UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
DEPARTAMENTO DE COMPUTAÇÃO

Gabriel Gonçalves Motta

Utilização de agrupamento como método de pré-processamento em problemas de regressão linear.

SÃO CARLOS -SP
2021

Gabriel Gonçalves Motta

Utilização de agrupamento como método de pré-processamento em problemas de regressão linear.

Trabalho de conclusão de curso apresentado ao Departamento de Computação da Universidade Federal de São Carlos, para obtenção do título de bacharel em Engenharia de Computação.

Orientador: Prof. Dr. Murilo Coelho Naldi

São Carlos-SP
2021

DEDICATÓRIA

Dedico este trabalho à minha mãe e namorada por serem meu alicerce durante toda minha graduação e aos meus amigos por compartilharem comigo todos os momentos dessa jornada.

RESUMO

Para que dados oriundos de situações do mundo real possam ser explorados em uma análise, é preciso realizar um pré-processamento de modo a facilitar a aplicações de aprendizado de máquina. Muitas técnicas de pré-processamento estão disponíveis hoje em dia, este trabalho tem como objetivo mostrar os impactos de utilizar o agrupamento como uma delas, mas especificamente em problemas de regressão linear. Foram realizados experimentos com duas bases: a primeira descrevendo dados de imóveis de Ames, uma cidade de Iowa nos Estados Unidos e a segunda contendo informações sobre COVID-19. Observou-se que, em ambos os casos, realizar o agrupamento antes do processo de regressão melhora o desempenho do regressor, com a intensidade dessa melhora dependendo da natureza da base. Apesar da melhora estar presente, indica-se usar o agrupamento em conjunto com outras técnicas de pré-processamento.

Palavras-chave: Regressão Linear. Agrupamento. Pré-processamento. Bases do Mundo Real. Imóveis de Ames. COVID-19.

ABSTRACT

For real-world data to be explored, pre-processing is needed in order to ease machine learning applications. Nowadays, various pre-processing techniques are available, this paper aims to show the impacts in using clustering as one of them, more specifically in linear regression problems. Two experiments were carried out, using two different databases: the first one describing data from a series of properties from Ames, a city from Iowa, in the United States of America and the second one containing information about COVID-19. It was observed that in both cases, clustering before applying the regression model improves regressor performance, based on database nature. Besides the improvement, it is recommended to use clustering alongside other pre-processing techniques.

Keyword: Linear Regression. Clustering. Pre-processing. Real-world Databases. Ames Properties. COVID-19.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE SIGLAS

COVID-19 - Coronavirus Disease 2019

MMQ- Mínimos Múltiplos Quadrados

MAPE - Mean Absolute Percentage Error

MSE - Mean Square Error

MAE - Mean Absolute Error

DBSCAN - Density-based Spatial Clustering of Applications with Noise

HDBSCAN- Hierarchical Density-based Spatial Clustering of Applications with Noise

CD - Core Distance

MRD - Mutua Reachability Distance

MST - Minimum Spanning Tree

UTI - Unidade de Tratamento Intensivo

SUMÁRIO

1 INTRODUÇÃO

Entender o mundo que o cerca sempre foi uma necessidade do ser humano e hoje não é diferente, em um mundo onde cada vez mais dados são gerados, é de suma importância entendê-los e transformá-los em informações relevantes.

Em situações como essas que a análise de dados e os modelos estatísticos mostram seu valor. Imprescindíveis para o bom funcionamento de empresas e governos, esses tipos de estudos são realizados desde o século XVII, quando estatísticos, como John Graunt, começaram a perceber o real valor de coletar informações numéricas sobre populações e suas economias (ROONEY, 2009).

O processo de análise de dados se beneficiou muito com o avanço da tecnologia e o aumento do uso dos computadores. Em 1890, o inventor Herman Hollerith, com a ajuda de uma máquina de cartões perfurados, conseguiu acelerar o tempo de realização do censo dos Estados Unidos de 7 anos para 18 meses (FOOTE, 2018). Inovações como essa, serviram de alicerce para tecnologias de suma importância, como *Big Data* e nuvem, presentes no dia a dia de todos que têm contato com tecnologia.

Por mais avançada que esteja a tecnologia acerca da análise de dados atualmente, para que se possa obter conjuntos de dados mais enxutos, descritivos e que forneçam melhores resultados, o pré-processamento é de suma importância.

Em adição a isso, é necessário saber que pré-processar uma base de dados do mundo real, pode solucionar problemas como atributos que não foram extraídos da melhor maneira possível, ou até mesmo ajudar a escolher atributos que forneçam mais informações para a sua análise. Ao realizar esse processo, é possível ter um maior entendimento da natureza dos dados, podendo inclusive auxiliar na construção de modelos mais robustos (FAMILI et al., 1997).

O agrupamento (do inglês *clustering*), técnica utilizada para agrupar amostra de dados que contém características similares, já é utilizada como alternativa para diminuir os requerimentos computacionais necessários para processar bases de dados muito pesadas, como no caso de Kelly e White (1993) que utilizaram agrupamento para analisar arquivos de imagens de um mapeador temático (satélite utilizado para gerar uma representação visual de informações geográficas de uma

região, como o relevo), tendo um impacto no tamanho dos arquivos de imagem de até 7 vezes.

