Universidade Federal de Minas Gerais
Departamento de Ciência da Computação

DCC057 Mineração de Dados

Trabalho Prático 2

Alexander Thomas Mol Holmquist

26 de fevereiro de 2021

1 Introdução

É muito difícil escolher um imóvel adequado aos desejos do cliente. Este é talvez o maior desafio, no que se diz a clientes compradores, no ramo de corretoria. Como dito na proposta deste trabalho, mas não ficou explícito, o objetivo aqui é ajudar os profissionais corretores a disporem opções interessantes para os consumidores, em frente a um conjunto de restrições. Primeiro restringimos o banco de dados às casas que obedecem às restrições impostas. Depois, agrupa-se os dados em um número de grupos escolhido automaticamente. Então, o corretor deve analisar os centroides de cada grupo, e criar uma descrição atrativa e que deixe clara a distinção entre esses grupos, para o comprador.

O processo, que denominou-se ”Corretoria Automatizada”, é ilustrado pela figura abaixo. Uma nota paralela, o repositório do projeto, contendo todo o desenvolvimento, pode ser encontrado em [Git (2021)]. O notebook ”mining.ipynb”deve guiar o leitor, em passos simples, pelo desenvolvimento do projeto.

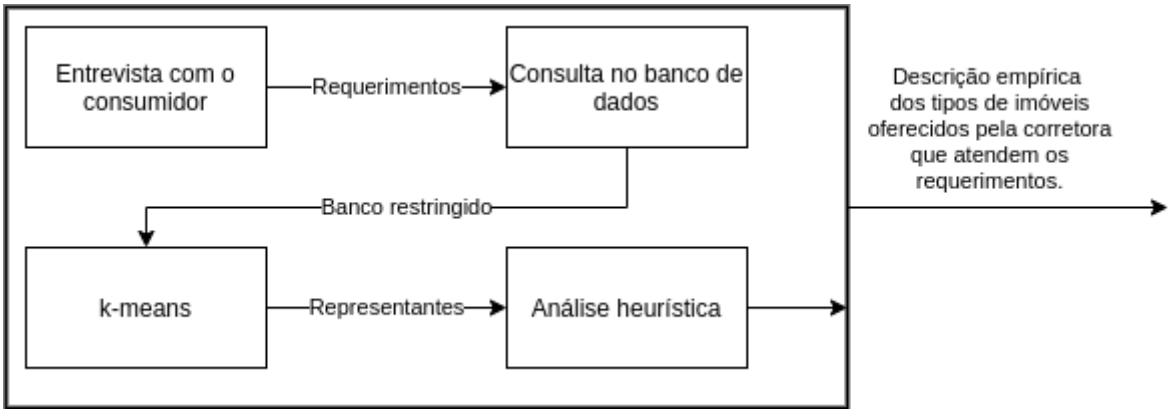

Figura 1: Diagrama de Corretoria Automatizada

Como a proposta não deixou claro o objetivo do projeto, um exemplo é dado a seguir. Imagine-se que um cliente busca um imóvel com preço entre 500 mil e 700 mil dólares, que seja novo, e possua boas condições gerais de uso (bem cuidado). O corretor então poderia fazer uma consulta no banco de dados, da forma:

```
SELECT * FROM KCDS WHERE
  price BETWEEN 500000 AND 700000 AND
  yr_built >= 2000                AND
  condition >= 4;
```

Note que aqui ”KCDS”significa ”King County housing dataset”[KCD (2020)]. O KCDS, conhecido amplamente, oference uma gama muito mais ampla que o ”Boston housing dataset”, mencinado na proposta. Por exemplo, neste existe o atributo ”waterfront”, que diz se o lar possui ou não uma vista frontal com elementos aquáticos.

O próximo passo é aplicar o algoritmo de agrupamento k-means para encontrar representantes (os centroides) para estes dados. Uma vez que se tem os representantes, é trabalho do corretor analisá-los, e propor descrições apresentáveis e que distinguam bem as diferentes categorias geradas pelo algoritmo.

