Modelos de aprendizagem de máquina para a gestão estratégica de bicicletas compartilhadas

Machine learning models for shared bicycle strategic management

DOI:10.34117/bjdv7n1-151

Recebimento dos originais: 10/12/2020
Aceitação para publicação: 09/01/2021

Johnattan Douglas Ferreira Viana
Mestrando do Programa de Pós Graduação em Ciência da Computação – PPgCC
Universidade do Estado do Rio Grande do Norte (UERN) - Mossoró/RN - Brasil
Universidade Federal Rural do Semi-Árido (UFERSA) - Mossoró/RN – Brasil
Endereço:Rua Francisco Mota Bairro, 572 - Pres. Costa e Silva, Mossoró - RN, 59625-900
e-mail johnattandouglas@gmail.com

Thalia Katiane Sampaio Gurgel
Mestranda do Programa de Pós Graduação em Ciência da Computação – PPgCC
Universidade do Estado do Rio Grande do Norte (UERN) - Mossoró/RN – Brasil
Universidade Federal Rural do Semi-Árido (UFERSA) - Mossoró/RN – Brasil
Endereço:Rua Francisco Mota Bairro, 572 - Pres. Costa e Silva, Mossoró - RN, 59625-900
e-mail thaliasampaio8@gmail.com

Lenardo Chaves e Silva
Doutor. Programa de Pós Graduação em Ciência da Computação – PPgCC
Universidade Federal Rural do Semi-Árido (UFERSA) - Pau dos Ferros/RN – Brasil
Endereço:BR-226, S/N, Pau dos Ferros - RN, 59900-000
e-mail lenardo@ufersa.edu.br

Sebastião Emidio Alves Filho
Doutor. Programa de Pós Graduação em Ciência da Computação – PPgCC
Universidade do Estado do Rio Grande do Norte (UERN) - Mossoró/RN – Brasil
Endereço:Av. Prof. Antônio Campos - Pres. Costa e Silva, Mossoró - RN, 59610-210
e-mail sebastiao.alves@gmail.com

Carlos Heitor Pereira Liberalino
Doutor. Programa de Pós Graduação em Ciência da Computação – PPgCC
Universidade do Estado do Rio Grande do Norte (UERN) - Mossoró/RN – Brasil
Endereço:Av. Prof. Antônio Campos - Pres. Costa e Silva, Mossoró - RN, 59610-210
e-mail heitorliberalino@gmail.com

Álvaro Alvares de Carvalho César Sobrinho
Doutor. Universidade Federal do Agreste de Pernambuco (UFAPE) - Garanhuns/PE – Brasil
Endereço: Av. Bom Pastor, S/N - Boa Vista, Garanhuns - PE, 55292-270
e-mail alvaro.alvares@ufape.edu.br

RESUMO
Os Sistemas de Bicicletas Compartilhadas são cada vez mais comuns nas cidades, evidenciando a relevância de abordagens que auxiliem na tomada de decisão e gestão estratégica das empresas responsáveis por esses sistemas. Neste trabalho, é apresentado uma versão estendida de um estudo de caso no sistema Capital BikeShare (Washington, D.C.) com o objetivo de realizar uma análise estatística para identificar a relação entre o clima e o serviço de aluguel de bicicletas, além de aplicar algoritmos de aprendizagem de máquina para classificar a quantidade de aluguéis. Como adicional desta versão, apresenta-se uma abordagem de agrupamentos das estações de aluguéis, que resultou em uma acurácia de 94,7%, maior do que a obtida anteriormente. Os resultados apresentados evidenciam que o agrupamento das estações, além de aumentar a acurácia dos modelos preditivos, também facilitam a aplicação desses modelos preditivos na prática. Dessa forma, os modelos preditivos apresentados são mecanismos eficientes que permitem gerenciar esses sistemas, adotando estratégias de negócios baseadas na previsão do tempo e nas estações do ano.

Palavras Chave. Sistemas de Bicicletas Compartilhadas. Mineração de Dados. Gestão Estratégica.

ABSTRACT
Shared Bicycle Systems are increasingly common in cities, highlighting the relevance of approaches that assist in decision making and strategic management of these systems. This work presents an extended version of a case study in the Capital BikeShare system (Washington, D.C.) aiming to make a statistical analysis to identify the relationship between climate and bicycle rental service, in addition to applying machine learning algorithms to classify the number of rentals. In this extended version, it is presented a rental station grouping approach, which resulted in an accuracy of 94.7%, greater than that obtained previously. The results obtained show that stations grouping, in addition to increasing the accurate predictive models, also facilitate the application of these predictive models in practice. Thus, the predictive models presented are efficient mechanisms that allow to manage these systems adopting business strategies based on the weather and the seasons.

Keywords. Shared Bicycle Systems. Data Mining. Strategic management.

1 INTRODUÇÃO

Apesar do progresso na aplicação da Mineração de Dados em projetos voltados para mobilidade urbana, o crescimento desenfreado das cidades trouxe consigo alguns problemas como a crescente quantidade de veículos e congestionamentos cada vez maiores e mais demorados. Neste contexto, a mobilidade urbana, além de ser uma temática desafiadora para a gestão pública, tem se destacado como uma das áreas de estudo mais relevantes no âmbito das Cidades Inteligentes [Georgescu et al., 2015], resultando em sistemas baseados na coleta de dados para auxiliar na gestão estratégica e tomada de decisão em cenários urbanos [Randhawa e Kumar, 2017].

As informações coletadas pelos sistemas de bicicletas compartilhadas

permitiram o surgimento de várias aplicações que podem realizar a integração de serviços, análise de dados e tomada de decisão em tempo real. Esses sistemas têm se destacado como uma alternativa relevante para reduzir congestionamentos no trânsito e melhorar a qualidade de vida das pessoas, por meio da prática de atividade física [Mossa et al., 2020]. Além dos benefícios práticos dos sistemas de bicicletas compartilhadas, os dados gerados por esses sistemas os tornam interessantes para a pesquisa científica, uma vez que variáveis como duração da viagem e os dados sobre geolocalização são explicitamente registrados [Viana et al., 2019]. Portanto, os dados gerados se destacam como recursos que podem ser utilizados para estudar a mobilidade urbana. A mineração de dados possibilita a análise dos registros desses sistemas e auxilia os gestores na tomada de decisão e no gerenciamento estratégico, uma vez que o correto entendimento sobre a dinâmica dos atores que envolvem a mobilidade urbana pode levar a decisões mais precisas.

