Filipe Portela, Manuel Santos, José Machado, António Abelha, and, Á.S., Rua, F.: Real-Time Decision Support in Intensive Medicine - An intelligent approach for monitoring Data Quality. International Journal of Medical and Bioengineering I, (2013) 2. Po*.

Rajaraman, A., & Ullman, J. D. (2011). Mining of Massive Datasets. *Lecture Notes for Stanford CS345A Web Mining*, *67*, 328. doi:10.1017/CBO9781139058452

Rotshtein, A. P., Posner, M., Rakityanskaya, A. B., Lev, M., & National, V. (2005). Football predictions based on a fuzzy model with genetic and neural tuning. *Cybernetics and Systems Analysis*, *41*(4), 619–630. doi:10.1007/s10559-005-0098-4

Santos, M. F. dos, & Azevedo, C. (2005). *Data Mining: Descoberta do Conhecimenta em Bases de Dados*. (FCA - Editora de Informática, Ed.).

Santos, M. Y., & Ramos, I. (2009). *Business Intelligence*. FCA - Editora de Informática.

Saravanan, N., & Ramachandran, K. I. (2009). Fault diagnosis of spur bevel gear box using discrete wavelet features and Decision Tree classification. *Expert Systems with Applications*, *36*(5), 9564–9573. doi:10.1016/j.eswa.2008.07.089

Sauter, V. L. (2011). *Decision Support Systems for Business Intelligence*. doi:10.1002/9780470634431

Shim, J. P., Warkentin, M., Courtney, J. F., Power, D. J., Sharda, R., & Carlsson, C. (2002). Past, present, and future of decision support technology. *Decision Support Systems*, *33*, 111–126. doi:10.1016/S0167-9236(01)00139-7

Simon, H. A. (1960). *The New Science of Management Decision*.

Simon, H. a. (1977). *The new science of management*.

Suzuki, a K., Salasar, L. E. B., Leite, J. G., & Louzada-Neto, F. (2010). A Bayesian approach for predicting match outcomes: The 2006 (Association) Football World Cup. *Journal of the Operational Research Society*, *61*(October 2015), 1530–1539. doi:10.1057/jors.2009.127

T. Larose, D., & Larose, C. D. (2014). *Discovering Knowledge in Data: An Introduction to Data Mining*

(2nd ed.). John Wiley & Sons, Inc. doi:10.1002/9781118874059

Tsakonas, a, & Dounias, G. (2002). Soft computing-based result prediction of football games. *The First International ...*, *3*(May), 15–21. Retrieved from http://www.researchgate.net/publication/2560104_Soft_Computing-Based_Result_Prediction_of_Football_Games/file/79e41509b9947b0861.pdf

Turban, E. (2010). *Decision Support and Business Intelligence* (Vol. 1968). Retrieved from http://prospero.murdoch.edu.au/search~S10?/rICT208/rict208/1,1,1,B/frameset~1838749&FF=rict208&1,1,

Turban, E., Sharda, R., & Aronson, J. (2008). Business intelligence: a managerial approach. *Tamu-Commerce.Edu*. doi:10.1109/HICSS.2012.138

Turban, E., Sharda, R., & Delen, D. (2011). *Decision Support and Business Intelligence Systems* (9th ed.). Prentice Hall.

Ulmer, B., & Fernandez, M. (2013). Predicting Soccer Match Results in the English Premier League, 5.

Vaishnavi, V., & Jr, W. K. (2007). *Design science research methods and patterns: innovating information and communication technology*. *Vasa*. Retrieved from http://medcontent.metapress.com/index/A65RM03P4874243N.pdf\nhttp://books.google.com/books?hl=en&lr=&id=sI2Y9Jh8tq8C&oi=fnd&pg=PP1&dq=design+science+research+methods+and+patterns+innovating+information+and+communication+technology&ots=k6bC04QyYN&sig=QX4ID6

Vercellis, C. (2009). *Business Intelligence: Data Mining and Optimization for Decision Making*. *Business Intelligence: Data Mining and Optimization for Decision Making*. doi:10.1002/9780470753866

W. Aha, D. (1997). *Lazy Learning*. Springer - Science + Business Media , B.V.

Witten, I. H., Frank, E., & Hall, M. a. (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (3rd ed.).

