Análise de padrões frequentes de mobilidade urbana

Pedro Gonçalves Pereira Lopes
Victor Antonio Bonilha dos Santos

[1]Centro Federal de Educação Tecnológica Celso Suckow da Fonseca - CEFET/RJ

pedro.lopes.1@aluno.cefet-rj.br, victor.bonilha@aluno.cefet-rj.br

Resumo. *O artigo realiza uma análise de padrões frequentes de mobilidade urbana de ônibus no Rio de Janeiro, utilizando dados coletados dos sistemas da Prefeitura do Rio de Janeiro, dentro do período de 29 de novembro a 31 de dezembro de 2021. O estudo emprega técnicas de pré-processamento de dados, remoção de outliers e discretização de variáveis contínuas, visando a otimização da qualidade e relevância das informações utilizadas.*

A abordagem metodológica adotada inclui a aplicação do algoritmo Apriori nos datasets resultantes, com o intuito de extrair padrões de associação significativos. As regras de associação obtidas são posteriormente analisadas e apresentadas de maneira visual através de gráficos elucidativos.

A pesquisa propicia insights valiosos sobre os comportamentos de mobilidade urbana, oferecendo uma compreensão mais aprofundada dos padrões de deslocamento e suas inter-relações. Os resultados apresentados não apenas contribuem para o entendimento da dinâmica do transporte público na cidade, mas também fornecem subsídios para melhorias e otimizações no planejamento urbano e na gestão do sistema de transporte. Este estudo representa, assim, um passo significativo na aplicação de técnicas de mineração de dados para análise de mobilidade urbana, destacando-se pela combinação de métodos avançados e um conjunto de dados representativo.

1. Introdução

No cenário urbano dinâmico do Rio de Janeiro, a gestão eficiente da mobilidade urbana torna-se um desafio crucial para garantir o bem-estar dos cidadãos e promover o desenvolvimento sustentável da cidade. Observa-se cada vez mais um aumento nos estudos relacionados a mobilidade urbana. Uma explicação para esse aumento é o aumento da população mundial e consequentemente o aumento populacional em áreas urbanas, culminando em um maior número de veículos nas cidades, provocando acidentes, congestionamento e poluição. [Chen et al., 2015; Ferreira et al., 2013]

Com a crescente disponibilidade de dados provenientes dos sistemas municipais, a mineração de padrões frequentes emerge como uma ferramenta essencial para extrair informações valiosas sobre os padrões de deslocamento de ônibus na região. Este artigo propõe uma análise abrangente desses padrões de mobilidade urbana, utilizando dados coletados entre 29 de novembro a 31 de dezembro de 2021, provenientes dos sistemas da Prefeitura do Rio de Janeiro.

Com esses dados em mãos, o objetivo definido foi de buscar padrões e tendências frequentes nesses dados a fim de encontrar possíveis melhorias para a mobilidade urbana no Rio de Janeiro. A princípio, foram traçados dois pontos que não poderiam passar

desapercebidos nessa busca: a análise precisa de alterações de velocidade dos ônibus, e os padrões associados ao aumento de poluição.

Ao adentrar o complexo cenário da mobilidade urbana e suas interações com fatores como velocidade, emissão de gases e características temporais, este trabalho busca não apenas compreender os padrões de deslocamento de ônibus, mas também fornecer subsídios valiosos para melhorias e otimizações no planejamento urbano e na gestão do sistema de transporte público. Assim, este estudo representa um passo significativo na aplicação de técnicas avançadas de mineração de dados para análise de mobilidade urbana, promovendo uma compreensão mais profunda dos desafios e oportunidades inerentes a esse contexto dinâmico.