Trabalhos como esse mostram que o agrupamento, ao criar grupos com objetos mais semelhantes, fornece elementos mais homogêneos para análise em comparação com o conjunto de dados original.

1.1 OBJETIVOS

Este trabalho visa mostrar que, após realizar um agrupamento em um conjunto de dados, por conta dos grupos conterem dados mais homogêneos, regressores treinados individualmente para cada grupo fornecem melhores resultados do que se fosse aplicado somente um regressor em todo o conjunto original.

1.2 MOTIVAÇÃO

Em vários momentos no cotidiano de uma empresa, por exemplo, é necessário realizar análises pontuais onde técnicas simples podem resolver o problema, como, por exemplo, na criação de um modelo de previsão de vendas de um determinado setor. Em situações como essas, um simples regressor pode solucionar tal tarefa de maneira satisfatória, no entanto, caso necessário, seria interessante dispor de outras técnicas, também simples, que ajudem a otimizar os resultados.

Este trabalho tem como motivação mostrar que ferramentas simples, como um regressor linear e um algoritmo de agrupamento combinados, por conta da homogeneidade dos grupos, podem potencializar os resultados de uma análise.

1.3 ESTRUTURA DO TRABALHO

Este trabalho mostra a análise de duas bases de dados, a primeira contendo dados sobre a venda de imóveis em Ames, no estado de Iowa, nos Estados Unidos. A segunda, contendo dados de relatórios sobre COVID-19 nos países durante a pandemia que está ocorrendo no mundo.

Com isso, no Capítulo 2 é apresentado o fundamento teórico das técnicas utilizadas para analisar essas duas bases, já no Capítulo 3, são dados mais detalhes sobre as bases, assim como são mostradas as ferramentas utilizadas para tal análise. No Capítulo 4 é descrito o experimento em si, para que no Capítulo 5 possam ser discutidos os resultados obtidos e listadas as conclusões obtidas ao fim do documento.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 ANÁLISE DE REGRESSÃO

Conceitualmente, “A análise de regressão é uma ferramenta estatística que utiliza a relação entre duas ou mais variáveis quantitativas de modo que uma variável pode ser predita pela outra, ou outras.” (KUTNER et. al. 2004 p.23).

Em situações em que essa relação pode ser representada por uma reta, pode-se chamar a análise de regressão linear. Nesse caso, tal relação é dada pela seguinte equação:

$$\text{(I)} \qquad Y = aX + b$$

onde X é a variável independente, Y é a variável dependente (variável a ser predita), e a e b são os parâmetros da reta. Sendo a a inclinação da reta (com a positivo em relações diretamente proporcionais e negativo em relações inversamente proporcionais) e b o ponto em que a reta cruza o eixo Y (quando X é igual a 0).

No entanto, ao analisar experimentos do mundo real, as observações podem não seguir uma relação perfeitamente linear, com isso, o modelo preditivo é descrito por:

$$\text{(II)} \qquad f(Y) = aX + b + \varepsilon$$

onde $f(Y)$ é o valor predito de Y pelo modelos e ε é a representação do erro obtido entre $f(Y)$ e Y.

Na maioria dos casos, existem mais do que uma variável independente que influencia a variável dependente. Nesses casos, a técnica é chamada de regressão linear multivariada e, segundo HEIJ (2004), é representada matricialmente. Sendo:

$$\text{(III)} \qquad Y = \begin{pmatrix} y_1 \\ \dots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{21} & \dots & x_{k1} \\ \dots & \dots & & \dots \\ 1 & x_{2n} & \dots & x_{kn} \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \dots \\ \beta_k \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_n \end{pmatrix}$$

onde os sub-índices n e k representam, respectivamente, o número de variáveis e o número de observações.

Com isso, a reta de uma regressão linear multivariada pode ser dada por:

$$(IV) \quad y = X\beta + \varepsilon$$

Em todos os cenários, o objetivo é encontrar uma reta em que o parâmetro ε seja minimizado. Para isso, muitas técnicas podem ser utilizadas, uma das mais conhecidas é o Método dos Múltiplos Quadrados (MMQ).

2.2 MÉTODO DOS MÍNIMOS QUADRADOS

Proposto por Adrien-Marie Legendre em 1805 o Método dos Mínimos Quadrados (MMQ) tem como objetivo encontrar uma equação que minimize a soma dos quadrados das diferenças entre um conjunto de observações de um ponto, reta ou curva feita com base neles (ROONEY, 2009).

No caso de uma relação linear entre as variáveis, utilizando a equação (II), o MMQ busca encontrar os valores de a e b onde a soma de ε seja minizada.

Segundo Porter et. al. (2009), os parâmetros a e b podem ser determinados pelas seguintes equações:

$$(III) \quad a = \overline{Y} - b\overline{X}$$

$$(IV) \quad b = \frac{\sum_{i=1}^{n} X_i(Y_i - \overline{Y})}{\sum_{i=1}^{n} X_i(X_i - \overline{X})}$$

onde $\overline{Y}$ e $\overline{X}$ são as médias amostrais de Y e X, respectivamente e X_i e Y_i são as coordenadas das amostras.