Zadeh, L. a. (1965). Fuzzy sets. *Information and Control*, *8*(3), 338–353. doi:10.1016/S0019-9958(65)90241-X

# ANEXO I – PUBLICAÇÕES CIENTÍFICAS

## DECISION SUPPORT SYSTEM FOR PREDICTING FOOTBALL GAME RESULT

**Autores:** João Gomes, Filipe Portela, Manuel Filipe Santos

**Conferência.** International Conference on Circuits, Systems, Communications and Computers (CSCC 2015)

**Jornal/Editora.** Computers, INASE

**Ano:** 2015

**Estado:** Publicado

**Abstract. T**here is an increase of bookmaker's number over the last decade, leading to the conclusion that the bet houses have obtained profitability in the detriment of its users. Based in this principle arises an opportunity to explore a set of artificial intelligence techniques in order to support the user betting decision. The development of this project aims to support bookmaker's users to increase their profits on bets related to football matches, suggesting to them which bet that they should carry out (home win, draw or away win). To this, it was collected several statistical information related to football games from the Premier League. It was developed a dataset and applied data mining techniques to create a model with good predictive capability. This model was then integrated in a decision support system which allows complement the machine intelligence with human perception. The model developed allowed to have profits of 20% in relation to an initial bankroll.

**Keywords.** Decision Support Systems, Data Mining, Football Games Prediction, Decision Support Systems for Football Betting, Knowledge Discovery in Database, Football Bets

## PREDICTING 2-WAY FOOTBALL RESULTS BY MEANS OF DATA MINING

**Autores:** João Gomes, Filipe Portela, Manuel Filipe Santos, António Abelha, José Machado

**Conferência.** European Simulation and Modelling Conference (ESM'2015)

**Ano:** 2015

**Estado:** Aguarda Publicação

**Abstract.** In the last decade, has been found an increase in the number of bookmakers, particularly in the online market (ebusiness). It is possible deducing that this activity is profitable for them and consequently damaging to their users. Nowadays, football is considered one of the most popular sports. Regarding the betting world it was acquired an outstanding position, which moves millions of euros during the period of a single football match. The

lack of profitability of football betting website users has been stressed as a problem. In accordance with the stated arises here, an opportunity to explore. This lack gave origin to this research proposal, which is going to address the possibility of existing a way to support the users on their online bets, in order to improve their results and profitability. A football match could be analysed from the perspective of several types of statistical data, which do not have a direct influence on the final match result. This research work has the aim of helping to improve the performance of online football bets, by providing users statistical data that may be important to take into account, at the time of doing their own bets. In this work it was possible introduce data mining models which are able to predict 2-way results (home team win / draw or visitor team win) with 96,2 % of sensitivity and a good level of accuracy (74.8%). These models are prepared to be the base of an Intelligent System.

**Keywords.** Decision Support Systems, Data Mining, Football Games Prediction, Decision Support Systems for Football Betting, Knowledge Discovery in Database, 2-way result, Football Bets, eBusiness, Intelligent Systems

# REAL-TIME DATA MINING MODELS TO PREDICT FOOTBALL 2-WAY RESULT

**Autores:** João Gomes, Filipe Portela, Manuel Santos

**Conferência.** Advancement on Information Technology International Conference (ADVCIT 2015)

**Jornal/Editora:** Jurnal Teknologi, Penerbit UTM Press

**Ano:** 2015

**Estado:** Aguarda Publicação

**Abstract.** Nowadays, football is considered one of the most popular sports. Regarding the betting world it has acquired an outstanding position which moves millions of euros during the period of a single football match. The lack of profitability of football betting users has been stressed as a problem. This lack gave origin to this research proposal, which it is going to analyze the possibility of existing a way to support the users on their online bets, in order to improve their results and profitability. Data mining models able to support the gamblers were induced in order to increase their profits in the medium/long term. Being conscience that the model can fail, sometimes, the results achieved by the models are encouraging and suggest that the system can help to increase the profits. The target attribute contains only two classes, "0" - victory of the home team or draw and "1" - away team win. The models are prepared to be induced in real-time. The entire process is executed automatically using online-learning. In terms of results the models achieved an accuracy upper than 75%. The results also show that the models are very good to predict class 0 with a specificity upper than 90%.