Em adição a esta introdução, o trabalho está organizado em mais 5 seções. Na seção 2, está a fundamentação teórica, onde serão apresentados conceitos gerais para o entendimento da metodologia aplicada. Logo depois, a seção 3 apresenta trabalhos relacionados ao tema, que serviram como inspiração para a concepção desse artigo. A seção 4 descreve detalhadamente a metodologia aplicada neste trabalho. A seção 5 apresenta os resultados esperados. Por fim, na seção 6 é apresentado as conclusões observadas após o término do trabalho.

2. Fundamentação Teórica

A mineração de padrões frequentes é um dos métodos essenciais para identificar informações valiosas sobre padrões que ocorrem com frequência em um grande volume de dados Weiss and Indurkhya [1998]. Uma regra de associação típica é representada por "A → B", onde "X"é o antecedente e "Y"é o consequente. Tais regras indicam uma tendência, ou seja, quando encontra-se o antecedente "A", qual a frequência que também encontra-se "B".

Uma regra de associação é caracterizada através das métricas de suporte, confiança e lift. O suporte de uma regra é a probabilidade de ocorrência do antecedente. A confiança é a probabilidade condicional de que o consequente ocorra, dado que o antecedente ocorreu. Já o lift mede a força da associação, indicando se a ocorrência do antecedente afeta positiva ou negativamente a ocorrência do consequente em comparação com sua ocorrência esperada ao acaso [Teoh and Rong, 2022]. Um valor de lift igual a 1 indica independência, enquanto um valor maior que 1 indica correlação positiva, e um valor menor indica correlação negativa.

Dentre os algoritmos de mineiração de padrões frequentes, o principal é o Apriori. Ele encontra associações frequentes entre itens em transações, assumindo que se um conjunto de itens é frequente, então todos os seus subconjuntos também são frequentes [Agrawal and Srikant, 2000].

3. Trabalhos relacionados

O presente trabalho foi fundamentado em uma base teórica composta pelas contribuições de dois artigos da área de mineração de dados. O artigo Frequent Pattern Mining: Current Status and Future Directions [Han et al., 2007] foi a primeira e principal fonte de extração de conhecimento sobre os conceitos de mineiração de dados e do algoritmo Apriori utilizado. Já o artigo Detecção de Anomalias Frequentes no Transporte Rodoviário Urbano

[Cruz et al., 2018] foi porta de entrada aos assuntos de mobilidade urbana e como aplicaros conceitos de extrações de padrões frequentes neste abaicenário.

4. Metodologia

Este trabalho tem como objetivo extrair padrões frequentes de mobilidade urbana, com foco nos fatores que levam às alterações de velocidade e emissão de gases poluentes. A seguir, são detalhadas as etapas do processo.

4.1. Pré-processamento

O pré-processamento consiste na discretização dos dados para que a extração de padrões frequentes funcione corretamente. Este processo resume-se à seleção dos dados, tratamento de outliers, normalização, e criação de intervalos.

Primeiramente, é feita a seleção de variáveis desejadas para estudo. Definidas por Carvalho, Diego, estas são:

- `DATE` (data da observação).
- `VELOCITY` (velocidade instantânea medida pelo GPS).
- `DISTANCE` (distância 3D em km percorrida entre a observação atual e anterior)
- `HEIGHT` (distância em km entre duas amostras).
- `VSPMode` (modo específico do veículo).
- `CO_2` (quantidade de dióxido de carbono em g/s).
- `CO` (quantidade de monóxido de carbono em g/s).
- `NO_x` (quantidade de óxidos de nitrogênio em g/s).
- `HC` (quantidade de hidrocarbonetos em g/s).

A partir da variável `DATE`, é possível extrair o dia da semana com a função `weekdays` [R Core Team]. Com a função `hour` do pacote lubridate [Grolemund and Wickham, 2011], foram colhidas as faixas de hora que cada amostra foi coletada. Com isso, foram criadas as novas variáveis `WEEKDAY` e `HOUR`. Por fim, fez-se a exclusão de todas as linhas contendo valores nulos.

