

FACULDADE PROFESSOR MIGUEL ÂNGELO DA SILVA SANTOS – FeMASS

CURSO DE GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

Processo de KDD com a utilização do algoritmo *K-Means* em um banco de dados sobre o futebol Europeu

POR:

JARDEL MARSAGLIA CASTELUBER

MACAÉ

2019

FACULDADE PROFESSOR MIGUEL ÂNGELO DA SILVA SANTOS – FeMASS

CURSO DE GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

Jardel Marsaglia Casteluber

PROCESSO DE KDD COM A UTILIZAÇÃO DO ALGORITMO *K-MEANS* EM UM BANCO DE DADOS SOBRE O FUTEBOL EUROPEU

Trabalho Final apresentado ao curso de graduação em Sistemas de Informação, da Faculdade Professor Miguel Ângelo da Silva Santos (FeMASS), para obtenção do grau de BACHAREL em Sistemas de Informação

Professor Orientador: Isac Mendes Lacerda, M.e

MACAÉ/RJ

2019

JARDEL MARSAGLIA CASTELUBER

PROCESSO DE KDD COM A UTILIZAÇÃO DO ALGORITMO *K-MEANS* EM UM BANCO DE DADOS SOBRE O FUTEBOL EUROPEU

Trabalho de Conclusão de Curso apresentado ao curso de graduação em Sistemas de Informação, da Faculdade Professor Miguel Ângelo da Silva Santos (FeMASS), para obtenção do grau de BACHAREL em Sistemas de Informação.

Aprovada em \_\_\_ de \_\_\_\_\_\_\_ de 2019

BANCA EXAMINADORA

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Isac Mendes Lacerda, M.e

Faculdade Professor Miguel Ângelo da Silva Santos (FeMASS)

1º Examinador

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Anselmo Pestana Ribeiro Costa, M.e

Faculdade Professor Miguel Ângelo da Silva Santos (FeMASS)

2º Examinador

**EPÍGRAFE**

“A persistência é o menor caminho do êxito”. (Charles Chaplin)

**RESUMO**

Na era da informação, obter conhecimento sobre dados das mais variadas fontes é de grande valia para organizações. Neste trabalho justificou-se a relevância do tema para o período atual da sociedade, foi explicado o conceito geral de Mineração de Dados, como funciona o processo de Descoberta de Conhecimento em Banco de Dados, foi descrita técnicas de mineração e onde suas abordagens se aplicam e também foi apresentado o algoritmo *K-Means* que serviu de suporte para o experimento aqui realizado demonstrando toda a implementação necessária para validar o conteúdo abordado ao longo do trabalho. Para a demonstração do processo de KDD presente no capítulo 3 foi utilizado o banco de dados *European Soccer Database* que forneceu mais de dez mil registros para alimentar o algoritmo, o que possibilitou obter resultados confiáveis.

Palavras-chave: Banco de Dados, Mineração de Dados, *K-Means*,

**ABSTRACT**

In the information age, gaining knowledge about data from variety of sources is of great value to organizations. This work justified the relevance of the theme to the current period of society, explained the general concept of Data Mining, how the process of Knowledge Discovery in Database works, described mining techniques and where their approaches apply. It was also presented the *K-Means* algorithm that supported the experiment performed here demonstrating all the implementation necessary to validate the content covered throughout the work. For the demonstration of the KDD process present in chapter 3, the European Soccer Database was used, which provided more than ten thousand records to feed the algorithm, which allowed to obtain reliable results.

Key words: Database, Data-mining, *K-Means*.

**LISTA DE FIGURAS**

[Figura 1: Etapas do processo de KDD. Fonte: Adaptação de Fayyad (1996) 19](#_Toc25835485)

[Figura 2: Etapas de KDD. Fonte: Elaboração própria 21](#_Toc25835486)

[Figura 3: Algoritmo *K-Means*. Fonte: Elaboração própria 29](#_Toc25835487)

[Figura 4: Primeiro passo *K-Means*. Fonte: Elaboração própria 30](#_Toc25835488)

[Figura 5: Segundo passo *K-Means*. Fonte: Elaboração própria 31](#_Toc25835489)

[Figura 6: Terceiro passo *K-Means*. Fonte: Do Autor 31](#_Toc25835490)

[Figura 7: Tabelas do Banco de Dados. Fonte: www.kaggle.com/hugomathien/soccer 33](file:///Z:\TCC%202%20Jardel%20Casteluber.docx#_Toc25835491)

[Figura 8: Bibliotecas utilizadas. Fonte: Elaboração própria 34](#_Toc25835492)

[Figura 9: Amostra da tabela *Player*. Fonte: Elaboração própria 34](#_Toc25835493)

[Figura 10: Amostra da tabela *Player\_Attributes*. Fonte: Elaboração própria 35](#_Toc25835494)

[Figura 11: Seleção de dados. Fonte: Elaboração própria 35](#_Toc25835495)

[Figura 12: Limpeza dos dados. Fonte: Elaboração própria 36](#_Toc25835496)

[Figura 13: Removendo de valores nulos. Fonte: Elaboração própria 36](#_Toc25835497)

[Figura 14: Representação do DF transformado. Fonte: Elaboração própria 36](#_Toc25835498)

[Figura 15: Configuração do *K-Means*. Fonte: Elaboração própria 37](#_Toc25835499)

[Figura 16: Centroides iniciais. Fonte: Elaboração própria 38](#_Toc25835500)

[Figura 17: Inicialização do *K-Means*. Fonte: Elaboração própria 38](#_Toc25835501)

[Figura 18: Novos centroides. Fonte: Elaboração própria 39](#_Toc25835502)

[Figura 19: Reinicialização do K-Means. Fonte: Elaboração própria 39](#_Toc25835503)

[Figura 20: Centroides durante iterações. Fonte: Elaboração própria 40](#_Toc25835504)

[Figura 21: Saída gerada pelo K-Means. Fonte: Elaboração própria 41](#_Toc25835505)

[Figura 22: Divisão dos clusters. Fonte: Elaboração própria 41](#_Toc25835506)

[Figura 23: Representação por cluster. Fonte: Elaboração própria 41](#_Toc25835507)

[Figura 24: Registros do cluster 0. Fonte: Elaboração própria 42](#_Toc25835508)

[Figura 25: Jogadores. Fonte: pesquisa do nome em www.google.com 42](#_Toc25835509)

[Figura 26: Registros do cluster 1. Fonte: Elaboração própria 43](#_Toc25835510)

[Figura 27: Jogadores. Fonte: Pesquisa do nome em www.google.com 43](#_Toc25835511)

[Figura 28: Registros do cluster 2. Fonte: Elaboração própria 43](#_Toc25835512)

[Figura 29: Jogadores. Fonte: Pesquisa do nome em www.google.com 43](#_Toc25835513)

[Figura 30: Registros do cluster 3. Fonte: Elaboração própria 44](#_Toc25835514)

[Figura 31: Jogadores. Fonte: Pesquisa do nome em www.google.com 44](#_Toc25835515)

[Figura 32: Amostra do grupo "meio-de-campo". Fonte: Elaboração própria 45](#_Toc25835516)

[Figura 33: Jogadores. Fonte: Pesquisa do nome em www.google.com 45](#_Toc25835517)

[Figura 34: Amostra do grupo "defesa". Fonte: Elaboração própria 45](#_Toc25835518)

[Figura 35: Jogadores. Fonte: Pesquisa do nome em www.google.com 46](#_Toc25835519)

[Figura 36: Amostra do grupo "ataque". Fonte: Elaboração própria 46](#_Toc25835520)

[Figura 37: Jogadores. Fonte: Pesquisa do nome em www.google.com 46](#_Toc25835521)

[Figura 38: Amostra do grupo "goleiro". Fonte: Elaboração própria 47](#_Toc25835522)

[Figura 39: Jogadores. Fonte: Pesquisa do nome no sitio www.google.com 47](#_Toc25835523)

[Figura 40: Comparação de atributos. Fonte: Elaboração própria 48](#_Toc25835524)

[Figura 41: Comparação entre atributos. Fonte: Elaboração própria 48](#_Toc25835525)

**LISTA DE TABELAS**

[Tabela 1: Comparação entre definições para KDD definidas por Fayyad (1996) e Adriaans e Zantinge (1997). Fonte: Autor. 22](#_Toc24831268)

**SUMÁRIO**

[1 INTRODUÇÃO 12](#_Toc24831371)

[1.1 OBJETIVOS 13](#_Toc24831372)

[1.1.1 OBJETIVO GERAL 13](#_Toc24831373)

[1.1.2 OBJETIVOS ESPECÍFICOS 13](#_Toc24831374)

[1.2 JUSTIFICATIVA 13](#_Toc24831375)

[1.3 METODOLOGIA DE PESQUISA 14](#_Toc24831376)

[2 MINERAÇÃO DE DADOS 17](#_Toc24831377)

[2.1 PROCESSO DE KDD - *knowledge discovery in database* 18](#_Toc24831378)

[2.1.1 ETAPAS SEGUNDO FAYYAD 18](#_Toc24831379)

[2.1.2 ETAPAS SEGUNDO ADRIAANS e ZANTINGE 21](#_Toc24831380)

[2.1.3 COMPARAÇÃO ENTRE AS DUAS ABORDAGENS 22](#_Toc24831381)

[2.2 MÉTODOS E TÉCNICAS DE MINERAÇÃO DE DADOS 23](#_Toc24831382)

[2.2.1 TÉCNICAS DE MINERAÇÃO DE DADOS 25](#_Toc24831383)

[2.3 K-Means 28](#_Toc24831384)

[3 DEMONSTRAÇÃO DO PROCESSO DE KDD 33](#_Toc24831385)

[3.1 Base de Dados 33](#_Toc24831386)

[3.2 Ambiente de Trabalho 34](#_Toc24831387)

[3.3 Seleção de Dados 34](#_Toc24831388)

[3.4 Pré-Processamento e Limpeza 35](#_Toc24831389)

[3.5 Transformação 36](#_Toc24831390)

[3.6 Mineração de Dados 37](#_Toc24831391)

[3.7 Interpretação dos Resultados 40](#_Toc24831392)

[4 CONSIDERAÇÕES FINAIS 49](#_Toc24831393)

[REFERÊNCIAS BIBLIOGRÁFICAS 51](#_Toc24831394)

[APÊNDICE 54](#_Toc24831395)

# INTRODUÇÃO

Há alguns anos a quantidade de dados gerados pelas mais variadas fontes de dados possíveis está muito além do que a capacidade humana consegue interpretar. Com a queda nos preços de *hardwares* responsáveis pelo armazenamento de dados, foi possível guardar informações históricas que antigamente era impraticável devido ao elevado preço de dispositivos de armazenamento em relação a capacidade que os mesmos ofereciam. Precisou-se criar novas estruturas de armazenamento e sistemas que pudessem ser capazes de gerenciar essa quantidade crescente de dados, além dos já conhecidos Sistemas Gerenciadores de Banco de Dados (SGBD).

Mas o que fazer com esses dados, como extrair informações no contexto onde eles estão inseridos?

Os meios tradicionais de análise de dados não eram adequados e um novo conceito foi se popularizado cada vez mais como meio de descobrir informações e conhecimento trabalhando sobre essa abundancia de dados, basicamente com a finalidade de guiar decisões humanas. Este conceito recebeu o nome de *Data Mining* (Mineração de Dados), que se fundamenta na ideia de adquirir conhecimento e descobrir tendências, baseadas na análise em busca de padrões nos dados armazenados.

O processo de Mineração de Dados é composto por etapas que definem um ciclo de *Knowledge Discovery in Database* (KDD), que é apresentado ao longo desse trabalho, tais como: Seleção, Pré-Processamento, Transformação, Mineração de Dados e Avaliação. Ao final de todas essas etapas, é gerado um conhecimento em cima dos dados e entrada.