Com os parâmetros e a e b calculados, tem-se a reta com a soma dos quadrados dos erros minimizada. Vários métodos podem ser utilizados para medir o quanto esse modelo representa de fato os dados observados, um desses métodos é o Erro Percentual Absoluto Médio, ou MAPE (*Mean Absolute Percentage Error*).

2.3 MAPE

Juntamente com o Erro Quadrático Médio, ou MSE (*Mean Square Error*) e Erro Absoluto Médio, ou MAE (*Mean Absolute Error*), MAPE é uma das ferramentas mais utilizadas para avaliar a qualidade de um algoritmo de regressão.

Em resumo, o MAPE é a média dos módulos do erro, dividido pelo valor real de Y em um determinado modelo de regressão. Formalmente, Myttenaere et. al (2016) demonstram que: dado X, um determinado vetor de variáveis explanatórias (a entrada do algoritmo de regressão), Y a variável alvo, o modelo e g o modelo de regressão, o MAPE de g é dado pela média de:

$$\text{(V)} \qquad \frac{|g(X)-Y|}{|Y|}$$

2.4 HDBSCAN

Criado por Campello et. al (2013), o HDBSCAN, ou *Hierarchical Density-based spatial clustering of applications with noise* (em português, Agrupamento Hierárquico Espacial Baseado em Densidade de Aplicações com Ruído) é um algoritmo de agrupamento que gera uma hierarquia de agrupamento baseado em densidade de onde somente os grupos mais significativos são extraídos. Tal extração de grupos é feita com base em um conceito chamado de estabilidade.

Segundo McInness et. al. (2019) o algoritmo do HDBSCAN é constituído por quatro etapas, sendo elas:

1. Transformação do Espaço
2. Construção da *Minimum Spanning Tree*
3. Construção da hierarquia dos Grupos
4. Extração dos Grupos

2.4.1 Transformação do Espaço

Para agrupar os dados, é preciso encontrar regiões de maior densidade de observações e agrupá-las entre várias outras observações localizadas de forma mais esparsa, sem que dados atípicos (*outliers)* ou erros de leitura possam influenciar de maneira muito ativa.

Para encontrar tais regiões mais densas, o algoritmo calcula para todos os pontos do conjunto, uma medida formalmente definida como *Core Distance* (chamado aqui de CD). Essa medida representa a distância de um ponto x para seu k-ésimo vizinho mais próximo. A CD é representada por:

$$\text{(I)} \qquad Core\ Distance\ (CD):\ core_k(x)$$

Para separar os pontos de baixa densidade, o conceito de *mutual reachability distance* (chamada aqui de MRD) é definido. A MRD entre dois pontos x_1 e x_2 é representada por:

$$\text{(II)}\ Mutual\ Reachability\ Distance\ (MRD):\ max\{core_k(x_1), core_k(x_2), \mathrm{d}(x_1, x_2)\}$$

onde, $core_k(x_1)$ é a CD do ponto x_1, $core_k(x_2)$ a CD do ponto x_2 e $\mathrm{d}(x_1, x_2)$ a distância simples entre os dois pontos.

Tendo todas as MRDs em mãos pode-se representar os dados como um grafo completo, definido formalmente como *Mutual Reachability Graph* (chamado aqui de MRG). Neste grafo, os vértices são os pontos observados e o peso de cada aresta é a MRD entre os respectivos pares de pontos. Com isso, regiões mais densas ficam ainda mais evidentes, e as observações esparsas ainda mais distantes.

2.4.2 Construção da *Minimum Spanning Tree*

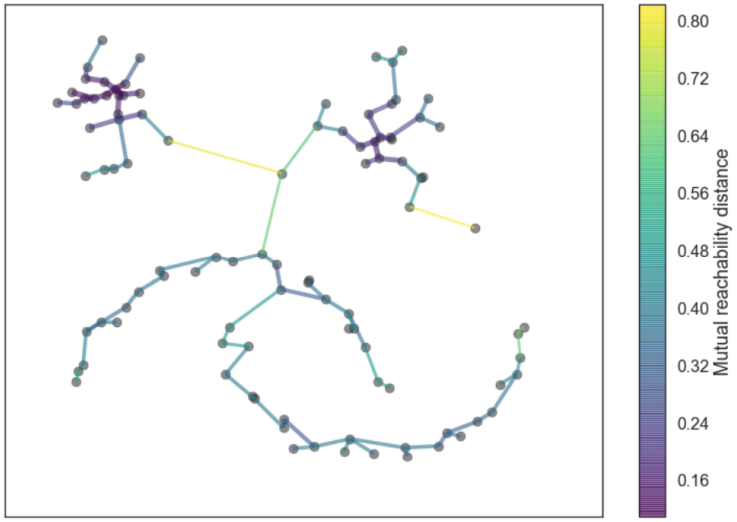
A *Minimum Spanning Tree* (em português, Árvore de Extensão Mínima), ou MST, é um subgrafo acíclico e conexo, que contém todos os vértices do grafo original, e cuja soma do peso das arestas é mínima.

Construído o MRG das observações, o algoritmo do HDBSCAN constrói a MST deste grafo. Deste modo, a árvore resultante vai ter como peso das arestas os

menores valores de MRDs calculados entre as observações.