O próximo passo foi a remoção de outliers. Para atingir o objetivo de discretização dos dados por meio de um agrupamento, a presença de outliers deve ser tratada a fim de gerar um agrupamento mais confiável [Hautamäki et al., 2005]. A detecção foi feita manualmente via cálculo de quartis, intervalo interquartil, e limites inferior e superior. Abaixo, pode-se verificar os boxplots das variáveis `VELOCITY` e `CO_2`.

Observa-se na Figura 1 a grande presença de outliers. Estes foram removidos, assim como para `HEIGHT` e `DISTANCE`. Para as emissões de gases, porém, observa-se na Figura 2 como os dados estão extremamente acumulados na média. Por este motivo, o tratamento de outliers não foi realizado nas variáveis de gases poluentes, pois foi concluído que haveria demasiada perda de dados importantes.

Em seguida, realizou-se a normalização min-max das variáveis `VELOCITY`, `HEIGHT` e `DISTANCE`. Esse método transforma os valores para uma escala, neste caso, de 0 a 1 [Saranya and Manikandan, 2013] da seguinte forma:

$$X_{\text{norm}} = \frac{X - \min(X)}{\max(X) - \min(X)} \tag{1}$$

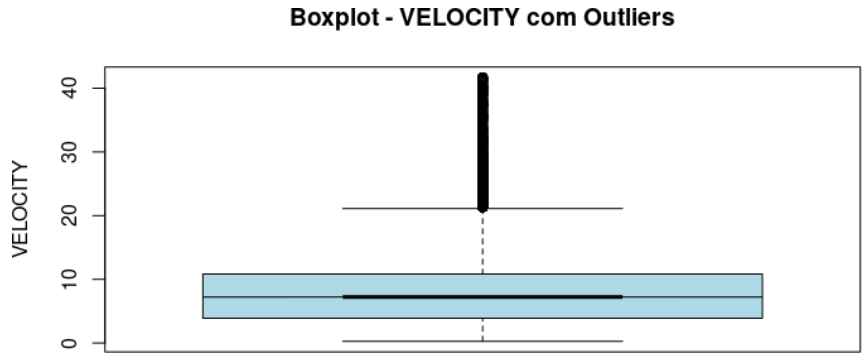

Figura 1. Boxplot de velocidade

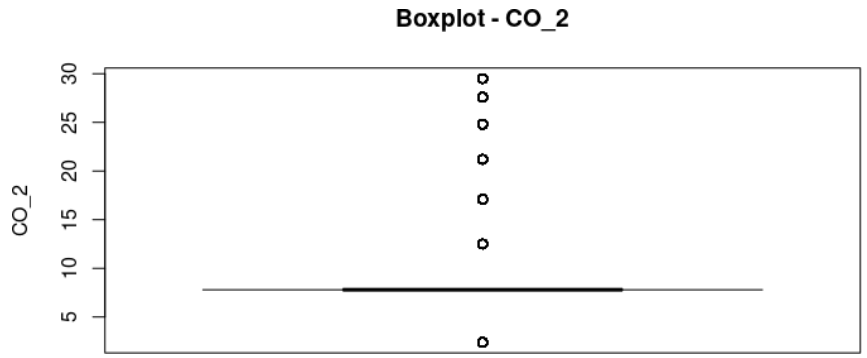

Figura 2. Boxplot de quantidades de dióxido de carbono

onde X_{norm} é o valor normalizado, X é o valor original, min(X) é o valor mínimo da variável, e max(X) é o valor máximo da variável. Os valores resultantes foram posteriormente arredondados em uma casa decimal.

Para as variáveis de emissão, entretanto, foi escolhido a normalização por z-score. Ela foi escolhida pois, com a grande concentração de dados em torno de um valor, a normalização por z-score normaliza os dados com base na média e no desvio padrão da variável [Shalabi et al., 2006] da seguinte forma:

$$Z = \frac{X - \mu}{\sigma} \tag{2}$$

onde Z é o valor normalizado, X é o valor original, μ é a média da variável e σ é o desvio padrão da variável. Estes dados também são arredondados para uma casa decimal.