A gestão estratégica pode ser compreendida como a combinação de dois conceitos: gestão e estratégia. O primeiro conceito, sinônimo de administração, é definido por Chiavenato [2014] como o processo de planejar, organizar, liderar e controlar o uso de recursos de uma organização. Já o segundo conceito é definido por Pasquale et al. [2011] como as ações que a organização deve realizar para alcançar seus objetivos. Dessa forma, no ambiente empresarial, a estratégia está relacionada com objetivos financeiros, produtivos, mercadológicos e competitivos. Nos sistemas de bicicletas compartilhadas (estudo de caso neste trabalho), o conceito de estratégia pode ser aplicada de acordo com os padrões encontrados nos registros de aluguéis dos usuários. Neste cenário, a administração estratégica se refere a melhor utilização dos recursos disponibilizados pelo sistema (bicicletas), e, quando bem praticada, otimiza a alocação de recursos, aumentando eficiência, eficácia, visibilidade, transparência e atendimento aos objetivos estratégicos das empresas que gerenciam esses sistemas.

Para garantir o bom funcionamento de um Sistema de Compartilhamento de Bicicletas é essencial redistribuir as bicicletas, repor o número de bicicletas em determinadas estação de acordo com a demanda [Liu et al., 2016]. No entanto, a demanda de bicicletas em cada estação de aluguel muda com frequência e esse é um problema desafiador [Xu et al., 2020]. Abordagens baseadas somente no monitoramento em tempo real podem não ser eficientes, já que leva tempo redistribuir as bicicletas quando ocorre um desequilíbrio. Prevendo as demandas futuras é possível distribuir um número viável de bicicletas em cada estação para melhorar a qualidade do serviço e a experiência do

usuário. Assim, são necessárias abordagens para que os fornecedores de bicicletas compartilhadas redistribuam as bicicletas entre as estações de forma proativa e econômica de forma a garantir o funcionamento efetivo do sistema.

Sob essa perspectiva, a Mineração de Dados pode ser utilizada para analisar grandes quantidades de registros a procura de padrões consistentes em situações em que as técnicas tradicionais de exploração e análise de dados não sejam suficientes. Dessa forma é possível sistematizar os registros, interpretando-os, não só para análise dos eventos passados, mas também para a predição de situações futuras, identificando tendências no hábito dos usuários.

Este artigo é uma versão estendida de "Análise Comparativa de Modelos Preditivos na Gestão Estratégica de Bicicletas Compartilhadas: Um Estudo de Caso", publicado nos Anais do LII Simpósio Brasileiro de Pesquisa Operacional [Viana et al., 2020], que analisa a acurácia de modelos preditivos para os serviços de aluguel de bicicletas utilizando informações de climáticas e sazonais e que tem as seguintes contribuições: (i) pré-processamento de uma base de registros, gerando 10 novas bases com características específicas; (ii) análise de correlação entre os atributos das bases geradas; (iii) aplicação de algoritmos de aprendizagem supervisionada, além de comitês homogêneos e heterogêneos, para classificação da quantidade de aluguéis conforme os registros contidos nas bases geradas; e, por fim, (iv) análise comparativa das diferentes técnicas de pré- processamento e algoritmos utilizados.

Esta versão estendida apresenta alguns adicionais, dentre eles, detalhamento da problemática, apresentação dos atributos com valores discrepantes e aplicação dos algoritmos individuais e em comitês na base de dados, usando uma abordagem de agrupamento por quadrante das estações de aluguéis.

2 TRABALHOS RELACIONADOS

Os sistemas de compartilhamento de bicicletas evoluíram rapidamente desde a década de 60 [Zhang et al., 2015]. Algumas pesquisas já abordaram a análise e mineração de dados desses sistemas a fim de auxiliar na tomada de decisão [Vogel et al., 2011]. Por exemplo, Caulfield et al. [2017] examinaram os padrões de uso de uma cidade da Irlanda, analisando a dinâmica de utilização de um desses sistemas. Mahmoud et al. [2015] usaram registros do sistema de bicicletas compartilhadas de Toronto para analisar fatores que influenciam o número de ciclistas na cidade.

Zhang e Mi [2018] estimaram os benefícios ambientais do compartilhamento de

bicicletas na cidade de Xangai. O'Mahony e Shmoys [2015] analisaram os registros de um sistema de Nova York, abordando o problema de distribuição equilibrada de bicicletas durante os horários de pico. Fishman et al. [2015] identificaram e quantificaram os fatores que influenciam a participação dos ciclistas nesses sistemas, propondo um modelo de regressão relacionado aos hábitos dos usuários de cidades da Austrália. Moncayo-Martínez e Ramirez-Nafarrate [2016] realizaram uma análise dos padrões de mobilidade usando clusterização para entender o comportamento dos usuários na Cidade do México. Chen e Jakubowicz [2015] construiu um modelo capaz de inferir padrões de comportamento de viagens, avaliando dados da cidade de Washington DC.

Borgnat et al. [2011] e Kaltenbrunner et al. [2010] usaram modelos estatísticos de predição para diferentes propósitos, em estações de Lyon e Barcelona, respectivamente. Borgnat et al. [2011] utilizaram esses modelos para predizer o número de aluguéis em uma determinada hora. Kaltenbrunner et al. [2010] também utilizaram modelos estatísticos de predição para indicar o número de bicicletas livres para aluguel.

Viana et al. [2019] enriqueceram um sistema de bicicletas compartilhadas com informações meteorológicas e sazonais. Os autores evidenciaram padrões de atividade dos ciclistas relacionados a informações de data e clima, além de identificar um conjunto de parâmetros que influenciam o fluxo de aluguel de bicicletas. Neste contexto, exploraram a relação entre esses parâmetros e padrões, a fim de apresentar modelos preditivos de regressão para previsão da quantidade de aluguel. Diferentemente, neste trabalho, foi utilizada a base disponibilizada pelos autores para aplicar algoritmos de classificação, além de utilizar uma abordagem de agrupamento para as estações.

3 METOLODOGIA

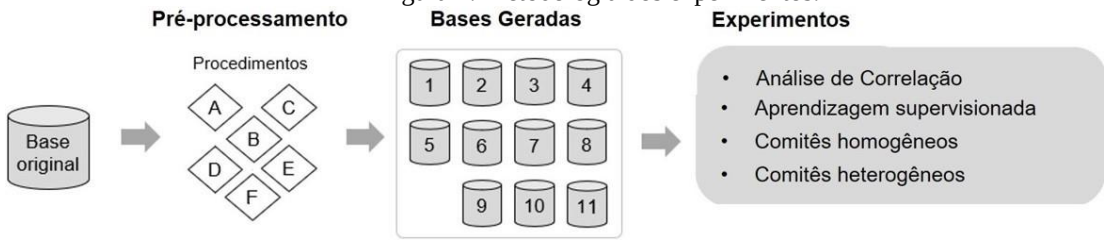
Nesse trabalho, foram utilizados, como estudo de caso, os registros do Capital BikeShare (CBS)[1]. Esse sistema está em funcionamento desde 2010 e atualmente possui mais de 500 estações e cerca de 4300 bicicletas. A base de dados original do ano de 2017 do CBS possui cerca de 3,75 milhões de registros. No entanto, para esse trabalho adotou-se, a base de dados disponibilizada por [Viana et al., 2019], que possui os registros dos aluguéis do CBS agrupados por hora e enriquecidos com informações de contexto. Essa base de dados possui 8737 registros e 17 atributos: *id*, *date*, *month*, *weekday*, *day*, *hour*, *season*, *workday*, *holiday*, *temperature*, *r_temperature*, *wind*, *humidity*, *dew_point*,

[1] Um sistema de bicicletas compartilhadas localizado em Washington D.C (capitalbikeshare.com).

pressure, *cut_description* e *qtd*.