O termo Mineração de Dados é utilizado para definir tanto o processo geral como uma etapa específica do processo de KDD. Neste trabalho, quando se está falando sobre o processo geral, utilizou-se a sigla MD e enquanto etapa, foi utilizada a terminologia sem abreviatura. Afim de oferecer a melhor compreensão do processo, o trabalho apresenta inicialmente o conceito de MD, em seguida é descrito o processo de KDD para então demonstrar a Mineração de Dados como etapa.

O foco deste trabalho será na etapa de MD para descoberta de conhecimento, utilizando o algoritmo de mineração de dados não supervisionado K-Means implementado na linguagem de programação Python, em um banco de dados público sobre o futebol Europeu, com o intuito de descobrir informações o que aos olhos humanos e seria praticamente impossível de descobrir sem auxílio da tecnologia.

## OBJETIVOS

### OBJETIVO GERAL

Demonstrar uma abordagem de KDD com utilização do algoritmo *K-Means* em um banco de dados sobre o futebol Europeu.

### OBJETIVOS ESPECÍFICOS

✓ Descrever e demonstrar o passo a passo do processo de KDD;

✓ Identificar e descrever uma base de dados escolhida ao tratamento do algoritmo;

✓ Implementar o algoritmo *K-Means* em *Python* e realizar as simulações;

✓ Analisar e descrever os resultados obtidos

## JUSTIFICATIVA

A informação sempre desempenhou um papel fundamental em todos os setores da sociedade. O acúmulo contínuo de dados abre espaço para extração de muito conhecimento implícito. Para isso, é necessário aplicar métodos para tais fins.

Organizar dados em grupos é uma das formas mais naturais de compreensão e aprendizagem (JAIN; DUBES, 1988). Porém, esse problema torna-se difícil à medida que a quantidade de objetos a serem agrupados aumentam, neste caso é necessário o auxílio de ferramentas computacionais para se trabalhar sobre dados.

O agrupamento de dados dispersos através de semelhanças compartilhadas entre si, denominado de “clusterização”, é útil para identificar perfis, grupos ou nichos e revelar informações implícitas. Esse novo conhecimento revelado é utilizado para guiar decisões sobre perguntas pertinentes ao campo no qual se aplicam os dados analisados.

O algoritmo *K-Means* estudado neste trabalho é responsável por fazer o agrupamento dos dados presentes em um banco de dados. A ideia é que este estudo abra caminhos para novas aplicações em outros bancos de dados, ao demonstrar como é o passo a passo para a execução do algoritmo.

Este estudo abre espaço para a aplicação de técnicas de mineração de dados em diversos setores econômicos e sociais, tornando-se muito atraente para descoberta de conhecimento por vários motivos:

* O volume de dados é enorme: MD aplica-se a grandes massas de dados. Empresas de telefonia, cartões de crédito, bancos, televisão por assinatura, comércio eletrônico, por exemplo, produzem grandes volumes de dados sobre seus negócios, que são passíveis de análises para gerar algum tipo de ação.
* Os dados estão organizados: Os *Data Warehouses* disponibilizam as mesmas ferramentas de consulta que o sistema de banco de dados operacional e mais facilidades: os dados podem ser estruturados em uma estrutura de um cubo que é uma forma que facilita a análise dos dados. Esta estrutura é chamada de OLAP – *On Line Analytical Process*, ou cubo OLAP. O acesso à informação na forma de cubos mostra-se muito eficiente sob o aspecto de versatilidade e velocidade, se comparado ao uso de tabelas (planilhas) de informações ou resultados de consultas a bancos de dados (THOMSEM, 2002).
* Disponibilidade de Recursos Computacionais: MD requer muitos recursos computacionais para operar seus algoritmos sobre grandes massas de dados e com a queda nos preços de *hardwares* nos últimos anos já é possível e praticável implementar esta tecnologia sem custos proibitivos;
* Bancos de Dados Distribuídos: A construção de banco de dados distribuídos aumenta ainda mais as possibilidades de atuação e assertividade da MD;
* Empresas exigem técnicas mais modernas de tomadas de decisão: Inicialmente empresas da área de finanças, seguros e telecomunicações sentiram necessidade e atualmente mais empresas também buscam informações sobre seus dados para traçar estratégias futuras;
* Informação é o Produto: Para empresas de serviços, é importante saber o que oferecer e para qual público. Para outras empresas, a venda de informações obtidas sobre dados de terceiros pode ser um produto.

## METODOLOGIA DE PESQUISA

Segundo Vergara (1998), a classificação básica de tipos de pesquisas se divide em dois critérios: quanto aos fins e quanto aos meios.

Quanto aos fins, Vergara (1998) define que uma pesquisa pode ser exploratória, descritiva, explicativa, metodológica, aplicada ou intervencionista, podendo não ficar restrito a um único tipo.

Uma pesquisa exploratória realiza-se em área na qual há pouco conhecimento acumulado e sistematizado, na qual não é comportada hipóteses que podem surgir durante ou ao final da pesquisa.

A pesquisa descritiva expõe características de determinada população ou determinado fenômeno. Pode também estabelecer correlações entre variáveis e definir sua natureza. Não tem compromisso de explicar os fenômenos que descreve, embora sirva de base para tal explicação.

Pesquisas de investigação explicativa tem como principal objetivo tomar algo inteligível, justificar lhe os motivos. Visa, portanto, esclarecer quais fatores contribuem, de alguma forma, para a ocorrência de determinado fenômeno.

Uma pesquisa metodológica estuda o que se refere a instrumentos de captação ou de manipulação da realidade. Associa-se a caminhos, formas, maneiras ou procedimentos para atingir determinado fim.

A pesquisa aplicada fundamenta-se na motivação pela necessidade de resolver problemas concretos. Tem finalidade prática, diferente da pesquisa pura, motivada basicamente pela curiosidade intelectual do pesquisador.

Uma investigação intervencionista tem como principal objetivo inferir na realidade estudada, para modificá-la, não se satisfazendo em apenas explicar.

Quanto aos meios, Vergara (1998) define que uma pesquisa pode ser de campo, laboratorial, documental, bibliográfica, experimental, *ex post facto*, participante, pesquisa-ação e estudo de caso, podendo assumir mais de um tipo simultaneamente.

Pesquisa de campo é investigação empírica realizada no local onde ocorre ou ocorreu um fenômeno ou que dispõe de elementos para explica-lo.

Pesquisa laboratorial é experiência que se realiza no local circunscrito, uma vez que seria inviável aplica-la no campo.

Investigação documental é a pesquisa que se realiza em documentos conservados no interior de órgãos públicos ou privados de qualquer natureza.

Pesquisa bibliográfica é o estudo sistematizado que se desenvolve com base em material publicado em meios de comunicação acessível ao público em geral.

Pesquisa experimental é investigação empírica na qual o pesquisador manipula e controla variáveis independentes e observa as variações que tal manipulação e controle produz em variáveis dependentes.

Investigação *ex post facto* refere-se a um fato já ocorrido. Aplica-se quando o pesquisador não pode controlar ou manipular variáveis, seja porque suas manifestações já ocorreram ou não são controláveis, distinguindo-se assim da pesquisa experimental.

A pesquisa participante não se esgota na figura do pesquisador. Dela toma-se parte pessoas implicadas no problema sob investigação, fazendo com que a fronteira pesquisador/pesquisado seja tênue.

Uma pesquisa-ação é um tipo particular de pesquisa participante e de pesquisa aplicada que supõe intervenção participativa na realidade social, sendo, portanto, intervencionista.

Estudo de caso define-se como o circunscrito a uma ou poucas unidades, entendidas como pessoa, família, produto, empresa, órgão público, comunidade ou país. Caracteriza-se pela profundidade e detalhamento, podendo ou não ser realizada no campo.

De acordo com o que foi definido por Vergara (1998), este trabalho de pesquisa define-se enquanto aos fins como:

Descritivo, pois apresenta-se as características e conceitos de MD, técnicas de Mineração de Dados e do algoritmo *K-Means*;

Explicativo, onde demonstra-se todo o processo de KDD para exemplificar o conteúdo teórico abordado neste trabalho, apresentando na prática uma implementação e demonstração de caso de uso do algoritmo *K-Means*;

Exploratório, pois a área de atuação desta pesquisa possui pouco conhecimento acumulado, e seus resultados podem servir de base para futuros trabalhos relacionados ao tema.

Quanto aos meios, segundo Vergara (1998), esta pesquisa será:

De laboratório, pois sua aplicação e simulações são feitas através de softwares computacionais;

Bibliográfica, onde o conteúdo necessário para embasamento desta pesquisa é extraído de material publicado em livros, artigos e trabalhos científicos disponível para o público em geral.

# Referencial Teórico

## MINERAÇÃO DE DADOS

MD é a exploração e análise, de forma automática ou semiautomática, de grandes bancos de dados com objetivo de descobrir padrões e regras. O objetivo principal do processo de MD é fornecer às empresas informações que as possibilitem montar melhores estratégias de marketing, vendas e suporte, melhorando seus negócios (BERRY; LINOFF, 1997).

Atualmente as organizações utilizam sistemas de gerenciamento de banco de dados (SGBD), e conseguem com eficiência organizar e armazenar dados obtidos em suas operações diárias, entretanto, boa parte ainda não consegue usar adequadamente todos estes dados para que se transformem em conhecimento ou informações, que possam ser utilizados em prol de suas organizações auxiliando as tomadas de decisões.

O conceito de MD vem sendo cada vez mais utilizado como forma de descoberta de informações, possibilitando extrair de grandes volumes de dados um resultado que melhore e facilite as escolhas, baseando-se nos dados minerados (HOSKING, 1997).

A aplicação de MD tem relação com descrições de conceitos e classes da discriminação dos dados pesquisados, por exemplo, a análise de uma base de dados de uma clínica ou hospital que classifica pacientes com maior risco para um determinado tipo de doença (idosos entre 60 a 80 anos têm maior probabilidade de ter alguma espécie de câncer, criança de até 12 anos têm maior probabilidade de ter problemas no sistema respiratório). Com frequência a MD vem sendo considerada uma mistura de pesquisas e estatísticas, bancos de dados e inteligência artificial. MD é a parte de um processo de pesquisa denominado Busca de Conhecimento em Banco de Dados (*Knowledge Discovery in Database* - KDD), descrito na seção seguinte, no qual possui própria metodologia para a preparação e exploração dos dados, interpretação dos seus resultados e assimilação dos conhecimentos minerados (HAN; KAMBER, 2011).

O processo de transformação da informação em conhecimento tem crescido exponencialmente. Devido à evolução dos computadores, a utilização do método de KDD tem possibilitado grandes avanços e uma maior agilidade e facilidade na obtenção de informação, proporcionando as pessoas analisarem massas de dados antes inacessíveis e obtendo conhecimento antes impossível de conseguir.

## PROCESSO DE KDD - knowledge discovery in database

Segundo Fayyad (1996), o termo “*knowledge discovery in database*” foi cunhado pela primeira vez em 1989, por Piatetsky-Shapiro, com objetivo de enfatizar que o conhecimento é o produto final de uma descoberta baseada em dados.

No meio profissional de tecnologia da informação em estudos estatísticos por análise de dados, chama-se o processo KDD de Mineração de Dados. No gabinete profissional de inteligência artificial é denominado de Descoberta de Conhecimento em Banco de Dados.

A definição para KDD é um processo, não trivial, de extração de informações implícitas, previamente desconhecidas e potencialmente úteis, a partir dos dados armazenados em um banco de dados (FAYYAD, 1996).

Prass (2007), em seu artigo (KDD – UMA VISÃO GERAL DO PROCESSO) faz uma análise da definição apresentada por Fayyad (1996) e aponta interpretações para os termos utilizados pelo autor. Segundo Prass (2007), “Não trivial”torna clara a existência de alguma técnica de busca ou inferência. “Previamente desconhecidos”indica que a informação deve ser nova para o sistema e de preferência também para o usuário. “Potencialmente úteis”deixa claro que esta informaçãodeve trazer consigo algum benefício, trazendo algum tipo de ganho para o usuário.

