LII Simpósio Brasileiro de Pesquisa Operacional
João Pessoa - PB, 3 a 5 de novembro de 2020

Análise Comparativa de Modelos Preditivos na Gestão Estratégica de Bicicletas Compartilhadas: Um Estudo de Caso

Johnattan Douglas Ferreira Viana
Thalia Katiane Sampaio Gurgel
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
{johnattandouglas, thaliasampaio8}@gmail.com

Lenardo Chaves e Silva
Programa de Pós Graduação em Ciência da Computação - PPgCC
Universidade Federal Rural do Semi-Árido (UFERSA) - Pau dos Ferros/RN -Brasil
lenardo@ufersa.edu.br

Sebastião Emidio Alves Filho
Carlos Heitor Pereira Liberalino
Programa de Pós Graduação em Ciência da Computação - PPgCC
Universidade do Estado do Rio Grande do Norte (UERN) - Mossoró/RN - Brasil
{sebastiao.alves, heitorliberalino}@gmail.com

RESUMO

Os Sistemas de Bicicletas Compartilhadas são cada vez mais frequentes nas cidades, evidenciando a relevância de abordagens que auxiliem na tomada de decisão e gestão estratégica desses sistemas. Esse trabalho apresenta um estudo de caso no sistema Capital BikeShare (Washington, D.C.) com o objetivo de fazer 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. A melhor acurácia obtida pelos modelos preditivos gerados foi de aproximadamente 85%, alcançada com a aplicação de um comitê homogêneo de Árvores de Decisão. Os resultados mostrados evidenciam que o entendimento dos hábitos dos usuários em conjunto com 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.

Tópicos (Mineração de Dados, Aprendizagem de Máquina)

ABSTRACT

Shared Bicycle Systems are increasingly frequent in cities, highlighting the relevance of approaches that assist in decision making and strategic management of these systems. This work presents 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. The best accuracy obtained by the generated predictive models was approximately 85%, employing homogeneous committee of Decision Trees. Results show that the understanding of users' habits with 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.

Paper topics (Data Mining, Machine Learning)

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 fazer a integração de serviços, análise de dados e tomada de decisão em tempo real. Esses sistemas têm se destacado como uma ótima alternativa para reduzir o congestionamento de tráfego [Hamilton e Wichman, 2018], além de impactar positivamente na saúde da população e no meio ambiente [Souza e Gomes, 2014]. Além dos benefícios práticos dos sistemas de bicicletas compartilhadas, os dados gerados por esses sistemas os tornam atraentes para a pesquisa, uma vez que variáveis como duração da viagem e os dados sobre geolocalização são explicitamente registrados [Viana et al., 2019]. Assim, os dados gerados se destacam como um recurso que pode ser utilizados um recurso para estudar a mobilidade urbana. Aliada a isso, 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 apuradas.

A gestão estratégica pode ser compreendida como a combinação de dois conceitos: Gestão e estratégia. O primeiro, sinônimo de administração, é definido por Chiavenato [2014] como sendo o processo de planejar, organizar, liderar e controlar o uso de recursos objetivando alcançar os objetivos de uma organização. Já o segundo é definido por Pasquale et al. [2011] como as ações que a organização deve realizar para atingir seus objetivos. Dessa forma, no ambiente empresarial, a estratégia está relacionada a objetivos financeiros, produtivos, mercadológicos e competitivos. Nos sistemas de bicicletas compartilhadas, como o estudo de caso deste trabalho, a 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, trazendo maior eficiência, eficácia, visibilidade, transparência e atendimento aos objetivos estratégicos desses sistemas.

Sob essa perspectiva, a Mineração de Dados pode ser utilizada para explorar 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.

Por essa razão, o intuito deste trabalho é analisar a acurácia de modelos preditivos para os serviços de aluguel de bicicletas utilizando informações de climáticas e sanzonais. Esse trabalho tem como seu principal objetivo realizar uma análise comparativa de diferentes técnicas de aprendizagem de máquina para identificar qual o melhor modelo preditivo capaz de classificar a quantidade de aluguéis em um sistema de bicicletas compartilhadas. Possuindo 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 e não-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.

