UNIVERSIDADE FEDERAL RURAL DO SEMI-ÁRIDO
UNIVERSIDADE DO ESTADO DO RIO GRANDE DO NORTE
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Uma Abordagem Para Predição de Demanda em Sistemas de Bicicletas Compartilhadas

MESTRANDO

Johnattan Douglas Ferreira Viana

ORIENTADOR

Dr. Lenardo Chaves e Silva

COORIENTADOR

Dr. Álvaro Alvares de Carvalho César Sobrinho

Mossoró-RN

2021

Johnattan Douglas Ferreira Viana

Uma Abordagem Para Predição de Demanda em Sistemas de Bicicletas Compartilhadas

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação - associação ampla entre a Universidade do Estado do Rio Grande do Norte e a Universidade Federal Rural do Semi-Árido, para a obtenção do título de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Lenardo Chaves e Silva
Coorientador: Prof. Dr. Álvaro Alvares de Carvalho César Sobrinho

Mossoró-RN

2021

© Todos os direitos estão reservados a Universidade do Estado do Rio Grande do Norte. O conteúdo desta obra é de inteira responsabilidade do(a) autor(a), sendo o mesmo, passível de sanções administrativas ou penais, caso sejam infringidas as leis que regulamentam a Propriedade Intelectual, respectivamente, Patentes: Lei n° 9.279/1996 e Direitos Autorais: Lei n° 9.610/1998. A mesma poderá servir de base literária para novas pesquisas, desde que a obra e seu(a) respectivo(a) autor(a) sejam devidamente citados e mencionados os seus créditos bibliográficos.

Catalogação da Publicação na Fonte.
Universidade do Estado do Rio Grande do Norte.

V614a Viana, Johnattan Douglas Ferreira
Uma Abordagem Para Predição de Demanda em Sistemas de Bicicletas Compartilhadas. / Johnattan Douglas Ferreira Viana. - Mossoró-RN, 2021.
114p.

Orientador(a): Prof. Dr. Lenardo Chaves e Silva.
Coorientador(a): Prof. Dr. Álvaro Alvares de Carvalho César Sobrinho.
Dissertação (Mestrado em Programa de Pós-Graduação em Ciência da Computação). Universidade do Estado do Rio Grande do Norte.

1. Sistemas de Bicicletas Compartilhadas. 2. BSRP. 3. Predição de Demandas. 4. Tomada de Decisão. 5. Gestão Estratégica. I. Silva, Lenardo Chaves e. II. Universidade do Estado do Rio Grande do Norte. III. Título.

O serviço de Geração Automática de Ficha Catalográfica para Trabalhos de Conclusão de Curso (TCC´s) foi desenvolvido pela Diretoria de Informatização (DINF), sob orientação dos bibliotecários do SIB-UERN, para ser adaptado às necessidades da comunidade acadêmica UERN.

Johnattan Douglas Ferreira Viana

Uma Abordagem Para Predição de Demanda em Sistemas de Bicicletas Compartilhadas

Dissertação apresentada ao programa de Pós-Graduação em Ciência da Computação (PPgCC) UERN/UFERSA para a obtenção do título de Mestre em Ciência da Computação.

APROVADA EM: 16 / 06 / 2021 .

Prof. Dr. Lenardo Chaves e Silva
Orientador e Presidente

Prof. Dr. Álvaro Alvares de Carvalho César Sobrinho
Coorientador - Universidade Federal do Agreste de Pernambuco - UFAPE

Prof. Dr. Francisco Milton Mendes Neto
Membro Interno - Universidade Federal Rural do Semi-Árido - UFERSA

Prof. Dr. Sebastião Emídio Alves Filho
Membro Interno - Universidade do Estado do Rio Grande do Norte - UERN

Profa. Dra. Carina Teixeira de Oliveira
Membro Externo - Instituto Federal de Educação, Ciência e Tecnologia do Ceará - IFCE

Aos meus pais.

Agradecimentos

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela disponibilização de bolsa nos meus primeiros meses no PPgCC (UERN/UFERSA).

À coordenação do PPgCC (UERN/UFERSA), pelo suporte acadêmico.

Ao prof. Lenardo Chaves e prof. Álvaro Sobrinho, pela orientação e críticas construtivas, que possibilitaram que esta pesquisa se tornasse o que é hoje.

Ao prof. Sebastião Filho, prof. Francisco Milton e profa. Carina de Oliveira, pela participação na banca examinadora.

À minha família. Em especial, aos meus pais, pela paciência e amor incondicional. A força que eles me ensinaram a ter foi essencial para superar os desafios deste período.

À minha psicóloga, Georgia Rebouças, pela ajuda em fazer eu me reconhecer como "capitão da minha alma e senhor do meu destino", fundamental para todas as escolhas acertadas que tomei nesse período pandêmico, tão incerto e desafiador.

Aos meus amigos de profissão, pelas palavras de motivação (que não foram poucas) de Adriana Wandermurem, Edwiges Barbosa, Graziele do Vale, Jhonatan Lisboa, Kátia Santana e Tainan Santana. Camila Santana e Hyago Paiva, especialmente, foram essenciais na revisão dos textos em inglês.

Aos amigos que fiz durante o mestrado, pelo apoio emocional e amizade preciosa. Caio Mateus, Cynthia Moreira, Daniela Sena, Elizeu Sandro, Geo Câmara, Ially Almeida, Israel Dias (meu irmão de orientação), Júlio Cartier, Lohana Aires, Otília Santos, Othon Braga, Antonio Queiroz, Rayana Souza, Rigoberto Fernandes, Thaiza Medeiros, Thalia Gurgel (minha dupla), Thiago Sobreira e Virgínia Maia, não teria conseguido sem vocês!

Às amizades que construí antes do mestrado, como professores da graduação, amigos do LAR (Laboratório de Redes de Computadores e Sistemas, do IFCE Aracati) e de longas datas. Dentre elas, Renan Braga, Eliezio Neto, Erivania Barbosa, Jaciana Barbosa, Igor Souza, Priscila Nogueira, Suelen Silva e Ylanna Fonseca, que me motivaram a alcançar esse sonho, hoje realizado.

Por fim, ao Universo, por ser como Ele é e por me permitir estar e ser também. Aquilo que é, já existia, e aquilo que há de ser, já existiu.