Na proposta havia sido dito que a busca era restrita a agrupamentos com 5 centroides. Como foi uma restrição imposta sem respaldo técnico ou de especialista, ela passou a ser desconsiderada. Buscaremos o número satisfatório de grupos, de acordo com a técnica automatizada da estatística gap.

2 Metodologia e ferramentas

Inicialmente, o plano era utilizar a plataforma *Lemonade* para realizar as operações de mineração de dados. Porém, visto que encontrou-se mais pesares que benefícios, transicionou-se para codificação manual dos ”fluxos de execução”em *python 3*, com o uso de bibliotecas clássicas como *scikitlearn*.

O plano de seguir à risca o processo de CRISP-DM (versão 1.0) foi um sucesso quase completo. Quase todos os passos recomendados pela referência[CDM (2000)] foram realizados, um por um. Por isso, o presente documento, a partir da próxima seção, é dividido de acordo com as fases do mesmo.

Cada fase é composta de tarefas. Cada tarefa tem uma descrição geral e uma saída, que pode ser usada em tarefas posteriores. Este relatório contém a visão futura, ou uma revisão, do processo CRISP-DM, no sentido de que é feita uma retrospectiva, em cada tarefa, já indicando o que o projeto conseguiu, ou não, atingir.

Nem todos os detalhes das tarefas foi realizado, a saber aqueles que se relacionam mais com uma situação real de negócio, que não é o presente caso.

3 Fase 1 – Entendimento do negócio

Esta fase foca em estabelecer, da forma mais clara possível, quais os objetivos do negócio, e então quais os objetivos da mineração de dados, e como ela pode ajudar o negócio a alcançar seus objetivos.

3.1 Determinar objetivos de negócio

1. Facilitar a pesquisa de corretores por imóveis que atendam as necessidades do cliente.
2. Sistematizar este processo.

O segundo objetivo é atingido pelo processo que se chamou de ”Corretoria Automatizada”, descrito na introdução. O primeiro objetivo não foi concretizado no desenvolvimento do projeto. Poderia ser mensurado, em uma aplicação, por um tempo probatório de três meses, em que a metodologia é utilizada. Se o tempo médio diminuir mais que a hipótese nula ou outros fatores indicarem, poder-se-ia concluir que houve sucesso.

3.2 Avaliar a situação

O principal fruto desta tarefa foi descobrir a origem do banco de dados que havia sido escolhido desde a proposta do projeto, e sua natureza pública e de confiança. Diversas tarefas de mineração de dados, que podem ser tomadas como exemplos de uso, foram encontradas[UC1 (2016), UC2 (2020), UC3 (2016)].

3.3 Determinar objetivos da mineração de dados

O objetivo da mineração de dados é, dado um banco de dados restringido pela consulta feita pelo corretor, agrupá-lo em K grupos. As métricas C-index e Dunn index seriam utilizadas para medir a qualidade do agrupamento final. Isto não foi feito porque houve complicações de memória no computador local, ao calcular o C-index, e não houve tempo de desenvolver um algoritmo especializado.

3.4 Produzir plano de projeto

Este plano é a proposta do projeto. Nota-se que o plano inicial não foi explicito em demonstrar os objetivos do projeto. Também falhou em detalhar cada passo a ser tomado.

4 Fase 2 – Entendimento dos dados

4.1 Coletar dados iniciais

Os dados são simplesmente o KCDS, que pode é disponível publicamente por várias fontes, como[KCD (2020)]. O local no repositório do GitHub é ”data/home_data.csv”.

4.2 Descrever os dados

Essa descrição foi dada por completo no arquivo ”doc/history/cdm-22.txt”. Aqui, fornecemos uma descrição mais rudimentar.

O banco de dados contém 21 colunas, e 21613 instâncias. O formato em que foi encontrado é CSV. Um arquivo YAML com propriedades básicas de cada atributo também pode ser encontrado em ”doc/history/data-description.yaml”.