O restante do trabalho está organizado como segue: Na Seção 2 são mostrados alguns trabalhos relacionados, que são voltados para análise de dados de sistemas de bicicletas compartilhadas; na Seção 3 é apresentada a metodologia utilizada em sua construção. Na Seção 4 são apresentados e discutidos os resultados. Por fim, a Seção 5 expõe as conclusões obtidas.

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. E 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 utilizou modelos estatísticos de predição para indicar o número de bicicletas livres para aluguel.

Viana et al. [2019] enriqueceu 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. Eles 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, será utilizada a base disponibilizada pelos autores para aplicar algoritmos de classificação.

3. Metolodogia

Nesse trabalho foi utilizado para o estudo de caso os registros do Capital BikeShare (CBS)[1], um sistema de bicicletas compartilhadas localizado em Washington D.C. Esse sistema está

[1] capitalbikeshare.com

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, listados e descritos na Tabela 1. Conforme descrito adiante, a variável *qtd* será tranformada na variável classe dessa base.

Tabela 1: Listagem dos atributos presentes na base de dados de Viana et al. [2019].

Nome	Descrição	Tipo
id	Identificação do registro	Numérico
date	Data	Numérico
month	Mês	Numérico
weekday	Dia da semana	Categórico
day	Dia do mês	Numérico
hour	Hora do dia	Numérico
season	Estação do ano	Categórico
workday	Dia útil (1 se verdadeiro)	Binário
holiday	Nome do feriado	Categórico
temperature	Temperatura	Numérico
r_temperature	Sensação térmica	Numérico
wind	Velocidade do Vento	Numérico
humidity	Umidade do ar	Numérico
dew_point	Ponto de condesação da água	Numérico
pressure	Pressão atmosférica	Numérico
cut_description	Informação climática adicional	Nominal
qtd	Quantidade de aluguéis realizados	Numérico

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]. O algoritmo de aprendizagem não-supervisionada utilizado foi o K-Means, enquanto que os algoritmos para aprendizagem supervisionada foram os seguintes: *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.

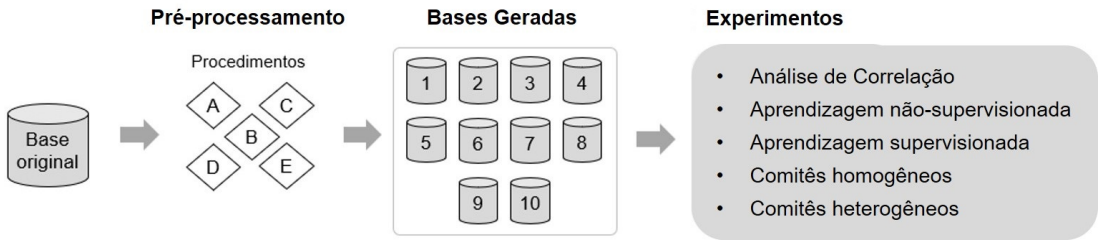


Figura 1: Metodologia dos experimentos.

[2]pandas.pydata.org

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

- **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.

- **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.

- **Procedimento C**: discretização da variável classe (*qtd*) utilizando *cut* e *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. Na Tabela 2 são especificadas as partições geradas por cada uma dessas abordagens e mostra a quantidade de registros presentes em cada partição.

Tabela 2: Partições usadas na categorização da variável classe (*qtd*).

Categorização	***cut***		***qcut***	
	Intervalo	**Quantidade**	**Intervalo**	**Quantidade**
Muito Baixa	(0, 398]	4894	(0, 51]	1751
Baixa	(398, 796]	2210	(51, 210]	1751
Média	(796 , 1193]	1102	(210, 446]	1740
Alta	(1193, 1591]	423	(446, 761)	1747
Muito Alta	(1591, 1988]	108	(761, 1988)	1748

- **Procedimento D**: normalização das variáveis meteorológicas. Essas variáveis foram normalizadas utilizando o método *MinMaxScaler* da biblioteca *Scikit Learn*[3], que ajusta os valores no intervalo entre 0 (menor valor) e 1 (maior valor).

- **Procedimento E**: aplicação da Análise de Componentes Principais (PCA) 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 *outliers*. Por outro lado, a redução da dimensionalidade nas outras bases geradas não se fez necessária, uma vez que essas já tem o tamanho reduzido.