"Caminho se conhece andando,
então de vez em quando é bom se perder.
Perdido fica perguntando,
vai só procurando e acha sem saber."
Chico César

Resumo

Os Sistemas de Bicicletas Compartilhadas (SBC) têm contribuído para a melhoria da mobilidade urbana e se tornaram cada vez mais comuns nas cidades. Embora esses sistemas forneçam um transporte econômico e sustentável, melhorando o trânsito nas cidades, para o seu funcionamento é necessário uma análise dos hábitos dos usuários para que as empresas fornecedoras desse serviço possam suprir a demanda de estações com alta procura e evitar o desperdício de recursos em estações pouco utilizadas, coletando bicicletas em estações superlotadas e redistribuí-las em estações que estão com baixo estoque ou mesmo vazias. Esse problema de rebalanceamento de bicicletas (do inglês *Bike Sharing Rebalancing Problem* - BSRP) requer estratégias eficientes de redistribuição, que dependem da modelagem e previsão de demanda de aluguéis, evidenciando a relevância de abordagens que auxiliem na tomada de decisão e gestão estratégica das empresas responsáveis por esses sistemas. Neste trabalho é analisado um conjunto de dados do Capital BikeShare (Washington, D.C.), explorando a relação de condições meteorológicas e sazonais nesse serviço. Modelos de regressão e classificação foram elaborados usando algoritmos clássicos da literatura como *Linear Regression* (LR), *Support Vector Machine* (SVM), *Gaussian Naive Bayes* (NB), *K-Nearest Neighbors* (KNN), *MultiLayer Perceptron* (MLP), *Random Forest* (RF) e *Decision Tree* (DT). O objetivo é predizer a quantidade de aluguéis, fornecendo um parâmetro para auxiliar na redistribuição eficiente de bicicletas nos SBC. A utilização de múltiplos algoritmos também foi experimentada para a construção de comitês homogêneos e heterogêneos. Os desempenhos dos modelos preditivos gerados foram analisados usando validação cruzada de 10 partes. Os testes estatísticos de Friedman e Nemenyi foram aplicados. Além disso, foram discutidas modelos descritivos utilizando *k-means* para agrupamento das estações de aluguéis, evidenciando tais agrupamentos como uma técnica útil na melhoria do desempenho dos modelos preditivos. O melhor modelo de regressão gerado utilizou o RF e alcançou coeficiente de determinação de 0,963, enquanto que o melhor modelo de classificação elaborado foi o DT, com acurácia de 96,21%. Um cenário de aplicação e uso dos potenciais modelos preditivos foi descrito, de forma a facilitar a interpretação dos resultados da predição para os gestores desses serviços e melhorar o processo de tomada de decisão nos SBC.

Palavras-chave: Sistemas de Bicicletas Compartilhadas, BSRP, Predição de Demandas, Tomada de Decisão, Gestão Estratégica.

Abstract

Shared Bicycle Systems (SBC) have contributed to the improvement of urban mobility and have become increasingly common in cities. However, for the proper functioning of these systems, it is necessary to analyze the users' habits so that it is possible to supply the demand for stations with high demand and avoid wasting resources in underused stations, collecting bicycles in overcrowded stations and redistributing them in stations that are low on stock or even empty. The rebalancing bicycles process (*Bike Sharing Rebalancing Problem* - BSRP) requires efficient redistribution strategies, which depends on the modeling and forecasting of rental demand, evidencing the relevance of approaches that assist in decision-making and strategic management of the companies responsible for these systems. In this work, a data set from Capital BikeShare (Washington, D.C.) is analyzed, exploring the relationship of meteorological and seasonal conditions in this service. Regression and classification models were developed using classic literature algorithms as *Linear Regression* (LR), *Support Vector Machine* (SVM), Gaussian Naive Bayes (NB), *K-Nearest Neighbors* (KNN), *MultiLayer Perceptron* (MLP), *Random Forest* (RF) and *Decision Tree* (DT). The objective is to predict the number of rentals, providing a parameter to assist in the efficient redistribution of bicycles in SBC. The use of multiple algorithms has also been tried to build homogeneous and heterogeneous committees. The performances of the generated predictive models were analyzed using 10-part cross-validation. Friedman and Nemenyi statistical tests also were applied. Moreove, descriptive models using *k-means* to group rental stations were also discussed, highlighting such groupings as a useful technique in improving the performance of predictive models. The best regression model was generated with RF and has determination coefficient of 0.963, while the best classification model was developed generated with DT and has an accuracy of 96.21%. An application scenario for potential predictive models has been described to facilitate the interpretation of the prediction results for the managers of these services and to improve the decision-making process in SBC.

Keywords: Shared Bicycle Systems, BSRP, Demand Prediction, Decision Making, Strategic Management.

Lista de Ilustrações

Figura 1 – Processo de Descoberta do Conhecimento. Fonte: Düsing (2000). 23
Figura 2 – Representação da metodologia da pesquisa. 43
Figura 3 – Quantitativo de aluguéis por tipo de usuário. 52
Figura 4 – Quantitativo de aluguéis por dia do mês. 53
Figura 5 – Quantitativo de aluguéis por hora. 54
Figura 6 – Relação do clima com o quantitativo de aluguéis em cada estação do ano. 55
Figura 7 – Quantidade de aluguéis em cada feriado. 56
Figura 8 – Relação entre distância e duração do aluguel. 56
Figura 9 – Matriz de correlação (Base 6). 58
Figura 10 – Relevância dos atributos nos modelos preditivos gerados pelo algoritmo RFR. 59
Figura 11 – Acurácia dos classificadores aplicados nas Bases 3 a 10. 62
Figura 12 – Divisão em quadrantes e limite territorial de Washington D.C. 67
Figura 13 – Aplicação do método cotovelo. 69
Figura 14 – Agrupamento das estações por proximidade com 8 centróides (Base 12). 69
Figura 15 – Agrupamento das estações por proximidade com 13 centróides (Base 13). 70
Figura 16 – Agrupamento das estações por proximidade com 20 centróides (Base 14). 70
Figura 17 – Gráfico Quantil-Quantil da distribuição normal do RFR isolado na base 13. 71
Figura 18 – Relação entre valores reais e previstos pelo RFR em cada base com estações agrupadas. 73
Figura 19 – Aplicação dos algoritmos de classificação nas bases com agrupamentos. 74
Figura 20 – Abordagem para utilização dos modelos preditivos gerados. 79
Figura 21 – Proposta de integração dos modelos preditivos por meio de um aplicativo móvel. 81
Figura 22 – Visualização em mapa de calor com ênfase em um agrupamento. 82
Figura 23 – Aplicativo do *Capital BikeShare* com informações em tempo real. . . . 83

