Relatório 3

Paulo Henrique Ferreira C. Mendes

26 de maio de 2018

Introdução e Contextualização

O presente relatório tem por objetivo descrever os datasets da Matemática tendo como objeto de estudo o programa de pós-graduação oferecido pela Universidade de Brasília. O programa de Pós-Graduação em Matemática da UnB, iniciado em 1962, oferece Mestrado e Doutorado em Matemática nas subáreas de Álgebra, Análise, Geometria e Matemática Aplicada (Probabilidade, Física-Matemática e Computação). O corpo docente mantém um programa ativo de pesquisa, participa regular e ativamente de forma destacada em reuniões científicas e em corpos editoriais de revistas científicas, além de manter intercâmbio científico com diversas instituições do país e do exterior[1].

O desenvolvimento do relatório se deu em cima dos arquivos oferecidos pela discilina DataScience4All com intuito de fomentar o ensino de ciência de dados e em contra-partida produzir resultados reais no âmbito acadêmico para fornecer análises reais e úteis no que diz respeito a produção científica da UnB. O datasets focam as principais áreas dentro da Matemática quem que houve desenvolvimento de pesquisas de pós-graduação, os professores envolvidos, o nível de participação em eventos nacionais e internacionais e a presença da produção científica em periódicos internacionais.

O primeiro dataset analisado foi o de perfis, um dos mais ricos em informações sobre os pesquisadores. Todos os pesquisadores são indexados de acordo com seu ID do Currículo Lattes o que permite acessar cada um pontualmente assim como fazer relações com os demais datasets que também utilizam o mesmo identificador. Dentro de cada ID é possível ler informações como nome e o resumo do Currículo Lattes além das áreas e sub-áreas de atuaçao, orientações dadas pelo pesquisador e sua produção bibliográfica o que facilita o entendimento do grau de participação e produção científica de cada um.

As áreas de desenvolvimento científico são estudadas com auxílio do dataset de Áreas de Atuação onde as produções científicas desenvolvidas na pós-graduação em Matemática são agrupadas de acordo com a área do conhecimento em que melhor se encaixam. As seis áreas encontradas mostram a diversidade das pesquisas e a possibilidade do pós-graduando atuar em diversos segmentos do saber.

O dataset de publicação aparece como validação das informações já estudadas em trabalhos anteriores com auxílio dos datasets OASIS e BDTD que também tem como foco principal as produções científicas da Universidade de Brasília. Com várias fontes em mãos é possível utilizar as ferramentas da linguagem R a fim de fortificar os resultados qualitativos no que diz respeito às dissertações e seus autores que foram aparecendo a medida em que foram analisados, principalmente, os arquivos .json com informações do Currículo Lattes.

Referencial Teórico

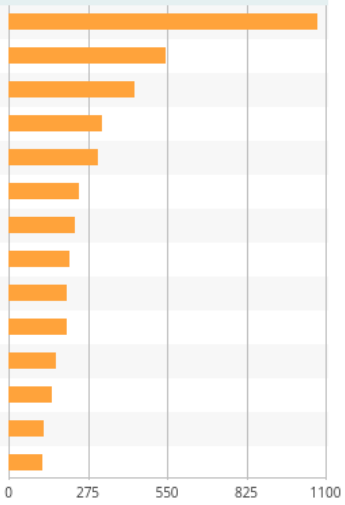
O Currículo Lattes é uma plataforma online desenvolvida pelo CNPq e tem por objetivo principal unificar as informações pessoais, acadêmicas e profissionais dos pesquisadores brasileiros em um único perfil, o Lattes é utilizado amplamente em todo o país e conta com mais de um milhão de perfis que podem ser consultados no domínio do CNPq. Os dados principais que podem ser encontrados a respeito de determinado pesquisador são principalmente: nome do autor, data de publicação, data de defesa das teses, palavras-chave, orientador, resumo, área de conhecimento, instituição. Dadas essas informações o Lattes se mostra como uma boa fonte de pesquisa e aplicação de conceitos e ferramentas de ciência da Ciência visto que as várias metodoligias de estudos deste ramo do saber podem ser aplicadas para extrair informações específicas sobre determinada área ou grupo de pesquisadores.