Utilizou-se a biblioteca Pandas[2] para aplicar cinco procedimentos de pré-processamento (isoladamente e em conjunto) na base original, resultando em 10 novas bases de dados. Para os experimentos utilizaram-se algoritmos da biblioteca *Scikit Learn* [Pedregosa et al., 2011]: *C- Support Vector Classification* (SVC), *Gaussian Naive Bayes* (NB), *K-Neighbors Classifier* (KNN), *Multi-layer Perceptron classifier* (MLP), *Random Forest Classifier* (RFC) e *Decision Tree Classifier* (DT). Além disso, também foram aplicados comitês de algoritmos usando um conjunto dos mesmos classificadores (comitês homogêneos) e de diferentes classificadores (comitês heterogêneos) [Kuncheva, 2014]. A metodologia do presente trabalho é resumida na Figura 1.

Figura 1: Metodologia dos experimentos.



Na etapa de pré-processamento foi verificado que não existiam registros duplicados nem discrepantes. Ademais, realizaram-se os seguintes procedimentos na base original:

a. **Procedimento A**: remoção de registros incompletos. Alguns atributos foram removidos da base, a saber: *id*, que não é útil para os propósitos deste trabalho; *date*, uma coluna redundante, já que outras colunas (ex. *month*, *day*) foram derivadas dela; *holiday*, que possuía 8449 valores faltosos, já que só possuía valor se o dia fosse feriado; e *cut_description*, que possuía todos os registros vazios.

b. **Procedimento B**: conversão da variável *season* para um valor numérico. Esse atributo que era categórico com o nome das estações do ano foi tratado para ser representando com valores numéricos, em um novo atributo (*SeasonN*). Dessa forma, "Winter" foi substituído por 1, "Spring" por 2, "Summer" por 3, e "Fall" por 4.

c. **Procedimento C**: discretização da variável classe (*qtd*) utilizando *cut* e

[2] pandas.pydata.org

qcut. Para isso, transformou-se essa variável numérica em valores discretizados que representam a quantidade de aluguéis: "Muito baixa", "Baixa", "Média", "Alta"e "Muito Alta". Para isso, utilizou-se o método *cut* e *qcut* da biblioteca Pandas. O primeiro segmenta e classifica valores de dados em partições (equidistância) e o segundo gera partições que possuem aproximadamente a mesma quantidade de registros. As especificações de cada partição gerada por meio dessas abordagens são descritas em Viana et al. [2020].

d. **Procedimento D**: normalização das variáveis meteorológicas. Essas variáveis foram normalizadas utilizando o método *MinMaxScaler* da biblioteca *Scikit Learn*[3], que substitui por 0 o menor valor do conjunto (*Min*), e por 1 o maior valor (*Max*), calculando o restante dos valores entre esse intervalo, de acordo com a Equação 1, onde *x* é o valor a ser normalizado.

$$x = \frac{x - \text{Min}}{\text{Max} - \text{Min}} \quad (1)$$

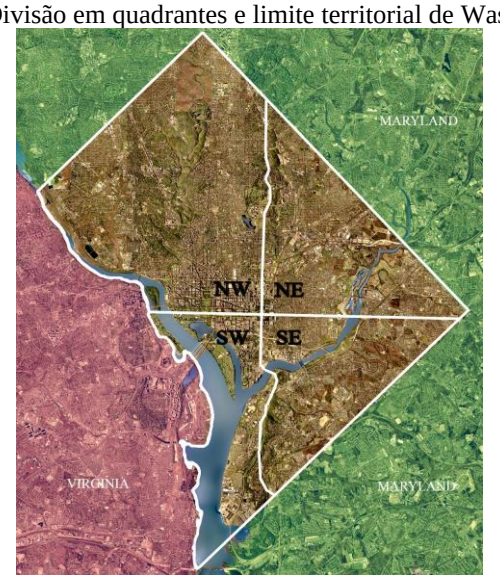
e. **Procedimento E**: aplicação da Análise de Componentes Principais (PCA, do inglês *Principal Component Analysis*) nas bases normalizadas, uma técnica de redução de dimensionalidade baseada na variância dos dados [Shlens, 2014]. Esse procedimento foi usado para redução de 25% e 50% da dimensionalidade. Assim, a partir das bases que possuíam 12 atributos (excluindo a variável classe), geraram-se bases que possuíam 9 e 6 atributos, respectivamente. No entanto, a aplicação da PCA na base original não foi proveitosa pois essa técnica é sensível a valores discrepantes. Por outro lado, a redução da dimensionalidade nas outras bases geradas não se fez necessária, uma vez que essas já tem o tamanho reduzido.

f. **Procedimento F:** agrupamento das estações de aluguéis. Conforme é apresentado na Figura 2, para definir os agrupamentos das estações foram consideradas a divisão por quadrantes (já existentes na estrutura da cidade de Washington DC): Noroeste (NW), Nordeste (NE), Sudeste (SE) e Sudoeste (SW). Esses quatro quadrantes foram usados para agrupar as 484 estações de bicicletas por meio da posição geográfica na qual o aluguel acontece. Cerca de 0,1% da base original (3836 registros) foi desconsiderada por possuir valores

[3] scikit-learn.org

nulos para as posições geográficas de inicio/fim do aluguel, impossibilitando identificar o quadrante onde o aluguel ocorreu.

Figura 2: Divisão em quadrantes e limite territorial de Washington, D.C.



Fonte: *United States Geological Survey* Wikipedia [2020]

Os procedimentos supracitados foram usados individualmente e em conjunto para gerar onze novas bases de dados. Na Tabela 1 são apresentadas as bases que foram geradas, especificando os procedimentos aplicados em cada uma.

Tabela 1: Procedimentos aplicados em cada base gerada.