Os procedimentos supracitados foram usados individualmente e em conjunto para gerar dez bases. Na Tabela 3 são mostradas as bases que foram geradas, especificando os procedimentos aplicados em cada uma.

[3] scikit-learn.org

Tabela 3: Procedimentos aplicados em cada base gerada.

	A	B	C		D	E	
			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

4. Resultados e Discussões

Nesta seção é descrita a análise de correlação das bases, bem como a utilização de abordagens de aprendizagem não-supervisionada (K-Means) e supervisionada com algoritmos clássicos de classificação (SVC, NB, KNN, MLP, RFC e DT), além da aplicação de comitês de algoritmos.

A análise de correlação foi realizada em cada uma das bases criadas para uma breve análise da relação entre as variáveis. Correlação é uma análise descritiva que utiliza uma medida entre duas variáveis, assim, 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 quer dizer que a correlação é 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]. As Bases de 1 a 6 apresentaram matrizes de correlação similares. Então, para reduzir espaço e evitar redundância, na Figura 2 é mostrada somente a matriz de correlação da Base 6, colorida conforme um mapa de calor que possui um intervalo de cores no qual a cor mais clara representa a correlação, e a escura a correlação inversa. Valores próximos de zero estão colorido em roxo, indicando a inexistência de correlação.

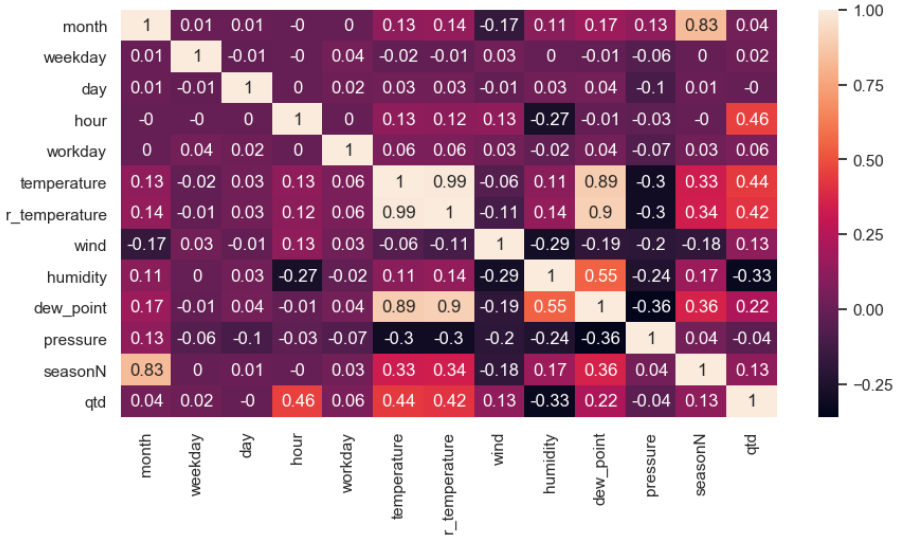


Figura 2: Matriz de Correlação da Base 6.

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 começar 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 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.

Mesmo que a aprendizagem não-supervisionada não seja ideal para a tarefa de classificação, uma vez que não utiliza a variável classe, ela foi aplicada com o mero intuito de experimentação, apenas para visualizar possíveis agrupamentos nos registros. Para tal, utilizou-se a Base 2, onde a variável *qtd* não foi discretizada. Para isso, foi necessário definir o número de *clusters* que se desejava criar. O número ideal de *clusters*, nesse caso, foi 4. Esse valor foi obtido por meio do Método Cotovelo (do inglês *Elbow Method*). No entanto, com a aplicação do K-Means para geração de 4 *clusters*, o algoritmo não evidenciou agrupamentos bem definidos na Base 2. Uma das razões para a ineficiência desse algoritmo pode ser o fato de que nesta base as instâncias já foram agrupadas por hora, uma vez que a variável classe dessa base é o total de aluguéis realizados em cada hora da base de dados original. Essa redução de instâncias e o agrupamento por esse atributo pode ter encoberto padrões e agrupamentos relacionados às instâncias de cada aluguel individualmente.

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. A média desses valores é mostrada na Tabela 4, com aproximação de 3 casas decimais. Em negrito, está destacada a melhor acurácia obtida para cada uma das bases. A árvore de decisão foi o algoritmo com melhor desempenho nas bases onde a redução de dimensionalidade não foi aplicada. Nas outras, a rede neural se destacou.