**Keywords.** Decision Support Systems, Data Mining, Football Games Prediction, Decision Support Systems for Football Betting, Knowledge Discovery in Database, Soccer Bets, Football Bets

# ANEXO II – TESTES PROTÓTIPO

| Jornada | Equiva Visitada | Equipa Visitante | Resultado Final | R3S | R2SC | R2SF |
|---|---|---|---|---|---|---|
| 1 | Burnley | Chelsea | 1 : 3 | 138 | 107 | 138 |
| | Newcastle | Man City | 0 : 2 | 157 | 113 | 157 |
| | Liverpool | Southampton | 2 : 1 | 137 | 137 | 107 |
| | Arsenal | Crystal Palace | 2 : 1 | 126 | 126 | 103 |
| | Leicester | Everton | 2 : 2 | 330 | 136 | 159 |
| | QPR | Hull | 0 : 1 | -100 | -100 | -100 |
| | Stoke | Aston Villa | 0 : 1 | 430 | 185 | -100 |
| | West Brom | Sunderland | 2 : 2 | -100 | 177 | 125 |
| | West Ham | Tottenham | 0 : 1 | -100 | 127 | -100 |
| | Man Utd | Swansea | 1 : 2 | -100 | -100 | -100 |
| 6 | Stoke | Newcastle | 1 : 0 | 115 | 115 | 28 |
| | West Brom | Burnley | 4 : 0 | 80 | -100 | 16 |
| | Arsenal | Tottenham | 1 : 1 | -100 | 106 | 16 |
| | Chelsea | Aston Villa | 3 : 0 | 20 | 20 | 2 |
| | Crystal Palace | Leicester | 2 : 0 | -100 | -100 | 44 |
| | Hull | Man City | 2 : 4 | 49 | 11 | -100 |
| | Man Utd | West Ham | 2 : 1 | 29 | 29 | 6 |
| | Southampton | QPR | 2 : 1 | 53 | 53 | 11 |
| | Sunderland | Swansea | 0 : 0 | 220 | 43 | 49 |
| | Liverpool | Everton | 1 : 1 | -100 | -100 | 23 |
| 11 | Swansea | Arsenal | 2 : 1 | -100 | 300 | 85 |
| | Sunderland | Everton | 1 : 1 | 230 | -100 | 78 |
| | Tottenham | Stoke | 1 : 2 | -100 | 105 | -100 |
| | West Brom | Newcastle | 0 : 2 | 255 | 71 | 255 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | QPR | Man City | 2 : 2 | 355 | 12 | 164 |
| | Burnley | Hull | 1 : 0 | 155 | -100 | 43 |
| | Man Utd | Crystal Palace | 1 : 0 | 31 | 31 | 5 |
| | Southampton | Leicester | 2 : 0 | 43 | 43 | 9 |
| | West Ham | Aston Villa | 0 : 0 | 265 | -100 | 16 |
| | Liverpool | Chelsea | 1 : 2 | -100 | 33 | -100 |
| **16** | Everton | QPR | 3 : 1 | -100 | 41 | 9 |
| | Swansea | Tottenham | 1 : 2 | -100 | 59 | -100 |
| | Man Utd | Liverpool | 3 : 0 | 83 | 83 | 21 |
| | Arsenal | Newcastle | 4 : 1 | 40 | 40 | 9 |
| | Burnley | Southampton | 1 : 0 | -100 | -100 | 100 |
| | Chelsea | Hull | 2 : 0 | 15 | 15 | 1 |
| | Crystal Palace | Stoke | 1 : 1 | -100 | 55 | 38 |
| | Leicester | Man City | 0 : 1 | 43 | 9 | 43 |
| | Sunderland | West Ham | 1 : 1 | 220 | 45 | 48 |
| | West Brom | Aston Villa | 1 : 0 | 115 | -100 | 28 |
| **21** | Man Utd | Southampton | 0 : 1 | -100 | 103 | -100 |
| | Arsenal | Stoke | 3 : 0 | 47 | 47 | 10 |
| | Crystal Palace | Tottenham | 2 : 1 | -100 | -100 | 66 |
| | Burnley | QPR | 2 : 1 | -100 | -100 | 29 |
| | Chelsea | Newcastle | 2 : 0 | 18 | 18 | 1 |
| | Everton | Man City | 1 : 1 | 275 | 22 | 101 |
| | Leicester | Aston Villa | 1 : 0 | -100 | -100 | 29 |
| | Swansea | West Ham | 1 : 1 | 235 | 70 | 31 |
| | West Brom | Hull | 1 : 0 | -100 | -100 | 23 |
| | Sunderland | Liverpool | 0 : 1 | 105 | 26 | -100 |
| **Total Lucro** | | | | 1914 | 513 | 626 |

| | | | |
|---|---|---|---|
| **Taxa de Acerto** | 62% | 72% | 80% |