Agora que os dados estão normalizados, foi preciso categorizá-los em intervalos. O modo de divisão foi escolhido baseado nas frequências das variáveis.

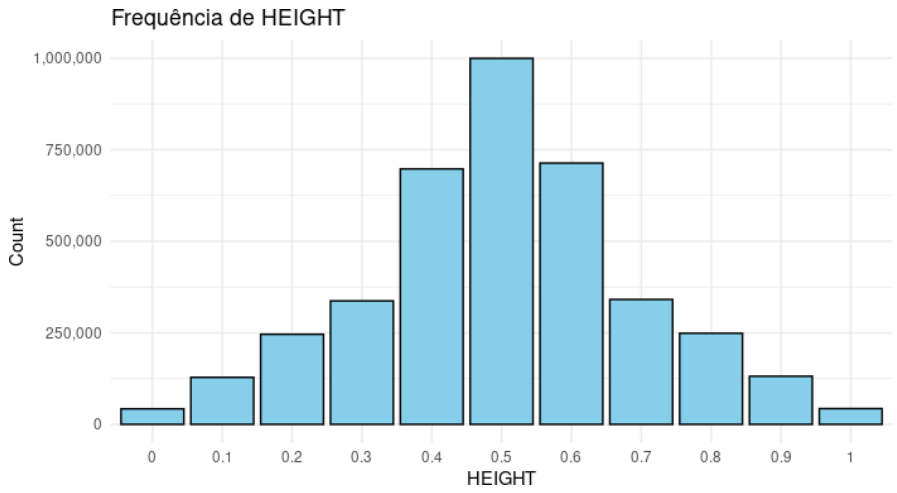

Figura 3. Frequência de altura

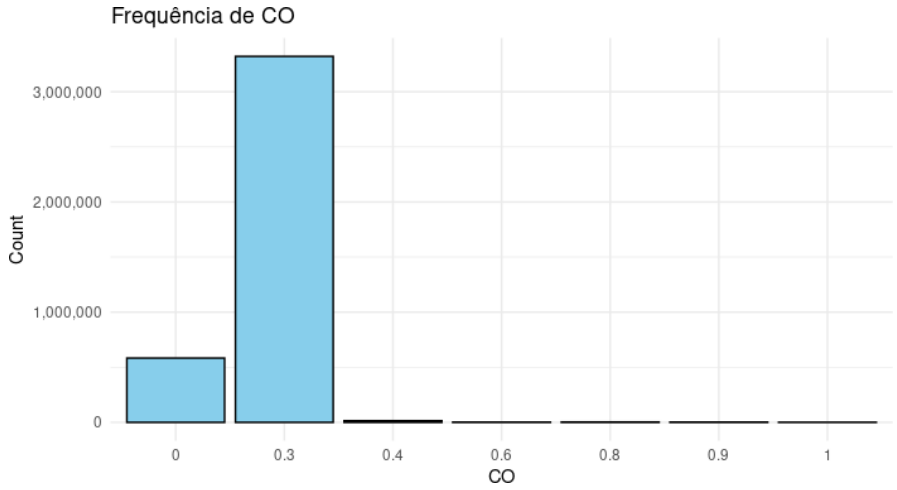

Figura 4. Frequência de quantidades de monóxido de carbono

Para as variáveis `VELOCITY`, `HEIGHT` e `DISTANCE`, que obtiveram uma distribuição de frequências similar à normal, foram feitos os intervalos com binning de largura igual, um método de binning que divide cada intervalo em larguras iguais [Jishan et al., 2015]. Ele é feito de forma:

$$\text{Intervalo} = [\min(X),\ \min(X) + \Delta),\ \ldots,\ \max(X) - \Delta,\ \max(X)] \tag{3}$$

onde $\min(X)$ e $\max(X)$ são os valores mínimo e máximo de X, respectivamente, e $\Delta = \frac{\max(X) - \min(X)}{n}$ é a largura de cada bin. O número de intervalos escolhido foi 3, criando as categorias "Baixa","Média", e "Alta".