Tabela 4: 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

Na Figura 3 são representas as acurácia dos algoritmos aplicados em cada base gerada.

Os gráficos do lado esquerdo são das bases onde a discretização da variável classe foi feita utilizando *cut*, e as do lado direito utilizando *qcut*. 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 causa da perda de informação que acontece ao diminuir o número de variáveis.

Na Base 5, na qual os algoritmos isolados obtiveram melhor desempenho, foram aplicados os comitês homogêneos 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.

Na Tabela 5 são sumarizados os resultados dos experimentos na Base 5.

Tabela 5: Acurácia e desvio padrão na Base 5 dos comitês homogêneos de tamanhos 5, 10 e 20.

T	**SVC**	**NB**	**KNN**	**MLP**	**DT**
5	0,674 ± 0,053	0,639 ± 0,107	0,683 ± 0,042	0,760 ± 0,020	0,831 ± 0,026
10	0,675 ± 0,052	0,636 ± 0,112	0,684 ± 0,039	0,766 ± 0,019	0,838 ± 0,026
20	0,676 ± 0,052	0,632 ± 0,113	0,684 ± 0,039	0,763 ± 0,020	0,846 ± 0,022

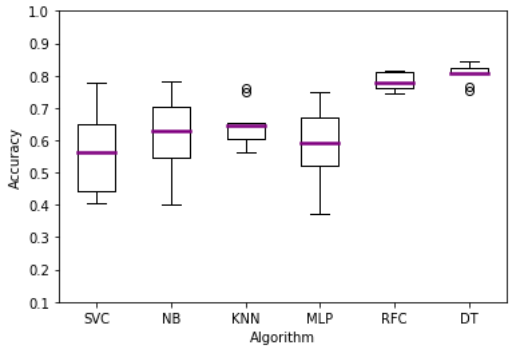
Para compor os comitês heterogêneos foram usados os três algoritmos com melhor desempenho na aplicação dos comitês homogêneos (KNN, MLP e DT). Esses algoritmos foram combinados para compor os seguintes comitês heterogêneos: KNN e MLP; KNN e DT; MLP e DT; KNN, MLP e DT. Para tal, aplicou-se o *Stacking*, 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]. Na Tabela 6 são detalhados os resultados obtidos para cada um desses comitês (nos tamanhos 1, 5, 10 e 20) com cada meta-classificador usado.

Tabela 6: Acurácia e desvio padrão na Base 5 dos comitês heterogêneos para cada meta-classificador.

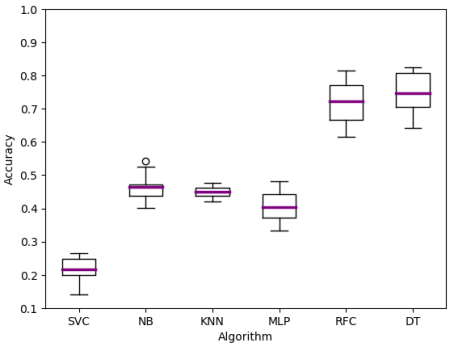
Meta-classificador	**T**	**KNN e MLP**	**KNN e DT**	**MLP e DT**	**KNN, MLP e DT**
MLP	1	0,540 ± 0,081	0,749 ± 0,088	0,749 ± 0,088	0,749 ± 0,088
	5	0,555 ± 0,088	0,749 ± 0,088	0,749 ± 0,088	0,749 ± 0,088
	10	0,609 ± 0,069	0,749 ± 0,088	0,749 ± 0,088	0,749 ± 0,088
	20	0,617 ± 0,087	0,749 ± 0,088	0,749 ± 0,088	0,749 ± 0,088
DT	1	0,554 ± 0,096	0,749 ± 0,088	0,749 ± 0,088	0,749 ± 0,088
	5	0,566 ± 0,075	0,749 ± 0,088	0,749 ± 0,088	0,749 ± 0,088
	10	0,611 ± 0,072	0,749 ± 0,088	0,749 ± 0,088	0,749 ± 0,088
	20	0,671 ± 0,053	0,749 ± 0,088	0,749 ± 0,088	0,749 ± 0,088