A Figura 2.1 representa essa MST, onde as arestas são ordenadas por cor, com base no seu peso (ditado pela MRD entre as observações).

Figura 2.1: Minimum Spanning Tree com Peso das Arestas Ordenado.



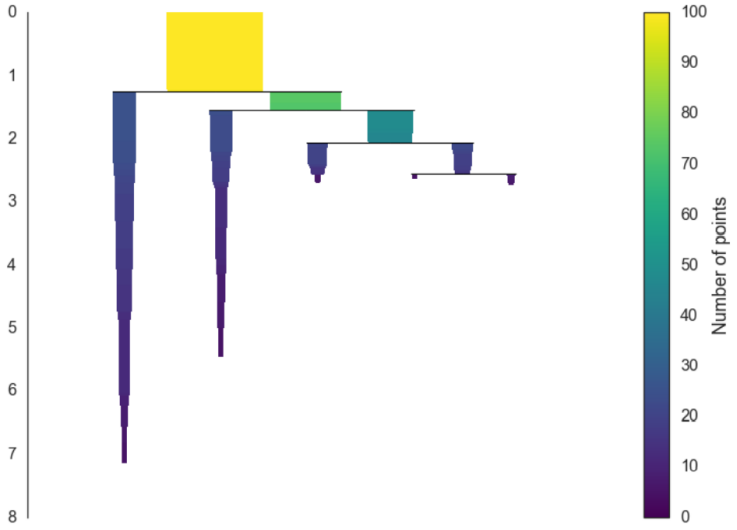
Fonte: McInes et. al. (2019)

2.4.3 Construção da Hierarquia dos Grupos

Dada a MST, o próximo passo é ordenar as arestas com base nos seus pesos e, iterando sobre elas, cortar as arestas da MST, da maior para a menor. Para que o conjunto de elementos que resulta após o corte seja considerado um possível grupo, é preciso que ele tenha um número de elementos mínimo (passado como parâmetro ao algoritmo), caso contrário, isso será entendido apenas como um grupo perdendo pontos (que serão categorizados como observações atípicas).

Desse modo, podemos representar esse processo por um dendrograma, onde apenas divisões de grupos com elementos acima do mínimo são mostradas. A Figura 2.2 mostra um dendrograma com um tamanho mínimo de 5. Na figura, a largura da faixa representa o número de elementos no grupo, com isso, conforme elementos vão saindo deste grupo, a largura dessa linha vai diminuindo.

Figura 2.2: Dendrograma com Tamanho Mínimo de *Cluster* Igual a Cinco.

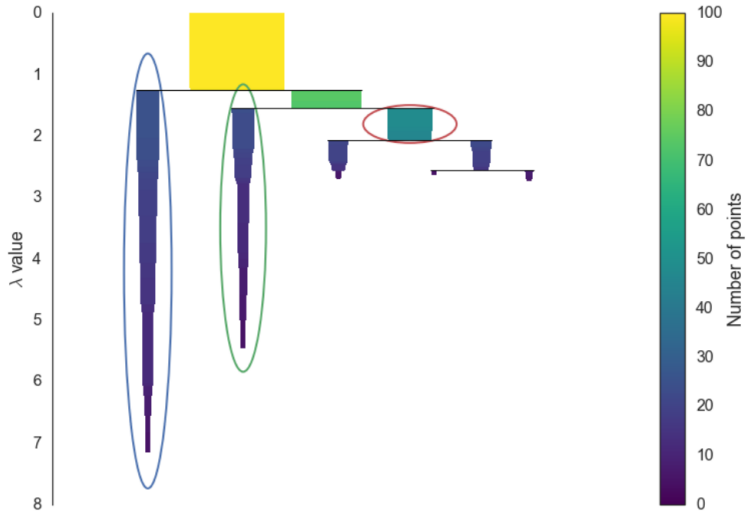


Fonte: McInes et. al. (2019)

2.4.4 Extração dos Grupos

O objetivo é escolher grupos que persistam por mais tempo no processo de eliminação de arestas. Com o dendrograma em mãos, pode-se extrair os grupos ao selecionar as faixas com maior área pintada. Porém, para extrair grupos planos, precisa-se definir que, ao selecionar uma faixa, não se pode selecionar nenhum dos filhos. A Figura 2.3 mostra o processo realizado no dendrograma da Figura 2.3.

Figura 2.3: Extração de Grupos pelo Método da Maior Área.



Fonte: McInes et. al. (2019)

Para fins de formalização, define-se o seguinte algoritmo, onde λ é dado por:

$$\text{(III)} \quad \lambda = \frac{1}{distância}$$

Para um dado grupo pode-se definir $\lambda_{nascimento}$ e λ_{morte} como, respectivamente, o momento em que ele se dividiu e se tornou o próprio grupo e o momento em que o ele se dividiu em outros dois (caso isso tenha ocorrido). Do mesmo modo, pode-se definir λ_p como o momento em que um dado ponto saiu de um grupo (momento esse, que ocorreu entre o $\lambda_{nascimento}$ e o λ_{morte} deste grupo).