O processo de KDD engloba uma sequência de cinco passos: Seleção, Pré-processamento/Limpeza, Transformação, Mineração de Dados e Interpretação/Avaliação dos resultados. Na literatura sobre KDD existem definições apresentadas sob a perspectiva de diferentes autores. Optou-se por fazer um estudo da descrição das etapas baseando-se em duas obras sobre o tema e em seguida é apresentada uma comparação entre ambas as definições.

### ETAPAS SEGUNDO FAYYAD

A figura 1 representa o fluxo definido por Fayyad (1996) que cada etapa do processo deve seguir.

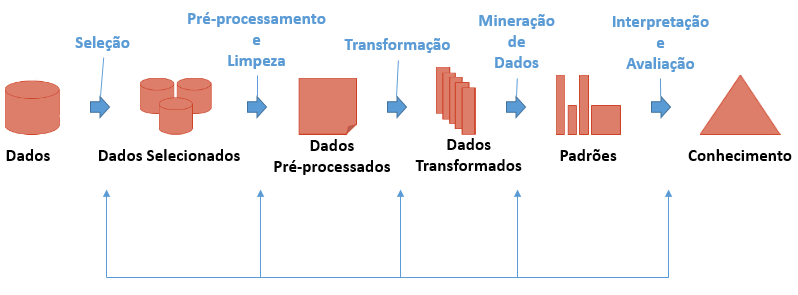


Figura 1: Etapas do processo de KDD. Fonte: Adaptação de Fayyad (1996)

A análise do problema a ser resolvido é o início do processo de descoberta de conhecimento em banco de dados. Compreender o problema é fundamental para definição, da maneira correta, dos objetivos do processo de KDD e suas análises, afim de propor uma solução.

* **Seleção de Dados**

Após definidos os objetivos, inicia-se a seleção de dados disponíveis no banco de dados. São selecionados elementos com relevância para a solução do problema na formação desse subconjunto de dados. Nessa etapa, é importante ter atenção na escolha correta de dados uma vez que posteriormente, os algoritmos de mineração de dados serão aplicados justamente em cima desse subconjunto (BATISTA, 2003).

Esta etapa pode ser crítica uma vez que os dados podem não estar disponíveis em formato compatível ao processo de KDD.

Na fase de seleção é escolhido o conjunto de dados, contidos em um domínio, onde são encontradas as variáveis e registros que farão parte da análise. É um processo complexo, pois os dados podem vir de diversas fontes diferentes (planilhas, *data wharehouses* ou até mesmo escrito em papel) e também podem estar sem um formato padronizado.

* **Pré-Processamento / Limpeza**

Pré-processamento e limpeza dos dados é parte crucial no processo, pois trata-se de aplicar estratégias especificas, pois, uma boa eficiência dos algoritmos de mineração requer um tratamento especial em seus dados de entrada. Nesta etapa que são realizadas tarefas a fim de eliminar dados redundantes, inconsistentes ou incompletos. Com objetivo de reduzir ruídos ao longo do processamento (REZENDE, 2003).

* **Transformação**

Ao atingir a fase de transformação, os dados armazenados no banco de dados podem ainda não ser suficientes ou não se encontrarem em formato exigido pelos algoritmos na etapa de mineração.

É nesse momento então que são geradas as demais informações necessárias e estruturados os dados para formato compatível aos algoritmos. Os dados podem ser obtidos de fontes distintas para um único repositório de formato legível ao algoritmo utilizado. Sendo assim, Silva (2000) recomenda que já se tenha definido a técnica de mineração e o algoritmo minerador que serão utilizados para a partir de então transformar os dados para o formato adequado.

* **Mineração de Dados**

É na etapa de mineração de dados que são definidos os algoritmos mineradores que realizarão a busca pelo conhecimento implícito e útil do banco de dados, com objetivo encontrar padrões.

A etapa de Mineração de Dados é que recebe maior destaque em todo o processo. Nesta fase é aplicado o algoritmo de mineração apropriado, que tem como entrada os dados do repositório gerado através da etapa anterior, com a finalidade de se obter algum resultado proveitoso, que será enfim interpretado e avaliado na última fase (REZENDE, 2003).

* **Interpretação dos Resultados**

Por último vem a fase de Interpretação e Avaliação dos resultados obtidos após aplicação do algoritmo de mineração. A principal meta desta etapa é melhorar a compreensão do conhecimento adquirido, geralmente em forma de relatórios demonstrativos, com documentação e explicação das informações relevantes descobertas no processo de KDD. Caso o resultado não seja satisfatório, o que acontece com certa frequência, o processo pode retornar a qualquer uma das etapas anteriores ou até mesmo ser iniciado do zero para que sejam testados outros parâmetros a fim de se obter um resultado diferente (BATISTA, 2003).

Conclui-se que após finalizada a fase de processamento dos algoritmos, é então feita a leitura e interpretação dos resultados gerados. Nessa fase aprimora-se a compreensão do conhecimento obtido, tanto através de relatórios demonstrativos quanto documentos com informações relevantes.

### ETAPAS SEGUNDO ADRIAANS e ZANTINGE

Essa abordagem do processo de KDD tem fundamento na necessidade das organizações de constantemente precisar de novas informações sobre seus dados, então o processo é repetido sempre que há necessidades de novas informações. A Figura 2 ilustra a composição desse processo (ADRIAANS; ZANTINGE, 1997).

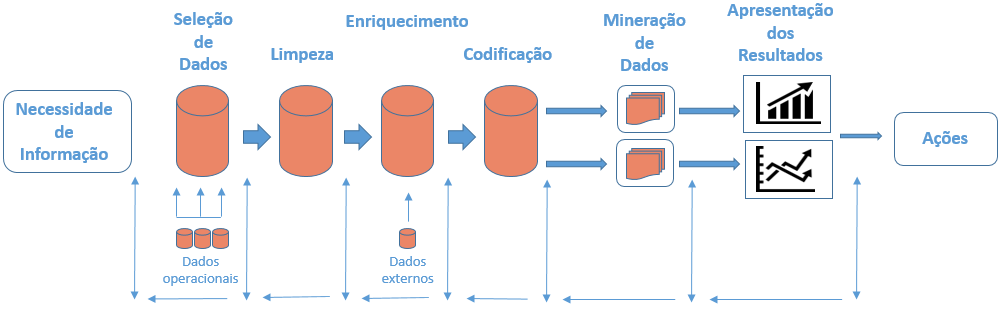


Figura 2: Etapas de KDD. Fonte: Elaboração própria

* **Seleção dos Dados**

Etapa onde ocorre uma análise de todos os dados operacionais da base de dados e são selecionados somente aqueles que são relevantes para chegar aos objetivos do processo. Novas seleções podem ser feitas quando houver uma nova iteração, podendo incluir dados anteriormente descartados.

* **Limpeza**

Nesta etapa são utilizadas várias operações de limpeza nos dados, como remoção de dados duplicados e correção nos dados, por exemplo. Pode ser executada inúmeras vezes, já que não é possível prever todos os problemas de qualidade existentes no banco de dados com antecedência.

* **Enriquecimento**

Pode-se incluir ou transformar algumas informações no banco de dados se porventura for de interesse para alcançar o objetivo. Um exemplo é converter uma data de nascimento para a idade de uma pessoa.

* **Codificação**

A maneira como os dados são armazenados nos bancos de dados pode não representar a melhor forma para a utilização no processo de KDD. Normalmente os dados têm sua representação apropriada ao contexto da aplicação. Atributos com valores literais, por exemplo, podem não ser adequados para certos algoritmos de mineração, fazendo-se necessário adequá-los através de normalização dos valores dentro de um determinado intervalo.

* **Mineração de Dados**

Esta é a etapa onde os dados são manipulados para que seja extraído o conhecimento. Com o uso de uma ferramenta de consulta SQL (*Structured Query Language*) é possível ter uma visão geral dos dados para então prosseguir para uma análise menos trivial. Neste primeiro momento, cerca de 80% do conhecimento é extraído e já pode conter alguma informação interessante (ADRIAANS; ZANTINGE, 1997).

* **Apresentação dos Resultados**

Ao término da etapa de mineração de dados são geradas informações num formato específico de acordo com a técnica utilizada. Deve-se levar em conta que os dados podem estar codificados ou mesmo que o método utilizado na etapa de mineração gere informações com alguma representação ou formalismo muito específico. Os resultados devem ser exibidos de forma clara para que seja possível um entendimento trivial e uma fácil interpretação das informações, mesmo para quem não fez parte do processo de KDD.

### COMPARAÇÃO ENTRE AS DUAS ABORDAGENS

A Tabela 1 apresenta uma comparação entre as duas abordagens descritas nas seções anteriores.

Tabela 1: Comparação entre as duas definições. Fonte: Elaboração própria.

|  |  |  |
| --- | --- | --- |
| **Etapas** | **FAYYAD** | **ADRIAANS e ZANTINGE** |
| **Definição do objetivo** | O objetivo é definido para solucionar o problema que gerou a necessidade de implementação de um processo de descoberta de conhecimento | O objetivo surge da constante necessidade das organizações em obterem novas informações sobre seus dados. |
| **Entendimento dos dados** | Ainda durante a fase de definição do objetivo é realizado o entendimento do domínio e avaliação das possibilidades de sucesso de acordo com os dados disponíveis. | Não existe a etapa de compreensão dos dados. É pressuposto que o objetivo proposto pelo usuário já está vinculado aos dados a serem selecionados. |
| **Seleção dos dados** | É feito um subconjunto de dados selecionados a partir da (s) base (s) de dados disponíveis. Este subconjunto conterá apenas aqueles dados relevantes para a solução do problema. | Ocorre uma análise dos dados e são selecionados apenas os relevantes com a finalidade de reduzir o volume do conjunto de dados. |
| **Limpeza** | Realizada na etapa de pré-processamento, porém o autor não define como deve ser conduzida a tarefa de limpeza. | Utiliza algumas operações de limpeza nos dados. É executada inúmeras vezes, já que é impossível prever todos os problemas de qualidade |
| **Transformação** | Utiliza algumas técnicas para formatação dos dados, para adequá-los ao algoritmo de mineração. Recomenda-se já se ter definido a técnica de mineração e o algoritmo minerador. | Trata de como incorporar dados externos à base (enriquecimento) e como transformar os dados (codificação) para que possam ser utilizados na etapa de mineração. |
| **Mineração dos dados** | Considerada a fase mais importante do processo, onde é escolhida a técnica de mineração, definido o algoritmo minerador e realizada a busca pelo conhecimento no banco de dados. | Definição da técnica e dos algoritmos, porém, é garantido que 80% do conhecimento pode ser extraído por consultas SQL e complementado com a utilização de técnicas e algoritmos avançados. |
| **Resultados** | Ocorre a interpretação dos resultados com objetivo de melhorar a compreensão do conhecimento. Caso não tenha alcançado os objetivos, o processo é retomado. | Os resultados são interpretados com o objetivo de melhorar a compreensão, pois os dados podem estar codificados, dependendo do método utilizado na etapa de mineração. |

Após análise das definições apresentadas pelos autores e observação na base de dados utilizada no experimento desse trabalho, chegou-se à conclusão que o modelo definido por Fayyad (1996) seria o mais adequado para tratar o tipo de problema em questão e ele foi escolhido para nortear a demonstração de MD apresentada no capítulo 3.

## MÉTODOS E TÉCNICAS DE MINERAÇÃO DE DADOS

Tradicionalmente as técnicas de MD estão divididas em algoritmos de aprendizado supervisionado (preditivo) e não-supervisionado (descritivo). Embora o limite dessa classificação seja muito tênue (alguns métodos descritivos podem ser preditivos e vice-versa), ela ainda é interessante para fins didáticos (HAN; KAMBER, 2006).

