UNIVERSIDADE FEDERAL DE SERGIPE
CAMPUS ALBERTO CARVALHO
DEPARTAMENTO DE SISTEMAS DE INFORMAÇÃO

RAFAEL MENESES SANTOS

ANÁLISE DA VIABILIDADE DA MINERAÇÃO DE ESTATÍSTICAS PÚBLICAS DO CAMPEONATO BRASILEIRO DE FUTEBOL

ITABAIANA
2014

UNIVERSIDADE FEDERAL DE SERGIPE
CAMPUS ALBERTO CARVALHO
DEPARTAMENTO DE SISTEMAS DE INFORMAÇÃO

RAFAEL MENESES SANTOS

ANÁLISE DA VIABILIDADE DA MINERAÇÃO DE ESTATÍSTICAS PÚBLICAS DO CAMPEONATO BRASILEIRO DE FUTEBOL

Trabalho de Conclusão de Curso submetido ao Departamento de Sistemas de Informação da Universidade Federal de Sergipe, como requisito parcial para a obtenção do título de Bacharel em Sistemas de Informação.

Orientador: Dr. METHANIAS COLAÇO RODRIGUES JÚNIOR
Coorientador: Msc. ANDRÉ VINICIUS RODRIGUES PASSOS NASCIMENTO

ITABAIANA
2014

Meneses Santos, Rafael.

Análise da Viabilidade da Mineração de Estatísticas Públicas do Campeonato Brasileiro de Futebol / Rafael Meneses Santos – Itabaiana: UFS, 2014. 56 f.

Trabalho de Conclusão de Curso em Bacharel em Sistemas de Informação – Universidade Federal de Sergipe, Curso de Sistemas de Informação, 2014.

1. Mineração de Dados. 2. Inteligência Artificial. 3. Sistemas de Informação. I. Análise da Viabilidade da Mineração de Estatísticas Públicas do Campeonato Brasileiro de Futebol.

RAFAEL MENESES SANTOS

ANÁLISE DA VIABILIDADE DA MINERAÇÃO DE ESTATÍSTICAS PÚBLICAS DO CAMPEONATO BRASILEIRO DE FUTEBOL

Trabalho de Conclusão de Curso submetido ao corpo docente do Departamento de Sistemas de Informação da Universidade Federal de Sergipe (DSIITA/UFS) como parte dos requisitos para obtenção do grau de Bacharel em Sistemas de Informação.

Itabaiana, 18 de março de 2014.

BANCA EXAMINADORA:

Prof. Methanias Colaço Rodrigues Júnior, Doutor
Orientador
DSIITA/UFS

Prof. André Vinícius Rodrigues Passos Nascimento, Mestre
Coorientador
DSIITA/UFS

Prof. Andrés Ignácio Martínez Menéndez, Mestre
DSIITA/UFS

SANTOS, Rafael Meneses. **Análise da Viabilidade da Mineração de Estatísticas Públicas do Campeonato Brasileiro de Futebol.** 2014. Trabalho de Conclusão de Curso – Curso de Sistemas de Informação, Departamento de Sistemas de Informação, Universidade Federal de Sergipe, Itabaiana, 2014.

RESUMO

O mercado esportivo vem crescendo de forma constante nos últimos anos. Esportes como futebol, beisebol e basquete movimentam bilhões de reais em todo o mundo. Em um ambiente que envolve tanto risco, o planejamento deve ser feito cuidadosamente, sendo necessário buscar novas formas de conseguir informações que possam ajudar no processo decisório. Na última década, vem surgindo como uma nova tendência, o uso da mineração de dados para extrair informações de diversas fontes de dados dentro dos esportes. Estatísticas de treinos, das partidas e mineração de comentários em redes sociais, são alguns dos exemplos que podem ser usados. Os esportes mais desenvolvidos nessa área no momento são o basquete e o beisebol, sendo que no futebol, na maioria dos casos, as equipes ainda buscam formas tradicionais de análise dos seus dados. Diante desse contexto, este trabalho teve como objetivo verificar se o uso de algoritmos de Mineração de Dados em estatísticas públicas do Campeonato Brasileiro de Futebol é viável. Para isso foi feita uma comparação entre algoritmos mineração de dados selecionados, descrevendo os padrões descobertos. Os resultados mostraram que o processo é viável e cada algoritmo pode trazer padrões diferentes.

Palavras-chave: Mineração de Dados, Estatísticas Esportivas, Futebol

SANTOS, Rafael Meneses. **Feasibility Analysis of Mining Public Statistics of the Brazilian Football Championship**. 2014. Course Conclusion Paper – Information Systems Course, Information Systems Department, Federal University of Sergipe, Itabaiana, 2014.
.

ABSTRACT

The sports market has been growing steadily in the last years. Sports such as football, baseball and basketball move billions of dollars worldwide. In an environment that involves high-risk decisions, planning must be done carefully, the search for new ways of getting information to apply in the decision making process is required. In the last decade, data mining has emerged as a way to extract information from multiple data sources from sports. Statistics from training activities, matches and mining comments on social networks, are examples that can be used. Basketball and baseball are the most developed sports in this area whereas football, in most cases, teams still seek traditional forms of data analysis. Given this context, this paper aims to determine whether the use of Data Mining algorithms on publicly available statistics of the Brazilian Football Championship is feasible. For that, a comparison of selected data mining algorithms was done, describing discovered patterns. The results showed that the process is feasible and each algorithm can find its own kind of patterns.

Key-words: Data Mining, Sports Statistics, Football

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizado de Máquina
CSS	Cascading Style Sheets
DOM	Document Object Model
DW	Data Warehouse
ETL	Extract, Transform e Load
KDD	Knowledge Discovery in Databases
HTML	HyperText Markup Language
MD	Mineração de Dados
OLAP	On-line Analytical Processing
RNA	Rede Neurais Artificiais
SQL	Structured Query Language
URL	Uniform Resource Locator

SUMÁRIO

1. INTRODUÇÃO

Com a constante evolução do hardware usado para o armazenamento de dados, é possível armazenar uma quantidade maior de dados por um custo cada vez menor. Esses dados são usados para auxiliar tanto as operações diárias, como o planejamento estratégico das empresas (HAN; KAMBER; PEI, 2011).

Quando existe uma pequena quantidade de dados é possível, através de métodos manuais, investigar esses dados e extrair uma informação útil para o usuário. Mas nem sempre esse cenário ocorre. Conforme essa quantidade de dados cresce, o entendimento da mesma decresce rapidamente. No meio de tantos dados armazenados, possivelmente existe uma grande quantidade de informações úteis escondidas. A análise dessa imensa quantidade de dados de forma manual acaba se tornando inviável, o que levou à pesquisa de técnicas computacionais para auxiliar essas atividades de análise. Essas técnicas acabaram formando uma nova área de pesquisa que ficou conhecida como Mineração de Dados (WITTEN; FRANK; HLL, 2011).

A Mineração de Dados (MD) pode ser definida como o processo de analisar os grandes bancos de dados de forma automatizada ou semi-automatizada em busca de padrões relevantes (HAN, KAMBER, PEI, 2011; WITTEN, FRANK, HLL, 2011). É uma das etapas que compõem o processo conhecido como Descoberta de Conhecimento em Banco de Dados (KDD – do inglês Knowledge Discovery in Databases) (FAYYAD; PIAETSKY-SHAPIRO; SMYTH, 1996). Além disso, é uma área interdisciplinar que engloba outras áreas como: Estatística, Aprendizado de Máquina, Inteligência Artificial, Banco de Dados, Reconhecimento de Padrões, entre outras (GORUNESCU, 2011).

Umas das áreas que vêm atraindo interesse dos pesquisadores é o esporte. Grandes bancos de dados são gerados a partir de dados estatísticos capturados durante as partidas, além de outras atividades que envolvem as negociações das equipes esportivas. Essas equipes podem ser empreendimentos multimilionários, sendo que uma decisão gera despesas milionárias. Essas decisões envolvem riscos muito altos e podem causar prejuízos para as equipes. Para diminuir esse risco, é necessário que existam informações úteis para apoiar as decisões (SCHUMAKER; SOLIEMAN; CHEN, 2010).

O uso de Mineração de Dados nesta área cresceu consideravelmente nos últimos anos. Os esportes que mais se destacam são o basquete e o beisebol nos quais existem alguns

casos de sucesso. Mesmo assim, a maioria das equipes tomam decisões a partir do instinto ou de informações passadas por analistas, que muitas vezes não conseguem analisar todos os dados armazenados (SCHUMAKER; SOLIEMAN; CHEN, 2010).

Diante desse contexto, a proposta deste trabalho é verificar se a Mineração de Dados aplicada às estatísticas públicas de partidas do Campeonato Brasileiro de Futebol é viável. No trabalho, serão usadas técnicas de Mineração de Dados em estatísticas públicas encontradas na *Web* e, em seguida, serão verificados os resultados desses processamentos.

1.1. Motivação

Segundo Schumaker, Solieman e Chen (2010), é possível encontrar cinco níveis de uso de dados estatísticos dentro de uma equipe esportiva. Esses níveis são:

- Nível 1 – Não usam dados estatísticos;
- Nível 2 – Dados são capturados, mas especialistas tomam decisões a partir da intuição e de instinto;
- Nível 3 – Especialistas usam dados históricos no processo decisório;
- Nível 4 – Especialistas usam estatísticas no processo decisório;
- Nível 5– Especialistas usam Mineração de Dados no processo decisório;

Grande parte das equipes está classificada entre os níveis 2 e 4. De acordo com Schumaker, Solieman e Chen (2010), equipes que usaram Mineração de Dados no beisebol, conseguiram melhores resultados nos campeonatos, em comparação aos anos anteriores nos quais as análises dos dados eram feitas apenas por analistas.

Tradicionalmente, a comissão técnica das equipes de futebol busca informações através de olheiros, notícias, observações feitas durante os treinos, etc. Todas essas fontes tem algo em comum, a subjetividade. Esse tipo de informação não é suficiente para melhorar a qualidade da tomada de decisão das equipes. É preciso buscar algo mais objetivo, nesse caso, as estatísticas baseadas nos dados gerados durante as partidas e treinos.

Segundo Anderson e Sally (2013), treinadores, jogadores, dirigentes, casas de apostas, jornalistas, torcedores e acadêmicos buscam cada vez mais pelos números do futebol. Apesar disso, ainda existe uma grande resistência por boa parte das equipes, as quais continuam a planejar suas ações de forma tradicional, através da intuição e opiniões subjetivas.

Graças a essa nova percepção, algumas empresas estão surgindo, principalmente na Europa, com o propósito de fornecer os mais diversos dados relacionados a eventos das partidas (ANDERSON, SALLY, 2013; KUPER, SZYMANSKI, 2011). Esses dados são analisados por especialistas, que buscam informações que possam trazer vantagem competitiva para as equipes, mas devido à complexidade e quantidade de dados que envolvem o futebol, o uso de computadores e algoritmos específicos faz-se necessário para auxiliá-los nessa análise.

Sendo assim, surge o interesse em verificar o uso de Mineração de Dados no futebol. Nesse trabalho, algoritmos de Mineração de Dados serão avaliados em estatísticas públicas do Campeonato Brasileiro de Futebol. A partir do resultado, será possível verificar quais algoritmos possuem os maiores índices de acertos na previsão de resultados, além de padrões encontrados durante a execução dos mesmos. Tais resultados podem trazer novos meios de conseguir informações para melhorar as decisões das equipes.

1.2. Trabalhos Relacionados

Durante a pesquisa bibliográfica, alguns trabalhos similares foram encontrados, sendo que os que mais se destacaram foram os trabalhos de Costa (2010), Farias (2008) e Nunes e Sousa (2006).

Costa (2010) fez uma análise de ações ofensivas que resultavam em finalizações durante um Campeonato Europeu de Seleções de 2008. Foram analisadas, de forma manual, 297 ações de seleções classificadas entre seleções de sucesso e de insucesso dentro de um critério estabelecido pelo autor. O trabalho destaca-se pelo grande nível de detalhes observados na partida que envolve desde o começo da ação ofensiva, até a finalização e pelo resultado, o qual constatou um grande número de padrões.

Já o trabalho de Farias (2008), teve como objetivo a construção e comparação de modelos estatísticos para previsão de resultados de partidas de futebol. Ele trabalhou com resultados das partidas do Campeonato Brasileiro de Futebol do ano de 2006 e chegou a prever o resultado de 70% das partidas de uma rodada, sendo que as demais variavam entre 30 e 60%.de acerto.

No trabalho de Nunes e Sousa (2006), foram aplicados algoritmos de mineração de dados em duas bases de dados, uma do campeonato português de 2004 a 2005 e a outra, de partidas de algumas seleções européias, com objetivo de buscar padrões novos e úteis. As

bases continham dados de eventos temporais da partida. Os eventos podiam ser de três tipos: substituições, gols e aplicações de cartões. Os autores usaram um algoritmo de classificação e outro de regras de associação. No final, concluíram que, apesar de encontrem diversos padrões, nenhum deles era novo ou continham conhecimento relevante.

1.3. Justificativa

Como pôde ser visto na seção anterior, alguns trabalhos relacionados apresentaram bons resultados de acordo com o que foi proposto por eles. A partir dessa análise, foram encontrados alguns pontos relevantes que diferenciam o presente trabalho dos demais.

Primeiro, foi priorizado o uso de uma base de dados que possua alguns anos de registros históricos armazenados, enquanto em alguns trabalhos relacionados eram usadas bases com no máximo um ano de atividades, o que, segundo Witten, Frank e Hall (2011), dificulta a obtenção de resultados de qualidade.