Ele descreve a venda na área de King County, em Washington, EUA. O período registrado é maio de 2014 a maio de 2015.

Cada atributo é contemplado na simples relação abaixo. O significado de algumas palavras de nicho específico pode ser encontrado em [Glo (2017)].

- id: identificador único para o imóvel
- date: dia em que o imóvel foi vendido
- price: preço de venda do imóvel
- bedrooms: número de quartos
- bathrooms: razão entre o número de banheiros e ”bedrooms”
- sqft_living: tamanho da casa, em pés quadrados
- sqft_lot: tamanho do terreno, em pés quadrados
- floors: quantidade de andares da casa
- waterfront: ”1”se a casa tem uma vista frontal com uma região aquática, ou ”0”se não tem
- view: nota de 0 a 4 do quanto a vista da casa é boa
- condition: nota de 1 a 5 da condição geral e de conservação da casa
- grade: nota de 1 a 13 da qualidade de construção da casa
- sqft_above: área do imóvel, sem contar o porão
- sqft_basement: área do porão

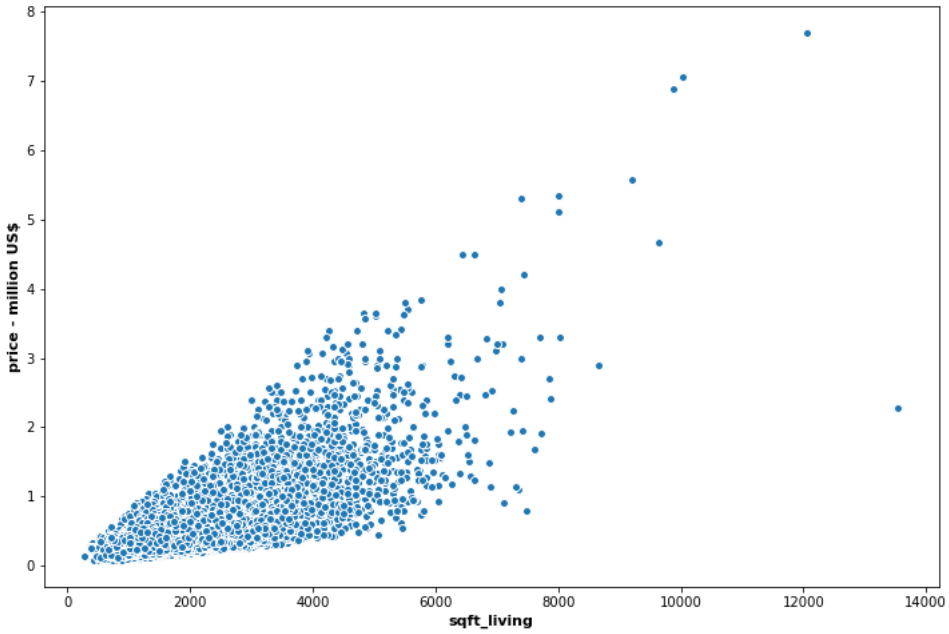

Figura 2: O preço do imóvel tem alta correlação com o tamanho do espaço de vivência (sqft_living)

- yr_built: ano em que o imóvel foi construído
- yr_renovated: ano em que o imóvel foi reformado, ou ”0”, se nunca foi
- zipcode: código postal
- lat: latidude
- long: longitude
- sqft_living15: tamanho médio da área de vivência das 15 casas vizinhas mais próximas, em pés quadrados
- sqft_lot15: tamanho médio do terreno das 15 casas vizinhas mais próximas, em pés quadrados

4.3 Explorar os dados

A partir desta tarefa, o projeto se prolongou excessivamente na direção de exploração dos dados. Apesar disso, abaixo são dadas descrições e ilustrações que ajudarão o leitor a se familiarizar com os dados em mãos.