	A	**B**	**C**		**D**	**E**		**F**
			cut	***qcut***		**25%**	**50%**	
Base 1	X							
Base 2	X	X						
Base 3	X	X	X					
Base 4	X	X		X				
Base 5	X	X	X		X			
Base 6	X	X		X	X			
Base 7	X	X	X		X	X		
Base 8	X	X	X		X		X	
Base 9	X	X		X	X	X		
Base 10	X	X		X	X		X	
Base 11	X	X	X		X			X

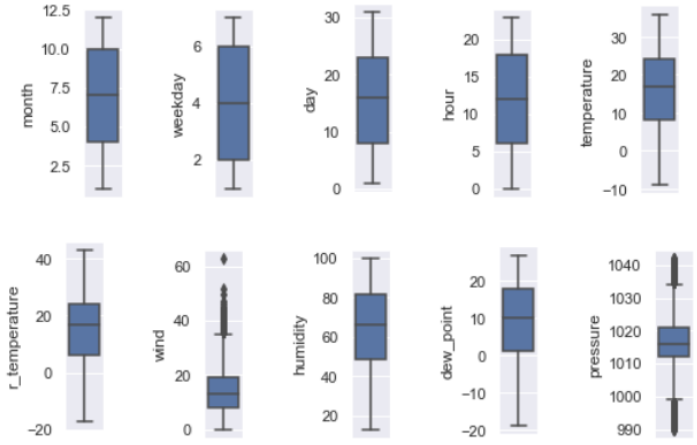
4 RESULTADOS E DISCUSSÕES

Nesta seção é apresentada a análise de correlação, bem como a utilização da aprendizagem supervisionada com algoritmos clássicos de classificação (SVC, NB, KNN, MLP, RFC e DT), além da aplicação de comitês homogêneos e heterogêneos.

Não foram identificados valores discrepantes nos atributos *month*, *weekday*, *day* e *hour*. Assim, o valor de *month* está sempre entre 1 e 12; *weekday* entre 1 e 7; *day*, entre 1 e 31; e *hour* entre 0 e 23. Ou seja, não existem valores fora da faixa comum para esses quatro atributos.

No que diz respeito aos atributos de clima, foram identificados valores discrepantes nos atributos *wind* e *pressure* (Figura 3). O atributo *pressure* tem relação com a *temperature*, já que quanto maior é a temperatura, menor é a pressão; e quanto maior é a pressão, menor é a temperatura. Isso ocorre porque, sob baixas temperaturas, o ar fica mais pesado e comprime o ar que está abaixo, elevando, assim, a pressão atmosférica. No que diz respeito a variável *wind*, é possível que em condições de ventos extremos aconteçam falhas de leitura ou viés substancial se baseado na observação humana [Dupuis e Field, 2004]. Como esses dois atributos praticamente não possuem correlação significativa com a quantidade de aluguéis (nas bases 4 e 6, *pressure* com correlação de 0,04 e *wind* com correlação de 0,13), esses valores discrepantes presentes nas bases de dados não foram tratados.

Figura 3: Intervalo de valores dos atributos da base de dados 1



4.1.ANÁLISE DE CORRELAÇÃO

A análise de correlação foi realizada em cada uma das bases de dados criadas para uma breve análise da relação entre as variáveis. A correlação entre esses atributos considera dois vetores aleatórios x e y de tamanhos n com médias $\bar{x}$ e $\bar{y}$ respectivamente. O coeficiente de

correlação (ρ) entre essas variáveis, neste caso, foi calculada por meio dos coeficientes de Pearson (Equação 2), onde d_i é a diferença entre cada posição de x e y [Becker, 2018].. Quanto mais próximo de 1 essa medida está, mais o valor de uma variável interfere no valor da outra (ou seja, quanto maior uma, maior a outra). Por outro lado, se o valor está próximo de –1, existe uma correlação inversa (quanto maior um, menor o outro). Contudo, quando o coeficiente de correlação é 0, não há correlação entre as variáveis [Bussab e Morettin, 2010].

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (2)$$

As matrizes de correlação das bases de dados foram coloridas conforme um intervalo de cores, no qual: a cor mais clara representa a correlação e a escura a correlação inversa. Valores próximos de zero estão coloridos em roxo, indicando a inexistência de correlação.

As matrizes de correlação geradas para as bases de dados de 1 a 6 apresentam diferenças sutis (Figura 4). Como esperado, a estação do ano infere nas outras variáveis climáticas, e, além disso, tem uma relação bastante significativa com o mês do ano, já que, em Washington DC, as estações são bem definidas e costumam iniciar e terminar nas mesmas datas. Observou-se uma relação entre *temperature* e *r_temperature*; *wind* e *humidity*; e *dew_point* e *pressure* que tem correlação com as outras variáveis climáticas. Em relação a variável classe (*qtd*), é possível observar que ela tem uma correlação positiva com *hour*, *temperature* e *r_temperature*, enquanto tem uma correlação negativa com o atributo *humidity*. Em outras palavras, subtende-se que conforme o clima está mais quente e menos úmido, o número de aluguéis aumenta.

Essa correlação entre temperatura e a quantidade de aluguéis também foi identificada por Mahmoud et al. [2015] no sistema de compartilhamento de bicicletas em Toronto. Em termos de gestão estratégica, por exemplo, essas informações podem ser utilizadas para planejar, em conjunto com ferramentas de previsão do tempo, um calendário de manutenção nas bicicletas em épocas do ano com baixo fluxo de aluguéis (época de quadros chuvosos ou de extremo frio).

Em relação às bases de dados geradas com a PCA com redução de 25% (7 e 9) e 50% (8 e 10) não foi visualizada correlação entre os atributos, o que era esperado, já que novos atributos (totalmente diferentes dos originais) foram gerados com a aplicação dessa técnica.

4.2.APRENDIZAGEM SUPERVISIONADA

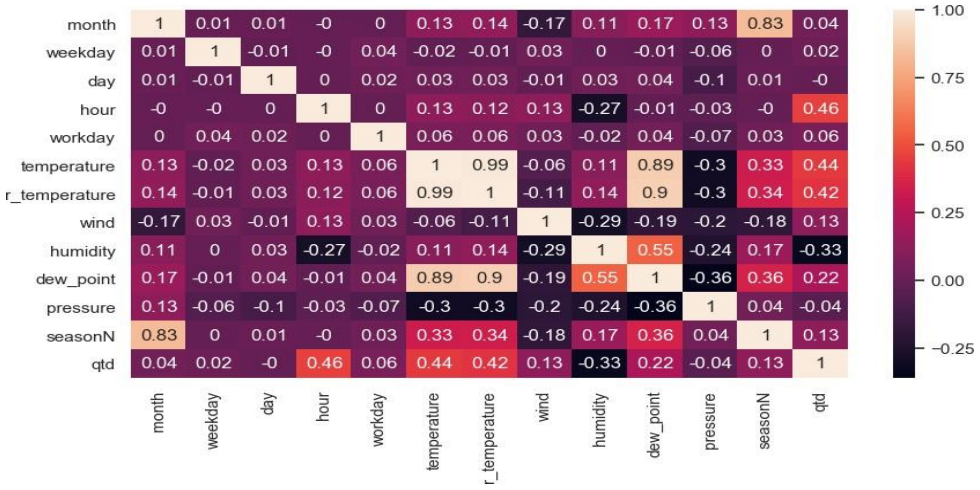
Para os algoritmos de aprendizagem supervisionada utilizou-se validação cruzada *k-fold* de modo que a mesma base é usada para treinamento e teste do algoritmo [James et al., 2013]. Assim, dividiu-se a base em 10 partes, usando 9 para treino e a parte remanescente para teste, repetindo o experimento 10 vezes e calculando a média da acurácia e do desvio padrão para cada algoritmo. As médias desses valores são exibidas na Tabela 2, com aproximação de 3 casas decimais. Em negrito, está destacada a melhor acurácia obtida para cada uma das bases. Nas bases de dados de 3 a 6, a árvore de decisão foi o algoritmo com melhor desempenho. Já nas bases de dados onde a redução de dimensionalidade foi aplicada, o classificador com melhor acurácia foi o MLP. E, na base de dados 11, o melhor classificador foi o RFC. A aplicação dos algoritmos na base de dados 11 será discutida na Seção 4.4.