Outro ponto encontrado foi o uso de poucos atributos no processo de mineração. A busca por uma quantidade adequada de atributos para serem usados pelo modelo de mineração tem uma grande influência nos resultados, determinando assim o nível de qualidade dos mesmos. Sendo assim, neste trabalho, a quantidade de atributos também será levada em conta na seleção das fontes de dados.

Além dos pontos apresentados anteriormente, este trabalho também fará uma comparação entre os principais algoritmos usados atualmente na Mineração de Dados, sendo que alguns deles não foram usados nos trabalhos encontrados. Com isso, é possível ter resultados mais diversos e um novo olhar diante desse contexto.

1.4. Objetivo Geral

O objetivo do projeto é analisar a viabilidade da extração e mineração de estatísticas públicas do Campeonato Brasileiro de Futebol, ou seja, verificar se existem padrões nas estatísticas extraídas e verificar o nível de acertos nas previsões de resultados. A partir disso, será possível avaliar os algoritmos selecionados, destacando os pontos relevantes entre eles.

1.4.1. Objetivos Específicos

O objetivo geral pretende ser alcançado através dos seguintes objetivos específicos:

- Fazer um levantamento bibliográfico sobre técnicas de Mineração de Dados e Análise de Estatísticas nos Esportes, com foco direcionado ao futebol e à Mineração de Dados;
- Pesquisar e analisar fontes de dados que possam ser usadas para aplicar técnicas de Mineração de Dados;
- Modelar e implementar um modelo de banco de dados para armazenar dados referentes às partidas de futebol;
- Projetar e implementar um ambiente de extração das estatísticas;
- Identificar algoritmos de Mineração de Dados que possam ser aplicados;
- Criar visões necessárias para a execução de algoritmos selecionados;
- Executar os algoritmos de Mineração de Dados do Microsoft SQL Server Analysis Services (SSAS) e analisar os resultados obtidos.

1.5. Metodologia

No projeto, será utilizada uma metodologia de pesquisa bibliográfica, na qual será feito um estudo dos temas: Mineração de Dados e Análise de Estatísticas no Esporte com foco no futebol e na Mineração de Dados. A partir disso, será feito um estudo de caso no Campeonato Brasileiro de Futebol com estatísticas públicas que serão obtidas através da Internet. Esses dados serão processados por algoritmos de Mineração de Dados e com isso, será possível analisar a questão principal do trabalho.

Na primeira fase do projeto, será feito um estudo sobre técnicas de Mineração de Dados e Análise de Estatísticas no Esporte, principalmente no futebol. Além disso, fontes públicas de estatísticas do Campeonato Brasileiro de Futebol que podem ser encontradas na Internet serão selecionadas e analisaremos a que mais se adéqua ao nosso estudo. Serão avaliados critérios como a quantidade de registros disponíveis, quantidade de atributos e a consistência desses dados disponibilizados.

Na próxima fase, será feito um estudo de caso direcionado ao Campeonato Brasileiro de Futebol. Esse estudo consiste na definição de objetivos e planejamento das etapas seguintes. Posteriormente, será implementado um sistema que irá extrair, limpar e armazenar os dados da fonte de dados da Internet, também conhecido como processo de

ETL[1]. Em seguida, será criado um projeto de Mineração de Dados na ferramenta SSAS para que os algoritmos possam ser usados no processamento dos dados obtidos. Esses algoritmos serão selecionados nessa fase e para cada algoritmo selecionado, uma ou mais visões podem ser criadas para adequar os dados que serão usados como entrada.

Por fim, os algoritmos de Mineração de Dados serão executados e os resultados obtidos serão compilados em um relatório final. Esse relatório irá conter a relação de algoritmos que foram selecionados durante etapas anteriores, padrões encontrados em cada um e a porcentagem de acertos na previsão de resultados, que apontará o algoritmo com melhor precisão na previsão de resultados das partidas.

1.6. Estrutura do trabalho

O trabalho foi divido em 5 capítulos. O primeiro capítulo apresentou uma introdução ao trabalho, mostrando os objetivos que devem ser alcançados, motivações, justificativas e os trabalhos relacionados que foram encontrados. No capítulo 2, é apresentada a fundamentação teórica utilizada para desenvolver o trabalho. São discutidos os conceitos encontrados no campo da Mineração de Dados e Análise de Estatísticas Esportivas. Em seguida, no capítulo 3, é abordado o processo de mineração de dados realizado no trabalho, que envolve desde a etapa inicial de seleção de base de dados até a execução dos algoritmos na ferramenta SSAS. Também é apresentada uma visão geral de ferramenta SSAS e dos algoritmos que serão utilizados no trabalho. O capítulo 4 traz os resultados encontrados pelo SSAS. São apresentados inicialmente os padrões encontrados pelos algoritmos individualmente e logo em seguida, a comparação entre a precisão dos algoritmos. O capítulo 5 apresenta as conclusões que puderam ser extraídas desse estudo, bem como sugestões de possíveis trabalhos futuros.

[1] O processo ETL, do inglês *Extract, Transform and Load*(Extrair, Transformar e Carregar), serve para organizar os dados através de etapas de extração, transformação e carregamento de dados, tornando os mesmos adequados para o uso no processo de Mineração de Dados.

2. REVISÃO BIBLIOGRÁFICA

Nesse capítulo, serão abordados os conceitos necessários para a realização do trabalho. Começamos apresentando, na seção 2.1, o campo da Mineração de Dados, detalhando os pontos relevantes dessa área e o processo de KDD o qual faz parte. Na seção 2.2, veremos uma visão geral sobre a análise de estatísticas esportivas, direcionando o foco para o futebol.

2.1. Mineração de Dados

De acordo com Witten, Frank e Hall (2011), a convergência entre computação e comunicação tem influenciado a sociedade de modo que a mesma vive cercada de informações. Essas informações são encontradas na sua forma primitiva: o dado. Esses dados são caracterizados como registros de fatos e somente se tornam informações quando se sabem os padrões e relações que existem entre eles.

A quantidade de dados armazenados atualmente é imensa e em constante crescimento. A facilidade em armazenar dados nos permite guardar dados que antes seriam descartados. Com o surgimento da Internet esse cenário ficou mais evidente. Somos sobrecarregados com informações e muitas vezes não fazemos o uso devido delas.

É estimado que a quantidade dados armazenados no mundo dobra a cada 20 meses (WITTEN; FRANK; HALL, 2011). Diante desse cenário, podemos perceber uma tendência que fica cada vez mais evidente. Conforme a quantidade de dados cresce nosso entendimento dos mesmos decresce. Dentro dessa imensidão de dados, existem informações potencialmente úteis que dificilmente são descobertas e usadas. Ferramentas poderosas e versáteis são necessárias para auxiliar a descoberta de informações nessas bases de dados. Dessa necessidade, surgiu a Mineração de Dados. (HAND; MANNILA; SMYTH, 2001).

Witten, Frank e Hill (2011) definem Mineração de Dados como sendo o processo de analisar os bancos de dados de forma automatizada ou semi-automatizada em busca de padrões úteis. Pode ser vista como uma evolução natural da tecnologia da informação. É uma das etapas do processo de KDD (Knowledge Discovery in Database ou Descoberta de Conhecimento em Banco de Dados), o qual será detalhado na Seção 2.1.4. Como se trata de uma área interdisciplinar, a Mineração de Dados incorporou técnicas de outros domínios,

como por exemplo, estatística, aprendizado de máquina (AM), banco de dados, recuperação de informação, etc. (HAN; KAMBER; PEI, 2011). Essa natureza interdisciplinar contribui significantemente para o sucesso da Mineração de Dados.

A Mineração de Dados pode ser aplicada para qualquer tipo de dado desde que os dados tenham um significado para a aplicação alvo. São vários os exemplos de fontes de dados que podem ser mineradas, entre eles podemos citar: bancos de dados, data warehouses, dados transacionais, textos, dados multimídias, etc. (HAN; KAMBER; PEI, 2011).

Outro ponto importante a ser definido são os conceitos de conceito, instância e atributo. O conceito refere-se a aquilo que pretende ser aprendido, sendo que o resultado do conceito é denominado descrição do conceito. O conjunto usado pelo processo de mineração é composto por instâncias. Uma instância pode ser definida como um registro dentro de uma fonte de dados. Cada instância é composta por valores que representam cada aspecto da instância. Esses valores são chamados de atributos (WITTEN; FRANK; HALL, 2011).

2.1.1. Tipos de Atributos

Cada instância que faz parte do processo de mineração de dados é formada por um conjunto de valores que representam características de cada instância. Essas características também são conhecidas como atributos. Pensando em um modelo de tabela, cada instância representa uma linha nessa tabela, enquanto os atributos representam colunas. Esses atributos podem ser de quatro tipos: nominal, ordinal, intervalar e racional (WITTEN; FRANK; HALL, 2011).

Atributos nominais são aqueles que não possuem relação entre si, ou seja, servem apenas para rotular e nomear estados dos atributos. Por exemplo, um atributo que represente o estado civil de uma pessoa pode conter os estados: solteiro, casado, divorciado e viúvo. Não é possível fazer uma relação entre esses estados, como por exemplo, medir a distância, ordenar ou aplicar operações matemáticas entre eles (HAN; KAMBER; PEI, 2011).

Os atributos ordinais são atributos que possuem relação de ordem entre eles, porém, não existe a noção de distância entre os mesmos. A temperatura pode ser considerada um atributo ordinal, tendo como estados: frio, morno e quente. É possível compará-los e perceber qual deles é o mais quente ou frio, mas não é possível calcular a distância entre eles (WITTEN; FRANK; HALL, 2011).

Os atributos intervalares são valores numéricos que são medidos em uma escala de unidades com tamanhos iguais e não possuem um ponto zero definido. Eles podem ser ordenados, além de permitirem o uso de operações de adição e subtração de valores que representem intervalos (HAN; KAMBER; PEI, 2011). Um exemplo de atributo intervalar é o ano. Podemos analisar a diferença e a média entre anos, mas não faz sentido somá-los ou multiplicá-los porque não existe um ponto inicial, ou seja, um zero definido (WITTEN; FRANK; HALL, 2011).

Atributos com valores numéricos e ponto zero definido são chamados de atributos racionais. É possível encontrar relações de ordem e aplicar operações matemáticas em atributos racionais. Um atributo que mede uma distância percorrida é um exemplo de atributo racional (WITTEN; FRANK; HALL, 2011).

Em muitos casos, atributos podem ser considerados de vários tipos diferentes. Na maioria dos casos, quando analisamos atributos dentro do contexto de mineração de dados, classificamo-los em discretos ou contínuos.

Um atributo discreto possui um número finito de valores, podendo ou não ser representado por um valor numérico. Já atributos contínuos, são representados por números e podem ter uma quantidade infinita de estados (HAN; KAMBER; PEI, 2011).

2.1.2. Métodos de Mineração de Dados

De acordo com Gorunescu (2011) o processo de Mineração de Dados consiste em construir um modelo de mineração que represente a fonte de dados que será minerada para resolver um problema. Esses problemas podem ser resolvidos por métodos que são divididos em dois grupos:

1. Métodos preditivos: São métodos que usam variáveis selecionadas para predizer outras variáveis. Os principais exemplos são: Classificação, Regressão, Detecção de Anomalias;
2. Métodos descritivos: São aqueles que revelam padrões nos dados, facilmente interpretados pelo usuário. Os principais exemplos são: Agrupamento, Regras de Associação, Padrões de Sequência.

2.1.2.1. Classificação

O processo de classificação consiste em separar instâncias de uma fonte de dados em um conjunto de categorias ou classes, com base em seus atributos. Para a classificação seja feita, são necessários quatro componentes fundamentais (GORUNESCU, 2011):

- Classe: Representa a variável do modelo que representa a classe, também conhecida como rótulo ou *label*;
- Preditores: São variáveis independentes do modelo que representam os atributos da instância a ser classificada;
- Conjunto de treinamentos: São instâncias já classificadas que são usadas para treinar o modelo;
- Conjunto de testes: Contem instâncias novas que serão classificadas pelo modelo com o objetivo de verificar precisão do modelo de mineração.

Para construir o modelo de classificação, também conhecido como classificador, é preciso executar um processo de treinamento com um conjunto de dados classificados corretamente. Em seguida, é feita a comparação com o conjunto de casos de teste para verificar a qualidade do classificador construído.

2.1.2.2. Descoberta de Regras de Associação e Sequência

Os métodos de regras de associação buscam regras e medidas de dependências entre instâncias, também conhecidas como itens, que ocorrem com freqüência em determinadas situações. Essas situações são denominadas transações. Para que o algoritmo possa ser executado, é preciso que a fonte de dados forneça uma forma de identificar as transações e os itens que fazem parte da mesma.

Quando a ordem de ocorrência dos itens nas transações é relevante para o contexto do problema é precisar usar métodos de descoberta de padrões de sequência para que os padrões de sequência possam ser encontrados (GORUNESCU, 2011).

2.1.2.3. Regressão

Na estatística, a análise de regressão significa o uso de um modelo matemático que estabelece uma relação entre valores de múltiplas variáveis. Enquanto métodos de classificação predizem categorias ou classes discretas de uma instância, os métodos de regressão criam modelos que predizem valores contínuos (HAN; KAMBER; PEI, 2011).

2.1.2.4. Agrupamento (*Clustering*)

Os métodos de agrupamento são usados para separar as instâncias da fonte de dados em grupos (*clusters*) que possuam similaridades entre seus atributos. A principal diferença entre os métodos de agrupamento e os de classificação é a falta de um atributo de classe previamente definido. As classes serão definidas a partir dos clusters gerados (GORUNESCU, 2011).