Primeiramente, veja os gráficos que mostram a correlação entre dois atributos muito importantes e ”price”, que contém o preço de venda do imóvel (figuras 2 e 3).

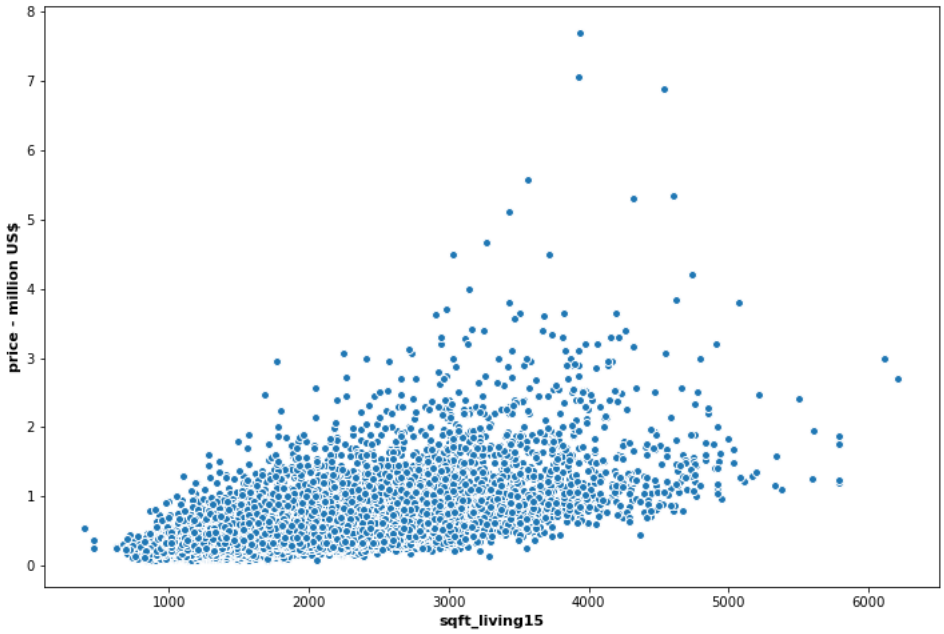

Figura 3: O preço do imóvel também tem alta correlação com o tamanho das casas vizinhas (sqft_living15)

4.4 Verficar a qualidade dos dados

Estranho que ”bathrooms”e ”floors”têm valores decimais. Não se conseguiu descobrir o porquê de ”floors”ter valores decimais, e por isso foi decidido não alterar. ”bathrooms”é a razão entre o número de banheiros e o número de quartos. Foi decidido não alterar este atributo também, em um ato de confiança na qualidade dos dados apresentados pela entidade pública, no que se refere a especialidade técnica.

5 Fase 3 – Preparação dos dados

5.1 Selecionar dados

O atributo ”id”foi removido, obviamente seu uso não apresenta qualquer vantagem. O mesmo acontece para a coluna ”date”, que representa a data da venda do imóvel. O período coberto por ”date”é somente de maio de 2014 a maio de 2015. Além disso, este não é uma característica representativa para um imóvel.

Os dados geográficos (lat, long, zipcode) foram removidos por simplicidade, mas poderiam ser usados em uma situação em que, a exemplo, o cliente pede por um imóvel próximo do seu local de trabalho.

Por fim, tem-se 16 colunas restantes.