Para as variáveis de emissão, foi feita a discretização por z-score. Este método foi escolhido devido à grande concentração de dados em torno da média. Ele dividiu os valores por meio do desvio padrão, agrupando valores abaixo de um desvio padrão na

categoria "Baixa", valores acima de um desvio padrão na categoria "Alta", e os valores do meio na categoria "Média".

Nota-se que a variável `VSPMode` não sofreu nenhuma alteração. Isso ocorreu pois ela já se encontrava discretizada, uma vez que o modo específico do veículo já é dividido em intervalos [Khan and Frey, 2016].

Tabela 1. Definições de VSPMode

VSPMode	Definição (kW/ton)
1	$VSP < -2$
2	$-2 \leq VSP < 0$
3	$0 \leq VSP < 1$
4	$4 \leq VSP < 7$
5	$7 \leq VSP < 10$
6	$10 \leq VSP < 13$
8	$13 \leq VSP < 16$

4.2. Algoritmo Apriori

Após o pré-processamento, executou-se o algoritmo Apriori em conjuntos de duas variáveis selecionadas. Isto ocorreu devido à quantidade de regras de associação geradas ao utilizar o dataset inteiro, sendo essa uma forma de filtrar os dados para obter apenas as regras de associação mais esperadas. Os parâmetros de suporte e confiança foram extremamente baixos, a fim de exibir absolutamente todas as regras de associação do conjunto.

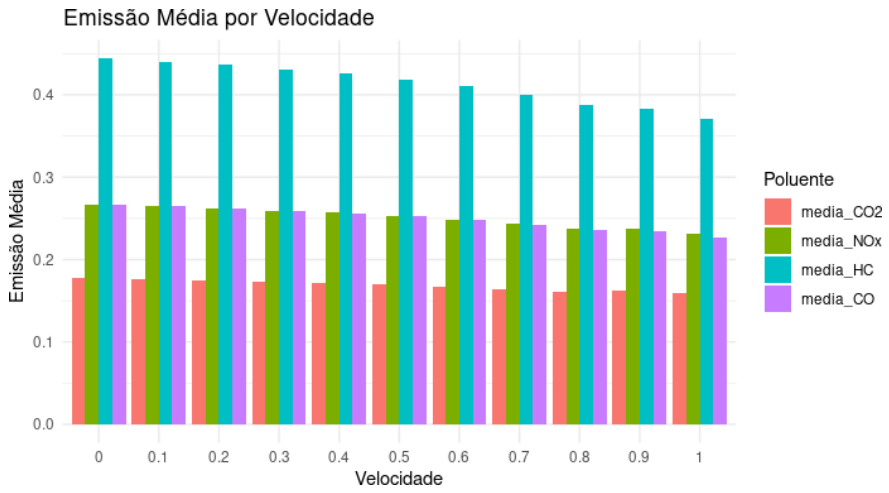

Figura 5. Médias de quantidades de emissões por faixas de velocidade

A Figura 5 apresenta as médias de quantidades de emissões por faixas de velocidade. Com isso, é possível visualizar uma leve tendência de queda de média de quantidade de emissões em medida do aumento de velocidade. Outras variáveis que também apresentaram essas tendências correlacionais por meio de gráficos foram selecionadas juntamente para a execução do algoritmo.

5. Avaliação experimental

Essa seção apresenta os resultados das regras de associação por meio de tabelas. A força de lift foi o fator considerado na escolha das regras a exibir.