Lista de Tabelas

Tabela 1 – Resumo estatístico das variáveis climáticas na base de dados. 45
Tabela 2 – Partições usadas na categorização da variável classe (*qtd*). 47
Tabela 3 – Comparação de desempenho entre regressores. 59
Tabela 4 – Comparação de desempenho entre classificadores nas Bases 3 a 10. . . 61
Tabela 5 – Comparação de desempenho de comitês homogêneos na Base 5. 64
Tabela 6 – Acurácia e desvio padrão na Base 5 dos comitês heterogêneos para cada meta-classificador. 66
Tabela 7 – Desempenho dos regressores nas bases de dados com estações agrupadas: (a) individualmente e (b) em comitês. 72
Tabela 8 – Comparação de desempenho entre classificadores nas Bases 11 a 14. . . 75
Tabela 9 – Comparação de desempenho de comitês homogêneos na Base 14. 76
Tabela 10 – Acurácia e desvio padrão na Base 14 dos comitês heterogêneos para cada meta-classificador*. 77

Lista de Quadros

Quadro 1 – Representação de uma matriz de confusão. 27
Quadro 2 – Fichamento dos trabalhos relacionados. 36
Quadro 3 – Listagem dos atributos presentes na base de dados enriquecida com os registros agrupados por hora. 45
Quadro 4 – Procedimentos aplicados em cada base gerada. 48
Quadro 5 – Parâmetros utilizados nos modelos de regressão. 49
Quadro 6 – Parâmetros utilizados nos modelos de classificação. 50
Quadro 7 – Detalhamento de aluguéis em cada região geográfica (quadrante) na Base 11. 68

Lista de Abreviaturas e Siglas

API	*Application Programming Interface*
AdaTC	*Adaptive Transition Constraint*
AUC	*Area Under the Curve*
BIC	*Bayesian Information Criterion*
BSRP	*Bike Sharing Rebalancing Problem*
BT	*Boosted Trees*
CD	*Critical Difference*
CV	*Cross Validation*
DCHR	*Department of Human Resources*
DP	Desvio Padrão
DT	*Decision Tree*
FN	Falsos Negativos
FP	Falsos Positivos
FPR	*False Positive Rate*
GBFS	*General Bikeshare Feed Specification*
GBM	*Gradient Boosting Machine*
GBRT	*Gradient Boosting Regression Tree*
IMSE	*Increased Mean Square Error*
INP	*Increased Node Purity*
ISBT	*Inter Station Bike Transition*
JSON	*JavaScript Object Notation*
KDD	*Knowledge Discovery from Databases*
KNN	*K-Nearest Neighbors*

LR	*Linear Regression*
MAE	*Mean Absolute Error*
MD	Mineração de Dados
MLP	*Multi-Layer Perceptron*
MSWK	*Meteorological Similarity Weighted K-nearest neighbor*
NB	*Naive Bayes*
NBRM	*Negative Binomial Regression Model*
NE	Nordeste
NPS	*National Park Service*
NW	Noroeste
PCA	*Principal Component Analysis*
PeMS	*Caltrans Performance Measurement System*
PRM	*Poisson Regression Model*
Q-Q	Quantil-Quantil
R^2	R-quadrado
RF	*Random Forest*
RFC	*Random Forest Classifier*
RFR	*Random Forest Regressor*
RMSE	*Root Mean Square Error*
ROC	*Receiver Operating Characteristic*
SAE	*Stacked Autoencoder*
SBC	Sistemas de Bicicletas Compartilhadas
SE	Sudeste
SGPR	*Gaussian Process Regressor*
SOM	*Self-Organizing Mapping*
SOM-RT	*Self-Organizing Mapping network with a Regression Tree*

SSE	*Sum Squared Error*
STI	Sistemas de Transporte Inteligente
SVC	*Support Vector Classifier*
SVM	*Support Vector Machine*
SVR	*Support Vector Regressor*
SW	Sudoeste
TICs	Tecnologias de Informação e Comunicação
TPR	*True Positive Rate*
URL	*Uniform Resource Locator*
USGS	*United States Geological Survey*
VN	Verdadeiros Negativos
VP	Verdadeiros Positivos
WCN-MC	*Weighted Correlation Network and Monte Carlo simulation*
XGBoost	*Extreme Gradient Boosting*
XGBTree	*Extreme Gradient Boosting Tree*

Sumário

1 INTRODUÇÃO . . . 17
1.1 Problemática . . . 18
1.2 Justificativa . . . 20
1.3 Objetivos . . . 20
1.4 Produção Científica . . . 21
1.5 Organização do Documento . . . 22

2 REFERENCIAL TEÓRICO . . . 23
2.1 Mineração de Dados . . . 23
2.2 Aprendizagem Supervisionada . . . 24
2.2.1 Métricas de Avaliação para Regressão . . . 26
2.2.2 Métricas de Avaliação para Classificação . . . 27
2.2.3 Algoritmos . . . 29
2.2.4 Comitês Homogêneos e Heterogêneos . . . 32
2.3 Aprendizagem Não-supervisionada . . . 33
2.4 Considerações Finais . . . 34

3 TRABALHOS RELACIONADOS . . . 35
3.1 Previsão por Cidade . . . 37
3.2 Previsão por Regiões . . . 38
3.3 Previsão por Estação . . . 39
3.4 Considerações Finais . . . 41

4 METODOLOGIA . . . 43
4.1 Base de Dados . . . 43
4.2 Pré-processamento . . . 46
4.3 Novas Bases Geradas . . . 48
4.4 Análises em Nível de Cidade e Regiões . . . 48
4.5 Modelos Preditivos . . . 49
4.6 Considerações Finais . . . 51

5 ANÁLISE EM NÍVEL DE CIDADE . . . 52
5.1 Identificação de Padrões . . . 52
5.2 Análise de Correlação . . . 57
5.3 Análise Comparativa de Modelos de Regressão . . . 58
5.4 Análise Comparativa de Modelos de Classificação . . . 60

5.4.1 Aplicação de Comitês 63
5.5 Considerações Finais 65

6 ANÁLISE EM NÍVEL DE REGIÕES 67
6.1 Agrupamento por Quadrante 67
6.2 Agrupamento por Proximidade 68
6.3 Aplicação dos Algoritmos 71
6.3.1 Regressão 71
6.3.2 Classificação 74
6.4 Considerações Finais 78