Tendo estes valores em mãos, resta definir a estabilidade de um grupo como sendo:

$$\text{(IV)} \quad \sum_{p \in grupo} (\lambda_p - \lambda_{nascimento})$$

Ou seja, a estabilidade de um grupo é o somatório das distâncias que correspondem o quanto um ponto perdura em um dado grupo.

Dados os conceitos descritos anteriormente, o processo de extração de grupos acontece da seguinte maneira:

- Seleciona-se os vértices folha na MST;
- Avança-se na MST na ordem inversa a qual ela foi ordenada;
    - Se a soma das estabilidades dos grupos filhos é maior do que a estabilidade de um grupo: a estabilidade do grupo é definida como sendo a soma das estabilidades dos seus filhos;
    - Se a soma das estabilidades dos grupos filhos é menor do que a estabilidade de um grupo: esse grupo é marcado como um dos selecionados e seus descendentes são ignorados
- Uma vez que se chega no vértice raiz, retorna-se os grupos planos selecionados.

3 MATERIAIS E MÉTODOS

3.1 CORPUS

Neste item, são descritas de maneira mais detalhada as duas bases escolhidas como objeto de estudo deste documento.

3.1.1 **Base de Imóveis de Ames**

Essa base[1], organizada por De Cock (2011), contém dados reais sobre imóveis residenciais do município de Ames, no estado de Iowa, nos Estados Unidos. A base pertencente a um grupo de corretores de Ames, contém informações sobre 1460 propriedades, possuindo 80 variáveis que descrevem características sobre elas.

A escolha dessa base foi pautada por se tratar de uma base que se mostra um ótimo exemplo para a aplicação de um regressor linear. Ademais, é uma base que contém dados reais, um dos principais pré-requisitos desta análise.

As variáveis da base, podem ser categorizadas em quatro tipos, sendo:

- 23 variáveis nominais: como o bairro em que o imóvel se encontra ou a zona de classificação do imóvel (comercial, industrial, residencial, etc);
- 23 variáveis ordinais: como a qualidade da construção da garagem (ordenada de *Excellent* (Excelente) até *Poor* (Ruim);
- 14 variáveis discretas: como a quantidade de quartos na propriedade e
- 20 variáveis contínuas: como a área do imóvel (medida em metros quadrados).

Tais variáveis servem de insumo para predizer o preço que tais imóveis foram vendidos e para esse propósito ela será utilizada. O objetivo do experimento é identificar as variáveis que mais influenciam no preço de venda no imóvel e utilizá-las para predizer o mesmo utilizando um regressor linear.

Por conta da base também conter os preços praticados na realidade, é possível comparar os valores obtidos no regressor com os dados reais e determinar o índice de erro (utilizando MAPE).

---

[1] Disponível em: kaggle.com/c/house-prices-advanced-regression-techniques

3.1.2 **Base COVID-19**

Essa base[2] pertence ao *Our World in Data* (do inglês, Nosso Mundo em Dados), um projeto do *Global Change Data Lab* (do inglês Laboratório de Dados de Mudanças Globais), que tem como objetivo organizar e disponibilizar dados, além de também desenvolver análises sobre problemas mundiais como pobreza, fome, guerras etc, criando conhecimento acerca desses assuntos, de modo que eles se tornem mais acessíveis e compreensíveis.

A escolha dessa base foi feita por se tratar de um assunto muito relevante nos dias de hoje, além dela também disponibilizar informações em um nível de detalhe muito satisfatório. A base centraliza informações de 194 países acerca dos dados referentes à pandemia, onde grande parte de suas informações são atualizadas diariamente. Com isso, ela contém os dados de todos os dias desde o primeiro caso reportado em cada país.

A base completa possui 62 variáveis, todas organizadas em:

- Casos confirmados: variáveis relacionadas ao número de casos de COVID-19 registrados. Atualizadas diariamente;
- Mortes confirmadas: variáveis relacionadas ao número de mortes causadas por COVID-19 registradas. Atualizadas diariamente;
- Vacinação: variáveis relacionadas ao número de pessoas vacinadas. Atualizadas diariamente;
- Testagem: variáveis relacionadas ao número de testes de COVID-19 realizados. Atualizadas semanalmente;
- Hospitalizações e UTI: variáveis relacionadas ao número de hospitalizações e entradas na UTI por conta da COVID-19. Atualizadas semanalmente;
- Taxa de reprodução: estimativa da Taxa de Reprodução[3] da COVID-19. Atualizada diariamente;
- Medidas políticas: variável baseada nas restrições adotadas pelos governos à pandemia. É baseada em 9 indicadores, como fechamento

[2] Disponível em: github.com/owid/covid-19-data/tree/master/public/data
[3] Um fator que estima a velocidade que a doença se espalha. Se for menor do que 1, espera-se que o número de casos diminua, se for maior do que 1, espera-se que o número de casos aumente.

de escolas, ambientes de trabalho e fronteiras. Atualizada diariamente;

- Dados Socioeconômicos: variáveis relacionadas a características socioeconômicas dos países, como população e densidade populacional. Além disso, também contém informações relevantes à estudos sobre COVID-19, como idade média da população e percentual de fumantes.