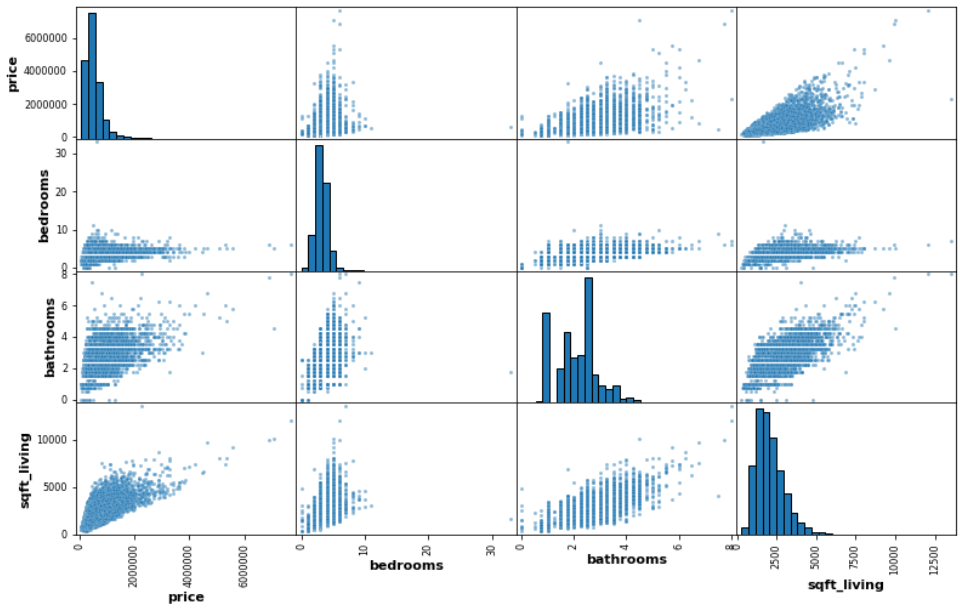

Figura 4: Os dados são multi-modais

5.2 Limpar dados

Os dados obtidos na fonte já vieram sem nenhum valor faltante. Isso parece ter sido resultado de uma imputação. De qualquer forma, mesmo que se tenha feito imputação para resultar nos dados que temos em mãos, isso não afeta significativamente os resultados de agrupamentos, a não ser que feito incorretamente.

5.3 Construir dados

Em nenhum momento se fez necessário criar atributos derivados ou transações artificiais. Também não houve necessidade de integrar tabelas, e por isso a tarefa 3.4 (Integrate data) não é mencionada.

5.4 Formatar os dados

A coluna ”yr_renovated”foi alterada: os valores ”0”foram substituídos pelo ano de construção da respectiva casa. Vale notar que todo o banco de dados original (exceto a coluna ”date”, que foi removida) continha valores numéricos. Mesmo as variáveis categóricas estavam representadas ordinalmente.

Decidiu-se por normalizar os dados pelo método Min-Max, isto é, mapear cada atributo para o intervalo [0, 1]. Não utilizou-se z-score ou quantis. A rejeição de z-score, e também da lei de potência, é devida ao comportamento assimétrico dos dados, em atributos muito importantes, como ”price”, ”condition”, ”sqft_living”. Alguns dos atributos têm comporamento gaussiano, outros têm comportamento logarítmico, mas não é possível aplicar uma distribuição só para todos (ver figura 4). A técnica de normalização por quantis foi rejeitada por simplicidade.

Além disso, a técnica min-max garante que atributos com muita variância, como ”price”, serão tratados de forma similar a atributos com pouca variância, a exemplo de ”waterfront”,