2.1.2.5 Análise de Anomalias (*Outliers*)

Em muitos casos, uma fonte de dados pode conter instâncias que não seguem o comportamento mais generalizado. Essas instâncias são conhecidas como *outliers*. Muitas técnicas de mineração descartam os *outliers* por se tratarem de exceções que podem causar problemas ao modelo de mineração. Mas em outros casos, os eventos raros que fogem dos padrões são justamente o que são procurados (HAN; KAMBER; PEI, 2011).

Um dos exemplos clássicos da análise de *outiliers* é o uso para detecção de atividades fraudulentas em cartão de crédito. Nesse tipo de análise, cada compra feita por um cliente é analisada em busca de anormalidades como, por exemplo, compras feitas em locais diferentes dos habituais, compras caras e em grandes quantidades, etc.

2.1.4. O Processo de KDD

Segundo Fayyad, Piatetsky-Shapiro e Smyth, KDD é um processo não trivial de identificar padrões compreensíveis, novos, válidos e potencialmente úteis em uma fonte de dados. Ele é não trivial porque exige uma busca e inferência, ou seja, não é apenas o cálculo de quantidades predefinidas.

No KDD, os dados são um conjunto de fatos, e os padrões são representados em forma de linguagem específica ou através de visualizações. Os padrões descobertos devem ser validados com novos dados, averiguando o grau de similaridade encontrado. Busca-se também que o padrão seja algo novo, tanto para o sistema como para o usuário, e potencialmente útil, trazendo benefícios para o usuário.

O processo de KDD é interativo e iterativo, envolvendo alguns passos que devem ser acompanhados de decisões do usuário (COLAÇO JUNIOR, 2004). Os principais passos do KDD, ilustrados na figura 1, são descritos a seguir:

- Seleção de Dados: Essa etapa consiste na seleção de fontes de dados que envolvam o problema que será trabalhado;
- Pré-processamento: Nessa etapa, as inconsistências que podem ocorrer devido à integração de fontes de dados diferentes são eliminadas. Por exemplo, um atributo que exista em duas fontes de dados pode conter valores nominais em uma fonte, enquanto em outra, pode ter valores intervalares. É preciso que essa inconsistência seja resolvida nessa etapa, definindo um único de tipo de atributo que possa receber valores das duas fontes de dados distintas;
- Transformação: Essa etapa tem como objetivo transformar os dados coletados em um formato que atenda os requisitos dos algoritmos;
- Mineração de Dados: Essa é a etapa principal do processo. Nela, os algoritmos de mineração de dados serão aplicados nos dados que foram preparados;
- Interpretação: Nessa etapa, todos os padrões encontrados serão analisados e a partir disso, será possível verificar qual deles poderá ser usado pelo usuário na tomada de decisão.

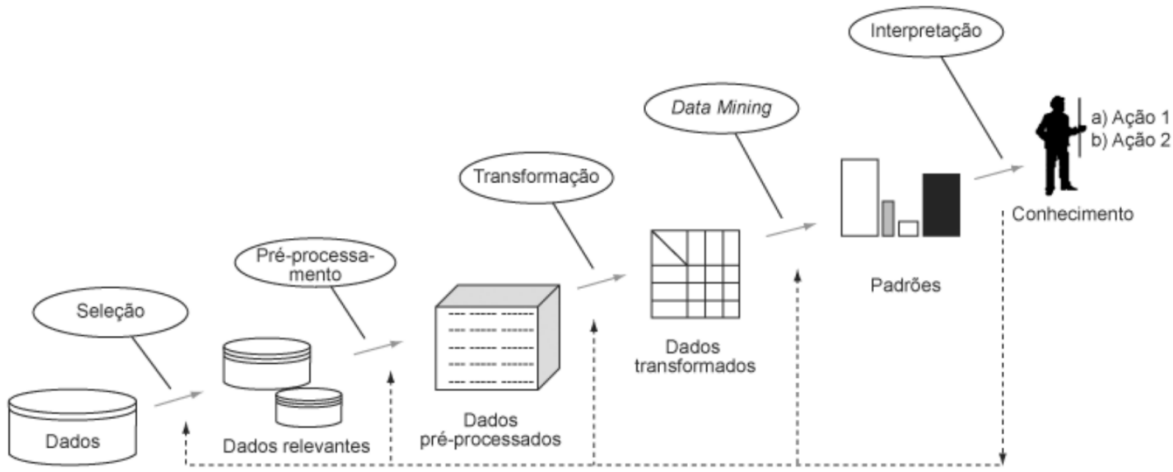


Figura 1: KDD (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996)

2.2. Análise de Estatísticas Esportivas

Grandes quantidades dados existem em todos os domínios dos esportes. Esses dados podem vir de fontes como, por exemplo, desempenho individual do jogador, decisões técnicas e gerenciais, eventos que ocorrem durante as partidas, etc. A tarefa principal não é coletar esses dados, e sim, definir quais dados coletar e o que fazer com eles. Através da descoberta de novos meios para extrair conhecimento desses dados, as organizações esportivas tem o potencial de conseguir uma vantagem competitiva em relação aos seus adversários (SCHUMAKER; SOLIEMAN; CHEN, 2010).

No princípio, acreditava-se que os especialistas (técnicos, gerentes e olheiros) poderiam converter efetivamente os dados coletados em conhecimento útil. Conforme o escopo e a quantidade de dados foram crescendo, percebeu-se que métodos novos deveriam ser buscados para que as organizações pudessem entender melhor os dados que tinham. Primeiro, buscaram contratar analistas de estatísticas esportivas para criar medidas de desempenho e melhor os critérios usados na tomada de decisão. O segundo passo foi buscar métodos práticos para extrair conhecimento útil usando técnicas de mineração de dados (SCHUMAKER; SOLIEMAN; CHEN, 2010).

Segundo Schumaker, Solieman e Chen (2010), a primeira parte do problema é identificar as medidas de desempenho, pois algumas métricas existentes podem ser usadas de forma incorretas, ou pior, não ter relação com o resultado da partida. A segunda parte consiste em achar padrões interessantes dentro dos bancos de dados. Esses padrões podem ser tendências de jogadas do adversário, predições de resultado através de uma base histórica ou até mesmo relações entre desempenho e contusões.

A adoção de técnicas de mineração de dados por organizações esportivas não aconteceu da noite para o dia. O passo inicial no beisebol veio do escritor esportivo e estatístico Bill James. Em 1977, James iniciou a sua publicação anual conhecida como “Bill James Baseball Abstracts”. Nessas publicações ele começou a questionar medidas de desempenho tradicionais do beisebol, comentando possíveis problemas existentes no uso de medidas imprecisas.

No início, ele acabou vendendo apenas 50 cópias de sua publicação, mas, mesmo assim, continuou publicando-a anualmente. Consequentemente, a venda dessas publicações começou a crescer. Em 1982, Bill James iniciou uma discussão no meio acadêmico sobre novas medidas de desempenho. Essas medidas ficaram conhecidas como sabermétricas. No início, a sabermétrica foi vista pela comunidade esportiva como uma tentativa fracassada da

academia de tentar formular as estatísticas das partidas, e por isso, foi ignorada durante muitos anos.

Alguns entusiastas começaram a testar a sabermétrica para avaliar jogadores amadores e profissionais. Esses entusiastas começaram a aplicar o conhecimento adquirido em apostas e perceberam uma melhoria nos resultados das previsões. Mesmo com o crescente sucesso da sabermétrica, organizações esportivas continuavam ignorado-a. Apenas em 2002, o uso de sabermétricas começou a ser aplicado em clubes profissionais de beisebol.

O Oakland A's, comandado por Billy Beane, foi o primeiro caso de sucesso a usar sabermétricas e algoritmos para escolha de jogadores em um clube profissional de beisebol. Graças a essa combinação, Beane pôde encontrar jogadores com sabermétricas similares aos melhores jogadores, mas com um custo de contração bastante inferior. Com essa equipe, o Oakland A's, tendo umas das folhas de pagamento mais baixas do campeonato, conseguiu por 5 anos consecutivos fazer parte do *playoff* da Liga Americana de Beisebol. Esse caso foi relatado no livro Moneyball que foi adaptado para o cinema (SCHUMAKER; SOLIEMAN; CHEN, 2010).

Já o futebol, vem enfrentando sua própria revolução dos números. Segundo Anderson e Sally (2013), o futebol é um esporte que sempre foi decido por atletas bem preparados e técnicos intransigentes que confiavam apenas na própria intuição e não aceitavam sugestões que fossem contrárias ao jeito de tradicional de se fazer as coisas. Esse cenário mudou bastantes nos últimos anos.

Charles Reep pode ser considerado o primeiro analista da história do futebol. Nascido em 1904 em Cornwall, cidade localizada na Inglaterra, foi um estudante de contabilidade e tenente-coronel da Força Área Real. Após assistir a uma palestra de Charles Jones, capitão do Arsenal, decidiu aplicar a contabilidade ao futebol, criando um sistema para anotar cada lance que ocorria em um jogo (ANDERSON; SALLY, 2013).

Enquanto assistia às partidas de futebol, Reep anotava em seu caderno eventos das mesmas da forma mais detalhada possível. Ele registrou mais de 2200 partidas durante sua vida, dedicando cerca de 80 horas a analisar cada uma delas.

De acordo com Anderson e Sally (2013), os dados coletados por Reep foram utilizados como base para um artigo científico, publicado em 1968, denominado "Habilidade e Sorte no Futebol" (Skill and Chance in Association Football), escrito por Reep e Bernard Benjamin, estatístico-chefe do Departamento Geral de Registro Civil. Esse artigo revelou

padrões que podem ser previstos nos lances de uma partida. Foi possível observar que diversos aspectos dos jogos seguiam padrões numéricos. Alguns dos padrões foram:

- As equipes marcavam, em média, um gol a cada nove finalizações;
- A maioria das jogadas termina após zero ou um passe completo, enquanto 91,5% nunca atingiam quatro passes certos;
- Cerca de 30% das bolas recuperadas na área adversária resultavam em finalizações;
- Cerca de metade dos gols resultavam de bolas recuperadas na área adversária.

As contribuições de Charles Reep foram extremamente importantes para o reconhecimento da análise esportiva no futebol. No entanto, a adoção desse tipo de análise pelas equipes teve início apenas nas últimas duas décadas.

Atualmente, organizado pela Faculdade Sloan de Administração do Instituto de Tecnologia de Massachusetts, existe um congresso dedicado à análise esportiva, o Congresso de Análise Esportiva (SLOAN ANALYTICS CONFERENCE, s.d.). É um evento anual que reúne mais de duas mil pessoas, entre elas, treinadores, dirigentes e executivos das grandes equipes de diversos esportes, com o objetivo de discutir as novas tendências na área de análise de estatísticas. Entretanto, dessas duas mil, menos de 5% representam o futebol (ANDERSON; SALLY, 2013).

Nos últimos anos, surgiram empresas como, por exemplo, a Opta Sports, Amisco, Prozone, Match Analysis, entre outras, que fornecem os mais diversos dados em quantidades cada vez maiores. Esses dados vão desde estatísticas das partidas, dados médicos detalhados, registros de treinos, etc. Os clubes também armazenam dados relacionados à venda de materiais oficiais, bilheteria dos jogos e consumo de alimentos e bebidas nos estádios (ANDERSON; SALLY, 2013).

Obter essas informações é apenas o primeiro passo, sendo necessária a análise dos mesmos por especialistas e algoritmos de mineração de dados.

3. ESTUDO DE CASO

O presente capítulo irá detalhar os passos para o desenvolvimento do estudo de caso. Primeiro, na seção 3.1, os objetivos do estudo de caso serão descritos. Essa etapa é fundamental, pois traz uma visão geral do que precisa ser feito e serve como base para as demais etapas. Em seguida, na seção 3.2, é apresentada a etapa de planejamento. Essa etapa consiste em selecionar as fontes de dados utilizadas no estudo de caso, construir uma *wrapper* para a extração dos dados e introduzir os conceitos da ferramenta SSAS usada no estudo. A seção 3.3 trata da execução e validação do estudo.

3.1. Definição de Objetivo

Foi definida como estudo de caso, a mineração de dados de estatísticas públicas do Campeonato Brasileiro de Futebol para verificar a precisão dos algoritmos selecionados. Para isso, é preciso realizados os seguintes objetivos:

- Seleção de uma fonte de dados estatísticos pública da *Web*;
- Extração de dados da fonte selecionada através de um ambiente ETL;
- Selecionar algoritmos da ferramenta SSAS que possam prever os dados;
- Criar uma visão comum a esses algoritmos;
- Criar modelos de mineração na ferramenta SSAS;
- Executar e ajustar os modelos de mineração para serem aplicados na fonte de dados, com intuito de melhorar a probabilidade de previsão de resultados.

3.2. Planejamento

Na fase de planejamento será traçado o caminho necessário para alcançar os objetivos descritos anteriormente.

Primeiro, é preciso definir qual será a fonte de dados pública que será usado para realizar o estudo de caso. Em seguida, será feita a implementação do ambiente de ETL responsável pela extração dos dados para o banco, adequando-os para que os mesmos tenham qualidade e possam ser usados pelos algoritmos. Durante essa fase é preciso

implementar um *wrapper* para a extração dos dados na Web. No final, a ferramenta SSAS e os algoritmos serão discutidos.