Tabela 2: Acurácia e desvio padrão dos algoritmos individuais em cada base.

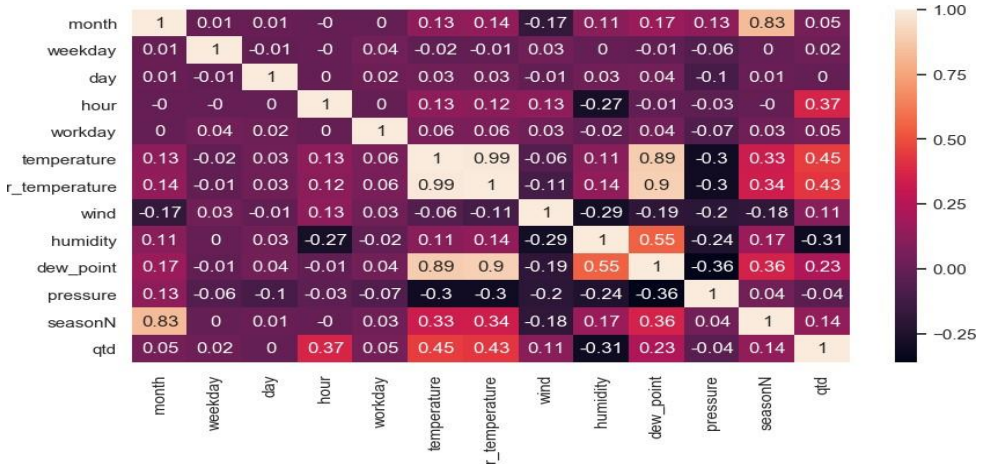
BASE	SVC	NB	KNN	MLP	RFC	DT
3	0,562 ± 0,129	0,631 ± 0,117	0,643 ± 0,061	0,591 ± 0,106	0,779 ± 0,028	**0,807** ± 0,026
4	0,217 ± 0,042	0,464 ± 0,041	0,450 ± 0,017	0,405 ± 0,046	0,722 ± 0,064	**0,746** ± 0,062
5	0,676 ± 0,052	0,628 ± 0,117	0,684 ± 0,043	0,774 ± 0,023	0,778 ± 0,030	**0,807** ± 0,026
6	0,489 ± 0,017	0,464 ± 0,041	0,550 ± 0,042	0,721 ± 0,059	0,712 ± 0,052	**0,744** ± 0,057
7	0,720 ± 0,036	0,650 ± 0,51	0,672 ± 0,037	**0,823** ± 0,022	0,717 ± 0,031	0,672 ± 0,044
8	0,654 ± 0,053	0,629 ± 0,067	0,611 ± 0,048	**0,661** ± 0,041	0,638 ± 0,050	0,581 ± 0,035
9	0,584 ± 0,036	0,463 ± 0,027	0,521 ± 0,047	**0,738** ± 0,060	0,603 ± 0,059	0,567 ± 0,066
10	0,496 ± 0,033	0,447 ± 0,031	0,450 ± 0,022	**0,525** ± 0,320	0,467 ± 0,040	0,435 ± 0,044
11	0,910 ± 0,015	0,748 ± 0,049	0,903 ± 0,016	0,933 ± 0,009	**0,938** ± 0,013	0,934 ± 0,009

De modo geral, a discretização por meio do método *cut* resultou em melhores resultados na acurácia dos algoritmos aplicados. Além disso, é possível perceber que a normalização diminuiu a variação da acurácia nos modelos gerados. É notável também que a PCA com 25% obteve melhor desempenho que a PCA com 50%, o que aconteceu por razão da perda de informação que acontece ao diminuir a quantidade de atributos.

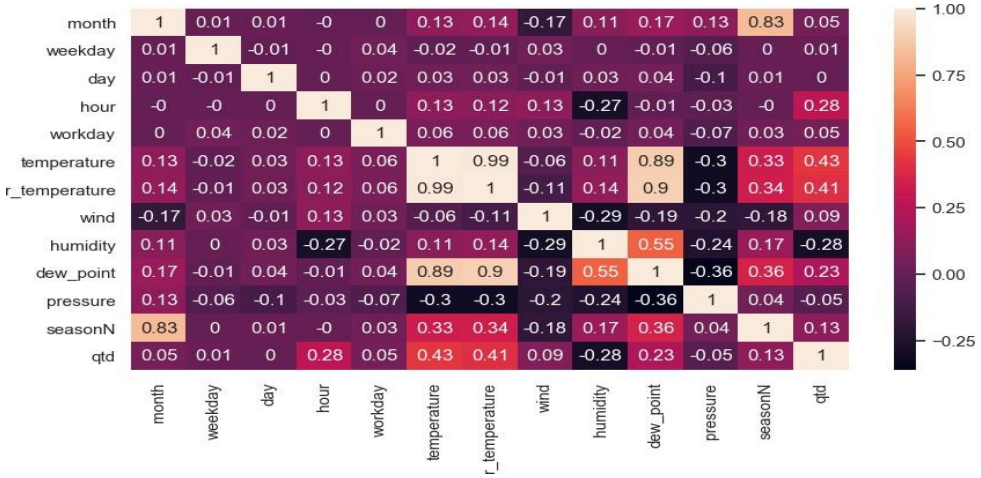
Figura 4: Correlação das bases geradas por cada método de discretização.



(a) Bases 1 e 2 (sem discretização).



(b) Bases 3 e 5 (discretização com método *cut*).



(c) Base 4 e 6 (discretização com método *qcut*).