No entanto, neste estudo, serão utilizadas as informações pertencentes apenas aos grupos de Mortes Confirmadas e Dados Socioeconômicos. O objetivo é, utilizando as variáveis socioeconômicas disponíveis, predizer o número de mortes que um país terá por conta da COVID-19.

3.2 FERRAMENTAS

O trabalho foi todo desenvolvido na plataforma *Google Colaboratory*[4], uma ferramenta que permite executar *notebooks* na linguagem *Python* em nuvem, com máquinas disponibilizadas pela própria plataforma.

No trabalho foram utilizadas 4 bibliotecas de *Python,* sendo elas:

- Pandas[5]: Utilizada para leitura das bases e manipulação das mesmas durante a análise;
- Sklearn[6]: Utilizada para a importação do regressor linear e para normalizar as bases;
- Numpy[7]: Utilizada para manipulação de Dados organizados em séries;
- HDBSCAN[8]: Utilizada para importar o método de agrupamento.

[4] Disponível em: colab.research.google.com
[5] Disponível em: pandas.pydata.org
[6] Disponível em: scikit-learn.org
[7] Disponível em: numpy.org
[8] Disponível em: hdbscan.readthedocs.io

4 DESENVOLVIMENTO

O procedimento, aplicado igualmente para as duas bases citadas anteriormente, consiste de quatro etapas:

1. Pré-processamento dos dados: momento onde são preparados os dados da base total para que o regressor possa ser aplicado;
2. Regressão base total: etapa em que é aplicado um regressor linear na base total, obtendo um primeiro valor de MAPE;
3. Agrupamento e regressão dos grupos: momento onde são agrupados elementos semelhantes, utilizando um algoritmo de agrupamento, e depois aplicado um regressor linear em cada um desses grupos. Após isso, os grupos, com seus valores preditos, são agrupados novamente ao formato da base total e é calculado um segundo valor de MAPE (utilizando os valores preditos durante o agrupamento);
4. Comparação dos resultados: etapa em que os dois MAPEs são comparados e é analisado o percentual de melhora.

4.1 EXPERIMENTO IMÓVEIS DE AMES

O objetivo neste experimento é predizer o preço de venda dos imóveis (variável *SalePrice*), com base nas suas características.

4.1.1 Pré-Processamento dos Dados

Após a importação da base, para que o regressor linear pudesse ser aplicado, foi necessário realizar um tratamento para transformar as variáveis ordinais em variáveis discretas, ou seja, variáveis com valores numéricos. Todas as 23 variáveis ordinais passaram por esse mapeamento, como exemplo, pode-se citar: *ExterQual* (descreve a qualidade dos materiais utilizados na parte externa no imóvel), *KitchenQual* (descreve a qualidade de construção da cozinha), *BsmtQual* (descreve a qualidade de construção do porão). O mapeamento foi feito seguindo a Tabela 4.1.

4.1.2 Regressão Base Total

Para aplicação do regressor linear, é preciso escolher quais variáveis da base serão utilizadas como variáveis independentes da reta. Um dos métodos para determinar isso é calculando o índice de correlação com a variável *SalePrice* (do inglês Preço de Venda), que corresponde ao valor que o imóvel foi vendido e é a variável dependente da regressão.

Tabela 4.1: Mapeamento de variáveis ordinais para discretas.

Valor Original	**Valor Mapeado**	**Significado**
Ex	5	*Excellent* (Excelente)
Gd	4	*Good* (Bom)
TA	3	*Typical/Average* (Típico/ Mediano)
Fa	2	*Fair* (Aceitável)
Po	1	*Poor* (Ruim)
NA	0	*Not Applicable* (Não Aplicável)

Fonte: Elaborada pelo autor.

Foram escolhidas as 10 variáveis com os maiores índices de correlação. As variáveis, assim como seus índices de correlação estão descritas na Tabela 4.2.

Tendo as variáveis em mãos, foi aplicada a função LinearRegression da biblioteca scikit-learn, que contém o regressor linear. O regressor, que levou em consideração apenas as 10 variáveis citadas acima, retornou um MAPE de 13,8%. Ou seja, em média, o regressor fornece resultados com um erro de 13,8% em comparação com o valor real.

4.1.3 Agrupamento e Regressão dos Grupos

Após, foram agrupados os elementos levando em consideração sua similaridade. O objetivo era, ao aplicar o HDBSCAN, fornecer como parâmetro um tamanho mínimo de grupo que fizesse que o algoritmo agrupasse os dados em uma estrutura que fosse interessante para a análise. Tal estrutura seria uma média de

25% de *outliers*, ao mesmo tempo que não houvesse grupos muito pequenos.

Tabela 4.2: Variáveis com maior correlação com *SalePrice*.

Nome da Variável	**Significado**	**Índice de Correlação**
OverallQual	Qualidade geral da casa	0.790982
GrLivArea	Área da sala de estar	0.708624
ExterQual	Qualidade dos materiais externos	0.682639
KitchenQual	Qualidade da cozinha	0.659600
BsmtQual	Qualidade do porão	0.644019
GarageCars	Quantidade de carros que cabem na garagem	0.640409
GarageArea	Área da garagem	0.623431
TotalBsmtSF	Área do porão	0.613581
1stFlrSF	Área do térreo	0.605852
FullBath	Quantidade de Banheiros	0.560664

Fonte: Elaborada pelo autor.