Os algoritmos de aprendizado supervisionados ou preditivos fazem previsões se baseando em um conjunto de exemplos. Cada exemplo usado para treinamento é rotulado com o valor de seu interesse, esses tipos de algoritmos procuram padrões nesses rótulos de valor. Pode-se usar qualquer informação que seja relevante à proposta do processo. Cada algoritmo procura diferentes tipos de padrões, e depois de encontrar o melhor possível, este será usado para fazer previsões para dados de testes sem rótulos (HAN; KAMBER, 2006).

O objetivo da predição é descobrir valores futuros, por vezes desconhecidos, para seus estudos. Afim de chegar em um resultado preditivo, Dallanora (2009) aponta para necessidade de uma pesquisa sobre os dados históricos do banco de dados. Entretanto, para que isso ocorra de maneira precisa, é importante que o banco minerado esteja íntegro e tenha passado por algum processo de limpeza e filtragem das informações dispensáveis.

A mineração de dados preditiva ou ações de predição, consistem na generalização de exemplos passados com respostas conhecidas, possibilitando o reconhecimento da classe para um novo dado.

Para Rezende (2005), existem dois principais tipos de técnicas preditivas, a classificação e a regressão.

A classificação consiste em predizer um valor categórico como, por exemplo,  
predizer se um determinado cliente é bom ou mau pagador.

Na técnica de regressão, o atributo a ser previsto é um valor contínuo como, por exemplo, predizer o lucro ou a perda em um empréstimo. É possível afirmar que quanto maior for o número de variáveis empregadas na mineração ao longo do processo de predição, menores serão as chances de ter uma previsão errônea.

Já os algoritmos de aprendizado não-supervisionados cumprem a tarefa de aprendizagem de máquina para inferir uma função de descrever a estrutura oculta a partir de dados não marcados (uma categorização ou classificação não está incluída nas observações). Uma vez que os exemplos fornecidos estão sem rótulos, não é possível avaliar com precisão a estrutura que é produzida pelo algoritmo.

A partir da avaliação de dados, o processo de descrição busca encontrar padrões descritivos em um subconjunto de dados, onde toda informação obtida influencia na tomada de decisão dentro de uma empresa, por exemplo, de investimentos em campanhas a mudança do foco estratégico da mesma

Segundo Dallanora (2009), não existirá uma nova informação, mas sim a melhor organização de informações já existentes, facilitando a visão e interpretação dos dados.

Na técnica descritiva temos como exemplo quando em uma empresa tem a possibilidade de criar um perfil para seus clientes, sendo informado ao algoritmo todos os valores de entrada de dados relevantes para criação desse perfil. A partir deste ponto, o algoritmo minerador fará o agrupamento dos clientes com base nas características em comum encontradas, podendo traçar diferentes perfis de consumidores.

O que difere as técnicas de aprendizado supervisionado e não-supervisionado é o fato de que os métodos não-supervisionados não precisam ter os registros pré-categorizados, ou seja, não é necessário um atributo alvo. Geralmente estes métodos usam alguma medida de similaridade entre os atributos. As tarefas de associação e agrupamento são consideradas como não-supervisionadas. Já no aprendizado supervisionado, as técnicas são providas com um conjunto de dados que contém uma variável alvo pré-definida e a categorização dos registros é feita em relação a esta variável (MCCUE, 2007).

### TÉCNICAS DE MINERAÇÃO DE DADOS

* **Classificação**

Uma das mais usuais entre as técnicas de mineração de dados, a técnica de Classificação, visa identificar a qual classe um determinado registro pertence (CAMILO; SILVA, 2009).

A classificação é uma tarefa da mineração de dados que consiste em avaliar os dados processados, classificando-os de acordo com as suas características. Para classifica, criam-se classes caracterizadas, e os dados processados são relacionados a essa classe por meio das peculiaridades (SILVA, 2000).

A técnicas de classificação podem ser definidas como não-supervisionadas e supervisionadas. Tem seu uso adotado para prever valores de variáveis passíveis de classificação. É possível, por exemplo, criar um modelo que faz a classificação de clientes de um banco como de risco ou especiais. Neste cenário, é criado um modelo para classificar qual categoria cada registro pertence: de risco ou especial (CAMILO; SILVA, 2009).

Dentre as principais finalidades de utilizar os algoritmos de classificação temos a identificação de eventuais transações bancarias suspeitas, classificação de clientes para aprovação de crédito e diagnostico a partir de sintomas, por exemplo.

* **Estimativa ou Regressão**

Diferente da classificação, a estimação é usada quando o registro é identificadopor um valor numérico e não um categórico (CAMILO; SILVA, 2009).

Dessa forma, podemos estimar o valor de determinada variável a partir dos valores das demais. Por meio dessa técnica, pode-se estimar valores gastos em mercados, telefonia e iluminação com base em seus consumos, por exemplo.

No âmbito da regressão existem ainda duas classificações, a linear e a não-linear. São chamadas de lineares, quando a relação entre as variáveis base e a resposta segue um comportamento onde x e y tem uma função linear. Já as não-lineares possuem a relação entre as variáveis base e a resposta que não segue um comportamento linear. A relação entre as variáveis pode se modelar com base em uma função polinomial. Entretanto, vale ressaltar que mesmo a regressão não-linear pode ser convertida a linear. Podendo ser também conhecida como Predições Numéricas, a partir das técnicas desenvolvidas por Sir Frances Galton (1822-1911) (CAMILO; SILVA, 2009).

* Associação

Associação é uma das técnicas mais conhecidas, sendo bastante explorada para explicar o problema da Análise da Cesta de Compras.

Agrawal e Srikant (1994), apresentam como principal objetivo da associação, a apresentação de particularidades com relação mútua entre os dados. Descrevendo que uma regra de associação identifica o quanto um conjunto de itens em um banco de dados se relaciona com a presença de um outro conjunto de itens no mesmo registro.

“ O objetivo das regras de associação é localizar tendências que podem ser utilizadas para compreender e explorar padrões comportamentais dos dados” (CASTANHEIRA, 2008, p.92).

Essa técnica se mostra eficiente quando consideramos dados sobre compras em supermercado, avaliando a porcentagem de clientes que na compra produto A, tende a comprar produto B. Dessa forma é possível um planejamento mais eficaz de disposição de produtos visando otimização nas vendas e no lucro.

* Análise de *Outliers*

Na técnica de *Outliers*, avalia-se o registro que não possui comportamento como da maioria (AMO, 2004). O termo *Outliers* pode ser traduzido em “discrepante”.

Apesar de ser descreditado por outros métodos de mineração, além de por vezes o minerador decidir por não o utilizar, temendo qualquer interferência, seu valor se mostra em aplicações como na detecção de fraudes. Podendo identificar uso incomum em cartões de crédito ao localizar gastos além do padrão de consumo, indicando possivelmente uma transação indevida.

A identificação dessas discrepâncias é feita na verificação de valores encontrados e analisando se tratam de valores plausíveis para determinada categoria (AMO, 2004).

* Agrupamento

As técnicas de agrupamento, método utilizado para as simulações desse trabalho de pesquisa, são consideradas como não-supervisionadas. Um conjunto de registros, agrupamentos (ou *cluster*) é gerado, contendo os registros mais semelhantes entre si. Geralmente as medidas de similaridades usadas são medidas de distâncias tradicionais (CAMILO; SILVA, 2009)**.**

Trata de um conjunto D de dados com n registros e *k* é o número de agrupamentos desejados, os algoritmos de particionamento organizam os objetos em k agrupamentos, tal que *k <= n*. Os algoritmos de agrupamento mais comuns são: *K-Means* e *K-Medoids* (AMO, 2004).

Os registros agrupados em um *cluster* são considerados similares aos elementos do mesmo *cluster* e não-similares aos elementos de outros *cluster*s. Como trabalha com o conceito de distância (similaridade) entre os registros, normalmente é necessário fazer a transformação dos diferentes tipos de dados (categóricos, ordinais, intervalos e binários) para uma escala comum, [0.0, 1.0] por exemplo (CASTANHEIRA, 2008).

Os algoritmos de agrupamento podem ser classificados nas seguintes categorias:

1. **Método Hierárquico**: A principal ideia dos métodos hierárquicos é agrupar por meio de aglomeração ou da divisão dos registros do conjunto. A forma produzida por estes métodos é um dendograma (gráfico em formato de árvore).
2. **Método Baseado em Grade:** Estes métodos têm sua estrutura baseada em formato de grades e dividem os registros nas células desta grade. Tem como característica o processamento bem rápido, se tratando de tempo de execução.
3. **Método Baseado em Modelo:** Os métodos desta categoria criam um modelo para cada agrupamento e tentam identificar o modelo mais apropriado para cada objeto. Este método tem como princípio que os dados estão gerados por uma série de probabilidade de distribuições.

De acordo com as técnicas apresentadas, a mais apropriada para a simulação apresentada nesse trabalho é a técnica de Agrupamento, pelo método baseado em modelo, utilizando o algoritmo não-supervisionado *K-Means*, que é descrito na seção seguinte.

## K-Means

Em 1957, Stuart Lloyd propôs um algoritmo simples e iterativo que se mostrou eficiente para encontrar locais mínimos para um dado problema. (LIBERTY, 2016)

O termo *K-Means* surgiu na literatura no ano de 1967 cunhado por James MacQueen, baseando-se no trabalho realizado por Lloyd em 1957, em seu artigo *Some methods for classification and analysis of multivariane observations.*

O método pode apresentar duas versões, a primeira conhecida como clássica e proposta por Lloyd, e a segunda, versão recente proposta por Hartigan e Wong (SLONIM; AHARONI; CRAMMER, 2013).

Sobre essas versões, a mais reconhecida é a mais recente, de Hartigan e Wong, apresentada no trabalho de 1979 *A K-Means Clustering Algorithm*, pois seu conjunto de mínimos locais é um subconjunto dos mínimos locais de Lloyd, ele geralmente apresenta melhores resultados em relação aos outros. Vale ressaltar que a escolha entre eles depende do *dataset* que se deseja *clusterizar* (CARDOSO, 2008).

O algoritmo *K-Means* é um método heurístico clássico da literatura e que possui um algoritmo de aprendizagem que organiza N objetos da base de dados em K grupos e cada um destes representa um *cluster*. K-Means é um algoritmo escalável e simples que pode ser alterado de forma fácil para trabalhar com fluxos de dados maiores em grandes bancos de dados (NALDI, 2011).

Conforme Linden (2009), o *K-Means* é uma heurística de agrupamento não hierárquico que tem sua técnica aplicada para minimizar a distância entre os elementos de um conjunto de k centros. O *K-Means* é iniciado com a escolha dos K elementos que compõe os centroides iniciais. Depois esses centroides são separados e é calculada a distância de cada elemento, agrupando esses elementos ao grupo que possui uma menor distância, recalculando o centroide. A Figura 3 representa em forma de pseudocódigo o comportamento do algoritmo *K-Means* e em seguida é mostrado seu passo a passo, segundo Cardoso (2008).

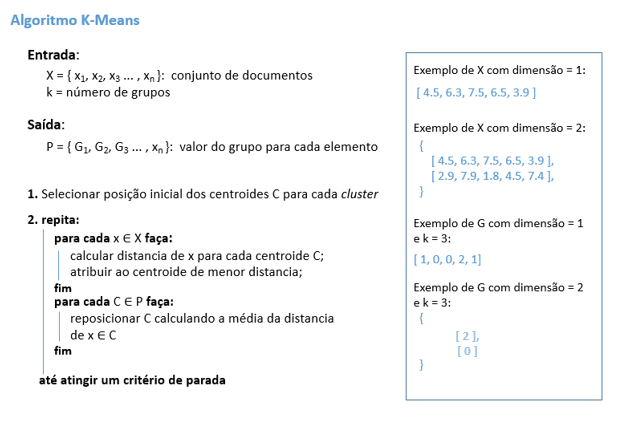


Figura 3: Algoritmo K-Means. Fonte: Elaboração própria

Para facilitar a compreensão do algoritmo *K-Means*, o conceito de clusterizaçãode dados deve ser melhor compreendido. Este conceito ganhou mais visibilidade quando reforçaram sua diferença em relação a Classificação, sendo a clusterização já tendo sido definida como um tipo de classificação.