3.2.1. Seleção de Objetos

Para selecionar a fonte de dados, foi preciso avaliar um conjunto de sites que divulgam resultados das partidas do Campeonato Brasileiro de Futebol. Dois critérios de seleção foram estabelecidos para verificar qual desses sites era o mais adequado: a quantidade de partidas armazenadas e quantidade de atributos por jogadores. Depois de pesquisar os principais sites, chegou-se à conclusão que o site Uol Esporte (UOL ESPORTE, s.d.) fornecia os dados mais completos, de acordo com os critérios preestabelecidos.

O Uol Esporte possui dados históricos dos campeonatos de 2010 a 2012, contendo 1140 partidas e mais de 25 mil registros de estatísticas de jogadores. No Quadro 1, é possível ver as estatísticas disponibilizadas pelo site.

Assistência	Bolas perdidas	Bolas recebidas
Bolas recuperadas	Cartões	Cartões Amarelos
Cartões Vermelhos	Cruzamentos	Cruzamentos aproveitamento
Cruzamentos conquistados	Cruzamentos errados	Defesas
Desarmes	Desarmes completos	Desarmes incompletos
Dribles	Dribles certos	Dribles errados
Escanteios cedidos	Escanteios conquistados	Faltas cometidas
Faltas recebidas	Finalizações	Finalizações certas
Finalizações erradas	Finalizações na trave	Gols
Gols contra	Impedimento	Passes
Passes aproveitamento	Passes certos	Passes errados
Recuos	Tempo jogado	Virada de jogo

Quadro 1: Estatísticas fornecidas pelo Uol Esporte.

Para armazenar os dados encontrados, foi criado um banco de dados segundo modelo Entidade-Relacionamento da figura 2:

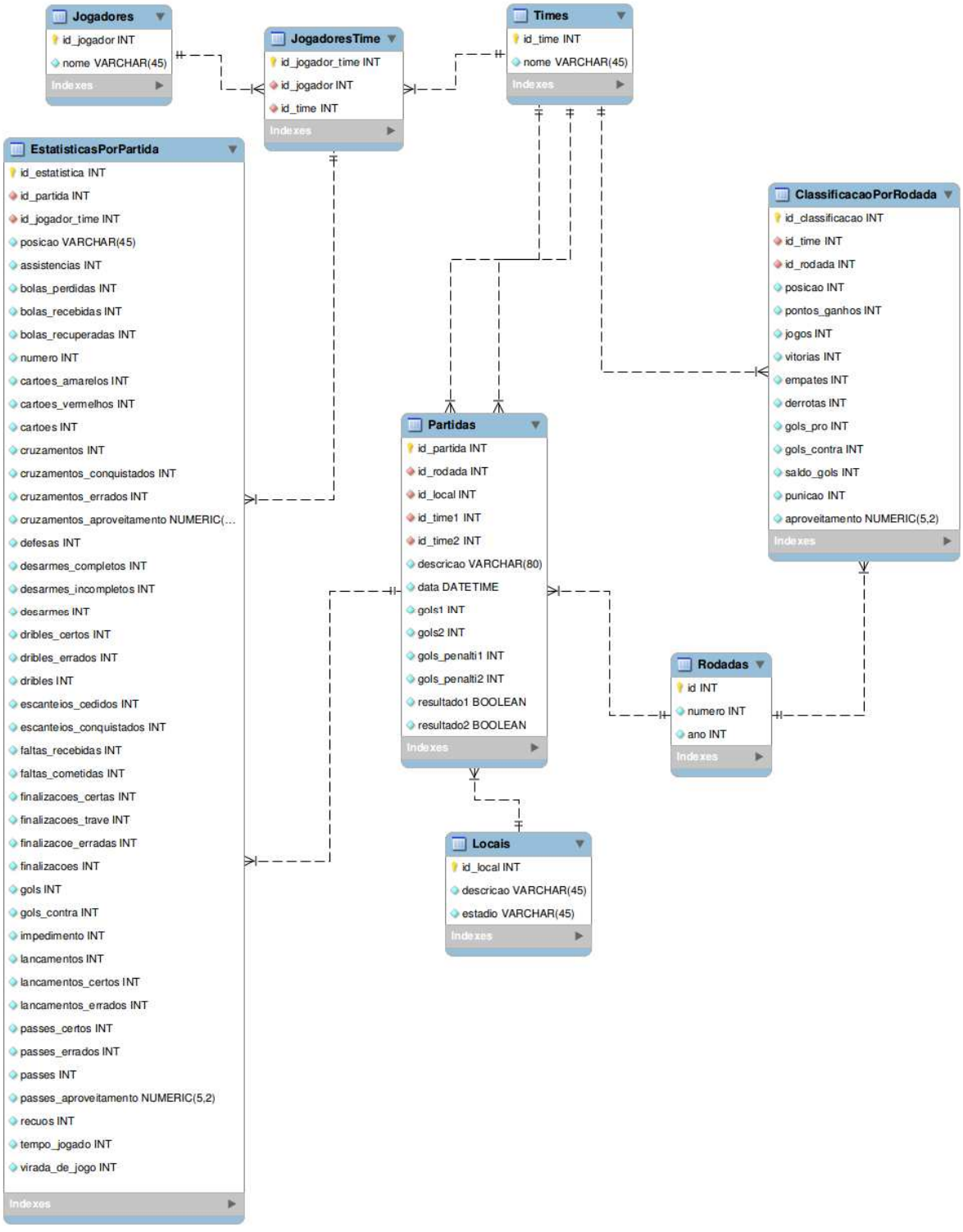


Figura 2: Modelo Entidade-Relacionamento do Banco de Dados.

3.2.2. Implementação do Ambiente de ETL

Para extrair os dados da Internet e armazená-los no banco de dados é preciso criar um *wrapper*. A extração de dados na *Web* divide-se duas categorias: a extração de

informação através da linguagem natural ou por dados estruturados. Programas que extraem dados de fontes estruturadas são chamados de *wrappers* (LIU, 2011).

Muitas páginas da internet são criadas normalmente através de dados obtidos de bancos e formadas segundo uma estrutura predefinida. Segundo Liu (2011), a criação de uma *wrapper* para extrair esses dados pode ser feita segunda três abordagens:

- Abordagem manual: Nela o programador observa estrutura da página em busca de padrões e em seguida escreve o programa para extrair os dados da página alvo. Essa abordagem não é indicada para extração de muitas páginas diferentes, pois dificilmente páginas vão usar estruturas similares o que torna o processo bastante complicado;
- Indução de *wrapper*: Usa técnicas de aprendizado de máquina a partir de uma abordagem de treinamento supervisionado;
- Extração automática: É uma abordagem que treinamento não supervisionado. É a mais usada quando é preciso extrair dados de muitas páginas diferentes.

Para esse trabalho, abordagem manual será usada devido à simplicidade de sua implementação e de estarmos trabalhando com extração de um único site. Para auxiliar nessa implementação será usada à biblioteca Jsoup (JSOUP, s.d.). O Jsoup é um parser HTML (*HyperText Markup Language*) que facilita o acesso à estrutura DOM (Document Object Model) de uma página HTML.

Para implementar o *wrapper,* é necessário observar a estrutura das páginas da fonte escolhida para que um processo de extração possa ser criado. O processo criado consiste de duas etapas. Primeiro, para cada ano, a página que lista os jogos das rodadas é acessada. O URL (*Uniform Resource Locator*) que identifica cada rodada segue o padrão indicado no quadro 2. O trecho [ano] deve ser substituído por um valor que represente o ano o qual serão extraídas as rodadas.

http://esporte.uol.com.br/futebol/campeonatos/brasileiro/[ano]/serie-a/estatisticas/jogos/

Quadro 2: URL para acesso as rodadas.

O URL anterior fornece acesso a uma página HTML que possui dados relacionados às partidas das rodadas daquele ano. O quadro 3 mostra parte do HTML em que é possível

extrair dados necessários para acessar as páginas individuais de cada partida. Também é extraído nome dos times e placar da partida.

```
<tr class="odd">
        <td class="time time1">
                <a href="/futebol/campeonatos/brasileiro/2012/serie-
                a/estatisticas/jogos/botafogo-x-sao-paulo-20-05.jhtm">Botafogo</a>
        </td>
        <td class="brasao">
                <a href="/futebol/campeonatos/brasileiro/2012/serie-
                a/estatisticas/jogos/botafogo-x-sao-paulo-20-05.jhtm">
                                                </td>
        <td class="placar">
                <a href="/futebol/campeonatos/brasileiro/2012/serie-
                a/estatisticas/jogos/botafogo-x-sao-paulo-20-05.jhtm">4 x 2</a>
        </td>
        <td class="brasao">
                <a href="/futebol/campeonatos/brasileiro/2012/serie-
                a/estatisticas/jogos/botafogo-x-sao-paulo-20-05.jhtm">
        </td>
        <td class="time">
                <a href="/futebol/campeonatos/brasileiro/2012/serie-
                a/estatisticas/jogos/botafogo-x-sao-paulo-20-05.jhtm">São Paulo</a>
        </td>
</tr>
```

Quadro 3: Parte do HTML da página de rodadas do campeonato.

A segunda etapa consiste em acessar individualmente cada partida de cada rodada. Nas páginas das partidas, é possível extrair dados sobre os jogadores e suas estatísticas na partida. O quadro 4 mostra parte do código HTML que serve pra identificar cada jogador que participou da partida. Os jogadores estão separados em duas *tags* e através da classe CSS (*Cascading Style Sheets*) é possível identificar de qual time eles fazem parte.

```
<ul class="times">
        <li class="time1" style="height: 342px;">
                <div class="jog" jogador_id="milton-raphael-9333">
                <div class="jog" jogador_id="brinner-10878">
                <div class="jog" jogador_id="fabio-ferreira-5984">
                <div class="jog" jogador_id="lucas-6898">
                <div class="jog" jogador_id="marcio-azevedo-6703">
                <div class="jog saiu" jogador_id="andrezinho-7693">
                <div class="jog entrou" jogador_id="elkeson-8758">
                <div class="jog" jogador_id="jadson-11208">
                <div class="jog" jogador_id="maicosuel-6874">
                <div class="jog" jogador_id="renato-816">
                <div class="jog" jogador_id="vitor-junior-8811">
                <div class="jog" jogador_id="herrera-5490">
        </li>
        <li class="time2" style="height: 342px;">
                <div class="jog" jogador_id="fabio-170">
                <div class="jog" jogador_id="diego-renan-9169">
                <div class="jog" jogador_id="leo-7054">
                <div class="jog" jogador_id="mateus-8274">
                <div class="jog" jogador_id="amaral-5498">
                <div class="jog" jogador_id="charles-6951">
                <div class="jog saiu" jogador_id="marcelo-oliveira-6741">
                <div class="jog entrou" jogador_id="everton-8521">
                <div class="jog" jogador_id="montillo-10243">
                <div class="jog saiu" jogador_id="souza-267">
                <div class="jog entrou" jogador_id="fabinho-souza-8617">
                <div class="jog saiu" jogador_id="tinga-199">
                <div class="jog entrou" jogador_id="anselmo-ramon-8628">
                <div class="jog" jogador_id="wellington-paulista-3954">
        </li>
</ul>
```

Quadro 4: Parte do HTML da página de resultado de uma partida.

Em cada *tag* do código anterior, pode-se encontrar os dados estatísticos dos jogadores. O quadro 5 exemplifica parte do código HTML o qual é possível identificar os dados que devem ser extraídos. A *tag* <th> possui o nome da estatística e a *tag* <td> representa o valor da mesma.

```
<table class="fundamentos-1" cellspacing="0" cellpadding="0">
	<tbody>
		<tr class="odd">
			<th>Bolas perdidas</th>
			<td>1</td>
		</tr>
		<tr>
			<th>Bolas recebidas</th>
			<td>3</td>
		</tr>
		<tr class="odd">
			<th>Faltas recebidas</th>
			<td>0</td>
		</tr>
		<tr>
			<th>Gols marcados</th>
			<td>0</td>
		</tr>
		<tr class="odd">
			<th>Passes certos</th>
			<td>13</td>
		</tr>
	</tbody>
</table>
```

Quadro 5: Parte do HTML com dados estatísticos de um jogador.

O processo é executado para cada rodada de cada ano. Foi executado o teste para extração de uma rodada, em que cada partida foi verificada e comparada com a versão do site. Não foram encontradas diferenças entre os valores, indicando que a extração foi realizada corretamente.

3.2.3. Microsoft SQL Server Analysis Services (SSAS)

O Microsoft SQL Server Analysis Services (SSAS), também conhecido apenas por Analysis Services, é uma solução desenvolvida pela Microsoft que permite a criação e implantação de banco de dados OLAP (On-line Analytical Processing) que são usados para auxiliar o suporte à decisão. A base de uma solução no SSAS é composta por um modelo dados semânticos e uma instância de servidor no qual é possível criar, processar, consultar e gerenciar os objetos desse modelo (MICROSOFT, s.d.e).

A figura 3 mostra a arquitetura da camada semântica do Analysis Services. Os modelos são criados a partir de dados históricos coletados em bancos de dados transacionais, entre outras fontes, e são marcados através de metadados que possibilitam medir, manipular e comparar os dados em consultas sob demanda (ad-hoc) e relatórios customizados. Depois que um modelo é criado, ele pode ser implantando no servidor do Analysis Services o qual os usuários autorizados poderão se conectar e interagir com o modelo através de ferramentas como o Excel, Reporting Services, etc. (MICROSOFT, s.d.e).