Nota-se que a aplicação dos comitês heterogêneos não foi tão proveitosa quanto a dos homogêneos. 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 ≈75% no restante dos comitês foi inferior aos resultados obtidos

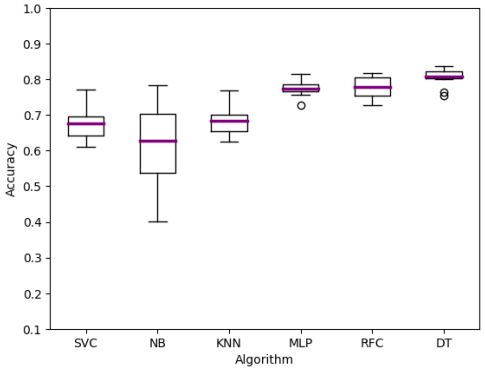
[4] `rasbt.github.io/mlxtend/user_guide/classifier/StackingClassifier/`



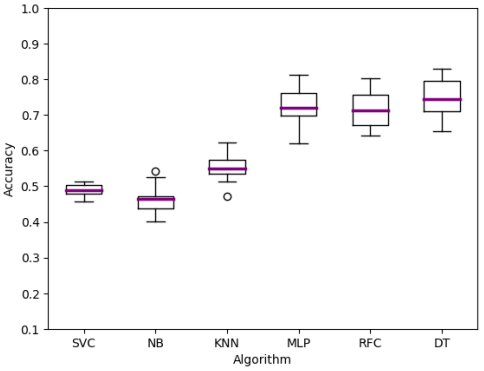
(a) Base 3 (*cut*).



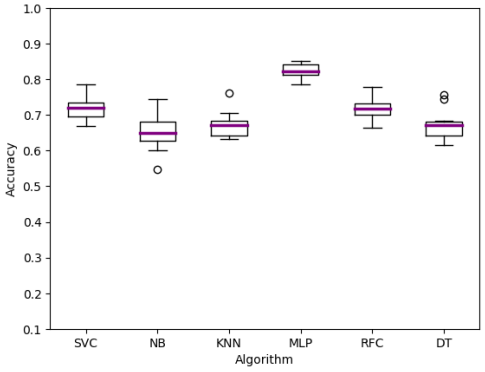
(b) Base 4 (*qcut*).



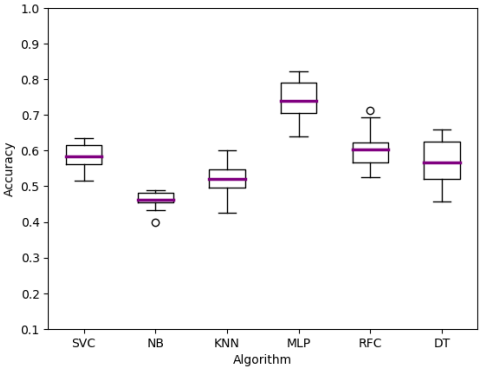
(c) Base 5 (*cut* e normalização).



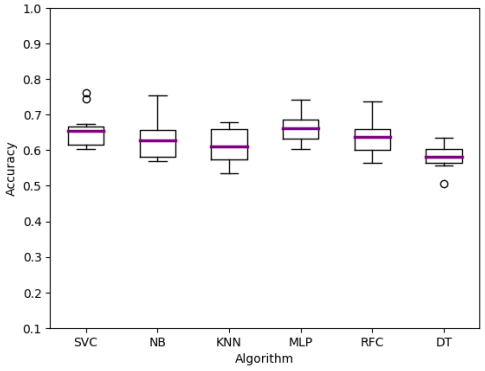
(d) Base 6 (*qcut* e normalização).



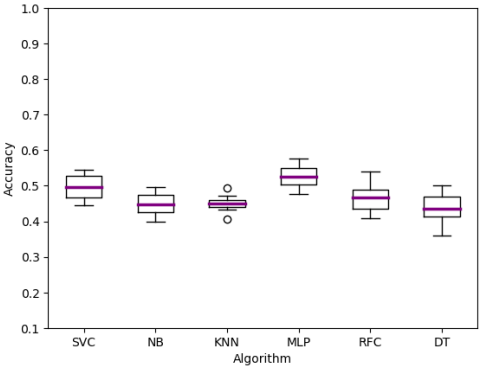
(e) Base 7 (*cut*, normalização e PCA 25%).