A análise de agrupamento ou clusterização é um método que divide conjuntos de dados em subconjuntos com características semelhantes sem informação previa sobre como devem ser. Enquanto a Classificação é o processo onde se define a qual grupo um dado deve ser melhor adequado. Nesse caso é necessário conhecer primeiramente os possíveis grupos ou suas características.

Cassiano (2014) diz que a clusterização pode ser entendida como o processo anterior a classificação. Considerando que na clusterização, teoricamente não há classes previamente  
conhecidas ou informações sobre o usuário não manipula os resultados do método, sendo assim definido como não supervisionado.

Dessa forma, a clusterização trata de um conjunto de técnicas numéricas que agrupam automaticamente, de tal forma que elementos de um mesmo *cluster* sejam mais parecidos entre si do que dos elementos de outros *clusters* (CASSIANO, 2014).

Existem diversos métodos de clusterização, e uma das categorias mais utilizadas é a dos Métodos Particionais, na qual o usuário define a quantidade de agrupamentos que deseja e então o método converge interativamente para a participação dentro das condições iniciais (JAIN, 2009).

Segundo Jain (2009), a implementação do algoritmo *K-Means* é fácil, eficiente e de bom desempenho, além de possibilitar variações mudando a função de medida usada no método, tornando o *K-Means* o algoritmo mais comum.

Para o autor, partindo de um conjunto de dados em uma matriz *‘n × d’*, sendo que cada um dos *‘n’* vetores horizontais representa uma observação com *‘d’* características, é possível interpretar estes vetores como pontos em um espaço *d*-dimensional.

Segundo Cardoso (2008), o algoritmo na versão de Lloyd, originalmente com a distância euclidiana, procede da seguinte forma:

* Primeiro Passo

Definir a quantidade K de *clusters* em que sua base deve ser dividida, depois escolhe-se K pontos que serão os centros iniciais dos *clusters*. Vale ressaltar que nessa fase é possível escolher os centros manualmente, aleatoriamente ou por algum algoritmo escolhido. A Figura 4 representa um exemplo de duas dimensões com K = 3.

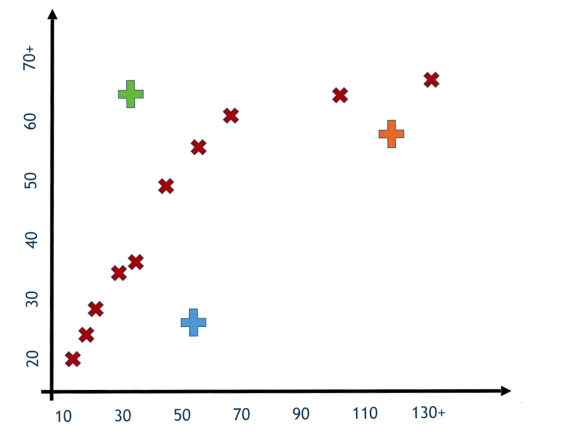


Figura 4: Primeiro passo K-Means. Fonte: Elaboração própria

* **Segundo Passo**

Neste passo o algoritmo calcula a distância euclidiana de cada centroide para todos registros analisados. O objetivo é agrupar os registros de acordo com o centroide de menor distancia calculado. A Figura 5 representa as distâncias de todos os registros para um *cluster*, as distâncias para todos os *clusters* e como os registros foram agrupados nesse passo.

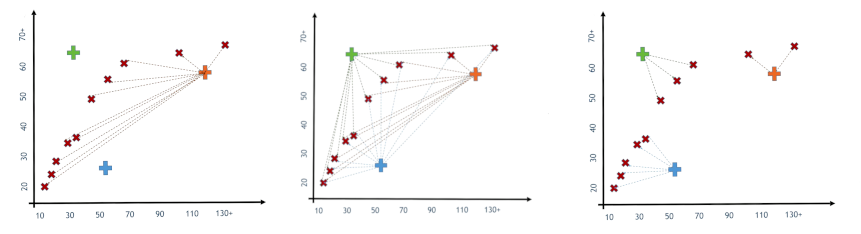


Figura 5: Segundo passo K-Means. Fonte: Elaboração própria

* **Terceiro Passo**

Os centroides são realocados para o novo centro de gravidade de seus membros atribuídos no passo anterior. É realizado um cálculo médio das distâncias do centroide para cada um de seus registros atribuídos. A Figura 6 representa o novo posicionamento dos centroides após o término deste passo.

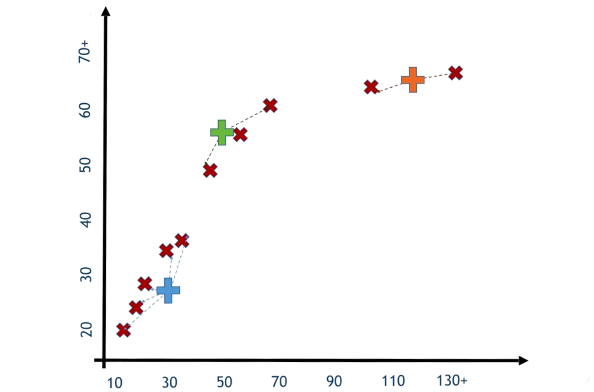


Figura 6: Terceiro passo K-Means. Fonte: Do Autor

* **Quarto Passo**

Os passos dois e três correspondem a uma iteração do algoritmo. Este passo é a última parte da iteração, aqui é checado o mecanismo de convergência que é o ponto de parada do algoritmo. Esse mecanismo checa se houve mudanças de registros entre os *clusters*, enquanto houver mudanças é sinal que o algoritmo ainda está trabalhando e um resultado ótimo ainda não foi encontrado e o algoritmo retorna ao segundo passo para uma nova iteração. O mecanismo de parada é atingido quando não há mais mudanças entre os *clusters* da última iteração para a atual. É possível também definir um número máximo de iterações, caso seja de interesse. Neste caso, a parada do algoritmo é acionada tanto pelo mecanismo de convergência quanto pelo número de iterações definido previamente.

# PROCESSO DE KDD COM USO DO *K-MEANS*

Neste capítulo é realizado o trabalho de MD sobre a base de dados do futebol europeu descrita a seguir. Foi utilizado o modelo de etapas de KDD definido por Fayyad no segundo capítulo para demonstrar o funcionamento de todo o processo, etapa por etapa, com objetivo de exemplificar a aplicabilidade do algoritmo *K-Means* no processo de clusterização de dados para obtenção de conhecimento.

## Base de Dados

O BD utilizado neste experimento está em formato SQLite. Nele contém informações sobre mais de 10.000 jogadores que atuaram em mais de 25.000 partidas ocorridas entre 2008 e 2016 em 11 países europeus. Na Figura 7 é possível visualizar de forma geral quais tabelas, e suas respectivas dimensões, estão presentes no BD.

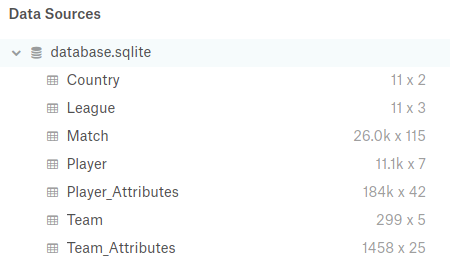


Figura 7: Tabelas do Banco de Dados. Fonte: www.kaggle.com/hugomathien/soccer

Após uma análise sobre os dados contidos no BD, chegou-se à conclusão que as tabelas *Player* e *Player\_Attributes* apresentam condições ideias para aplicação do algoritmo K-Means, pois seu volume de registros e a quantidade de atributos discretos para cada registro é suficientemente grande para representar confiabilidade nos resultados gerados pelo processo de KDD. Partindo desta perspectiva, foi definida uma pergunta para guiar esse experimento: É possível identificar em qual posição um jogador atua com base em suas habilidades?

Com o objetivo definido, iniciou-se o processo de KDD e a primeira etapa é apresentada após a descrição do ambiente de trabalho.

## Ambiente de Trabalho

Foi utilizado como instrumento para execução dos softwares um Desktop Asus com processador Intel Core i5-6600 3.5GHz, memória RAM de 16GB, SSD de 480GB com sistema operacional Ubuntu 19.04. Os códigos foram escritos em Python e o interpretador estava na versão 3.7. Para execução dos códigos foi utilizado o ambiente de desenvolvimento Jupyter Notebook. A Figura 8 mostra as bibliotecas necessárias para a execução do experimento e como elas foram nomeadas.

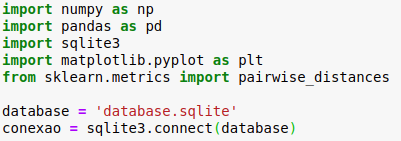


Figura 8: Bibliotecas utilizadas. Fonte: Elaboração própria

## Seleção de Dados

Nesta etapa foram selecionadas as tabelas de interesse *Player* e *Player\_Attributes,* e seus dados foram carregados do BD através de linguagem SQL. As Figuras 9 e 10 representam uma amostra dos dados.

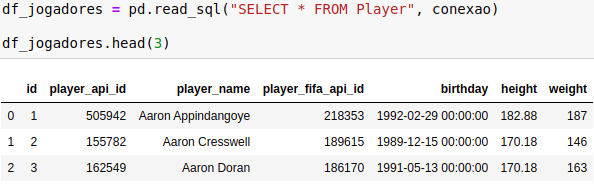


Figura 9: Amostra da tabela *Player*. Fonte: Elaboração própria

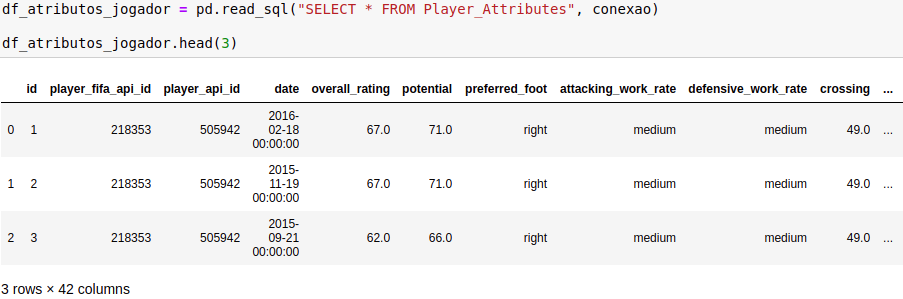


Figura 10: Amostra da tabela *Player\_Attributes*. Fonte: Elaboração própria

As duas tabelas selecionadas estão relacionadas através do atributo *player\_api\_id.* Foi criado um *Data Frame* (DF) usando essa relação associando o atributo *player\_name* da tabela *Player* com suas respectivas habilidades presentes na tabela *Player\_Attributes*. Na última parte dessa etapa foi descartado atributos do DF que não eram de interesse ao escopo, como representado na Figura 11.

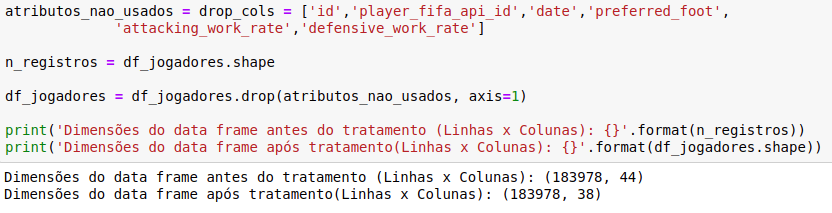


Figura 11: Seleção de dados. Fonte: Elaboração própria

Com todos os dados devidamente selecionados, seguiu-se para a fase de Pré-Processamento e Limpeza.

## Pré-Processamento e Limpeza

Esta é a etapa em que foi feito um tratamento nos dados selecionados com intuito de descartar registros com ruídos que impedem o funcionamento do algoritmo de mineração.