7 CENÁRIO DE APLICAÇÃO DOS MODELOS PREDITIVOS 79
7.1 Entradas 79
7.2 Análises 80
7.3 Saídas 81
7.4 Dados em Tempo Real 82
7.5 Considerações Finais 84

8 CONCLUSÕES 85

REFERÊNCIAS 87

APÊNDICE A – CORRELAÇÃO NAS BASES SEM REDUÇÃO DE DIMENSIONALIDADE (BASES 1 A 6) 94

APÊNDICE B – CORRELAÇÃO NAS BASES COM REDUÇÃO DE DIMENSIONALIDADE (BASES 7 A 10) . . . 97

APÊNDICE C – RESULTADOS DOS TESTES ESTATÍSTICOS . . 99

APÊNDICE D – CURVA ROC E AUC DOS CLASSIFICADORES INDIVIDUAIS APLICADOS NAS BASES DE DADOS 102

1 Introdução

Os Sistemas de Transporte Compartilhado (STC) promovem um transporte sustentável compartilhando por demanda (FRIEDMAN, 2021), aplicado com diversos tipos de veículos, como carros, táxis, bicicletas e até patinetes. Os STC são apresentados como um complemento aos serviços de transporte público tradicional como ônibus e metrô (JAPPINEN; TOIVONEN; SALONEN, 2013).

Um dos STC mais comuns são aqueles voltados para as bicicletas: os Sistemas de Bicicletas Compartilhadas (SBC). Em 2021, mais de dois mil SBC operam em diversas cidades no mundo, com mais de 9 milhões de bicicletas compartilhadas sendo utilizadas nesses sistemas (O'BRIEN *et al.*, 2021). Nesse sentido, a utilização das bicicletas como meio de transporte tem reduzido o congestionamento do trânsito (HAMILTON; WICHMAN, 2018), além de impactar positivamente na saúde da população e no meio ambiente (SOUZA; GOMES, 2014).

Os SBC são uma evolução dos sistemas tradicionais de aluguel de bicicletas, e tem como diferencial a automaticidade no processo de registro, empréstimo e devolução das bicicletas. Esses sistemas evoluíram rapidamente e já estão em sua quarta geração (MÁTRAI; TÓTH, 2016).

Esses sistemas são compostos por diversas estações de aluguéis espalhadas por cidades que possuem um número limitado de bicicletas estacionadas. Nessas estações, os usuários alugam uma bicicleta por um determinado período, retirando-a da estação e devolvendo-a na estação destino, que pode, inclusive, ser a mesma estação de origem. A praticidade dos sistemas incentiva novos ciclistas, uma vez que dispensa manutenção e preocupações com estacionamento, ajudando também a minimizar o uso de carros em viagens mais curtas (FISHMAN; WASHINGTON; HAWORTH, 2013). Além dos benefícios práticos dos SBC, os dados gerados por esses sistemas os tornam atrativos para pesquisas, uma vez que são registradas variáveis como hora de início e fim do aluguel e localizações de início e fim do trajeto. Ao contrário de outros serviços de transporte, os dados gerados por SBC viabilizam análises da mobilidade urbana, uma vez que são coletados dados de origem e destino de um único usuário. Desta forma é possível identificar padrões analisando e monitorando os dados coletados por cada usuário.

Essas análises só fazem sentido se estiverem voltadas para um gerenciamento estratégico desses sistemas. A gestão estratégica pode ser compreendida como a combinação de dois conceitos: gestão e estratégia. A gestão é definida por Chiavenato (2014) como o processo de planejar, organizar, liderar e controlar o uso de recursos visando alcançar os objetivos de uma organização. A estratégia é definida 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 aos objetivos financeiros, produtivos, mercadológicos e competitivos. Nos SBC, 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. No contexto estratégico, a Mineração de Dados (MD) permite aprimorar os serviços de compartilhamento de bicicletas, explorando dados de várias fontes (LIU *et al.*, 2016).

Perante o exposto, como pode-se explorar as informações desses sistemas para entender os hábitos dos usuários? Como os gestores desses sistemas podem analisar adequadamente os dados dos SBC públicos e privados, a fim de tomar melhores decisões no contexto estratégico e operacional dos sistemas? A correta análise desses dados e a aplicação da abordagem apresentada nesta pesquisa podem ajudar a responder a essas questões e auxiliar na tomada de decisões.

1.1 Problemática

Diversos problemas ainda precisam ser solucionados para efetivar e aumentar o uso dos SBC, dentre eles: empréstimos sem devolução; rebalanceamento de bicicletas (*Bike Sharing Rebalancing Problem* - BSRP); e diferentes frequências de uso das bicicletas, já que tanto o uso frequente quando a ociosidade podem afetar a vida útil das bicicletas e aumentar o custo de manutenção. Como a demanda que será necessária é incerta e se altera frequentemente, as ações de rebalanceamento são as mais desafiadores nos SBC (XU *et al.*, 2020). Em alguns momentos do dia, o estoque de bicicletas em determinadas estações se esgota, enquanto que em outras estações o estoque não é totalmente utilizado.

Como resultado, os fornecedores do serviço de aluguel de bicicletas devem distribuir as bicicletas de maneira eficaz e eficiente entre as diferentes estações (ALVAREZ-VALDES *et al.*, 2016; GHOSH *et al.*, 2017). Isto porque prever o número necessário de bicicletas compartilhadas em cada estação não é uma tarefa simples, uma vez que a demanda de bicicletas em cada estação de aluguel se altera ao decorrer do dia dependendo de fatores climáticos ou sazonais (feriados, acidentes de trânsitos, eventos que estejam acontecendo nas proximidades, etc). Portanto, algumas estações têm as bicicletas todas alugadas, enquanto existem bicicletas em outras estações que não foram utilizadas. Além disso, a experiência do usuário pode ser prejudicada quando a estação de origem está vazia, o que obriga o usuário a recorrer a outro meio de transporte ou a tentar encontrar uma bicicleta disponível em outra estação. Da mesma forma, se a estação de destino estiver cheia, o usuário deve esperar até que uma bicicleta seja retirada ou devolvê-la em outra

estação com pelo menos uma vaga de estacionamento disponível (GAST *et al.*, 2015). Soluções para esse problema somente baseadas no monitoramento em tempo real podem não ser eficientes, já que leva tempo para redistribuir as bicicletas quando ocorre um desequilíbrio (ANSERE; HAN; WANG, 2019).