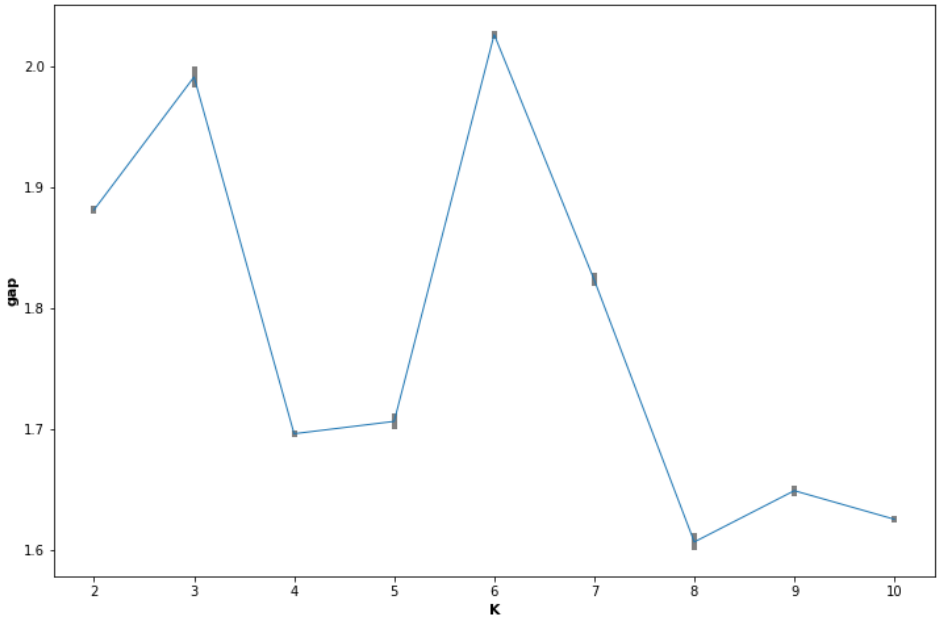

Figura 5: Distribuição da estatística gap

que é um atributo categórico binário. Assim, nosso modelo deve mostrar padrões mais interessantes, revelando a influência de atributos não tão determinantes na variância do conjunto completo dos dados.

6 Fase 4 – Modelagem

Esta fase é aqui resumida sem divisão por tarefas, por brevidade. O modelo aplicado foi o k-means, com o número de clusters verificado de forma interna e não supervisionada, através da estatística gap[gap (2001)]. O conjunto de dados mostrado na figura 6 é o resultado de uma redução de dimensionalidade com preservação de mais de 95% de variância. Esta redução ajuda a ignorar redundâncias.

O número de grupos considerado ideal foi três, como fica evidente pela distribuição da figura 5. As barras correspondentes a dois desvios padrões são quase imperceptíveis. Isto se deve provavelmente ao grande número de instâncias da tabela utilizada.

7 Fase 5 – Avaliação

Esta fase também será resumida nessa única seção. Para avaliar o projeto, devido à indisponibilidade de tempo para testes mais profundos, foi feito um teste somente, com o exemplo dado na Introdução. Este teste, assim como um passo a passo das principais etapas de preprocessamento, podem ser encontrados no *jupyter notebook* ”mining.ipynb”do repositório do projeto.

A consulta colocada na introdução, com os hipotéticos requerimentos de um cliente, resulta em uma tabela com apenas seis instâncias. Este número é muito pequeno para que os algo-