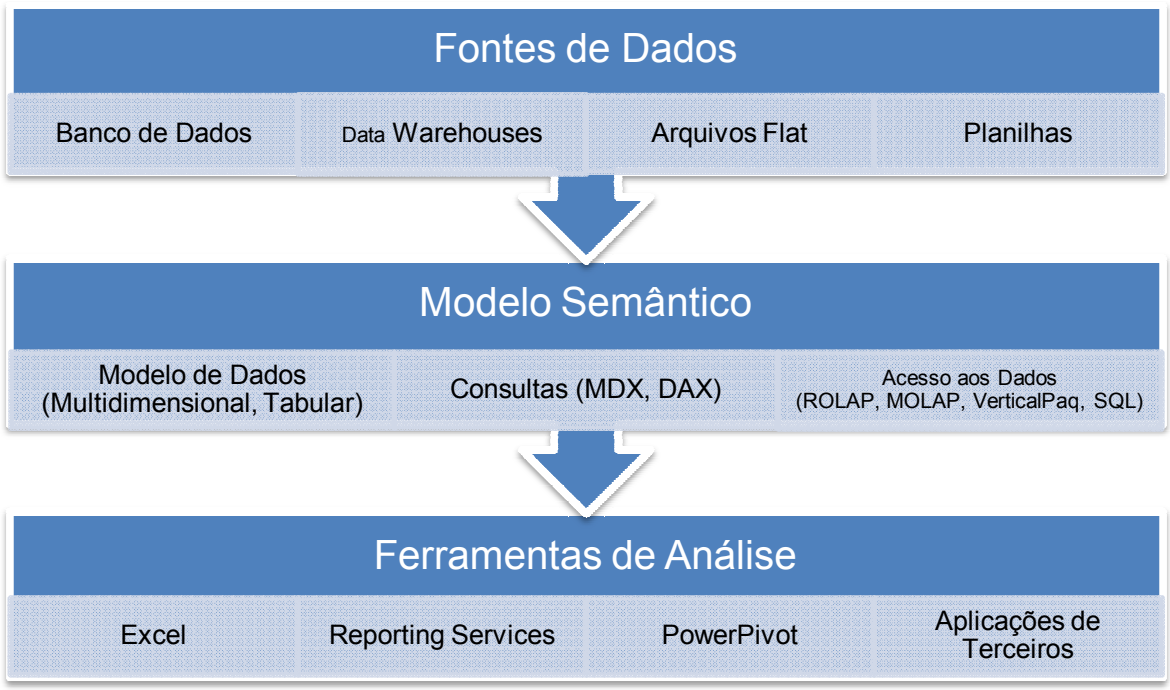


Figura 3: Diagrama da arquitetura da camada semântica do SSAS.

O SSAS também oferece uma plataforma integrada que permite realizar mineração dos dados da camada semântica. Nas próximas seções serão apresentados os algoritmos utilizados no trabalho. Foram selecionados algoritmos da ferramenta SSAS usados para predição de resultados discretos. Para exemplificá-los, foi utilizada uma visão do banco de dados AdventureWorks (CODEPLEX, s.d.) que traz registros classificados de pessoas que compram ou não bicicletas.

3.2.3.1. Algoritmo Microsoft Clustering

O algoritmo Microsoft Clustering é a versão do algoritmo de agrupamento (*clustering*) fornecida pela ferramenta SSAS. O algoritmo usa de técnicas iterativas para

agrupar registros de um banco de dados que contêm características similares em *clusters* ou grupos (MICROSOFT, s.d.a).

Com ele, é possível identificar relacionamentos entre os registros no banco de dados que dificilmente podem ser percebidos através de uma observação casual. É usado principalmente para entendimento dos dados, identificação de anomalias e predições (MICROSOFT, s.d.a).

Abaixo, seguem os passos que explicam o funcionamento do algoritmo:

1. Primeiro ele identifica os relacionamentos no banco de dados e cria uma série de *clusters* baseados nesses relacionamentos;
2. Após criar os *clusters*, o algoritmo calcula uma medida que pontua cada um dos *clusters*. Essa pontuação serve para medir o grau de similaridade entre as instâncias que fazem parte do cluster;
3. Em seguida, o algoritmo busca redefinir os *clusters* para que eles representem melhor os dados;
4. A pontuação entre os novos clusters e os antigos é comparada, sendo descartados os clusters com as piores pontuações;
5. Os passos são repetidos até que não seja mais possível melhorar os resultados das pontuações.

O SSAS permite que o seu algoritmo de agrupamento possa ser configurado para usar os algoritmos EM escalável, EM não escalável, K-Means escalável e K-Means não escalável (MICROSOFT, s.d.a).

Para preparar o modelo de mineração para o uso do Microsoft Clustering Algorithm é preciso fornecer os seguintes atributos:

- **Um atributo chave**: Cada modelo deve conter um atributo numérico ou textual que identifique unicamente cada registro. Chaves compostas não são permitidas;
- **Atributos de entrada**: Cada modelo deve conter ao menos um atributo de entrada que contenham valores que possam ser usados para construir os *clusters*;
- **Um atributo preditivo (Opcional)**: O algoritmo não necessita de um atributo preditivo para construir o modelo de mineração, mas se um atributo for informado, poderá ser usado também como entrada ou apenas para previsão (**PredictOnly**).

A principal diferença entre o Microsoft Clustering e os demais é que não é necessário declarar o atributo preditivo para poder construir um modelo de mineração, pois

ele trabalha estritamente com o relacionamento existente entre os atributos (MICROSOFT, s.d.a).

Depois que o modelo de mineração criado é processado, o SSAS disponibiliza uma janela para visualização dos perfis dos clusters. Para exemplificar usaremos como o exemplo o banco de dados AdventureWorks (CODEPLEX, s.d.). Nesse caso, será criado um modelo usando uma visão do banco de dados que representa os registros de pessoas que compram (Valor 1) ou não (Valor 0) bicicletas, indicados pelo atributo "Bike Buyer". A figura 4 mostra um exemplo de modelo processado. Nele, podemos ver que o Cluster 7 possui uma área maior para o valor 1 do atributo "Bike Buyer", indicando que a maioria daqueles registros são de pessoas que compraram bicicletas e também para o valor "0-1 Miles" do atributo "Commute Distance" (Distância para o trabalho). A partir dessa informação podemos perceber que existe uma relação que indica que pessoas que moram mais perto do trabalho tendem a comprar bicicletas.

Figura 4: Exemplo de perfis de *clusters* gerados pela ferramenta SSAS.

3.2.3.2. Algoritmo Microsoft Decision Trees

O SSAS também oferece a opção de criar árvores de decisão. O algoritmo Microsoft Decision Trees é um algoritmo de classificação e regressão usado para criar modelos de predição de atributos contínuos ou discretos (MICROSOFT, s.d.b).

Para os atributos discretos, o algoritmo faz predições baseadas nos relacionamentos entre os atributos de entrada, usando seus valores, ou estados, na previsão de um atributo previamente estabelecido como preditivo. No caso de atributos contínuos, é usada a técnica de regressão linear para montar a árvore de decisão (MICROSOFT, s.d.b).

O algoritmo constrói um modelo de mineração através da criação de divisões, as quais representam os nós na árvore de decisão. Cada nó possui um teste de condição para um atributo e o percentual de ocorrências de valores do atributo preditivo.

Para exemplificar uma árvore de decisão será usado o exemplo da seção anterior para um modelo de árvores de decisão. A figura 5 representa parte de uma árvore de decisão. Nela, são evidentes alguns padrões. Um deles seria [Number Cars Owned < 1] → [Age < 40] → [Region not "North Amercia"]. Esse padrão possui uma taxa de probabilidade muito alta, como é visto na figura 6, e indica que pessoas que não possuem carros, têm idade inferior a 40 anos e não moram na América do Norte tendem a comprar bicicletas com a probabilidade de 94,37%.

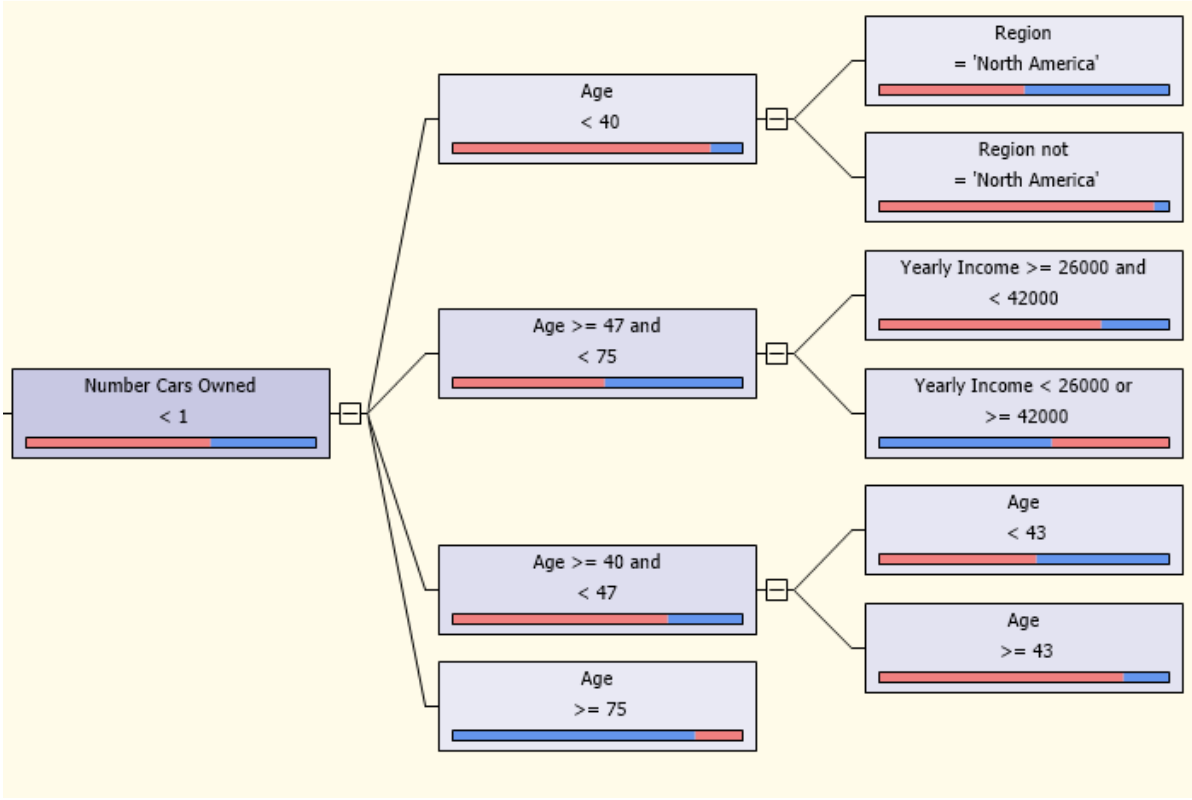


Figura 5: Exemplo de árvore de decisão gerado pela ferramenta SSAS.

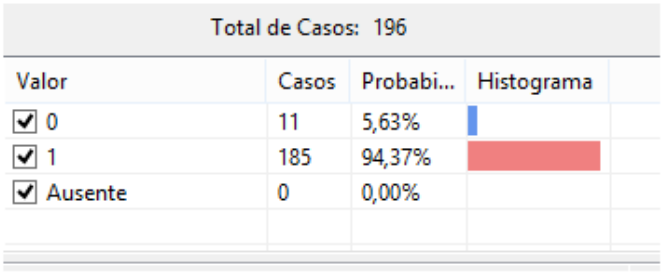


Figura 6: Detalhes do padrão.

Alguns requisitos devem ser considerados para que o modelo de mineração do algoritmo Microsoft Decision Trees possa ser construído. Esses requisitos são semelhantes aos do Microsoft Clustering, apresentados na seção 3.2.3.1. A diferença refere-se à necessidade de especificar um atributo preditivo no caso do Microsoft Decision Trees.

3.2.3.3. Algoritmo Microsoft Naive Bayes

O algoritmo Microsoft Naive Bayes é um algoritmo de classificação, baseado no Teorema de Bayes, usado em tarefas de predição. É o algoritmo que exige menos processamento de todos os algoritmos fornecidos pelo Analysis Services, sendo bastante usado para criação rápida de modelos de mineração. É aconselhável explorar os dados inicialmente com esse algoritmo, para em seguida buscar outros meios mais complexos e exatos de Mineração de Dados (MICROSOFT, s.d.d).

A grande diferença entre o Microsoft Naives Bayes e os outros algoritmos é o fato do Microsoft Naive Bayes necessita que todos os dados sejam discretos. Em casos de dados contínuos, os mesmos devem ser discretizados, ou seja, criando intervalos discretos entre os valores contínuos. Isso também vale para a parte de requisitos de atributos de entrada, na qual o Microsoft Naives Bayes se assemelha ao Microsoft Decision Trees, com a exceção de aceitar apenas atributos discretos ou discretizados como entrada.

Para criar o modelo de mineração, o algoritmo calcula a probabilidade de ocorrência de cada estado dos atributos de entrada para cada estado do atributo preditivo (MICROSOFT, s.d.d).

Seguindo o exemplo anterior que vem sendo usado na exemplificação dos modelos gerados pelos algoritmos, na figura 7 temos uma visualização disponibilizada pelo Analysis Services que mostra quais valores do modelo processado têm melhor pontuação para cada valor do atributo “Bike Buyer”. Mas uma vez, é reforçado pelo modelo de mineração que

pessoas que moram perto do trabalho, ou seja, possuem o valor “0-1 Miles” em “Commute Distance”, tendem a comprar bicicletas, além de outros padrões mostrados.