Para garantir o bom funcionamento de SBC é essencial redistribuir as bicicletas, o que inclui coletar as bicicletas devolvidas e realocá-las, geralmente quando as estações estão fechadas. Esses sistemas dependem da eficácia das operações de rebalanceamento, ou seja, os esforços de repor o número de bicicletas em determinada estação de acordo com sua demanda. Segundo Liu *et al.* (2016), duas questões precisam ser tratadas para solucionar este problema: a predição da demanda das estações e a otimização do roteamento dos veículos para realizar o balanceamento nas estações discrepantes. A predição depende da previsão das demandas de aluguel e devolução em cada estação, que pode ser afetado por vários fatores, como o tempo, a localização, as condições meteorológicas e as situações de trânsito. Já o problema de roteamento para realizar o rebalanceamento é agravado com as demandas discrepantes entre as estações, tornando as técnicas tradicionais de otimização de rotas inviáveis para sistemas com centenas de estações (REN *et al.*, 2020; DELL'AMICO *et al.*, 2014).

Prever precisamente a demanda de aluguel não é uma tarefa trivial, já que o fluxo de aluguel pode ser não-linear e incerto. Os resultados da previsão são dependentes de dados históricos, e além disso, podem ser influenciados por fatores meteorológicos e temporais (XU *et al.*, 2020).

Segundo Ren *et al.* (2020), o problema de rebalanceamento de bicicletas (*Bike Sharing Rebalancing Problem* - BSRP) é formulado como um problema de otimização dinâmico ou estático. Para o caso do problema estático, a demanda é prevista para todas as estações. Assim, o rebalanceamento pode ser feito de modo que um determinado número de bicicletas sejam retiradas e entregues em determinadas estações e que todas as estações tenham o número ideal de bicicletas disponíveis para aluguel com o número ideal de vagas disponíveis para devolução. Para o problema dinâmico, a operação de rebalanceamento é gerenciada com base nas informações em tempo real. Normalmente, o processo de rebalanceamento estático é implementado durante a noite, quando o sistema é relativamente estável devido ao baixo uso de bicicletas por clientes do serviço. O rebalanceamento estático é mais fácil de gerenciar e, por isso, é mais comumente utilizado nos SBC atuais (DELL'AMICO *et al.*, 2014).

Por outro lado, os processos de rebalanceamento dinâmico realizam a redistribuição das bicicletas com mais frequência durante o dia. Neste trabalho o foco está na implementação e avaliação de modelos preditivos capazes de auxiliar nas ações de rebalanceamento dinâmico dos SBC.

1.2 Justificativa

As informações coletadas pelos SBC permitiram o surgimento de várias aplicações que podem fazer a integração de serviços, análise de dados e tomada de decisão em tempo real. Assim, os dados gerados por esses sistemas se destacam como um recurso que pode ser utilizado para estudar a mobilidade urbana. Aliada a isso, a MD 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 mobilidade que acontece nesses sistemas pode levar a decisões mais apuradas.

A previsão de demanda nos SBC é fundamental para gerenciar ações de distribuição de um número adequado de bicicletas nas estações, amenizando o problema de desbalanceamento e melhorando a qualidade do serviço e a experiência do usuário. Assim, são necessárias abordagens para que os provedores de serviços de bicicletas compartilhadas redistribuam as bicicletas entre as estações garantindo o funcionamento efetivo do sistema.

Uma maneira promissora de melhorar a previsão da demanda de aluguéis é fazer uso de uma série de dados que estão direta ou indiretamente relacionados à esse serviço (LIU *et al.*, 2016). As informações de contexto permitem uma melhoria nos serviços de compartilhamento de bicicletas, possibilitando o desenvolvimento de soluções de redistribuição de bicicletas fazendo uso inteligente (e automatizado) dos dados.

Nesta pesquisa, explora-se dados de várias fontes relacionadas ao serviço de compartilhamento de bicicletas, como registros das viagens, origem e destino do usuário, bem como dados de contexto (dados climáticos e informações sazonais). Além disso, são aplicadas diferentes técnicas de aprendizagem de máquina, bem como avaliados o desempenho de algoritmos clássicos, para a geração de modelos preditivos voltados tanto para a classificação como para a regressão.

A abordagem proposta é genérica e pode ser aplicada nos outros tipos de STC com o intuito de extrair e descrever padrões de mobilidade. Dessa forma, é possível implementar avanços referentes à interpretação dos registros nessa modalidade de sistema de transporte já que a predição da demanda auxilia os gestores desses serviços na tomada de decisão quanto a tarefa de rebalanceamento.

1.3 Objetivos

O objetivo geral com esta pesquisa é propor uma abordagem baseada em predição de demanda de aluguéis para o auxílio a tomada de decisão em SBC.

Os objetivos específicos são os seguintes:

- identificar atributos da base de dados que influenciam na quantidade de bicicletas alugadas;
- enriquecer com informações de contexto a base de dados do Capital BikeShare, um dos SBC de Washington D.C.;
- realizar uma análise comparativa de modelos preditivos para a demanda de aluguéis de SBC em nível de cidade e regiões;
- aplicar modelos descritivos para agrupar as estações de aluguéis e avaliar o impacto desta técnica nos modelos preditos gerados;
- realizar testes estatísticos para garantir a efetividade dos resultados alcançados;
- apresentar um cenário de aplicação dos modelos gerados; e
- disponibilizar *scripts* dos experimentos e bases de dados geradas para estudos posteriores.

1.4 Produção Científica

Evidenciando a relevância desta pesquisa, ao decorrer do seu desenvolvimento, três trabalhos foram publicados, a saber:

- Viana, J. D. F.; Braga, O. C.; Silva, L. C.; Mendes Neto, F. M. *Analyzing Patterns of a Bicycle Sharing System for Generating Rental Flow Predictive Models.* In: Workshop de Computação Urbana (COURB), 3, 2019, Gramado. Anais eletrônicos do COURB. Porto Alegre: Sociedade Brasileira de Computação, 2019. p. 57-70. ISSN 2595-2706. **(Menção honrosa)**
- Viana, J. D. F; Gurgel, T. K. S. G.; Silva, L. C; Alves Filho, S. E; Liberalino, C. H. P. 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), 2020, João Pessoa. Anais eletrônicos do SBPO. Campinas, Galoá, 2020.
- Viana, J. D. F; Gurgel, T. K. S. G.; Silva, L. C; Alves Filho, S. E; Liberalino, C. H. P; Sobrinho, A. A. C. C. Modelos de Aprendizagem de Máquina para a Gestão Estratégica de Bicicletas Compartilhadas. In: Brazilian Journal of Development, v. 7, n. 1, p.2209-2227, 2021. Curitiba. ISSN 2525-8761. DOI: 10.34117/bjdv7n1-151