4.3.APLICAÇÃO DE COMITÊS HOMOGÊNEOS E HETEROGÊNEOS

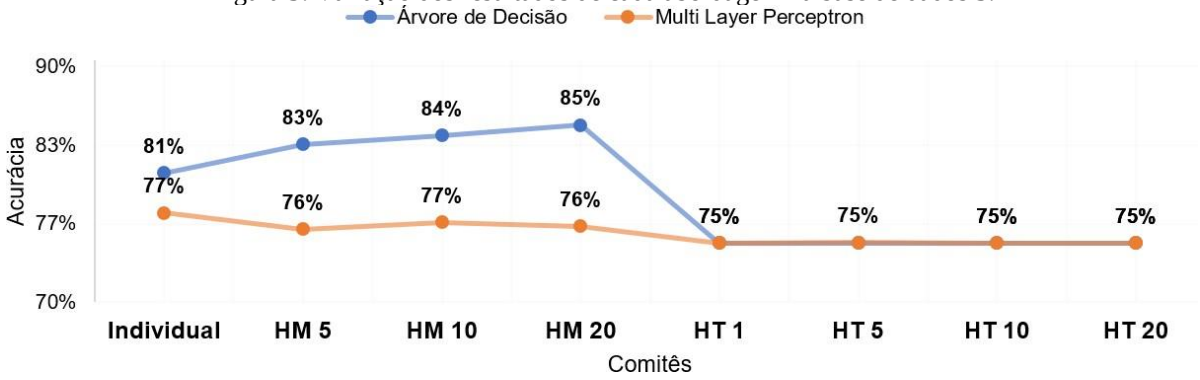
Na base de dados 5, na qual os algoritmos isolados obtiveram melhor desempenho, foram

aplicados os comitês homogêneos e heterogêneos. Os comitês homogêneos foram aplicados utilizando o *BaggingClassifier* (também da biblioteca *Scikit Learn*) que gera um conjunto de modelos utilizando um algoritmo de aprendizagem simples por meio da combinação por votos para classificação [Breiman, 1996]. Foram aplicados comitês homogêneos de tamanho $T = \{ 5, 10, 20\}$ para os algoritmos SVC, NB, KNN, MLP e DT.

Para compor os comitês heterogêneos foram usados os três algoritmos com melhor desempenho na aplicação dos comitês homogêneos (KNN, MLP e DT). Esses algoritmos foram combinados para compor os seguintes comitês heterogêneos: KNN e MLP; KNN e DT; MLP e DT; KNN, MLP e DT. Para tal, aplicou-se o *Stacking*, uma técnica de aprendizado de conjunto para combinar vários modelos de classificação por meio de um meta-classificador (nesse caso, os dois melhores: MLP e DT). Os modelos de classificação individual são treinados com base no conjunto de treinamento completo; então, o meta-classificador é ajustado com base nos resultados dos modelos de classificação individuais [Tang et al., 2015]. Essa técnica foi aplicada usando o *StackingClassifier* da biblioteca *Mlxtend*[4].

Os resultados detalhados destes experimentos podem ser conferidas em Viana et al. [2020]. Em resumo, a Figura 5 mostra a variação de desempenho das abordagens utilizadas para geração dos modelos preditivos na base de dados 5. Nota-se que a aplicação dos comitês heterogêneos não possui tanto impacto quanto a dos homogêneos.

Figura 5: Variação dos resultados de cada abordagem na base de dados 5.



Fonte: Viana et al. [2020]

[4] rasbt.github.io/mlxtend/user_guide/classifier/StackingClassifier/

O único comitê que variou de acurácia com a alteração do tamanho foi o KNN e MLP. Todos os outros permaneceram com a mesma acurácia, independente do meta-classificador usado. No entanto, esta acurácia de aproximadamente 75% no restante dos comitês foi inferior aos resultados obtidos anteriormente, portanto, a utilização dessa técnica nessa situação se mostrou ineficaz, uma vez que a acurácia não melhorou após o agrupamento dos comitês heterogêneos. Por outro lado, utilizando um comitê homogêneo de Árvores de Decisão de tamanho 20, obteve-se a acurácia de 85%, evidenciando a eficiência do algoritmo, tanto individualmente quanto em conjunto.

4.4.ABORDAGEM COM AGRUPAMENTOS DE ESTAÇÕES

Embora as análises anteriores auxiliem na tomada de decisão nesses sistemas, o quantitativo de aluguéis em nível de cidade pode não ser suficiente para solucionar o problema de desbalanceamento de bicicletas. De um ponto de vista operacional, é mais prático predizer a quantidade de aluguéis em determinadas regiões da cidade. Nesse sentido, realizaram-se agrupamentos das estações, de modo a alocar recursos conforme as necessidades de cada grupo.

Na Tabela 3 é detalhada a quantidade de estações em cada quadrante, bem como a duração média dos aluguéis que acontecem em cada um. A maior parte das estações encontram-se no quadrante NW, por conta do maior espaço territorial destinado a esse quadrante. Além disso, percebe-se que, embora o quadrante SW não possua o maior número de estações, esse quadrante possui, em média, aluguéis que duram mais. Por exemplo, os aluguéis no quadrante SW são cerca de 80% mais demorados em relação a aluguéis que acontecem no quadrante NE.

Tabela 3: Quantidade de estações e duração média dos aluguéis por quadrante.

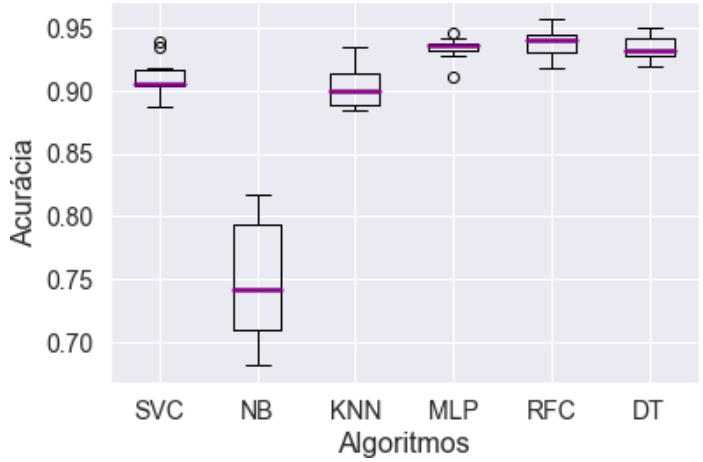
Quadrante	**Quantidade de Estações**	**Duração Média (minutos)**
NW	273 (53,4%)	18,2
NE	53 (11,0%)	14,7
SW	121 (25,0%)	26,48
SE	37 (7,6%)	15,67

Uma vez que a aplicação do método *cut* resultou em melhores acurácias nas bases de dados nais quais ele foi aplicado, na base de dados 11 (com o agrupamento das estações de aluguéis), aplicou-se esse método para discretização da quantidade de

aluguéis, utilizando os seguintes partições: 0 a 261, muito baixa; 262 a 520, baixa; 521 a 780, média; 781 a 1039, alta; 1040 a 1299, muito alta. Após a discretização, uma análise da correlação entre os atributos desta nova base mostrou resultados similares aos apresentados anteriormente (Seção 4.1).