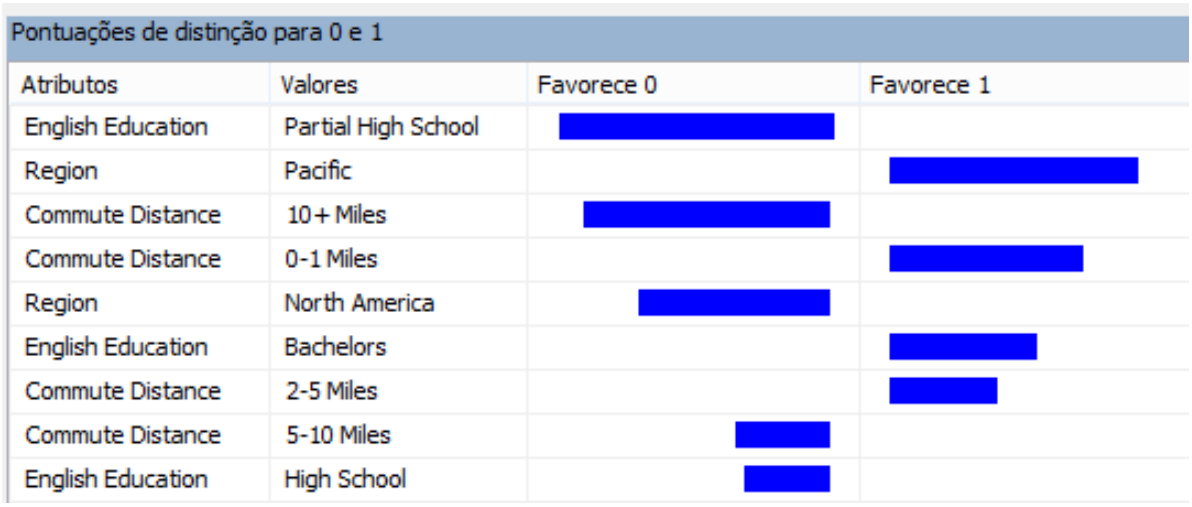


Figura 7: Visualização de Pontuações do Microsoft Naives Bayes.

3.2.3.4. Algoritmo Microsoft Neural Network

Por fim, temos o algoritmo Microsoft Neural Network. Esse algoritmo trabalha com o conceito de Redes Neurais Artificiais (RNA), no qual é usado um conjunto de treinamento contendo dados já classificado para calcular valores usados para tarefas de classificação, predição e regressão (MICROSOFT, s.d.d).

O algoritmo Microsoft Neural Network trabalha com o conceito de camadas para a criação de sua rede neural. Cada camada é formada por uma quantidade de atributos, chamados de neurônios, sendo que os neurônios de uma camada se conectam aos neurônios de outra camada. Essas ligações guardam valores conhecidos como pesos. É partir do cálculo desses pesos que a rede neural consegue identificar padrões (MICROSOFT, s.d.d).

São usadas três camadas de neurônios para formar a RNA usada pelo algoritmo Microsoft Neural Networks. Essas camadas são:

- Camada de entrada: Os neurônios da camada de entrada são definidos pelos valores dos atributos de entrada do modelo de mineração.
- Camada oculta: A camada oculta possui conexões com neurônios da camada de entrada e com neurônios da camada de saída. Os pesos dessas conexões são o que definem a relevância ou importância de cada atributo de entrada.

- Camada de saída: Os neurônios que pertencem à camada de saída correspondem aos valores dos atributos que pretendem ser previstos.

A figura 8, representa o modelo de mineração do algoritmo Microsoft Neural Networks aplicado ao exemplo do AdventureWorks. O Analysis Services disponibiliza uma visualização similar à do Microsoft Naive Bayes, da seção 3.2.3.3. Neste caso, o resultado mostrou-se um pouco diferente do apresentado com o Microsoft Naive Bayes. Apesar disso, foi possível ver uma relação entre os dois. Nesse caso, o valor "10+ Miles" do atributo "Commute Distance" apontou uma tendência muito forte para o valor 0 de "Bike Buyers", ou seja, pessoas que moram longe do trabalho tendem a não comprar bicicletas.

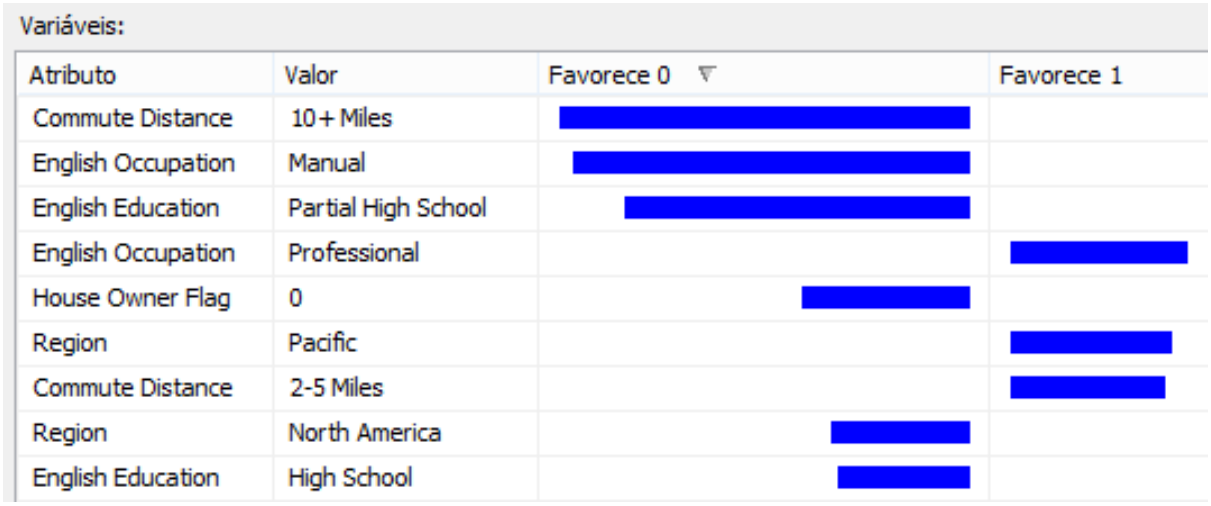


Figura 8: Visualização de pontuações do Microsoft Neural Network.

Os requisitos necessários para gerar o modelo de mineração do Microsoft Neural Networks são os mesmos do Microsoft Clustering, da seção 3.2.3.1.

3.3. Operação

Essa etapa consiste em apresentar a execução do estudo e validação de dados. Na etapa de execução, os dados serão extraídos do site Uol Esporte pelo *wrapper*. Em seguida, uma visão será criada, visando atender todos os requisitos dos algoritmos de predição selecionados. Ao final da etapa de execução, os dados serão usados na construção dos modelos de mineração para cada algoritmo selecionado.

Após a etapa de execução, a etapa de validação de dados será iniciada. O foco principal dessa etapa é buscar melhorias nos modelos de mineração e nos dados extraídos.

3.3.1. Execução

Dando início à etapa de execução, os dados foram extraídos pelo *wrapper* do ambiente ETL. Para cada rodada extraída, uma partida foi escolhida aleatoriamente com o intuito de comparar os dados presentes no ambiente físico com os dados contidos no site. Nenhuma diferença foi identificada entre esses dados comparados, indicando assim, uma extração bem sucedida.

A maioria das rodadas foi coletada com sucesso. Houve apenas um erro durante a extração de dados de duas partidas da 32ª rodada do campeonato de 2012, devido às mesmas não estarem disponível no site. Essas partidas são: Vasco x Internacional, que foi realizada na data 24/10/2012, e Figueirense x Botafogo, realizada também na mesma data.

Após coletar os dados, uma visão foi criada para atender os requisitos dos algoritmos de Mineração de Dados que foram selecionados. Os algoritmos Microsoft Clustering, Microsoft Decision Trees e Microsoft Neural Networks possuem requisitos compatíveis entre si, permitindo que uma visão possa ser usada pelos três. Entretanto, o algoritmo Microsoft Naives Bayes trabalha apenas com valores discretos ou discretizados. Diante desse contexto, é necessário criar uma visão apenas com valores discretos para que todos os algoritmos possam ser atendidos.

De acordo com os dados extraídos, foi criada uma visão com estatísticas dos times agregadas por partida. As estatísticas foram discretizadas em três tipos: maior, menor e igual. Esses tipos foram atribuídos através da comparação entre estatísticas agregadas por partida. Por exemplo, se time A conseguiu efetuar um total de 200 passes e o B, 150, o time A receberá o valor maior em seu atributo passes e o time B receberá o valor menor em passes. Já o valor igual é atribuído para times que possuam valores iguais para as mesmas estatísticas agregadas.

Além disso, a visão possui um atributo chave que identifica unicamente cada instância e um atributo preditivo discreto *label*, que indica vitória (V), empate (E) e derrota (D) do time. A figura 9 mostra parte do resultado de uma consulta SQL (Structured Query Language) realizada com essa visão.

	id_estatistica	cruzamentos	dribles	lancamentos	passes	labe
1	1	MAIOR	IGUAL	MAIOR	MENOR	E
2	2	MENOR	IGUAL	MENOR	MAIOR	E
3	3	MAIOR	MENOR	MAIOR	MAIOR	V
4	4	MENOR	MAIOR	MENOR	MENOR	D
5	5	MENOR	MENOR	MAIOR	MENOR	E
6	6	MAIOR	MAIOR	MENOR	MAIOR	E
7	7	MENOR	MENOR	MENOR	MENOR	E
8	8	MAIOR	MAIOR	MAIOR	MAIOR	E
9	9	MENOR	MENOR	MAIOR	MAIOR	V
10	10	MAIOR	MAIOR	MENOR	MENOR	D
11	11	MENOR	MENOR	MENOR	MENOR	V
12	12	MAIOR	MAIOR	MAIOR	MAIOR	D
13	13	MENOR	MENOR	MAIOR	MENOR	D
14	14	MAIOR	MAIOR	MENOR	MAIOR	V
15	15	MAIOR	MAIOR	MAIOR	MAIOR	D

Figura 9: Resultado de uma consulta à visão criada para o estudo.

Em seguida, foi criado o projeto no Analysis Services para iniciar a construção dos modelos mineração para cada algoritmo. O Analysis Services permite que os dados possam ser divididos em dois conjuntos: teste e treinamento. Para calcular a precisão dos algoritmos selecionados, 20% dos dados foram selecionados pelo Analysis Services para fazer parte do conjunto de teste. O cálculo de precisão do Analysis Services é demonstrado no quadro 6.

$$Precisão = \frac{Quantidade\ de\ Acertos}{Total\ de\ Casos\ de\ Teste}$$

Quadro 6: Fórmula para cálculo da precisão

3.3.2. Validação de dados

Na etapa de validação de dados, os resultados são avaliados em busca de possíveis melhoras na abordagem utilizada. Alguns pontos foram notados e precisaram ser modificados.

Observou-se primeiramente que os casos em que ocorrem empates sempre geravam dados divergentes para um mesmo valor preditivo, ou seja, para cada empate, haveria sempre uma instância com valores do tipo maior enquanto a outra, para as mesmas

estatísticas, os valores seriam do tipo menor. Esse tipo de situação não acrescia nada ao modelo de mineração e por isso todos os empates foram desconsiderados.

O outro ponto observado foi que algumas estatísticas estavam afetando o modelo negativamente por se tratarem de estatísticas diretamente ligadas ao resultado da partida. Para resolver isso, elas foram ignoradas. Essas estatísticas são: assistências, finalizações, finalizações certas, finalizações erradas, finalizações na trave, gols e gols contra.

Após esses ajustes, os modelos de mineração foram aplicados novamente aos dados e os resultados serão vistos no capítulo 5.

4. RESULTADOS

Nesse capítulo, os resultados encontrados serão apresentados. Para facilitar o entendimento, o presente capítulo foi organizados em 5 seções. A seção 4.1 apresenta os padrões e relações encontrados no processamento do modelo de mineração do algoritmo Microsoft Clustering. Seguindo a mesma estrutura, serão apresentados na seção 4.2, 4.3 e 4.4, padrões e relações encontrados pelos modelos de mineração dos algoritmos Microsoft Decision Trees, Microsoft Naives Bayes e Microsoft Neural Networks, respectivamente. Na seção final, seção 4.5, será apresentada a comparação entre a precisão de acertos de cada algoritmo, a qual permite identificar o algoritmo que obteve a melhor precisão.