1.5 Organização do Documento

Além da Introdução, apresentada neste capítulo, este documento está dividido em mais 7 capítulos. No Capítulo 2 é apresentada a fundamentação teórica, contendo as principais temáticas abordadas neste trabalho. No Capítulo 3 são discutidos os trabalhos relacionados. No Capítulo 4 é detalhada a metodologia aplicada no trabalho. Nos Capítulos 5 e 6 são discutidos e avaliados modelos preditivos de demanda de aluguel, respectivamente, em nível de cidade e regiões. Um cenário para aplicação dos modelos é descrito no Capítulo 7. No Capítulo 8 são apresentadas as conclusões alcançadas com esta pesquisa, suas limitações e são pontuados encaminhamentos para trabalhos futuros.

2 Referencial Teórico

Neste capítulo, são apresentados conceitos relacionados com o processo de Mineração de Dados (MD) (Seção 2.1). Além disso, são tratados conceitos referentes à aprendizagem supervisionada (Seção 2.2) e não-supervisionada (Seção 2.3), bem como algoritmos clássicos, tanto para a técnica de regressão quanto de classificação. Também são apresentadas algumas métricas de avaliação para os modelos oriundos de tais técnicas. Na Seção 2.4 são apresentadas as considerações finais deste capítulo.

2.1 Mineração de Dados

A MD agrupa contribuições de diversas áreas como estatística, otimização e inteligência artificial (LAYTON, 2017). Han, Pei e Kamber (2011), de maneira simples, definem a MD como "o processo de extrair conhecimento a partir de grandes quantidades de dados". Alguns autores tratam a MD como sinônimo de descoberta de KDD, enquanto outros se referem a ela apenas como uma etapa deste processo de descoberta do conhecimento.

Como mostrado na Figura 1, a seleção é a etapa na qual os dados relevantes para a análise são selecionados. No pré-processamento os dados são tratados de modo a aprimorar a confiabilidade da análise dos dados, aplicando técnicas de pré-processamento como limpeza (para tratar valores faltosos, ruídos, redundância, duplicação de registros e valores conflitantes), integração (junção de bases de dados diferentes), redução, transformação e discretização dos dados (HAN; PEI; KAMBER, 2011).

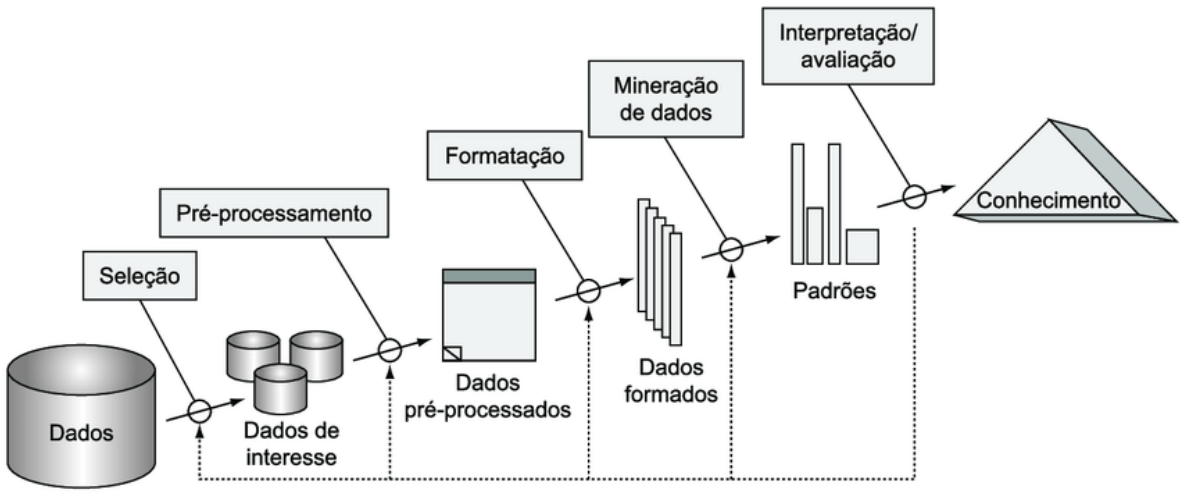


Figura 1 – Processo de Descoberta do Conhecimento. Fonte: Düsing (2000).

A etapa de formatação dos dados (também chamada por alguns autores de transformação) é realizada para facilitar a aplicação das técnicas de MD ajustando as variáveis (conversão entre medidas, alteração de escala, normalização, agregação, etc.). A MD é a principal etapa para extração do conhecimento, aplicando, para este fim, algoritmos que identificam padrões nos dados utilizados (LAYTON, 2017). Após essa etapa, os resultados alcançados são avaliados e interpretados, identificando os padrões relevantes. No fim do processo, é verificado se o conhecimento adquirido possui alguma validade para o contexto ao qual o KDD foi aplicado.

Como mostrado, o KDD se refere a uma sequência de etapas que envolve ajustes de parâmetros e possível interatividade entre essas etapas, executando novamente etapas realizadas anteriormente, com o intuito de refinar os conhecimentos extraídos. Por essa razão, é relevante destacar que determinadas técnicas adaptam-se melhor a determinadas situações e, portanto, não existe uma única maneira correta e efetiva para aplicação do KDD. Cada contexto no qual o KDD é aplicado exige um conjunto particular de parâmetros e técnicas específicas.

Com o KDD o conjunto de dados pode ser importado, organizado, integrado, transformado e visualizado (WICKHAM; GROLEMUND, 2017). Além disso, o KDD pode ser usado para gerar modelos preditivos e descritivos para identificar padrões por meio da aprendizagem de máquina. Modelos são gerados a partir da aprendizagem de máquina que faz com o que conhecimento seja adquirido por meio de dados. A aprendizagem de máquina é geralmente categorizada em três tipos: supervisionada, não-supervisionada e semi-supervisionada (VANDERPLAS, 2016).

No presente trabalho serão apresentados modelos preditivos de aprendizagem supervisionada e modelos descritivos nos quais é aplicada a aprendizagem não-supervisionada. No restante deste capítulo são abordados esses dois tipos de aprendizagem de máquina.