Após analisar os valores dos parâmetros do HDBSCAN, verificou-se que ao determinar como 13 o tamanho mínimo dos grupos, o algoritmo agrupava os dados na estrutura desejada. Os grupos obtidos são mostrados na Tabela 4.3.

Tabela 4.3: Grupos obtidos na Base de Imóveis de Ames.

Número do Grupo	**Quantidade de Elementos**
-1 (*Outliers*)	354
0	13
1	414
2	679

Fonte: Elaborada pelo autor.

Vale ressaltar que apesar dos *outliers* não serem de fato um grupo, eles foram tratados como tal no processo de treinamento do regressor individual. Ou seja, também foi treinado um regressor nos *outliers* de modo a calcular um valor de MAPE para eles.

Tendo feito o agrupamento, foi treinado um regressor linear para cada um dos 4 grupos (incluindo os outliers) e com isso foi predito um valor para o preço de venda. Após isso, os grupos foram desfeitos, todas as observações retornaram ao formato da base original e, utilizando os preços preditos pelos regressores do agrupamento, foi calculado na base total o segundo valor de MAPE, resultando em 13,2%.

4.1.4 Comparação dos Resultados

Com isso, foi verificado que após aplicar o agrupamento, a queda nos valores de MAPE calculados foi de:

$$\frac{MAPE\ Agrupado}{MAPE\ Original} - 1$$

$$= \frac{13{,}2\%}{13{,}8\%} - 1$$

$$= 0{,}956 - 1$$

$$= -4{,}6\%$$

Ou seja, os regressores dos grupos forneceram um MAPE 4,6% menor do que o original.

4.2 EXPERIMENTO COVID-19

O objetivo neste experimento é predizer o número máximo de mortes por COVID-19 em um dia que um país terá, com base em seus dados socioeconômicos.

4.2.1 Pré-Processamento dos Dados

Como dito anteriormente, apesar da base possuir informações diversas como quantidade de testes e população vacinada, nesse experimento foram usadas apenas as variáveis organizadas como Mortes Confirmadas e Dados Socioeconômicos

Todas variáveis organizadas como Dados Socioeconômicos, que serão utilizadas no experimento, estão descritas na Tabela 4.4.

Tabela 4.4: Variáveis de Dados Socioeconômicos da Base de COVID-19.

Nome da Variável	**Significado**
population	Número de habitantes do país
population_density	Densidade populacional
median_age	Idade média da população
aged_65_older	Percentual da população acima de 65 anos
aged_70_older	Percentual da população acima de 70 anos
gdp_per_capita	PIB per capita
cardiovasc_death_rate	Taxa de morte por doenças cardiovasculares
diabetes_prevalence	Porcentagem da população (entre 20 e 79 anos) com diabetes
hospital_beds_per_thousand	Leitos hospitalares por milhar de habitantes
life_expectancy	Expectativa de vida
human_development_index	Índice de Desenvolvimento Humano

Fonte: Elaborada pelo autor.

Para representar o número máximo de mortes por COVID-19 de um país, a variável escolhida de Mortes Confirmadas, foi a *new_deaths_smoothed_per_million* (do inglês, novas mortes suavizadas por milhão), que corresponde à média móvel dos últimos 7 dias de novas mortes por milhão de habitantes. A escolha dessa

variável foi feita para mitigar qualquer variação que pudesse ocorrer por conta *outliers* ou tamanho de população.

Como quer-se determinar o máximo dessa variável, o primeiro passo foi calcular esse indicador para cada país. Após esse cálculo, foram excluídos da análise países que, ou não possuíam mortes detectadas, ou não haviam reportado as mesmas aos órgãos responsáveis. Com isso, restaram 194 países para a análise.

Após isso, verificou-se quais das variáveis socioeconômicas possuíam observações suficientes para a análise. A partir disso, percebeu-se que a variável *hospital_beds_per_thousand* possuía mais de 25% de valores faltantes e, por conta disso, ela foi tirada da análise.

Com todas as variáveis em mãos, o próximo passo foi normalizar as variáveis. A normalização utilizada foi a L2, onde em cada linha, a soma dos quadrados dos valores é no máximo 1. Normalizar foi necessário pois as variáveis possíam valores em ordens de grandeza muito distantes e isso poderia influenciar no cálculo da correlação e do regressor linear.

4.2.2 Regressão Base Total

Tendo as variáveis selecionadas já normalizadas, foi treinado um regressor da função *LinearRegression* do scikit-learn no conjunto com as 10 variáveis restantes. O regressor retornou uma reta com um valor de MAPE de 34265%, o que é um péssimo valor, pois significa que, em média, o valor predito tem um erro de 34265% em relação ao valor real.

4.2.3 Agrupamento e Regressão dos Grupos

Aqui, o objetivo é o mesmo que no experimento com a base de imóveis de Ames, fornecer como argumento um tamanho mínimo de grupo que faça com que o HDBSCAN realize um agrupamento na estrutura desejada.