Na Figura 6 é apresentada a acurácia dos algoritmos aplicados na base de dados 11, a qual tem as estações agrupadas por quadrante. Em geral, os algoritmos obtiveram melhor desempenho nesta base. O RFC se destacou, com a acurácia de aproximadamente 93,82% ± 0,011. Dessa forma, o resultado alcançado pelo RFC (sem uso de comitês) ultrapassou os resultados alcançados pelo DT (93,24% ± 0,008) e MLP (93,36% ± 0,009), que foram destaques nas bases de dados anteriores.

Figura 6: Acurácia dos algoritmos aplicados na base de dados 11.

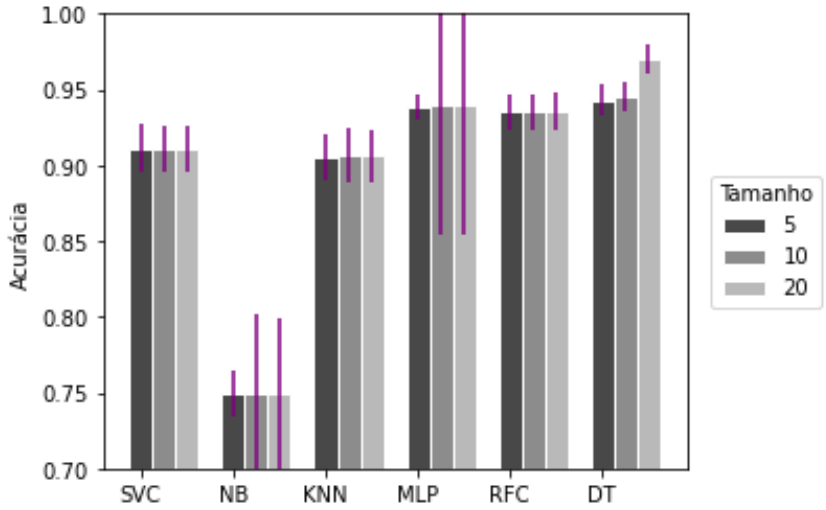


A aplicação dos comitês é apresentada na Figura 7. Os comitês homogêneos não foram proveitosos em todos os algoritmos. Como mostrado na Figura 7(a), somente MLP e DT tiveram um aumento significado da acurácia com a aplicação dessa técnica. Destaca-se, ainda, que os comitês homogêneos de DT, dos diversos tamanhos, apresentaram acurácias maiores do que as alcançadas pelos comitês homogêneos de RFC. A acurácia do MLP esteve entre os melhores, no entanto, este algoritmo demanda de um grande custo computacional.

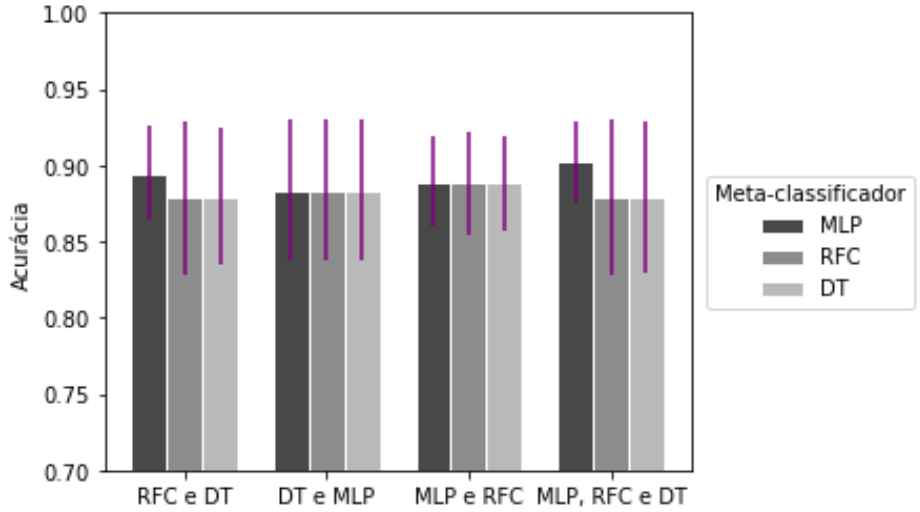
Adicionalmente, como exposto na Figura 7 (b), foram construídos comitês heterogêneos com a combinação dos algoritmos que obitveram melhores acurácias individualmente, usando cada um deles como meta-classificadores (MLP, RFC e DT), porém, não obtiveram acurácias maiores do que as já encontradas. O melhor resultado alcançado com os comitês heterogêneos foi de 90,21% ± 0,027, obtido com o comitê composto pelo MLP, RFC e DT, usando o MLP como meta-classificador.

Figura 7: Aplicação de comitês na base de dados com estações agrupadas.

(a) Comitês comitês homogêneos.



(b) Comitês heterogêneos.



5 CONCLUSÕES

Como mostrado, a mineração de dados aliada à prática da gestão estratégica se mostra um mecanismo eficiente que permite implementar avanços referente à interpretação dos registros nessa modalidade de sistema de transporte.

Com este trabalho evidenciou-se a possibilidade de gerar modelos preditivos para classificar a quantidade de aluguéis usando informações de contexto em sistemas de bicicletas compartilhadas. Nesse trabalho, foram fornecidas, aos gestores dessa modalidade de sistemas, informações precisas acerca do serviço de aluguel de bicicletas para auxiliá-los no entendimento do hábito dos usuários. Assim, o uso dos modelos preditivos apresentados auxiliam na tomada de decisão dos gestores desses sistemas. Na

classificação da quantidade de aluguéis, a melhor acurácia obtida foi de 94,7%, alcançada com a aplicação de um comitê homogêneo de Árvore na base de dados com as estações agrupadas. Dessa forma, os agrupamentos, além de melhorar a acurácia, também são mais úteis na solução prática do problema de desbalanceamento. A aplicação dos comitês heterogêneos não se mostrou efetiva, mesmo usando a Árvore de Decisão como meta-classificador. Como trabalhos futuros pretende-se aplica e avaliar a acurácia deses algoritmos em bases de dados com outras abordagens de agrupamentos.

REFERÊNCIAS

Becker, R. (2018). *The new S language*. CRC Press.