2.2 Aprendizagem Supervisionada

A aprendizagem supervisionada se refere a previsão de uma variável dependente a partir de uma lista de uma ou mais variáveis independentes. Por exemplo, quando pretende-se prever se um paciente possui ou não uma doença dada uma lista de sintomas, predizer o valor de uma residência baseando-se na sua localização ou, ainda, categorizar espécies de flores considerando as características da planta.

A premissa básica da aprendizagem supervisionada é que o conjunto de dados que se utiliza para treinar os modelos preditivos possua também a variável dependente resultante das variáveis independentes observadas. A variável dependente ainda é chamada, por alguns autores, como variável classe ou variável alvo. Considera-se que essas variáveis possuem as respostas ou as classes a serem previstas. Um modelo preditivo pode ser

usado para classificar futuros registros nos quais a variável dependente não é conhecida. A aprendizagem supervisionada pode ser aplicada de duas maneiras: regressão ou classificação.

A regressão é utilizada quando pretende-se prever um valor numérico da variável dependente, relacionando o comportamento de uma variável Y com outra X, de modo que $Y = a + bX$, onde X é a variável explicativa, Y é a variável dependente, a inclinação da linha é b e a interceptação, ou seja, o valor de Y quando $X = 0$, é representada por a (MONTGOMERY; PECK; VINING, 2012). O modelo de regressão é chamado de simples quando envolve uma relação causal entre duas variáveis. O modelo de regressão é multivariado quando envolve uma relação causal com mais de duas variáveis. Isto é, quando o comportamento de Y é explicado por mais de uma variável independe $X_1, X_2, \ldots, X_n$.

A classificação é a tarefa de aproximar uma função de mapeamento (f) de variáveis de entrada (X) para variáveis de saída discretas (y). As variáveis de saída são frequentemente chamadas de classes, rótulos ou categorias (ao decorrer deste documento utilizaremos o termo "classe"). Assim, a função de mapeamento f prevê a classe para uma determinada observação X_i (KOTSIANTIS *et al.*, 2007).

Para construir os modelos preditivos, seja pra regressão ou classificação, é necessário treinar e depois testar os modelos. Para isso, são usadas bases de dados diferentes para o treinamento e teste do modelo. Os dados de treinamento são os dados que serão usados pelo algoritmo para criação do modelo preditivo. Após o treinamento, os dados de teste são apresentados ao modelo preditivo, de modo que ele realize previsões utilizando a aprendizagem adquirida no treinamento. Com a etapa de teste é possível avaliar o desempenho do modelo (HAN; PEI; KAMBER, 2011).

A maneira que os dados são divididos para treinamento e teste influencia diretamente no desempenho do modelo preditivo. Os dois métodos mais utilizados na literatura para dividir os registros para treinamento e testes são o *hold-out* e a validação cruzada (*Cross-Validation* - CV). O *hold-out* divide a base de dados em dois grupos de registros, um que será utilizado exclusivamente para treino e outro que está reservado somente para teste. Uma divisão comum ao usar o método *hold-out* é usar 70% dos dados para treinamento e os 30% restantes dos dados para teste. Já na validação cruzada (KASSAMBARA, 2018), também conhecida como *k-fold*, a base de dados é dividida em um conjunto de dados de k partes e o treinamento e teste é executado iterativamente com cada uma destas partes. Cada iteração separa a k-ésima parte para testar o modelo treinado. Então, a parte de treinamento (k_i) é retornada ao conjunto de dados enquanto a próxima parte (k_{i+1}) é separada para treinamento. Assim, ao final do processo, todas as k partes são treinadas e testadas.

O método *hold-out* não foi escolhido porque com ele os registros usados para treinamento e teste podem ser muito parecidos, diminuindo o desempenho do modelo em decorrência de um *overfitting*, onde o modelo mostra-se adequado apenas para os

dados de treino (HAWKINS, 2004). Como o método da validação cruzada treina e testa o modelo com todos os dados disponíveis para evitar a variância, um resultado mais robusto é alcançado.

2.2.1 Métricas de Avaliação para Regressão

Para avaliar a qualidade dos modelos de regressão, algumas métricas são apresentadas na literatura, dentre elas:

- **R-quadrado ($\mathbf{R^2}$)**: também chamado de coeficiente de determinação, corresponde a uma medida de ajuste de um modelo estatístico linear generalizado em relação aos valores observados, variando de 0 a 1. Essa métrica representa o quanto o modelo pode explicar os valores observados. Quanto maior esta métrica, mais explicativo será o modelo, descrevendo melhor as amostras. Esse coeficiente é definido pela Equação 2.1, onde $\hat{y}$ é uma estimativa, y_i é a observação real, $\bar{y}$ é a média das estimativas e n é o número de pontos do experimento (AIT-AMIR; POUGNET; HAMI, 2015).

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{2.1}$$

- **erro médio absoluto (*Mean Absolute Error* - MAE)**: mede a magnitude média dos erros em um conjunto de previsões. Conforme mostrado na Equação 2.2, esta métrica se refere à média das diferenças das n amostras de teste entre a previsão ($\hat{y}_j$) e a observação real (y_j), em que todas as diferenças individuais têm peso igual (KASSAMBARA, 2018).

$$MAE = \frac{1}{n} \sum_{j=1}^{n} |y_j - \hat{y}_j| \tag{2.2}$$

- **raiz quadrada do erro quadrático médio (*Root Mean Square Error* - RMSE)**: é uma regra de pontuação quadrática que também mede a magnitude média do erro. Na Equação 2.3 representa-se essa métrica pela raiz quadrada das diferenças quadradas médias entre a previsão (S_i) e a observação real (O_i), onde n é o número de observações disponíveis para análise. O RMSE é uma boa medida de precisão, mas apenas para comparar erros de previsão de diferentes modelos ou configurações de modelo para uma determinada variável e não entre variáveis, pois depende da escala (NEILL; HASHEMI, 2018).

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (S_i - O_i)^2} \tag{2.3}$$

Tanto o MAE quanto o RMSE expressam o erro médio do modelo preditivo. Eles podem variar de 0 a ∞ e são indiferentes às direções de erro. Além disso, eles têm pontuações com orientação negativa. Portanto, por representarem taxas de erro, valores menores correspondem a modelos melhores (KASSAMBARA, 2018).