4.1. Algoritmo Microsoft Clustering

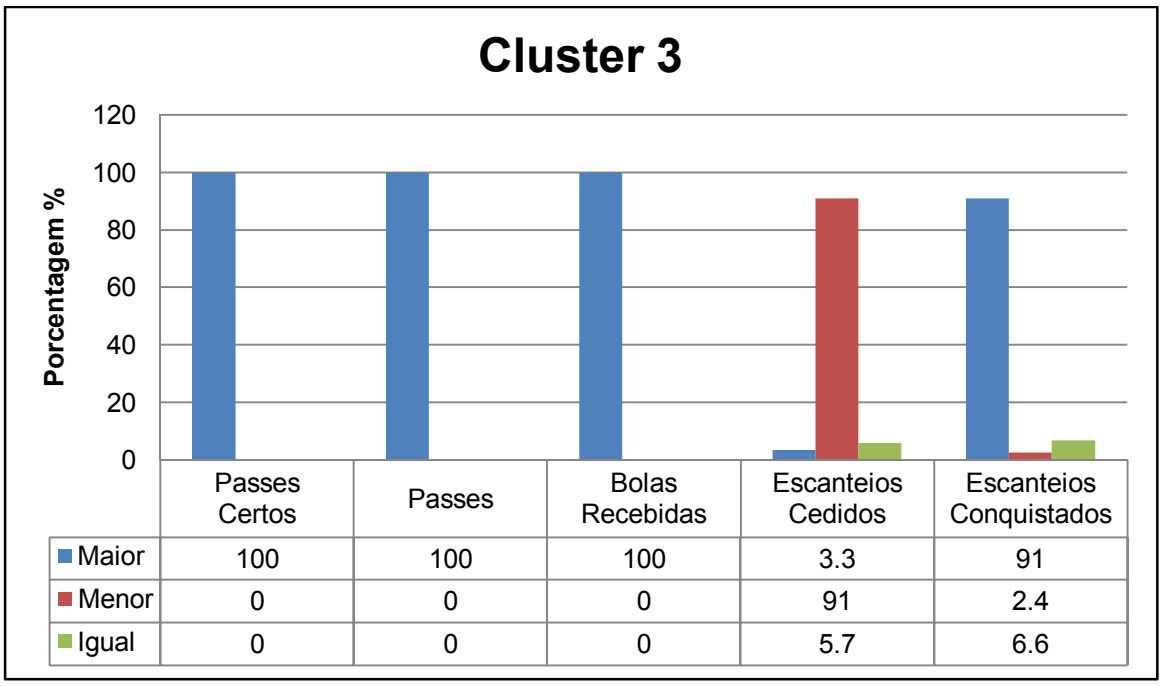
O modelo de mineração, após ser processado, gerou 10 *clusters* que foram analisados para identificar os *clusters* que possuíam os maiores índices de vitórias e derrotas. Na tabela 1, é possível ver a porcentagem de ocorrências de instâncias tanto para a classe de vitória, quanto para a classe de derrota nos 10 *clusters* criados pelo Analysis Services, além da quantidade de instâncias em cada *cluster*. Após essa análise, dois *clusters* foram selecionados por possuírem os melhores índices de vitórias e derrotas. Os *clusters* selecionados foram o Cluster 3 e o Cluster 4.

Tabela 1: Distribuição de vitórias e derrotas por *cluster*.

Nome do Cluster	Quantidade	Vitórias %	Derrotas %
Cluster 1	194	54,00	46,00
Cluster 2	181	42,90	57,10
Cluster 3	172	29,40	70,60
Cluster 4	158	68,00	32,00
Cluster 5	131	39,70	60,30
Cluster 6	125	62,80	37,20
Cluster 7	98	50,10	49,90
Cluster 8	81	62,00	38,00
Cluster 9	72	44,80	55,20
Cluster 10	77	44,90	55,10

.

No Cluster 3, com um total de 172 instâncias, 70,60 % das suas instâncias foram derrotadas, enquanto no Cluster 4, que possui 158 instâncias, 68,00 % ganharam a partida. A partir disso, foram gerados gráficos com os padrões mais freqüentes de estatísticas para cada *cluster*. No gráfico 1, que representa o Cluster 3, podemos ver que existe uma tendência muito expressiva que indica que os times que perdem a partida realizam um maior número de passes, recebem mais a bola, cedem menos escanteios e conquistam mais escanteios.



	Passes Certos	Passes	Bolas Recebidas	Escanteios Cedidos	Escanteios Conquistados
Maior	100	100	100	3.3	91
Menor	0	0	0	91	2.4
Igual	0	0	0	5.7	6.6

Gráfico 1: Padrões do Cluster 3.

Já no gráfico 2, que representa o Cluster 4, pode-se perceber que, por se tratar de uma classe inversa à do Cluster 3, alguns padrões também foram inversos. Por exemplo, no Cluster 3 um número de passes maior numa partida representa uma alta probabilidade de derrota, enquanto aqui, o oposto é encontrado, o número de passes menor indica vitória.

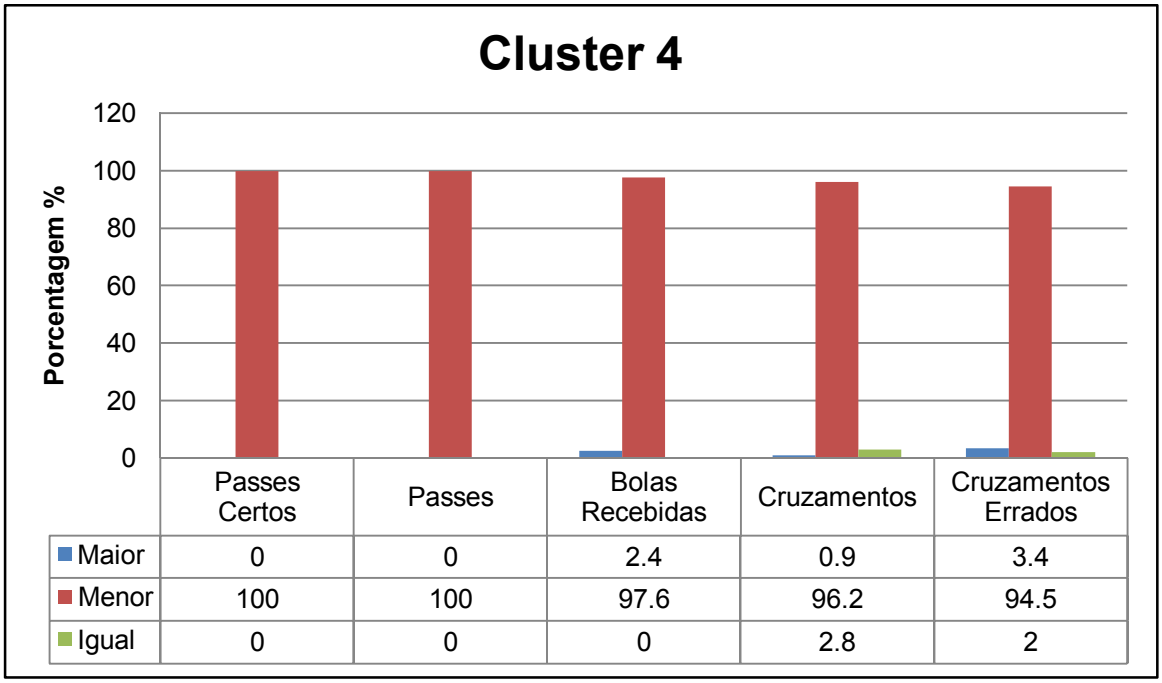


Gráfico 2: Padrões do Cluster 4.

4.2. Algoritmo Microsoft Decision Trees

Outro modelo de mineração analisado foi o modelo criado para o algoritmo Microsoft Decision Trees. O modelo gerou uma árvore de decisão que foi dividida em duas partes para ser apresentada no trabalho. Nela, a cor azul representa a classe derrota, enquanto a cor vermelha representa a classe vitória.

A primeira parte da árvore de decisão pode ser vista na figura 10. Nela podemos ver dois caminhos com grande distribuição de instâncias tanto para a classe vitória, quanto para a classe derrota. O primeiro caminho, representado por [Bolas Recebidas = 'MAIOR'] → [Dribles not = "MENOR"] → [Passes Certos = 'MENOR'], possui apenas 13 instâncias, sendo assim, não possuem um grau de influência adequado para a inferência de qualquer relação.

O segundo caminho, representado por [Bolas Recebidas = 'MAIOR'] → [Dribles = "MENOR"] → [Lancamentos Certos = 'Menor], possui um total de 93 instâncias, sendo que 83,78 % delas são da classe derrota. É seguro dizer, a partir dessa observação, que times que recebem mais a bola, driblam menos e fazem menos lançamentos certos possuem uma probabilidade alta de perder uma partida.

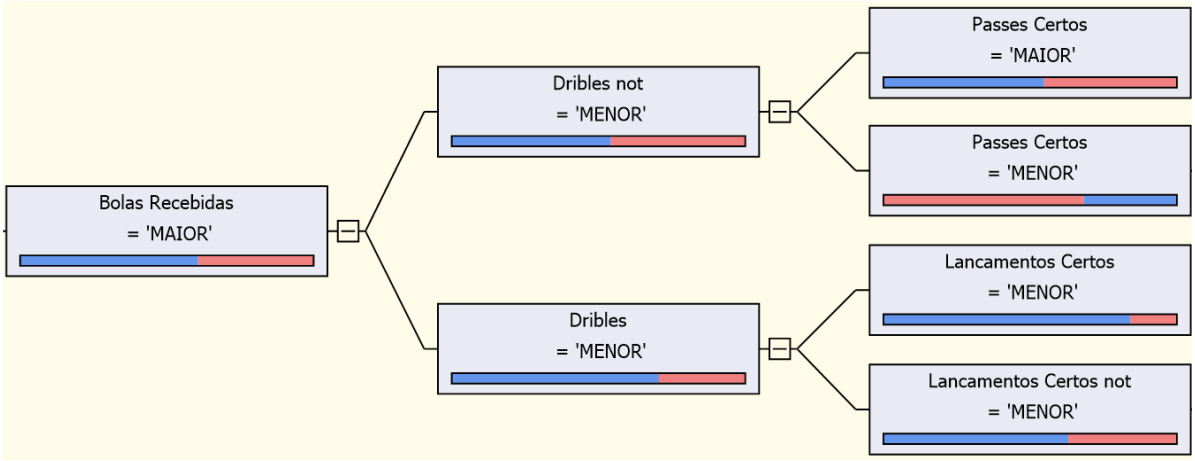


Figura 10: Primeira parte da árvore de decisão

A figura 11 representa a segunda parte da árvore de decisão gerada pelo Analysis Services. É possível ver um caminho o qual possui uma quantidade maior de distribuição de instâncias para a classe vitória. O caminho [Bolas Recebidas = 'MENOR'] → [Desarmes not = 'MENOR'] → [Bolas Perdidas not = 'MAIOR'] possui 202 casos, sendo que, 74,97 % desses casos são de instâncias da classe vitória.

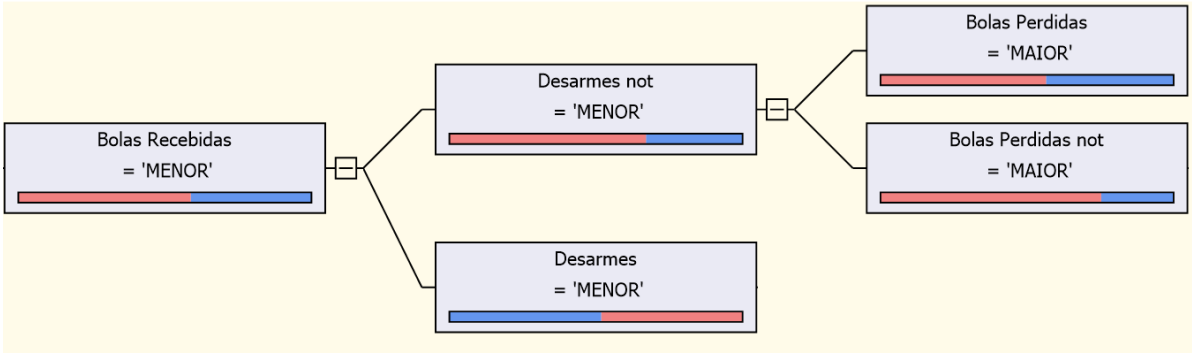


Figura 11: Segunda parte da árvore de decisão

4.3. Algoritmo Microsoft Naives Bayes

O modelo de mineração do algoritmo Microsoft Naives Bayes destacou quatro estatísticas que mais influenciaram na predição do resultado das partidas. Na figura 12, podemos perceber os times que receberam menos a bola (59,10%), desarmaram mais (58,10%) e realizaram menos passes (60,00%) ganharam mais partidas.

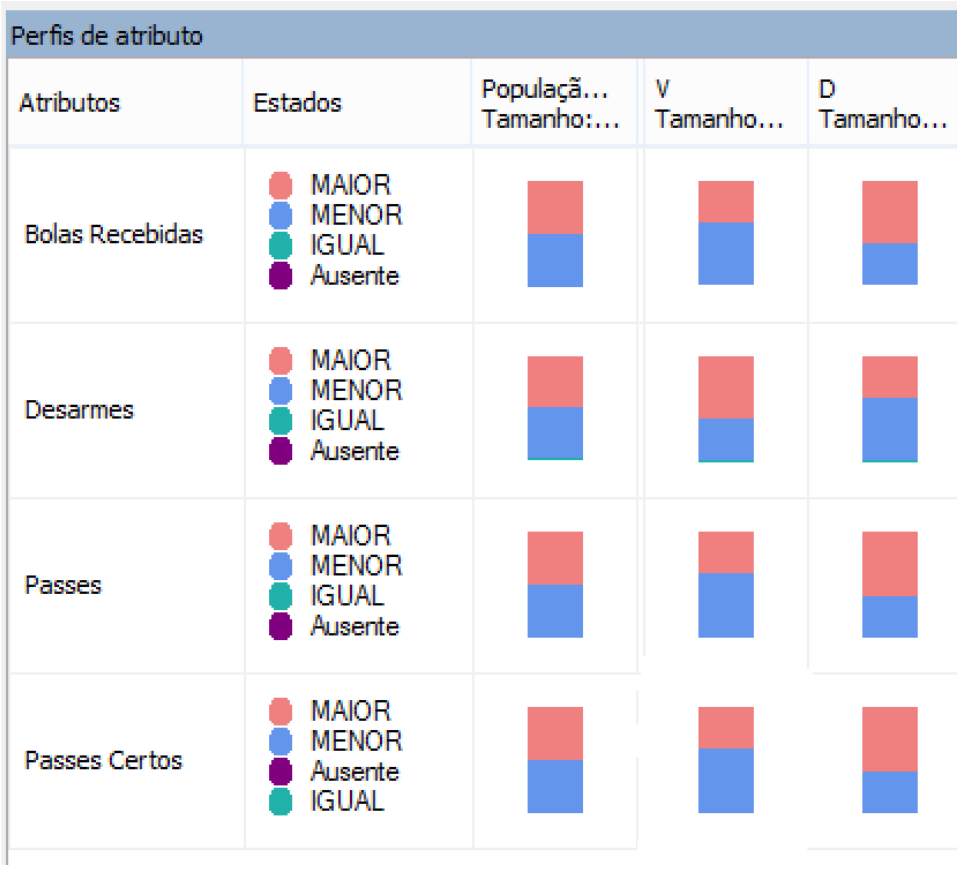


Figura 12: Perfis de atributo gerado pelo modelo do Microsoft Naive Bayes

4.4. Algoritmo Microsoft Neural Network

Finalizando as análises individuais, temos no gráfico 3, os padrões encontrados pelo modelo de mineração do Microsoft Neural Network e na tabela 2, informações sobre as probabilidades de ocorrência das instâncias tanto para classe que representa as vitórias, como para classe que representa as derrotas. Nota-se que times que fazem menos defesas e erram menos cruzamentos têm mais chances de ganhar uma partida, enquanto times que erram mais cruzamentos e recebem mais a bola tendem a perder a partida.