A título de refência neste contexto pode-se citar Jesús Mena-Chalco, professor e pesquisador de Ciência da Ciência que tem em um de seus trabalhos uma obra focada na investigação do relacionamento entre as orientações passadas por professores a seus alunos e os trabalhos que estes vieram a produzir. Este é apenas um exemplo para ilustrar as pesquisas em Ciência da Ciência e bases de dados como Crrículo Lattes que são capazes de avaliar como o conhecimento é disseminado através das pesquisas e como o conhecimento é produzido no âmbito universitário. De posse desses dados até mesmo as industrias podem extrair valor ao buscar novas tendências principalmente no campo da Ciência e Tecnologia tornando-se precursoras na produção de tal ou qual produto/serviço.

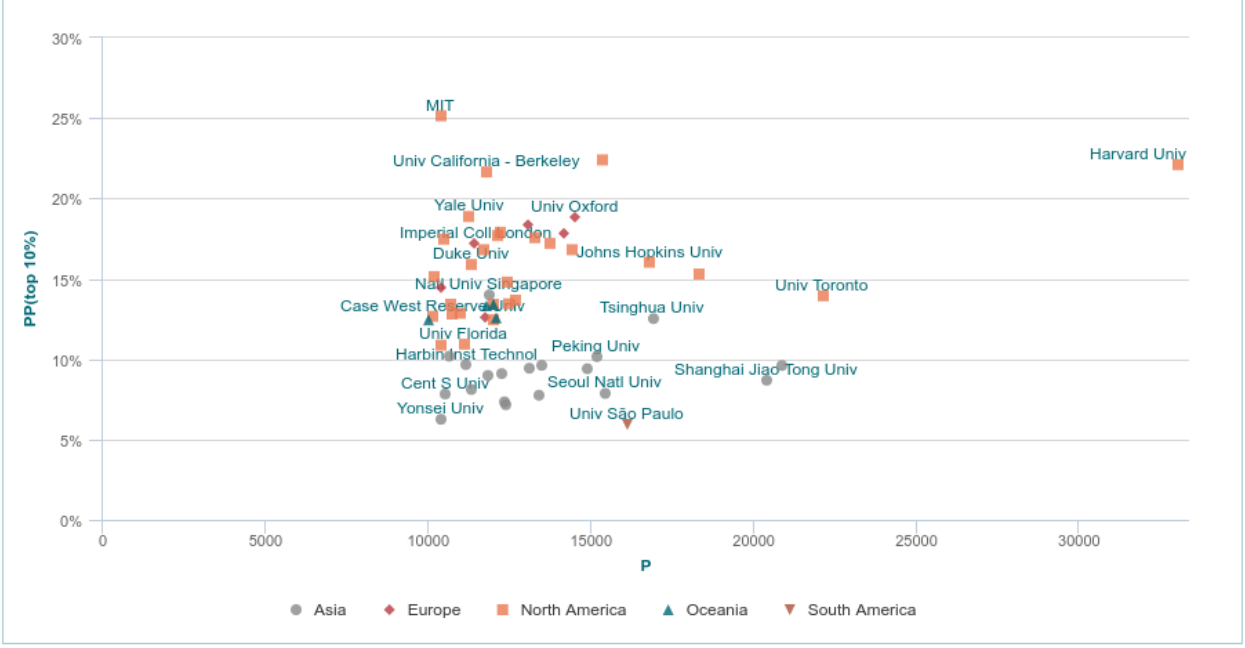
O cenário internacional conta com a CWTS Leiden Ranking que é um grande repositório de informações sobre o desempenho científico das mil maiores universidades do mundo. É possível filtrar os dados a partir de vários indicadores chamados de "bibliométricos". A mais conhecida é a exibição de lista tradicional onde é possível classificar as universidades de acordo com um indicador selecionado. São utilizadas basicamente duas perspectivas para ilustrar as produções das universidade: a visualização do gráfico mostra as universidades em um gráfico de dispersão, visando explorar o desempenho das universidades usando dois indicadores selecionados e a vista do mapa mostra as universidades e fornece uma perspectiva geográfica sobre as universidades e seu desempenho.