A Figura 12 representa a limpeza dos dados duplicados contidos no DF.

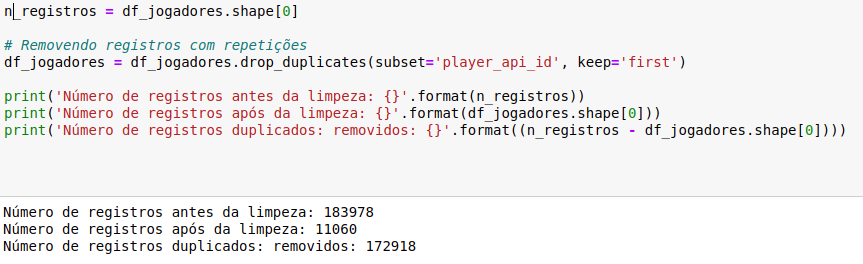


Figura 12: Limpeza dos dados. Fonte: Elaboração própria

O próximo passo foi descartar todos os registros que continham algum atributo sem nenhum valor atribuído (valor nulo). A Figura 13 demonstra como foi este processo.

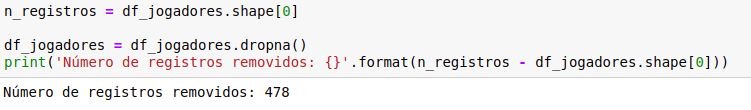


Figura 13: Removendo de valores nulos. Fonte: Elaboração própria

Como houve remoção de dados do DF, foi necessário reorganizar os *index* dos registros para manter sua integridade antes de passar para a próxima etapa.

## Transformação

Nesta fase foi criado um segundo DF contendo apenas atributos discretos para serem utilizados pelo algoritmo de mineração. A Figura 14 apresenta quais atributos foram removidos do DF original para compor o novo DF.

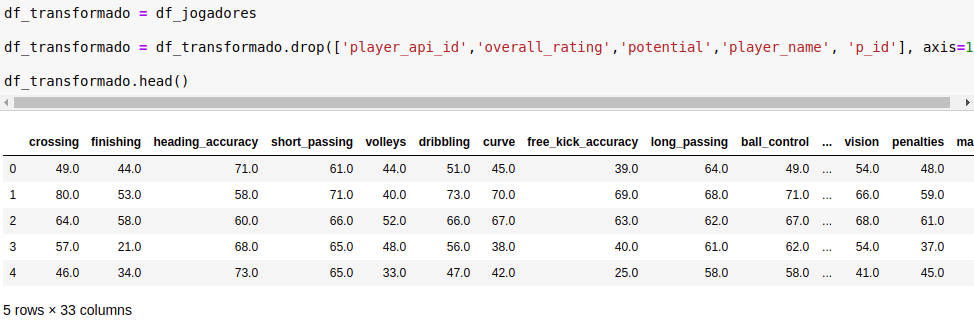


Figura 14: Representação do DF transformado. Fonte: Elaboração própria

O final da etapa de transformação foi concluído pois os dados estavam organizados e consistentes para todos os 33 atributos, que são: *'crossing', 'finishing', 'heading\_accuracy', 'short\_passing', ‘voltes', 'dribbling', 'curve', 'free\_kick\_accuracy', 'long\_passing', 'ball\_control', 'acceleration', 'sprint\_speed', 'agility', 'reactions', 'balance', 'shot\_power', 'jumping', 'stamina', 'strength', 'long\_shots', 'aggression', 'interceptions', 'positioning', 'vision', 'penalties', 'marking', 'standing\_tackle', 'sliding\_tackle', 'gk\_diving', 'gk\_handling', 'gk\_kicking', 'gk\_positioning', 'gk\_reflexes'.*

Neste momento do experimento existiam dois DFs. A função do primeiro foi servir como base para integração dos resultados gerados na etapa seguinte, de Mineração de Dados, com finalidade de analisar os resultados posteriormente. O segundo DF representava os dados transformados com o formato exigido pelo *K-Means.*

## Mineração de Dados

O algoritmo *K-Means* utilizado nessa etapa foi implementado pelo autor seguindo os passos descritos no referencial teórico e a representação do código-fonte está disponível no Apêndice A, não havendo necessidade de descrever seu passo a passo novamente para responder à pergunta desse experimento, pois seu passo a passo está descrito na seção 2.3.

No processo de configuração do algoritmo é escolhido um valor K, que é a quantidade de *clusters* que se deseja obter. Em tempo de execução, cada valor de K é nomeado de centroide e cada um destes necessita de valores iniciais associados para que a primeira iteração do algoritmo seja inicializada.

O processo de escolha do valor de K foi baseado na pergunta levantada para este experimento. Assumiu-se que um jogador pode atuar em 4 posições distintas: goleiro, defesa, meio-de-campo e ataque. Partindo dessa proposta, foi definido o número 4 para K e seus valores iniciais foram gerados aleatoriamente como representado na Figura 15.

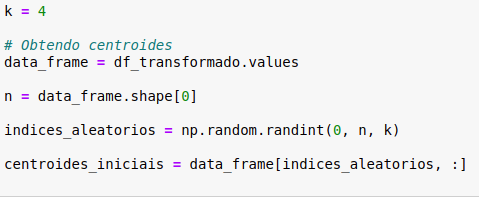


Figura 15: Configuração do K-Means. Fonte: Elaboração própria

A Figura 16 apresenta uma amostra com 2000 registros exibidos em um gráfico para mostrar em quais regiões os centroides aleatórios foram alocados. Os eixos X e Y do gráfico representam respectivamente valores dos atributos *crossing* e *finishing* para cada um dos 2000 registros.

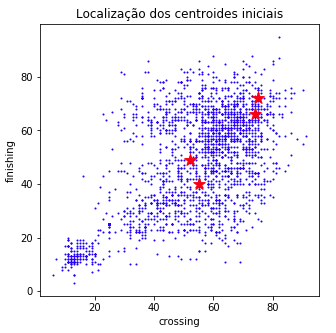


Figura 16: Centroides iniciais. Fonte: Elaboração própria

Com o DF preparado, K definido e centroides iniciais declarados, iniciou-se o processo de clusterização através do algoritmo K-Means e os resultados gerados foram armazenados nas variáveis ‘dados\_centroides’ e ‘clusters\_agrupados’ como representado na Figura 17.

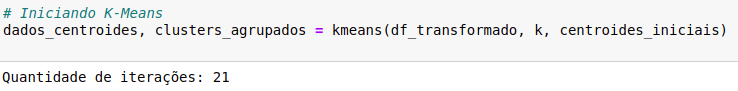


Figura 17: Inicialização do *K-Means*. Fonte: Elaboração própria

Para demonstrar a influência que a posição inicial dos centroides exerce sobre a quantidade iterações, foi realizada uma nova escolha dos centroides em que visualmente se mostraram mais balanceados em relação a dispersão dos dados e o algoritmo foi executado novamente. A Figura 18 mostra a representação gráfica do posicionamento dos novos centroides e na Figura 19 é possível observar a quantidade de iterações do *K-Means* para a nova execução.

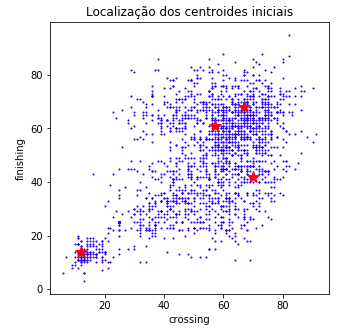


Figura 18: Novos centroides. Fonte: Elaboração própria

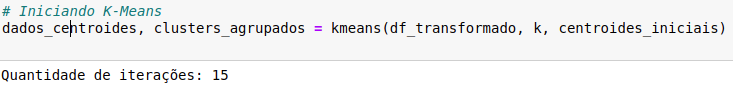


Figura 19: Reinicialização do *K-Means*. Fonte: Elaboração própria

Foi possível observar uma redução de 28,57% na quantidade de iterações quando os centroides são melhor posicionados, constatando que uma boa escolha dos centroides iniciais garante melhor performance do algoritmo. Não foi constatado que a escolha dos centroides iniciais exerce influência nos resultados finais dos registros associados a cada *cluster* para este experimento.

Para fins de melhor representatividade da movimentação dos centroides, foram utilizados os dados gerados na primeira execução do algoritmo. A Figura 20 mostra o estado de cada centroide até a sexta iteração, da sétima iteração em diante não foi mais possível perceber visualmente mudanças nas movimentações.

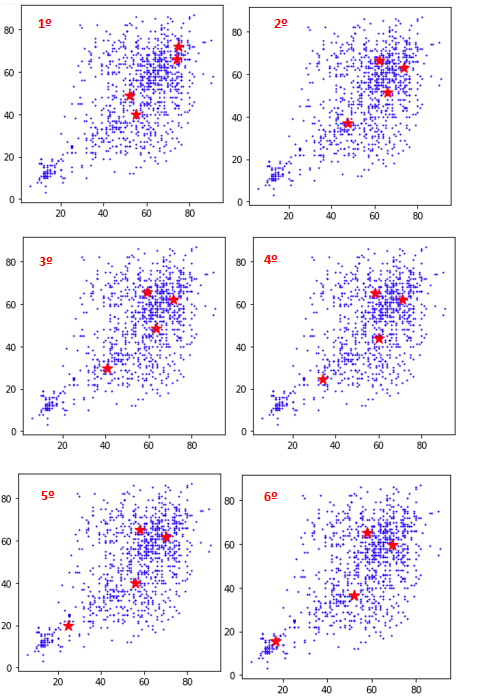


Figura 20: Centroides durante iterações. Fonte: Elaboração própria

## Interpretação dos Resultados

A análise dos resultados se iniciou com o processo de junção do output gerado pelo algoritmo *K-Means* com o DF tratado para esse experimento. Na Figura 21 é possível visualizar os 5 primeiros registros do DF e em qual *cluster* eles foram alocados.

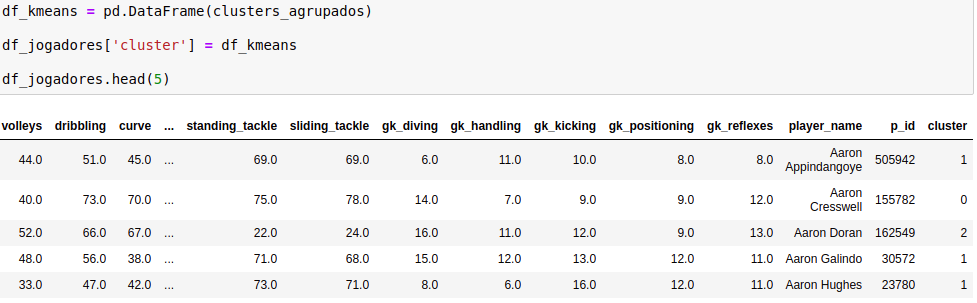


Figura 21: Saída gerada pelo *K-Means*. Fonte: Elaboração própria

O próximo passo foi desmembrar o DF em outros 4, onde cada novo DF é composto apenas por membros de seus respectivos *clusters*, como representado na Figura 22.

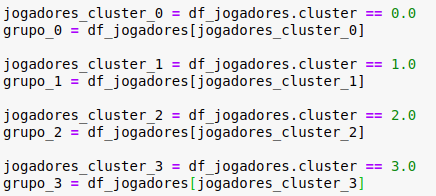


Figura 22: Divisão dos *clusters*. Fonte: Elaboração própria

A primeira análise foi realizada calculando a média de cada atributo processado pelo algoritmo para todos os *clusters* e a Figura 23 representa os resultados.