Com isso, ao analisar os valores dos parâmetros da função do HDBSCAN, verificou-se que 8 era o tamanho mínimo que retornava os grupos nas condições

desejadas. Os grupos obtidos são mostrados na Tabela 4.5.

Tabela 4.5: Grupos obtidos na Base de COVID-19.

Número do Grupo	**Quantidade de Elementos**
-1 (*Outliers*)	74
0	10
1	13
2	9
3	88

Fonte: Elaborada pelo autor.

Após o agrupamento foi aplicado o regressor linear para cada um dos 5 grupos (4 grupos formados em conjunto com os *outliers*) e foi predito um valor para a variável dependente desejada. Em seguida, os grupos foram desfeitos, as observações retornaram ao formato da base original e, utilizando os valores do regressor linear pós agrupamento, foi calculado o segundo valor de MAPE, que resultou em 1561%.

Apesar da melhora em comparação com o MAPE original, 1561% ainda é um valor de MAPE muito precário, ele significa que os valores preditos pelos regressores agrupados possuem um erro de 1561% em relação ao valor real.

4.2.4 Comparação dos Resultados

Com os valores obtidos, verificou-se que a queda nos valores de MAPE calculados foi de:

$$\frac{MAPE\ Agrupado}{MAPE\ Original} - 1$$

$$= \frac{1561\%}{34265\%} - 1$$

$$= 0,0455 - 1$$

$$= -95,5\%$$

O que significa que os regressores dos grupos forneceram um MAPE 95,5% menor do que o original.

4.3 DISCUSSÃO DOS RESULTADOS

Pôde-se observar que, ao aplicar essa técnica em bases que já se comportam bem com regressores lineares (como no caso da Base de imóveis de Ames), a melhora obtida de 4,6% não é tão expressiva, porém está presente.

No entanto, ao ser aplicada em uma base que, inicialmente, não apresentou bons valores de MAPE, a melhora se mostra muito mais expressiva (como no caso da Base de COVID-19). Apesar disso, somente essa técnica não fez com que os valores de MAPE obtidos fossem satisfatórios. O valor de 1561% de MAPE ainda é muito ruim um valor tão alto pode ser causado por uma seleção ruim de atributos (os Dados Socioeconômicos não indicam bem o número de mortes que um país por ter por COVID-19), ou até mesmo por uma característica da base de não poder ser descrita de maneira fiel por um modelo linear.

Para obter valores de MAPE satisfatórios em bases como a de COVID-19, seria preciso unir ao agrupamento, outras técnicas de pré-processamento, ou até mesmo avaliar a aplicação de uma outra espécie de modelo preditivo, como por exemplo uma rede neural.

CONCLUSÃO

Com os resultados obtidos nos dois experimentos, pode-se concluir que o agrupamento é uma estratégia de pré-processamento válida em problemas de regressão linear. O agrupamento, ao fornecer grupos mais homogêneos, faz com que modelos preditivos tenham um desempenho melhor. No caso de regressores lineares, os modelos aplicados nos grupos individualmente, fornecem retas com valores de MAPE mais baixos em comparação com a base original.

No entanto, verificou-se também que seu desempenho é bastante dependente da natureza da base. Em bases cuja distribuição já se mostra propícia à aplicação de um regressor linear, a melhora ocasionada pela aplicação de um agrupamento é presente, porém não tão evidenciada.

Contudo, em bases que inicialmente não apresentam bons resultados após a aplicação de um regressor linear, aplicar o agrupamento pode aprimorar consideravelmente os resultados obtidos pelo regressor. Ainda assim, é indicado que essa estratégia seja aplicada em conjunto com outras técnicas de pré-processamento (como redução de dimensionalidade, ou uma melhor seleção de atributos), para que o resultado seja o melhor possível.

REFERÊNCIAS

ROONEY, A. **The Story of Mathematics**. Inglaterra. Arcturus Publishing Limited, 2008.

FOOTE, K. **A Brief History of Analytics**. Disponível em: https://www.dataversity.net/brief-history-analytics/. Acesso em: 04/11/2021.

FAMILI, A. et. al. Data Preprocessing and Intelligent Data Analysis. **Intelligent Data Analysis.** Holanda. v.1, n.1. 1997.

KELLY, P. et. al. Preprocessing remotely-sensed data for efficient analysis and classification. In: AUTOMATIC OBJECT RECOGNITION. 3. 1993. Washington, EUA.

KUTNER, M. et. al. **Applied Linear Regression Models.** Estados Unidos. Richard D. Irwin INC, 1983.

HEIJ, C. **Econometric Methods with Applications in Business and Economics.** Estados Unidos. Oxford University Press, 2004.

MYTTENAERE, A. et. al. Mean Absolute Percentage Error for regression models. **Selected papers from the 23rd European Symposium on Artificial Neural Networks.** Bélgica. 2015.

MOULAVI, D. et. al. **Density-Based Clustering Based on Hierarchical Density Estimates.** Dept. of Computing Science, University of Alberta, Edmonton, Canada, 2013.

MCINNES, L. et. al. **hdbscan Documentation: Release 0.8.1.** Disponível em: https://hdbscan.readthedocs.io/en/latest/. Acesso em: 04/11/2021.

DECOCK, D. Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project. **Journal of Statistics Education.** v.19, n.3, 2011.