(f) Base 9 (*qcut*, normalização e PCA 25%).



(g) Base 8 (*cut*, normalização e PCA 50%).



(h) Base 10 (*qcut*, normalização e PCA 50%).

Figura 3: Acurácia dos algoritmos supervisionados aplicados em cada base.

anteriormente, portanto, a utilização dessa técnica nessa situação se mostrou ineficaz, uma vez que a acurácia não melhorou após o agrupamento dos comitês heterogêneos (Figura 4).

Na Figura 4 é mostrada a variação de desempenho das abordagens utilizadas para geração dos modelos preditivos na Base 5. Por outro lado, utilizando um comitê homogêneo de Árvores de Decisão de tamanho 20, obteve-se a acurácia de $\approx 85\%$, o melhor resultado encontrado com os experimentos. Esse resultado evidencia que é possível gerar modelos preditivos eficientes para predizer a quantidade de aluguéis usando informações de contexto em sistemas de bicicletas compartilhadas. O algoritmo de Árvore de Decisão foi eficiente, tanto individualmente quanto nos comitês homogêneos. Então, o modelo preditivo construído a partir desse algoritmo representa uma ferramenta que pode ser utilizada na gestão estratégica para auxiliar na tomada de decisão nos sistemas de bicicletas compartilhadas, principalmente, se utilizada pelos gestores para estimar a quantidade de aluguéis em determinada hora/dia.

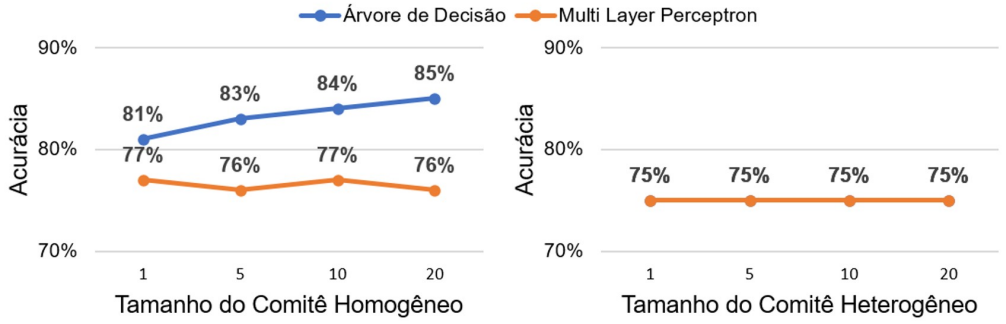


Figura 4: Variação dos resultados de cada abordagem na Base 5.

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.

5. Conclusões

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 fornecemos aos gestores dessa modalidade de sistemas informações precisas acerca do serviço de aluguel de bicicleta para auxiliá-los no entendimento do hábito dos usuários. Assim, o uso dos modelos preditivos apresentados auxiliam na tomada de decisão desses sistemas. Na classificação da quantidade de aluguéis, a melhor acurácia obtida foi de $\approx 85\%$, que foi alcançada com a aplicação de um comitê homogêneo de Árvores de Decisão. A aplicação dos comitês heterogêneos não se mostrou efetiva, mesmo usando a Árvore de Decisão como meta-classificador. Como trabalhos futuros pretende-se aplicar nessas mesmas bases alguns algoritmos que utilizam a abordagem de lógica difusa na classificação e comparar o seu desempenho com os resultados dos algoritmos apresentados neste trabalho.

Referências

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.

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.

Hamilton, T. L. e Wichman, C. J. (2018). Bicycle infrastructure and traffic congestion: Evidence from dc's capital bikeshare. *Journal of Environmental Economics and Management*, 87:72 – 93. ISSN 0095-0696. URL http://www.sciencedirect.com/science/article/pii/S0095069616300420.

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.

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.

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 Machine 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.

Souza, L. C. e Gomes, E. T. A. (2014). O uso da bicicleta como meio de transporte: Mobilidade urbana na cidade do recife. In *Anais do I Congresso Brasileiro de Geografia Política, Geopolítica e Gestão do Território*, p. 384–395. Letra1. URL http://www.editoraletra1.com/anais-congeo/arquivos/978-85-63800-17-6-p384-395.pdf.

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., 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.

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.