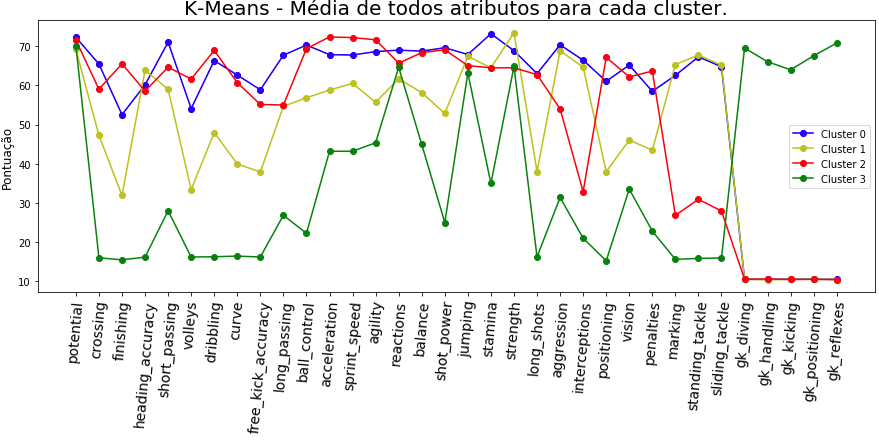


Figura 23: Representação por *cluster*. Fonte: Elaboração própria

Observou-se que o *cluster* 3 apresentou comportamento inversamente proporcional em relação aos demais *clusters* para os atributos *gk\_diving, gk\_handling, gk\_kicking, gk\_positioning e gk\_reflexes*, que são habilidades desejadas para jogadores que atuam na posição de goleiro, indicando que neste *cluster* foram agrupados os goleiros presentes no DF.

Para rotular os *clusters*, optou-se por ordenar os registros de cada grupo utilizando o atributo *overall\_rating* (não processado pelo *K-Means*), que representa a avaliação geral de cada jogador. O ordenamento aconteceu de forma decrescente, pois assim os melhores jogadores de cada *cluster* assumem as primeiras posições no DF, tornando-se mais fácil identificar o rótulo de cada agrupamento. As Figuras 24, 26, 28 e 30 representam os resultados para os *clusters* 0, 1, 2, 3 respectivamente e as Figuras 25, 27, 29 e 31 apresentam um breve resumo, obtido através da busca do nome do jogador no site de pesquisa www.google.com, para dois dos cinco primeiros jogadores de cada grupo.

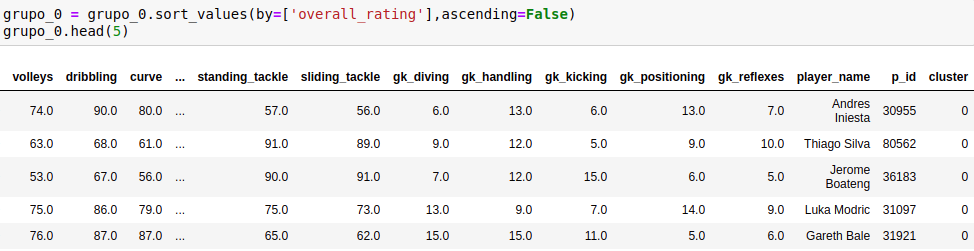


Figura 24: Registros do *cluster* 0. Fonte: Elaboração própria

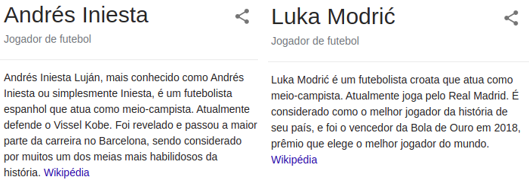


Figura 25: Jogadores. Fonte: pesquisa do nome em www.google.com

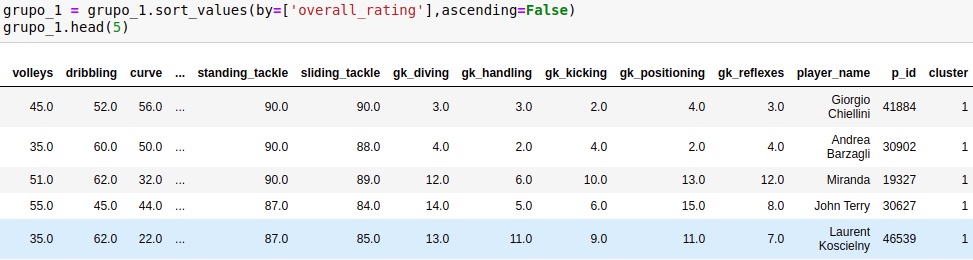


Figura 26: Registros do *cluster* 1. Fonte: Elaboração própria

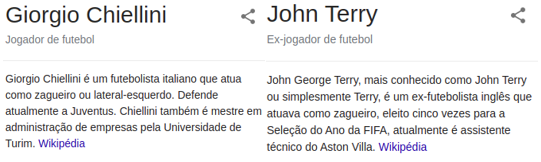


Figura 27: Jogadores. Fonte: Pesquisa do nome em [www.google.com](http://www.google.com)

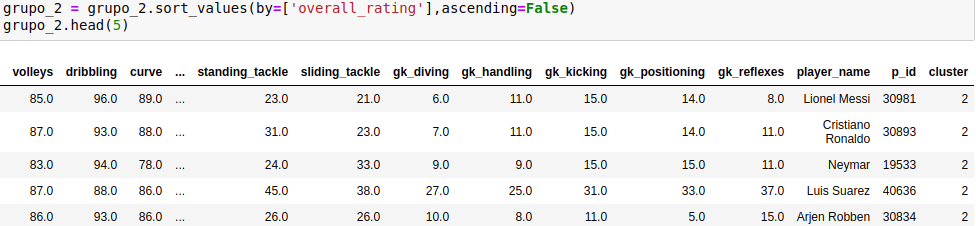


Figura 28: Registros do *cluster* 2. Fonte: Elaboração própria

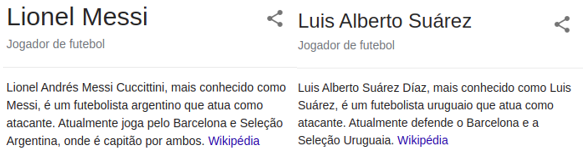


Figura 29: Jogadores. Fonte: Pesquisa do nome em [www.google.com](http://www.google.com)

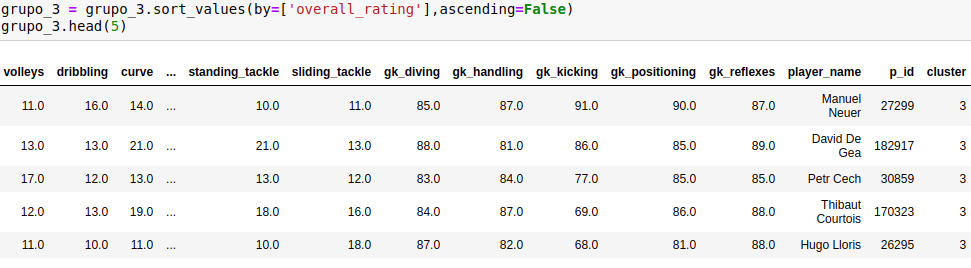


Figura 30: Registros do *cluster* 3. Fonte: Elaboração própria

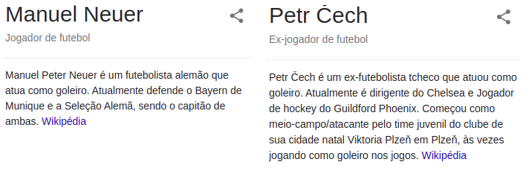


Figura 31: Jogadores. Fonte: Pesquisa do nome em [www.google.com](http://www.google.com)

Através das descrições reais apresentadas dos jogadores foi possível constatar que o *cluster* 3 agrupou os goleiros, como observado anteriormente e os *clusters* 0, 1 e 2 agruparam respectivamente jogadores que atuam no meio-de-campo, defesa e ataque.

Neste ponto do experimento é possível supor que já existe a resposta da pergunta proposta para esse processo de KDD, pois foi possível validar que os melhores jogadores de cada *cluster* estão relacionados a cada uma das quatro posições de atuação definidas para a escolha do valor K. Porém, para avaliar a eficiência do algoritmo e confirmar essa suposição, foi necessário observar registros de forma aleatória dentro de cada *cluster*. Foram criados novos DFs com nomes das variáveis fazendo referência ao rótulo do *cluster* e apenas alguns atributos necessários para posterior representação gráfica foram incluídos. As figuras representam cinco registros dispostos aleatoriamente e as figuras apresentam o resumo sobre dois jogadores de cada grupo.

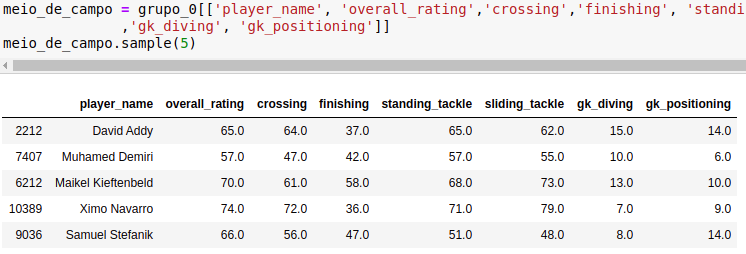


Figura 32: Amostra do grupo "meio-de-campo". Fonte: Elaboração própria

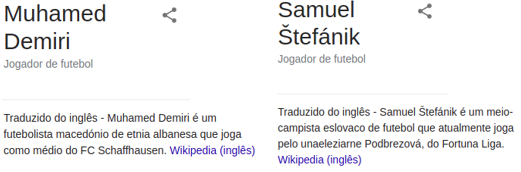


Figura 33: Jogadores. Fonte: Pesquisa do nome em [www.google.com](http://www.google.com)

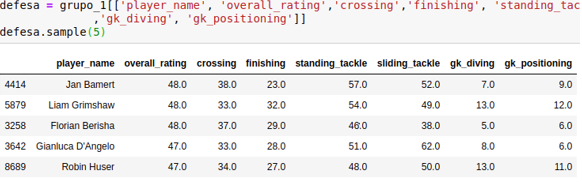


Figura 34: Amostra do grupo "defesa". Fonte: Elaboração própria



Figura 35: Jogadores. Fonte: Pesquisa do nome em [www.google.com](http://www.google.com)

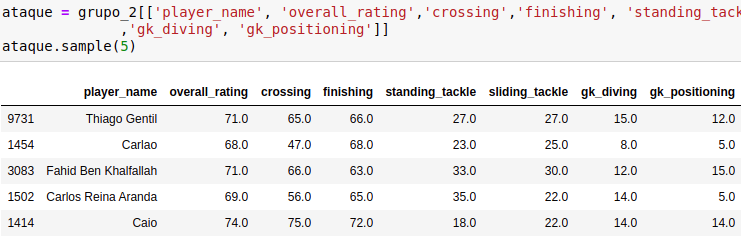


Figura 36: Amostra do grupo "ataque". Fonte: Elaboração própria

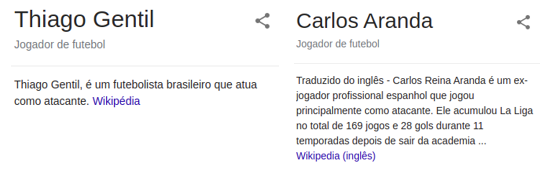


Figura 37: Jogadores. Fonte: Pesquisa do nome em [www.google.com](http://www.google.com)

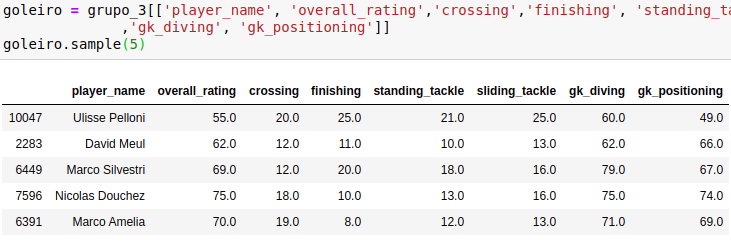


Figura 38: Amostra do grupo "goleiro". Fonte: Elaboração própria



Figura 39: Jogadores. Fonte: Pesquisa do nome no sitio [www.google.com](http://www.google.com)

Diante dos resultados foi possível confirmar a eficiência do algoritmo *K-Means*. Os jogadores escolhidos de forma aleatória realmente atuam ou atuaram nas posições que rotulam seus *clusters*. Concluiu-se que a pergunta levantada para esse experimento pode ser respondida através deste processo de KDD.