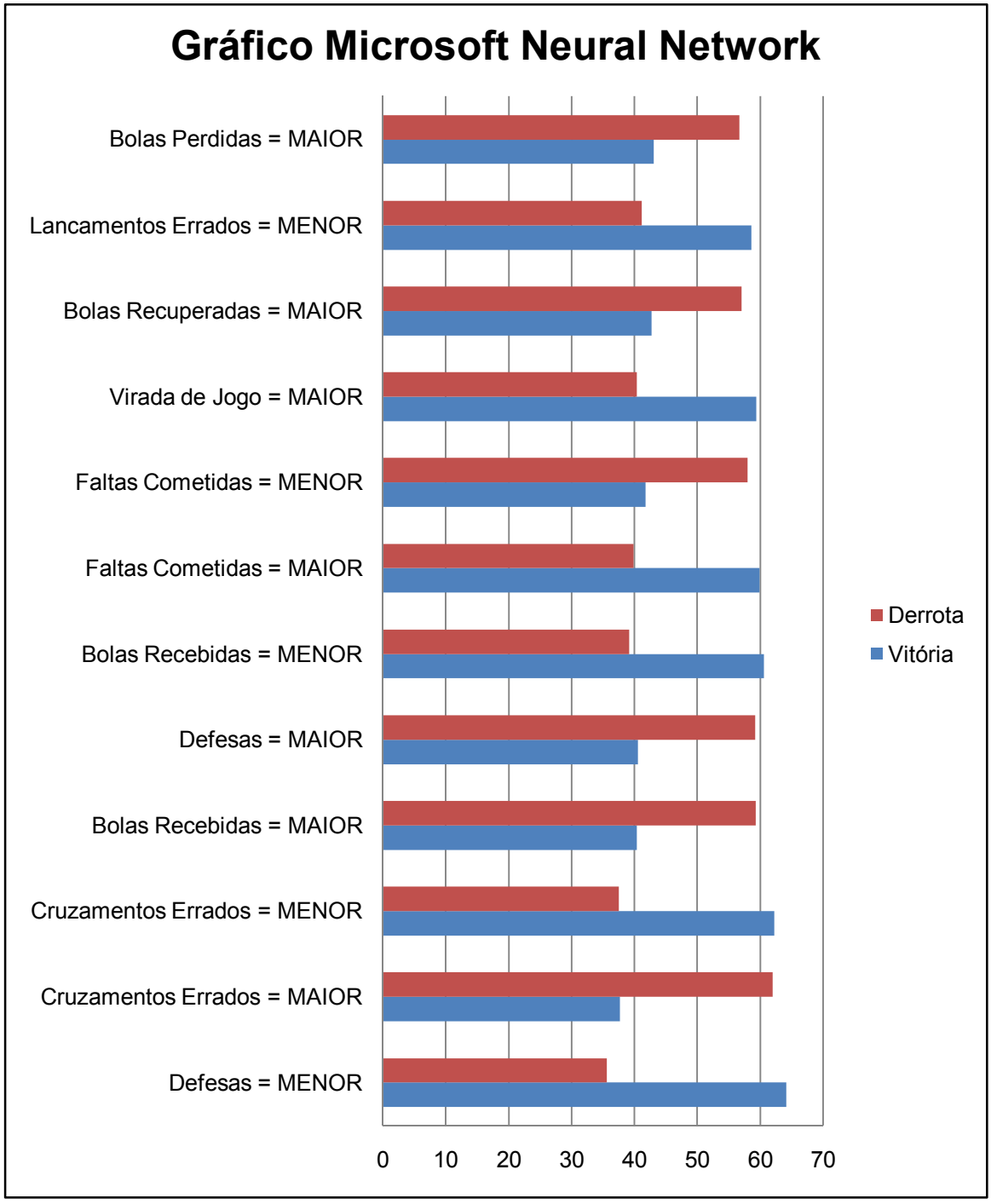


Gráfico 3: Padrões encontrados pelo Microsoft Neural Network

Tabela 2: Dados sobre padrões encontrados pelo Microsoft Neural Network

Nome da Estatística	Valor	Vitórias %	Derrotas %
Defesas	MENOR	64,21	35,57
Cruzamentos Errados	MAIOR	37,73	62,05
Cruzamentos Errados	MENOR	62,29	37,49
Bolas Recebidas	MAIOR	40,42	59,36
Defesas	MAIOR	40,56	59,22
Bolas Recebidas	MENOR	60,64	39,13
Faltas Cometidas	MAIOR	59,89	39,89
Faltas Cometidas	MENOR	41,79	57,99
Virada de Jogo	MAIOR	59,4	40,38
Bolas Recuperadas	MAIOR	42,74	57,04

4.5. Análise da Precisão dos Algoritmos Selecionados

A análise de precisão serve para medir a qualidade do modelo de mineração em prever casos reais que não fazem parte do conjunto de treinamento. Como foi visto na seção 3.3.1, 20% dos dados extraídos da fonte de dados *Web* foram separados em um conjunto de teste que seria usado futuramente na verificação de precisão dos modelos de mineração.

Na tabela 3, é possível ver a precisão dos modelos de mineração para predição de vitórias e derrotas. Como pode ser visto, o algoritmo Microsoft Naives Bayes obteve melhor precisão de acerto de vitórias com 66,86 %, e também obteve melhor precisão para derrotas com 61,21%.

Tabela 3: Tabela de precisão dos modelos de mineração

Nome do Algoritmo	% de acerto em vitórias	% de acerto em derrotas
Microsoft Clustering	65,59	52,10
Microsoft Decision Trees	61,96	55,95
Microsoft Naive Bayes	66,86	61,21
Microsoft Neural Networks	65,14	60,89

5. CONCLUSÃO

O uso de estatísticas para apoiar decisões táticas de equipes de futebol ainda está em fase inicial e necessita de muita pesquisa para que mais resultados venham a surgir. Ainda são poucos os trabalhos acadêmicos que trabalham nessa linha de pesquisa, mas os mesmos vêm mostrando que existem informações desconhecidas que podem trazer uma vantagem competitiva para as equipes.

O presente trabalho foi proposto com o objetivo de explorar o uso mineração de dados no contexto do Campeonato Brasileiro de Futebol. Com esse estudo foi possível verificar que o processo de mineração de dados é viável, pois com apenas uma única visão de dados foi possível conseguir diversas informações.

O uso de vários algoritmos também possibilitou observar diferentes tipos de padrões. Os algoritmos chegaram a conseguir um taxa de precisão de aproximadamente 67,00 %. Essa taxa pode se considerada boa, tendo em vista que as estatísticas selecionadas não estavam diretamente relacionadas ao resultado das partidas.

As estatísticas relacionadas ao número de passes, bolas recebidas, bolas recuperadas, desarmes e cruzamentos foram as que mais se destacaram, mostrando que possuem influência no resultado das partidas.

5.1. Trabalhos Futuros

Com base no trabalho desenvolvido, algumas sugestões para trabalhos futuros são:

- Buscar padrões em um nível maior de detalhe, como por exemplo, posição e jogador;
- Aplicar técnicas de mineração em base de dados mais detalhadas, que envolvam eventos temporais. Entretanto, como não foram encontradas bases desse tipo durante a pesquisa, é possível que no momento só existam bases proprietárias para essa situação;
- Criar um processo específico para mineração de dados de partidas de futebol. Um processo bem definido com tarefas claras e objetivas pode auxiliar bastante, eliminando boa parte do trabalho que existe em usar processos genéricos;

- Propor um modelo de mineração que, a partir de um conjunto de treinamento com jogadores com uma determinada característica (bons zagueiros, artilheiros, etc.), possa encontrar jogadores similares. Com uma ferramenta desse tipo, as equipes poderão agilizar o processo de pesquisa na hora de comprar jogadores.

3. REFERÊNCIAS BIBLIOGRÁFICAS

ANDERSON, C.; SALLY, D. **Os Números do Jogo - Porque Tudo o Que Você Sabe Sobre Futebol Está Errado**. Tradução André Fontenelle. [S.l.]: Paralela, 2013.

CODEPLEX. **AdventureWorks Databases**. Disponível em: <http://msftdbprodsamples.codeplex.com/>. Acesso em: 6 fev. 2014.

COLAÇO JÚNIOR, M. **Projetando sistemas de apoio à decisão baseados em data warehouse**. Rio de Janeiro: Axcel Books, 2004.

COSTA, C. F. S. **Análise das Acções Ofensivas com Finalização Resultantes de Jogo Dinâmico.** Dissertação (Mestrado em Treino Desportivo para Crianças e Jovens, Especialidade de Ciência do Desporto) - Faculdade de Ciências do Desporto e Educação Física. Coimbra: Universidade de Coimbra, 2010.

FARIAS, F. F. **Análise e Previsão de Resultados de Partidas de Futebol**. Dissertação (Mestrado em Estatística) – Instituto de Matemática. Rio de Janeiro: Universidade Federal do Rio de Janeiro, 2008.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **From data mining to knowledge discovery in databases**. AI magazine, v. 17, n. 3, p. 37 - 54, 1996.

GORUNESCU, F. **Data Mining**: Concepts, Models and Techniques. Criaova: Springer, 2011 (Intelligent Systems Reference Library (Book 12)).

HAN, J.; KAMBER, M.; PEI, J. **Data Mining**: Concepts and Techniques. 3. ed. Massachusetts: Morgan Kaufmann Publishers, 2011.

HAND, D. J.; MANNILA, H.; SMYTH, P. **Principles of Data Mining**. Massachusetts: A Bradford Book The MIT Press, 2001 (Adaptive Computation and Machine Learning Series).

JSOUP. **Java HTML Parser**. Disponível em: <http://jsoup.org/>. Acesso em: 12 fev. 2013.

KUPER, S.; SZYMANSKI, S. **Soccernomics**. New York: Nation Books, 2011.

LAROSE, T. D. **Discovering Knowledge in Data**: An Introduction To Data Mining. New Jersey: Wiley Publishing, 2005.

LIU, B. **Web Data Mining**: Exploring Hyperlinks, Contents, and Usage Data. 2. ed. Chicago: Springer, 2011.

MICROSOFT. **Microsoft Clustering Algorithm**. Disponível em: <http://technet.microsoft.com/en-us/library/ms174879%28v=sql.100%29.aspx>. Acesso em: 5 fev. 2014.

_____. **Microsoft Decision Trees Algorithm**. Disponível em: <http://technet.microsoft.com/en-us/library/ms175312%28v=sql.100%29.aspx>. Acesso em: 5 fev. 2014.

_____. **Microsoft Naive Bayes Algorithm**. Disponível em: <http://technet.microsoft.com/en-us/library/ms174806%28v=sql.100%29.aspx>. Acesso em: 4 jan. 2014.

_____. **Microsoft Neural Network Algorithm**. Disponível em: <http://technet.microsoft.com/en-us/library/ms174941%28v=sql.100%29.aspx>. Acesso em: 4 jan. 2014.

_____. **SQL Server Analysis Services**. Disponível em: <http://technet.microsoft.com/pt-br/library/bb522607.aspx>. Acesso em: 3 jan. 2014.

NUNES, S.; SOUSA, M. **Applying Data Mining Techniques to Football Data from European Championships.** In: 1ª Conferência de Metodologias de Investigação Científica, 2006, Porto, p. 4-16.

SCHUMAKER, R. P.; SOLIEMAN, O. K.; CHEN, H. **Sports Data Mining**. New York: Integrated Series in Information Systems, 2010.

SLOAN ANALYTICS CONFERENCE. **About the conference**. Disponível em: <http://www.sloansportsconference.com/?page_id=1851>. Acesso em: 5 fev. 2014.

UOL ESPORTE. **Campeonato Brasileiro de Futebol**. Disponível em: <http://esporte.uol.com.br/futebol/campeonatos/brasileirao/jogos/>. Acesso em: 15 mar. 2013.

WITTEN, I. H.; FRANK, E.; HALL, M. A**. Data Mining: Practical Machine Learning Tools and Techniques**. 3. ed. Massachusetts: Morgan Kaufmann Publishers, 2011.