Ranking de Universidades Brasileiras por número de publicações

	Universidade	P	P (top 10%)	PP (top 10%)
1	Univ São Paulo	1071	82	7,6%
2	Univ Campinas	545	32	6,0%
3	Univ Fed Rio de Janeiro	438	25	5,8%
4	Univ Fed Minas Gerais	325	29	8,9%
5	Fed Univ Pernambuco	312	26	8,2%
6	Fed Univ Rio Grande do Sul	247	16	6,5%
7	Fed de Santa Catarina	228	15	6,7%
8	Univ Estadual Paulista	213	14	6,5%
9	Univ Brasília	204	13	6,3%
10	Univ Fed Ceará	204	21	10,1%
11	Univ Fed Fluminense	166	17	10,0%
12	Fed Univ São Carlos	149	14	9,2%
13	Fed Univ Rio Grande do Norte	120	6	5,2%
14	Univ Fed Paraná	116	11	9,7%



Ranking de mundial de Universidades por número de publicações



Metodologia

O presente relatório tem como foco principal o resultados obtidos através da análise do datasets de ciência e pós-graduação da Matemática. Com esse intuito fez-se importante a utilização de uma metodologia mais específica afim de preparar todo o caminho até as conclusões. Trata-se do CRIPS-DM[2].

O Cross Industry Standard Process for Data Mining (CRISP-DM) é uma metodologia desenhada para mineração de dados focados em datasets com grandes volumes de dados. O CRISP-DM é baseado em etapas de estudo que visa dar o foco necessário proporcionando a melhor obtenção de informações possível.As fases são:

1- Entendimento do negócio: Busca entender de uma forma geral a função e o contexto dos dados. Foca em detalhes que são importantes para o cliente ou organização. É onde aparecem os primeiros "por quê's" e "como's".

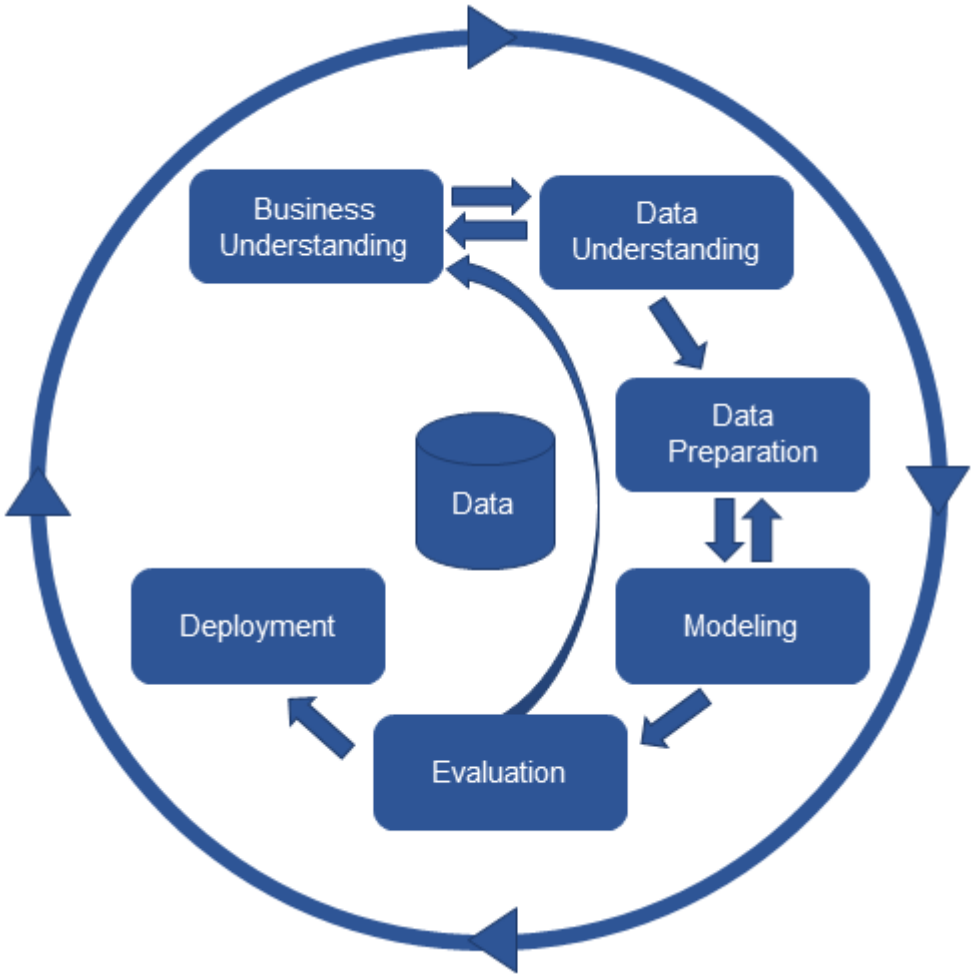
2- Compreensão dos dados: Foco em inspecionar, organizar e descrever todos os dados disponíveis. Verificar possíveis informações falsas ou lacunas presentes nos datasets.