Foi interessante também fazer uma comparação entre alguns atributos utilizando o DF que deu origem aos quatro grupos estudados. As cores aplicadas para representação nas Figuras 40 e 41 estão de acordo com o *cluster* que cada registro foi agrupado, sendo elas azul, amarelo, vermelho e verde para os rótulos meio-de-campo, defesa, ataque e goleiro respectivamente.

A Figura 40 representa uma comparação entre as habilidades consideradas desejáveis para goleiros e atacantes. Observando como os registros foram plotados, notou-se que os *clusters* de defesa, meio-de-campo e ataque se mostraram extremamente opostos ao agrupamento de goleiros, como era de se esperar.

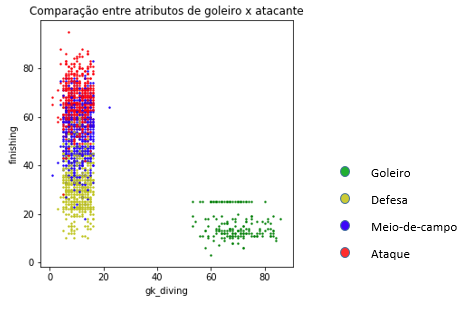


Figura 40: Comparação de atributos. Fonte: Elaboração própria

Uma outra comparação foi feita utilizando os atributos *finishing* e *standing\_tack*. O primeiro é desejável para atacantes e o segundo contempla tanto defensores quanto meio-campistas. O resultado é apresentado na Figura 41 e é possível visualizar de forma mais clara a disposição dos *clusters* agrupados pelo *K-Means*.

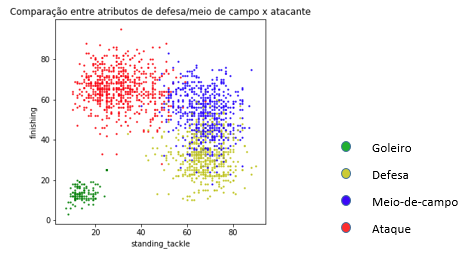


Figura 41: Comparação entre atributos. Fonte: Elaboração própria

A conclusão desse experimento foi que o algoritmo *K-Means* se mostrou eficiente para fazer a clusterização de dados. O processo de KDD demonstrado foi capaz de responder à pergunta levantada de acordo com o contexto do BD.

# CONSIDERAÇÕES FINAIS

O processo de KDD definido no capítulo 2 mostrou-se eficiente pois foi possível gerar um conhecimento sobre dados não categorizados. Suas etapas foram seguidas e o resultado gerado foi capaz de responder à pergunta levantada para o experimento.

Foi necessário realizar uma análise nos dados para identificar o melhor modelo para se aplicar neste trabalho. A base de dados tem um volume grande de registros que possibilitou simular uma atividade de Mineração de Dados e analisar em tempo real o desempenho do hardware do computador ao processar o experimento. O algoritmo *K-Means* em si não possui alta complexidade computacional (um dos motivos de sua popularidade) e o que foi observado é que o maior tempo de processamento foi dedicado à biblioteca de impressão gráfica.

O experimento realizado neste trabalho representa apenas um dos caminhos possíveis de se seguir na área de Mineração de Dados. Seguindo a estruturação proposta é possível aplicar esse experimento em infinitos contextos, visto que a cada dia estamos produzindo mais dados, uma vez que novos sistemas de informação surgem todos os dias.

Observa-se que cada vez mais investimentos serão aplicados na área de MD pois a informação gerada pelo processo tem um enorme potencial econômico, se bem aplicadas.

A linguagem de programação Python usada para implementação do trabalho possui vasta documentação para a área de MD, sendo uma das principais utilizadas por cientistas de dados. Há diversas bibliotecas que tratam desse contexto e o crescimento desproporcional destas bibliotecas específicas sobre outras que também fazem parte da esfera de bibliotecas do Python indica qual rumo o futuro da linguagem deve assumir. Esse crescimento implica em melhora da performance e ampliação de funcionalidades para as bibliotecas, que possibilitam resultados mais precisos. A performance dessa linguagem permitiu cientistas em 2019 capturarem pela primeira vez na história a imagem de um buraco negro. Duas bibliotecas usadas para obter essa imagem também foram utilizadas nesse trabalho, sendo elas a *Numpy* e *Matplotlib.*

Para um trabalho futuro, é possível propor um estudo que tem como objetivo melhorar a distribuição de políticas sociais para a população baseando-se na necessidade de cada cidadão. O projeto consiste em realizar uma pesquisa para elaboração de um formulário com perguntas que simbolize necessidades que levam ao bem-estar da sociedade. Cada pergunta terá um grau de concordância que cada pessoa deverá responder de acordo com seu perfil. O resultado dessa pesquisa serviria de entrada para um trabalho de MD voltado para o agrupamento de dados, posteriormente seria feito um estudo sobre cada grupo para traçar estratégias que resolvam as principais necessidades apontadas pelos membros associados a cada grupo afim de lhes proporcionarem políticas mais aplicadas com objetivo de melhorar índices sociais da população.

# REFERÊNCIAS BIBLIOGRÁFICAS

ADRIAANS, Pieter e ZANTINGE, Dolf. **Data Mining**. London, UK: Addison-Wesley, 1997.

AGRAWAL, R; SRIKANT, R. **Fast Algorithms for Mining Association Rules.** Proceedings of the 20th VLDB Conference. Chile, 1994. Disponível em:< AGRAWAL, R; SRIKANT, R. **Fast Algorithms for Mining Association Rules.** Proceedings  
of the 20th VLDB Conference. Chile, 1994.>. Acesso em: 15 set. 2019.

AMO, Sandra de. **Técnicas de Mineração de Dados.** Jornada de Atualização em Informática. Universidade Federal de Uberlândia, 2004.

BATISTA, Gustavo Enrique de A. P. A. B. **Pré-Processamento de Dados em Aprendizado de Máquina Supervisionado.** Tese de Doutorado (Doutorado em Ciência da Computação e Matemática Computação). USP – São Carlos, 2003. Disponível em:<http://www.teses.usp.br/teses/disponiveis/55/55134/tde-06102003-  
160219/publico/TeseDoutorado.pdf>. Acesso em: 18 set. 2019.

BERRY, M.J.A.; LINOFF, G. **Data Mining Techniques**. New York: John Wiley & Sons, Inc. 1997.

CAMILO, Cássio Oliveira; SILVA, João Carlos da. **Mineração de Dados: Conceitos,  
Tarefas, Métodos e Ferramentas**. Instituto de Informática. Universidade Federal de  
Goiás, 2009. Disponível em: < <http://www.inf.ufg.br/sites/default/files/uploads/relatoriostecnicos/RT-INF_001->

CARDOSO, G. S. et al. Clusterização K-Means: Uma proposta de melhoria. In: *Workshop de Trabalhos de Iniciação Científica e Graduação Bahia, Alagoas e Sergipe*.  
Salvador: **Anais** do ERBASE 2008.

CASSIANO, K. M. *Análise de Séries Temporais Usando Análise Espectral Singular  
(SSA) e Clusterização de Suas Componentes Baseada em Densidade*. Tese (Doutorado)— PUC-RIO, Rio de Janeiro, 2014.

CASTANHEIRA, Luciana Gomes. **Aplicação de Técnicas de Mineração de Dados em Problemas de Classificação de Padrões.** Laboratório de Computação Evolucionária da Universidade Federal de Minas Gerais, 2008.

DALLANORA, Julio Fernando. **Análise do Perfil de Clientes a partir de Técnicas de  
Mineração de Dados.** UNIJUI – Universidade Regional do Noroeste do Estado do Rio  
Grande do Sul, Ijuí, 2009.

FAYYAD, Usama; Piatetsky-Shapiro, G; Smyth, P. ***Data mining* to *Knowledge Discovery in Databases*.** American Association for Artificial Intelligence, 1996.

HAN, J; KAMBER, M. **Data Mining: Concepts and Techniques**. Elsevier, 2006.

HARTIGAN, J. A. e WONG M. A. **Algorithm AS 136: A K-Means Clustering Algorithm.** Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 28, No. 1(1979), pp. 100-108

HOSKING, J.R.M.; WALLIS, J.R. Regional frequency analysis: an approach based on L-moments. Cambridge University Press, Cambridge , 224 p., 1997.

JAIN, P. *Large Scale Optimization Methods for Metric and Kernel Learning*. Dissertação (Mestrado) — University Of Texas At Austin, Texas, 2009.

JAIN, Anil e DUBES, Richard. **Algorithms for clustering data.** Prentice-Hall, Inc. Upper Saddle River, NJ, USA ©1988

LIBERTY, Edo. **Algorithms in Data Mining.** 2016 Acesso em 25 de novembro, 2019 em: https://pdfs.semanticscholar.org/0074/4cb7cc9ccbbcdadbd5ff2f2fee6358427271.pdf

LINDEN, Ricardo. Técnicas de Agrupamento. **Revista de Sistemas de Informação da FSMA.** N. 4 p. 18-36. 2009.

MACQUEEN, J. **Some Methods For Classification And Analysis Of Multivariate Observations.** [*Proc. Fifth Berkeley Symp. on Math. Statist. and Prob.*, Vol. 1 (Univ. of Calif. Press, 1967)](https://projecteuclid.org/euclid.bsmsp/1200512974), 281-297

MCCUE, C. **Data Mining and Predictive Analysis** - Intelligence Gathering and Crime Analysis. Elsevier, 2007.

NALDI, M. C. **Técnicas de Combinação para o Agrupamento Centralizado e Distribuído de Dados**. Tese (Doutorado em Ciências de Computação e Matemática Computacional). Instituto de Ciências Matemáticas e de Computação – ICMC – USP, 2011.

PRASS, F. S. (2007). KDD – **Uma Visão Geral do Processo**. Acesso em 10 de junho, 2019 em: <http://fp2.com.br/blog/wp-content/uploads/2012/07/kdd_uma_visao_geral_do_processo.pdf>

REZENDE, Solange O. **Sistemas Inteligentes: Fundamentos e Aplicações.** Barueri, SP: Ed.Manole, 2003.

REZENDE, Solange O. **Mineração de Dados.** XXV Congresso da Sociedade Brasileira de  
Computação. UNISINOS – São Leopoldo, RS, 2005. Disponível em:<http://www.lbd.dcc.ufmg.br/colecoes/enia/2005/0102.pdf>. Acesso em: 27 set. 2019

SILVA, Denilson R. da. **Análise e Triagem de Padrões em Processamento de Descobertade Conhecimento em Base de Dados.** Dissertação de Mestrado (Mestrado em Ciência daComputação) - Pontifícia Universidade Católica do Rio Grande do Sul – RS, 2000.Disponível em: <http://www.pucrs.br/inf/pos/dissertacoes/arquivos/denilson.pdf>. Acesso  
em: 02 set. 2019.

SLONIM, N.; AHARONI, E.; CRAMMER, K. Hartigan’s k-means versus lloyd’s *k-means*– is it time for a change? In: *International Joint Conference on Artificial Intelligence*.[S.l.: s.n.], 2013. p. 1677–1684.

THOMSEN, E. **OLAP Solutions: Building Multidimensional Information Systems***.* John Wiley & Sons, Inc. New York, NY, USA, 2002.

VERGARA, Sylvia C. **Projetos e Relatórios de Pesquisa em Administração.** 2ª ed. São Paulo: Atlas, 1998.

# APÊNDICE

1. **Algoritmo K-Means**