Tabela 2. Correlações positivas com consequente CO2

lhs	rhs	lift
VELOCITY=Alta	CO_2=Alta	3.2362711
VELOCITY=Alta	CO_2=Baixa	1.4788650
VELOCITY=Média	CO_2=Alta	1.1647137
DISTANCE=Alta	CO_2=Alta	2.4171441
DISTANCE=Alta	CO_2=Baixa	1.6678641
DISTANCE=Média	CO_2=Baixa	1.3543179
DISTANCE=Baixa	CO_2=Alta	1.2414061
HEIGHT=Alta	CO_2=Alta	2.8287591
HEIGHT=Baixa	CO_2=Baixa	2.5287773
HEIGHT=Alta	CO_2=Baixa	1.1966191
WEEKDAY=Tuesday	CO_2=Alta	1.6491566
WEEKDAY=Sunday	CO_2=Baixa	1.1056733
WEEKDAY=Wednesday	CO_2=Baixa	1.1004634
VSPMode=7	CO_2=Alta	161.619030
VSPMode=6	CO_2=Alta	161.619030
VSPMode=5	CO_2=Alta	161.619030
VSPMode=8	CO_2=Alta	161.619030
VSPMode=4	CO_2=Alta	161.619030
VSPMode=3	CO_2=Alta	161.619030
VSPMode=1	CO_2=Baixa	6.735246
VSPMode=2	CO_2=Média	1.182956

Observando a Tabela 2, pode-se chegar a algumas conclusões. Por exemplo, os modos específicos de veículo mais altos têm grande ligação com altas emissões de CO2, os dias da semana mais e menos associados à emissão de CO2 foram respectivamente terça-feira e domingo, e que o aumento da elevação se correlaciona com o aumento da emissão de CO2.

Tabela 3. Correlações positivas com consequentes VELOCITY

lhs	rhs	lift
HEIGHT=Baixa	VELOCITY=Alta	1.1430910
HOUR=0	VELOCITY=Alta	2.0385133
HOUR=1	VELOCITY=Alta	2.0186419
HOUR=21	VELOCITY=Alta	1.9823013
HOUR=22	VELOCITY=Alta	1.9489127
HOUR=2	VELOCITY=Alta	1.7250357
HOUR=23	VELOCITY=Alta	1.5296647
HOUR=19	VELOCITY=Alta	1.4131498
HOUR=3	VELOCITY=Alta	1.3019089
HOUR=18	VELOCITY=Alta	1.2367475
HOUR=21	VELOCITY=Média	1.1894354
HOUR=2	VELOCITY=Média	1.1698873
HOUR=19	VELOCITY=Média	1.1423706
HOUR=1	VELOCITY=Média	1.1403393
HOUR=3	VELOCITY=Média	1.1313955
HOUR=18	VELOCITY=Média	1.1313754
HOUR=14	VELOCITY=Baixa	1.1227492
HOUR=15	VELOCITY=Baixa	1.1079991
WEEKDAY=Sunday	VELOCITY=Alta	1.3332247
WEEKDAY=Saturday	VELOCITY=Alta	1.2064658

Da Tabela 3, pode-se inferir que, por exemplo, quedas de elevação, a faixa horária de 18 a 3 horas, e finais de semana estão correlacionados positivamente a velocidades altas, enquanto os horários da tarde 14 e 15 horas estão correlacionados positivamente a velocidades mais baixas.

6. Conclusão

Em síntese, este estudo explorou padrões de mobilidade urbana no Rio de Janeiro usando mineração de dados. Foram identificadas associações entre altas velocidades de ônibus e maiores emissões de CO2, destacando a influência dos modos específicos de veículos. Além disso, observa-se que diminuições na elevação e certos períodos do dia estão fortemente correlacionados com velocidades mais baixas ou mais altas. Este trabalho avança na aplicação de técnicas de mineração de dados para insights na mobilidade urbana. Contribuições incluem subsídios para o planejamento urbano e otimização do transporte público. Limitações incluem a necessidade de dados mais precisos. O estudo sugere oportunidades para análises temporais mais detalhadas e modelos avançados.