3: Preparação dos dados: Preparar os datasets e definir os seus formatos visando sua futura apresentação e análise. Neste momento são escolhidos os atributos a serem minerados.

4- Modelagem::Fase em que são aplicadas as ferramentas e tecnicas de mineração de dados.

5- Avaliação:É o momento de acompanhamento dos resultados objetivos e a estudo da aplicação dos insights e conhecimentos obtidos.

6- Desenvolvimento:De posse dos resultados obtidos com a mineração dos dados é dado início às mudanças ou implementações necessárias visando os fins desejados.



A ferramenta principl foi linguagem R, utilizada ao longo do desenvolvimento do trabalho. Mais uma vez foram utilizadas as bibliotecas e funções para importação e limpeza de dados, todos em formato .json, além das manipulações visando uma melhor apresentação dos resultados obtidos e seu estudo. A produção dos gráficos foram executadas com auxílio da biblioteca ggplot2.A principais bibliotecas e funções utilizadas foram:

1-**library(jsonlite), library(dplyr),library(ggplot2)**

2-**fromJSON(),names(),head(),sapply(),lenght(),data.frame()**

rbind(),dim(),select(),glimpse(),filter(),ggplot()

Resultados e Análises

A mineração dos dados fornecidos pelos datasets em estudo seguem como referência o CRISP-DM, por isso, visando melhor desenvolvimento e apresentação dos resultados, os resutaldos foram agrupados de acordo com cada fase da metodologia em questão.

Fase 1 - Entendimento do Negócio

O primeiro dataset analisado foi o MatematicaRedeResearchers.json que contém todas as áreas em que existem algum tipo de produção científica, a nível de pós-graduação, produzida ou em produção até 2017 por membros

do Departamento de Matemática. Como é possível notar os temas desenvolvidos não são exculivamente da Matemática ou ainda das Ciências Exatas, existe uma flexibilização que permite ao pesquisador trabalhar em diferentes campos do saber. Outro dado interessante é o desenvolvimento de trabalhos na área da educação o que valida as informações apresentadas no site do departamento onde são apresentados diversos projetos nesse âmbito[4].

print(administração)

[1] 1

print(educação)

[1] 4

print(engBioMédica)

[1] 1

print(matemática)

[1] 26

print(ProbeEstat)

[1] 2

print(Psicologia)

[1] 2

Fase 2 e 3 - Compreensão e Preparação dos dados

Visando aprofundar as análises será dado foco ao grupo de quatro pesquisadores do Departamento de Matemática que desenvolveram trabalhos na área da Psicologia. Estes podem ser identifados pelo número de seu ID do Currículo Lattes, como já é conhecido temos dois resultados:

```
##  chr [1:2] "0556476746202406" "5874654544324539"
```

De posse dos IDs é possível fazer algumas observações, para tal foi adicionado o dataset MatematicaRede.profile.json que contém o perfil Lattes de todos os membros do departamento. De posse dessas informações é possível comprovar a existência dos pesquisadores com os ID que foram encontrados e ainda verificar se eles, mesmo sendo "nativos" da Matemática, possuem alguma pesquisa na área da Psicologia olhando suas áreas de atuação:

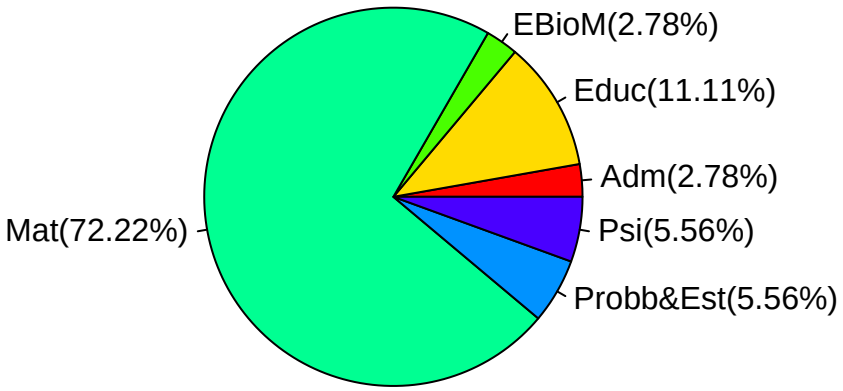
```
str(pesq1 <- data_frame(perfil$`0556476746202406`$areas_de_atuacao$area))
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    3 obs. of  1 variable:
##  $ perfil$`0556476746202406`$areas_de_atuacao$area: chr  "Matemática" "Educação" "Psicologia"
```

```
str(pesq2 <- data_frame(perfil$`5874654544324539`$areas_de_atuacao$area))
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    4 obs. of  1 variable:
##  $ perfil$`5874654544324539`$areas_de_atuacao$area: chr  "Matemática" "Educação" "Educação" "Psicolog
```

Ainda sobre as áreas de pesquisa é possível notar, como talvez já fosse esperado, que as sub-áreas da Matemáticas são majoritariamente as mais buscadas para desenvolvimento de pesquisas e isso pode ser verificado no gráfico a seguir que mostra a distribuição das pesquisas por área e diante do número pode-se afirmar que as sub-áres da Matemática são as de maior interesse e por consequência possuem os **resultados**



mais relevantes.

Fase 4 - Modelagem

Limitando o escopo de análise aos pesquisadores que se dedicaram à pesquisas envolvendo temas da Psicologia é interessante analisar que se em algum momento antes de adquirirem seus títulos eles se dedicaram de alguma forma ao estudo da Psicologia. Para isso é feita uma análise dos resumos dos seus currículos Lattes. Devido

```
resumo_pesq1 <- (perfil$`0556476746202406`$resumo_cv )
resumo_pesq2 <- (perfil$`5874654544324539`$resumo_cv)
resumo_pesq1
```

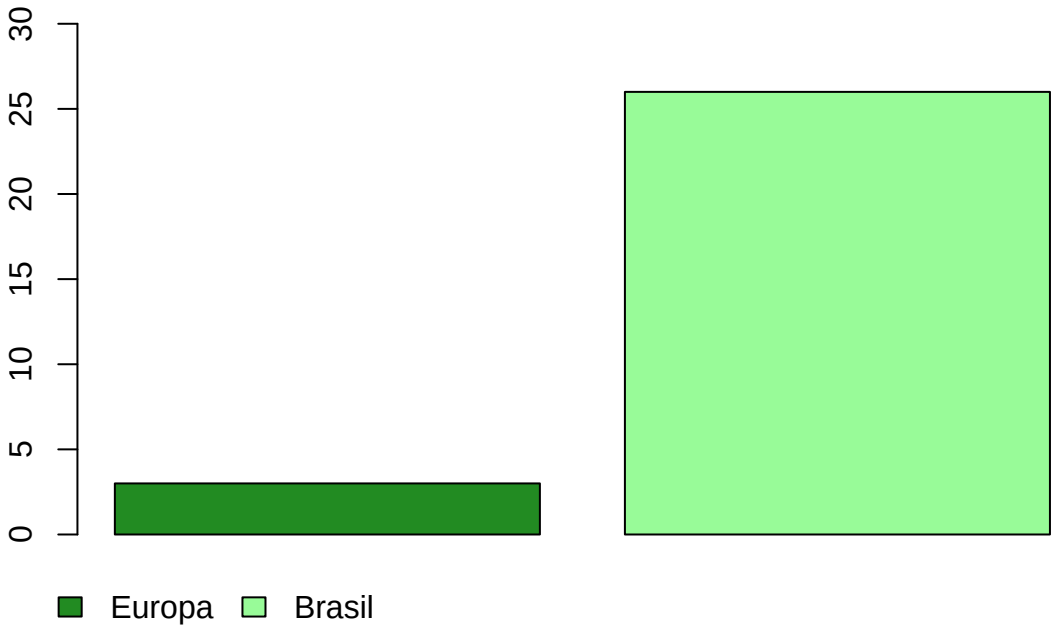
```
## [1] "Professor Associado I na Universidade de Brasília - UnB, com lotação no Departamento de Matemáti
```

```
resumo_pesq2
```