Borgnat, P., Robardet, C., Rouquier, J. B., Abry, P., Fleury, E., e Flandrin, P. (2011). Shared Bicycles in a City: A Signal Processing and Data Analysis Perspective. *Advances in Complex Systems*, 14(3):415–438. URL https://hal-ens-lyon.archives-ouvertes.fr/ ensl-00490325.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140. Bussab, W. O. e Morettin, P. A. (2010). *Estatística Básica*. Saraiva.

Caulfield, B., O'Mahony, M., Brazil, W., e Weldon, P. (2017). Examining usage patterns of a bike- sharing scheme in a medium sized city. *Transportation Research Part A: Policy and Practice*, 100:152 – 161. ISSN 0965-8564. URL http://www.sciencedirect.com/science/ article/pii/S0965856416304141.

Chen, L. e Jakubowicz, J. (2015). Inferring bike trip patterns from bike sharing system open data.

In *2015 IEEE International Conference on Big Data (Big Data)*, p. 2898–2900. Chiavenato, I. (2014). *Introdução à teoria geral da administração*. Editora Manole.

Dupuis, D. J. e Field, C. A. (2004). Large wind speeds: Modeling and outlier detection. *Journal of Agricultural, Biological, and Environmental Statistics*, 9(1):105–121. ISSN 10857117. URL http://www.jstor.org/stable/1400709.

Fishman, E., Washington, S., Haworth, N., e Watson, A. (2015). Factors influencing bike share membership: An analysis of melbourne and brisbane. *Transportation Research Part A: Policy and Practice*, 71:17 – 30. ISSN 0965-8564. URL http://www.sciencedirect.com/ science/article/pii/S0965856414002638.

Georgescu, M., Pavaloaia, V., Popescul, D., e Tugui, A. (2015). The race for making up the list of emergent smart cities. an eastern european country's approach. *Transformations in Business and Economics*, 14:529–549.

James, G., Witten, D., Hastie, T., e Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer Texts in Statistics. Springer.

Kaltenbrunner, A., Meza, R., Grivolla, J., Codina, J., e Banchs, R. (2010). Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system. *Pervasive and Mobile Computing*, 6(4):455 – 466. ISSN 1574-1192. URL http://

www.sciencedirect.com/science/article/pii/S1574119210000568. Human Behavior in Ubiquitous Environments: Modeling of Human Mobility Patterns.

Kuncheva, L. I. (2014). *Combining Pattern Classifiers: Methods and Algorithms*. Wiley Publishing, 2nd edition. ISBN 1118315235.

Liu, J., Sun, L., Chen, W., e Xiong, H. (2016). Rebalancing bike sharing systems: A multi-source data smart optimization. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, p. 1005–1014, New York, NY, USA. Association for Computing Machinery. ISBN 9781450342322. URL https://doi.org/ 10.1145/2939672.2939776.

Mahmoud, M., El-Assi, W., e Nurul Habib, K. (2015). Effects of built environment and weather on bike sharing demand: Station level analysis of commercial bike sharing in toronto.

Moncayo-Martínez, L. A. e Ramirez-Nafarrate, A. (2016). Visualization of the mobility patterns in the bike-sharing transport systems in mexico city. In *2016 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, p. 1851–1855.

Mossa, R. V., Ladewig, I., e Uvinha, R. R. (2020). Desafios da bicicleta como meio de transporte: o deslocamento de estudantes de dois colégios da rede pública no viário de curitiba. *Brazilian Journal of Development*, 6(6):33485–33505.

O'Mahony, E. e Shmoys, D. B. (2015). Data analysis and optimization for (citi)bike sharing. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, p. 687–694. AAAI Press. ISBN 0262511290.

Pasquale, P., Neto, C., Gomes, e Celso (2011). *Comunicação Integrada de Marketing: A Teoria na Prática*. Elsevier.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., e Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Ma chine Learning Research*, 12:2825–2830.

Randhawa, A. e Kumar, A. (2017). Exploring sustainability of smart development initiatives in india. *International Journal of Sustainable Built Environment*, 6(2):701 – 710. ISSN 2212-6090. URL http://www.sciencedirect.com/science/article/ pii/S2212609017300742.

Shlens, J. (2014). A tutorial on principal component analysis. *CoRR*, abs/1404.1100. URL http:

//arxiv.org/abs/1404.1100.

Tang, J., Alelyani, S., e Liu., H. (2015). *Data Classification: Algorithms and Applications.* Data Mining and Knowledge Discovery Series, CRC Press.

Viana, J. D. F., Gurgel, T. K. S., Chaves e Silva, L., Alves Filho, S. E., e Liberalino, C. H. P. (2020). Análise comparativa de modelos preditivos na gestão estratégica de bicicletas compartilhadas: Um estudo de caso. In *Anais do LII Simpósio Brasileiro de Pesquisa Operacional - SBPO*. Sociedade Brasileira de

Computação (SBC).URL https://proceedings.science/sbpo-2020/papers/

analise-comparativa-de-modelos-preditivos-na-gestao-estrategica-de-bicicleta

Viana, J. D. F., Braga, O., Silva, L., e Neto, F. M. (2019). Analyzing patterns of a bicycle sharing system for generating rental flow predictive models. In *Anais do III Workshop de Computação Urbana*, p. 57–70, Porto Alegre, RS, Brasil. SBC. URL https://sol.sbc.org.br/ index.php/courb/article/view/7468.

Vogel, P., Greiser, T., e Mattfeld, D. C. (2011). Understanding bike-sharing systems using data mining: Exploring activity patterns. *Procedia - Social and Behavioral Sciences*, 20:514 –

523. ISSN 1877-0428. URL http://www.sciencedirect.com/science/article/ pii/S1877042811014388. The State of the Art in the European Quantitative Oriented Transportation and Logistics Research – 14th Euro Working Group on Transportation & 26th Mini Euro Conference & 1st European Scientific Conference on Air Transport.

Wikipedia (2020). Washington, D.C. Online: https://pt.wikipedia.org/wiki/ Washington,_D.C.Acessado em 30/10/2020.

Xu, T., Han, G., Qi, X., Du, J., Lin, C., e Shu, L. (2020). A hybrid machine learning model for demand prediction of edge-computing-based bike-sharing system using internet of things. *IEEE Internet of Things Journal*, 7(8):7345–7356.

Zhang, L., Zhang, J., yu Duan, Z., e Bryde, D. (2015). Sustainable bike-sharing systems: characteristics and commonalities across cases in urban china. *Journal of Cleaner Production*, 97:124 –

133. ISSN 0959-6526. URL http://www.sciencedirect.com/science/article/ pii/S0959652614003448. Special Volume: Why have 'Sustainable Product-Service Systems' not been widely implemented?

Zhang, Y. e Mi, Z. (2018). Environmental benefits of bike sharing: A big data-based analysis. *Applied Energy*, 220:296 – 301. ISSN 0306-2619. URL http://www.sciencedirect. com/science/article/pii/S0306261918304392.