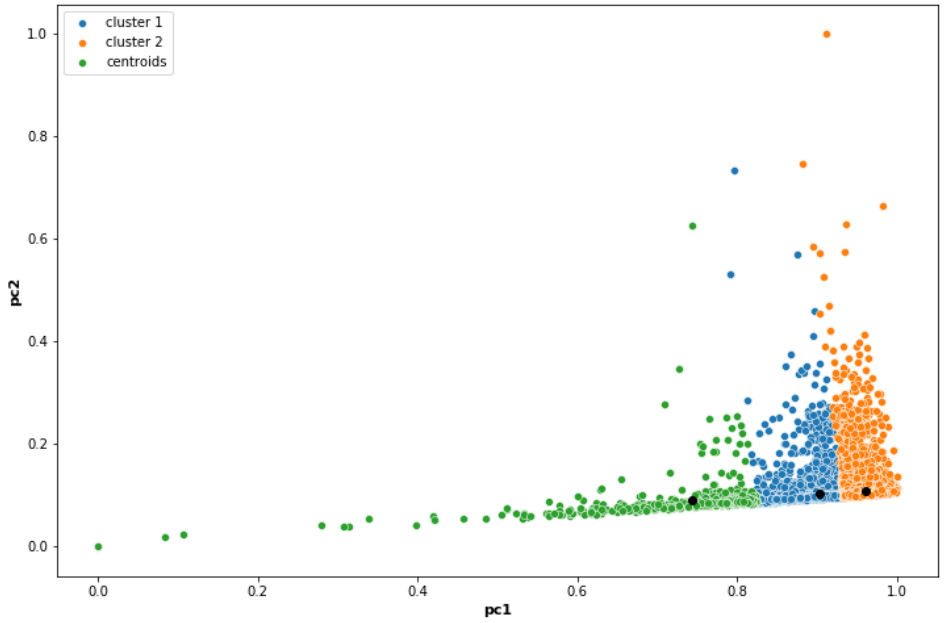

Figura 6: Agrupamento para dados em duas dimensões

ritmos propostos, especialmente a computação (randomizada) da estatística gap, dê resultado satisfatório.

Uma área de continuação deste trabalho seria, portanto, estabelecer o limiar mínimo do número de instâncias a partir do qual se usaria um método automatizado. Com um número de transações menor que o limiar, talvez seja mais conveniente que o corretor analise os dados um imóvel por vez. Tal limiar pode ser estabelecido através do uso prático da metodologia Corretoria Automatizada.

8 Fase 6 – Despacho

Esta fase do CRISP-DM também será sumarizada nesta seção. Muito se aprendeu com o desenvolver deste trabalho. Principalmente quanto à qualidade dos dados, o escritor está cada vez mais convencido de que praticamente nenhum banco de dados encontrado no dia-a-dia pode ser utilizado prontamente em um modelo.

Como o CRISP-DM aponta claramente, existem fases de preparação que não podem ser ignoradas, nem que seja para construir um memorial das experiências e das etapas de desenvolvimento de um projeto de mineração de dados, o que também se mostrou uma parte essencial de um projeto de mineração de dados estável e replicável.

Certas suposições tomadas na proposta inicial claramente foram péssimas estimativas; um exemplo é determinar que o número de grupos deve ser fixado em cinco, sem nem ao menos conhecer os dados. O plano do projeto deve ser feito com mais cuidado. Uma regra útil para evitar perda de tempo com atividades não produtivas é simplesmente não realizar o projeto, se o objetivo, tanto de mineração de dados quanto de negócio, não for claramente estabelecido.

Outro impedimento que deve ser reconhecido é a falta de recursos. O planejador deve verificar previamente a disponibilidade de todos os recursos envolvidos para o projeto. Ele

deve, ainda, ser pessimista, tendo em vista que muitos custos não previstos tendem a aparecer, mesmo nos projetos mais triviais.

Referências

2000. CRISP-DM 1.0. https://www.semanticscholar.org/paper/CRISP-DM-1.0%3A-Step-by-step-data-mining-guide-Chapman-Clinton/54bad20bbc7938991bf34f86dde0babfbd2d5a72. [Acessado em 01 de março de 2021].

2001. Estimating the number of clusters in a data set via the gap statistic. https://statweb.stanford.edu/~gwalther/gap. [Acessado em 01 de março de 2021].

2016. Caso de uso 1. https://rstudio-pubs-static.s3.amazonaws.com/155304_cc51f448116744069664b35e7762999f.html. [Acessado em 01 de março de 2021].

2016. Caso de uso 3. https://rstudio-pubs-static.s3.amazonaws.com/150743_fbe2be64165349798440e35351653b16.html. [Acessado em 01 de março de 2021].

2017. Glossário de termos dos dados. https://info.kingcounty.gov/assessor/esales/Glossary.aspx?type=r.

2020. Caso de uso 2. https://mlr3gallery.mlr-org.com/posts/2020-01-30-house-prices-in-king-county/. [Acessado em 01 de março de 2021].

2020. KCDS. https://github.com/rashida048/Datasets/blob/master/home_data.csv. [Acessado em 1 de março de 2021].

2021. Repositório GitHub. https://github.com/Yowgf/MD-TP2. [Acessado em 1 de março de 2021].