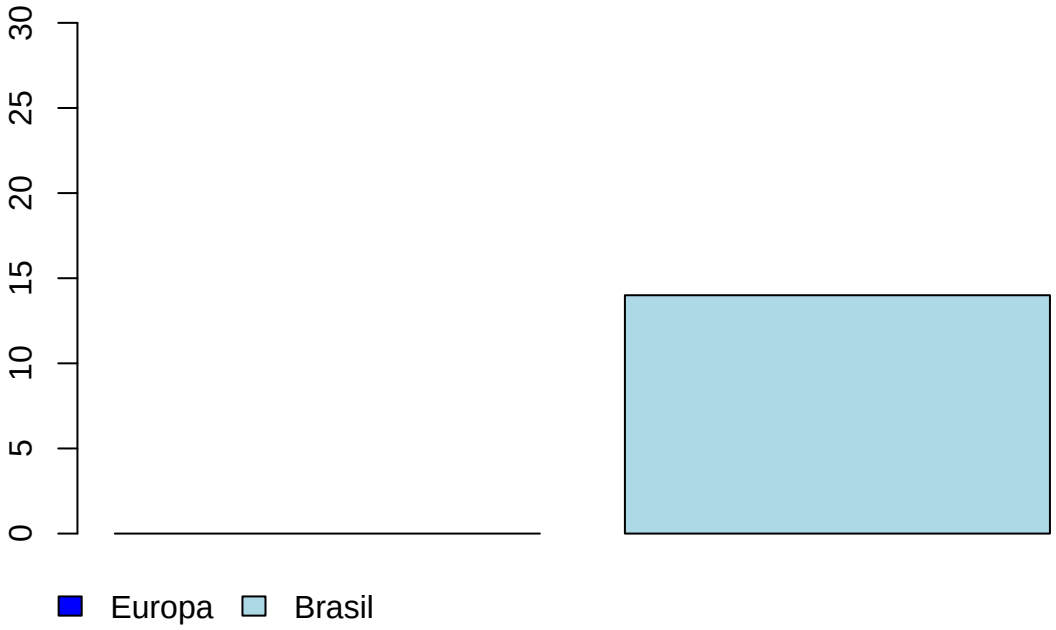
```
## [1] "Possui Licenciatura em Matemática (1995) e Especialização em Matemática (1998) pela Universidade
```

Outro dado interessante é analisar o nível de internacionalização das produções dos pesquisadores, isso pode ser analisado com a quantidade de participações em eventos internacionais e nacionais.Como é possível ver através dos gráficos a internacionalização das pequisas pode ser considerada baixa, porém existe forte participação desses pesquisadoes em âmbito nacional.

Pesquisador ID_0556476746202406



Pesquisador ID_5874654544324539



Fase 5 e 6 - Avaliação e Desenvolvimento

O simples estudo dos datasets da Matemática permitiu tirar conclusões acerca de seu programa de pós-graduação e validar as informações contidas no site do programa. Assim como havia sido afirmado anteriormente o Departamento é capaz de oferecer oportunidades de estudo em diversos campos do saber, ainda assim não deixa de privilegiar as sub-áreas das Ciências Exatas como a Álgebra e a Estatística. De acordo com os resultados obtidos é possível notar que as áreas de pós-graduação dentro da Matemática que não trabalham diretamente com sub-áreas das exatas não possuem o mesmo grau de internacionalização que as demais, porém por serem numericamente inferiores não tem impacto direto na nota do curso que é alta, outra informação que foi validade de acordo com o que está no site do programa.

Bibliografia

Pós-gradução-MAT.http://www.mat.unb.br/pagina/pos-sobre. Acessado em 26/05/2018

The CRISP-DM process model (1999), http://www.crisp-dm.org/

WIVES, Leandro Krug; PALAZZO, M. de Oliveira, José. Um estudo sobre Agrupamento de Documentos Textuais em Processamento de Informações não Estruturadas Usando técnicas de Clustering. Porto Alegre: CPGCC, 1999.CHAPMAN, P.; CLINTON, J.; KERBER, R.; KHABAZA, T.; REINARTZ, T.; SHEARER, C. & WIRTH, R. CRISPDM 1.0 step-by-step data mining guide. Technical report, CRISP-DM.

Pesquisa-MAT.http://www.mat.unb.br/Pesquisa-Publicacoes. Acessado em 26/05/2018