O estudo não só alcança seus objetivos ao analisar padrões, mas também oferece compreensão aprofundada da mobilidade urbana carioca. Essas descobertas informam decisões futuras para promover um desenvolvimento sustentável e melhor qualidade de vida para os cidadãos do Rio de Janeiro.

Agradecimentos

Um agradecimento especial ao professor Eduardo Ogasawara pelo conhecimento ensinado e ao Portal de periódicos CAPES pelo acervo cedido.

Referências

Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. *Proc. 20th Int. Conf. Very Large Data Bases VLDB*, 1215, 08 2000.

Carvalho, Diego. Como explorar os dados do moblab, 2022. URL https://www.diegocarvalho.org/en/post/busmobility/#fnref:1. [Online; accessed 8-December-2023].

Wei Chen, Fangzhou Guo, and Fei-Yue Wang. A survey of traffic data visualization. *IEEE transactions on intelligent transportation systems*, 16(6):2970–2984, 2015.

Ana Beatriz Cruz, João Ferreira, Diego Carvalho, Eduardo Mendes, Esther Pacitti, Rafaelli Coutinho, Fabio Porto, and Eduardo Ogasawara. Detecção de anomalias frequentes no transporte rodoviário urbano. In *Anais do XXXIII Simpósio Brasileiro de Banco de Dados*, pages 271–276, Porto Alegre, RS, Brasil, 2018. SBC. doi: 10.5753/sbbd.2018.22242. URL https://sol.sbc.org.br/index.php/sbbd/article/view/22242.

Nivan Ferreira, Jorge Poco, Huy T. Vo, Juliana Freire, and Claudio T. Silva. Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips. *IEEE transactions on visualization and computer graphics*, 19(12):2149–2158, 2013. ISSN 1077-2626.

Garrett Grolemund and Hadley Wickham. Dates and times made easy with lubridate. *Journal of Statistical Software*, 40(3):1–25, 2011. URL https://www.jstatsoft.org/v40/i03/.

Jiawei Han, Hong Cheng, Dong Xin, and Xifeng Yan. Frequent pattern mining: current status and future directions. *Data Mining and Knowledge Discovery*, 15(1):55–86, Aug 2007. ISSN 1573-756X. doi: 10.1007/s10618-006-0059-1. URL https://doi.org/10.1007/s10618-006-0059-1.

Ville Hautamäki, Svetlana Drapkina, Ismo Kärkkäinen, and Tomi Kinnunen. Improving k-means by outlier removal. volume 3540, pages 978–987, 06 2005. ISBN 978-3-540-26320-3. doi: 10.1007/11499145_99.

Syed Tanveer Jishan, Raisul Rashu, Naheena Haq, and Mohammad Rahman. Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique. *Decision Analytics*, 2, 12 2015. doi: 10.1186/s40165-014-0010-2.

Tanzila Khan and Henry Frey. Geospatial variation of real-world emissions from a passenger car. 06 2016.

R Core Team. weekdays: Extract parts of a posixt or date object. URL https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/weekdays. R Documentation.

C. Saranya and G. Manikandan. A study on normalization techniques for privacy preserving data mining. 2013. URL https://api.semanticscholar.org/CorpusID:16262497.

Luai Shalabi, Shaaban Zyad, and Basil Kasasbeh. Data mining: A preprocessing engine. *Journal of Computer Science*, 2, 09 2006. doi: 10.3844/jcssp.2006.735.739.

Teik Toe Teoh and Zheng Rong. *Association Rules*, pages 219–224. Springer Singapore, Singapore, 2022. ISBN 978-981-16-8615-3. doi: 10.1007/978-981-16-8615-3_13. URL https://doi.org/10.1007/978-981-16-8615-3_13.

S.M. Weiss and N. Indurkhya. *Predictive Data Mining: A Practical Guide*. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science, 1998. ISBN 9781558604032. URL https://books.google.com.br/books?id=xzVD8C2YpnQC.