2.2.2 Métricas de Avaliação para Classificação

Para a tarefa de classificação, o desempenho do modelo preditivo é avaliado de acordo com a quantidade de registros no conjunto de teste que são classificados por ele. A classe da variável dependente é comparada com a previsão do classificador para esse registro e, assim, utiliza-se uma matriz de confusão (Quadro 1) que quantifica e divide os registros classificados em: Verdadeiros Positivos (VP), aqueles que representam classificações corretas; Falsos Positivos (FP), aqueles que foram ditos como da classe, mas na verdade não são; Falsos Negativos (FN), aqueles que o classificador previu que não pertenciam a classe, mas na verdade pertenciam; e Verdadeiros Negativos (VN), aqueles que o classificador indicou que não pertenciam a classe e, de fato, não pertenciam.

Quadro 1 – Representação de uma matriz de confusão.

	Verdadeiro	**Falso**
Verdadeiro	VP	FP
Falso	FN	VN

A matriz de confusão é útil para analisar o desempenho do classificador em reconhecer e classificar os diferentes registros. VP e VN indicam a habilidade do modelo em classificar corretamente, enquanto FP e FN representam a capacidade de classificar erroneamente. Esses quatro quantitativos são utilizados para calcular métricas de desempenho de um classificador. Alguns exemplos de métricas são:

- **Acurácia:** porcentagem de registros no conjunto de teste que são classificadas corretamente pelo classificador (Equação 2.4). Quando o problema de classificação possuir mais de duas classes, é preciso calcular a métrica para cada uma delas. É possível aliar esta métrica ao desvio padrão, ou variância, que calcula a soma das diferenças quadradas da média para os registros considerando o tamanho da amostra.

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{2.4}$$

- **Precisão (P):** identifica quantas amostras foram classificadas positivamente, evidenciando a exatidão em classificar registros FP (Equação 2.5). Pode ser usada em uma situação em que os FP são considerados mais prejudiciais que os FN.

$$P = \frac{VP}{VP + FP} \tag{2.5}$$

- ***Recall* (R)**: também conhecida como Sensibilidade, destaca quão exata é a classificação para registros FN (Equação 2.6). Esta métrica é inversamente proporcional à precisão (GÉRON, 2019). Ela evidencia a exatidão em classificar registros VP.

$$S = \frac{VP}{VP + FN} \tag{2.6}$$

De acordo com Demsar (2006), alguns testes podem ser aplicados para melhorar a confiança na avaliação dos modelos preditivos quando aplicam-se vários classificadores em múltiplos conjuntos de dados. Alguns exemplos são os testes Friedman (1937) e Nemenyi (1963) que verificam se os resultados alcançados com as métricas supracitadas são estatisticamente diferentes entre si. A hipótese nula (H_0) do teste de Friedman considera que todos os algoritmos são equivalentes e suas classificações médias são iguais. Já a hipótese alternativa (H_1) considera que pelo menos um algoritmo tem desempenho superior aos demais. Buscando validar uma das hipóteses, o teste de Friedman ranqueia os algoritmos do melhor para o pior, para cada conjunto de dados separadamente, de modo que o algoritmo de melhor desempenho fique no topo do ranque. Assim, seja r_i^j a classificação no ranque do j-ésimo de k algoritmos no i-ésimo de N bases de dados, o teste de Friedman compara as classificações médias dos algoritmos, $R_j = \frac{1}{N}\sum_i r_i^j$. Sob a H_0, que afirma que todos os algoritmos são equivalentes e, portanto, suas classificações R_j devem ser iguais, a estatística de Friedman (Equação 2.7) é distribuída de acordo com χ_F^2 com $k-1$ graus de liberdade, quando N e k são grandes o suficiente (DEMSAR, 2006).

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right] \tag{2.7}$$

Se a hipótese nula for rejeitada, geralmente aplica-se um teste *post-hoc*, para quantificar a diferença nas métricas dos classificadores. Nesse sentido, o teste Nemenyi é utilizado quando todos os classificadores são comparados entre si. O desempenho de dois classificadores é significativamente diferente se as classificações médias correspondentes diferem pelo menos da diferença crítica (*Critical Difference* - CD) entre eles, calculada na Equação 2.8.

$$CD = q_\alpha\sqrt{\frac{k(k+1)}{6N}} \tag{2.8}$$

A CD é relevante para destacar se os desempenhos dos modelos de classificação são estatisticamente distinguíveis. Na Equação 2.8, o valor de q é baseado no intervalo estudentizado[1] dividido por $\sqrt{2}$, e, nesta pesquisa, o valor utilizado para α foi 0,1.

[1] A diferença entre os maiores e os menores dados de uma amostra calculada utilizando o desvio padrão da amostra (BEARD, 1967).

Quando vários classificadores são comparados, os resultados dos testes *post-hoc* podem ser representados visualmente (Apêndice C).

Existem outros estimadores para avaliar classificadores. Um deles é a curva de característica de operação do receptor (*Receiver Operating Characteristic Curve* - ROC). A curva ROC é um gráfico que mostra a capacidade do modelo de distinguir as classes de acordo com a taxa de VP (*True Positive Rate* - TPR) e a taxa de FP (*False Positive Rate* - FPR), utilizando diferentes valores de teste (METZ, 1978). A curva ROC geralmente é utilizada para problemas de classificação binária (HOO; CANDLISH; TEARE, 2017), no entanto, na presente pesquisa, adaptou-se a abordagem para gerar a curva ROC para multiclasses (LANDGREBE; DUIN, 2007).

Além da análise da curva ROC, utilizam-se indicadores de dimensão do efeito da curva. O mais empregado é a área sob a curva (*Area Under the Curve* - AUC). A AUC é o resultado da integração de todos os pontos durante o trajeto da curva e é um estimador do comportamento da acurácia do teste (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). O valor da AUC varia de 0,0 até 1,0. Quanto mais próximo de 1 o valor da AUC está, mais corretamente o modelo gera previsões para aquela classe específica. Como na presente pesquisa utilizou-se uma abordagem de multiclasse, foi necessário gerar a AUC para cada uma das classes em cada algoritmo aplicado nos experimentos (Apêndice D).

2.2.3 Algoritmos

Existe uma diversidade de algoritmos na literatura voltados para aprendizagem supervisionada. Abaixo são descritos os que foram utilizados nesta pesquisa.

Regressão Linear

Este é um algoritmo utilizado para problemas de regressão que trata dados dispersos de forma linear. O algoritmo gera uma linha de melhor ajuste no gráfico de dispersão, que representa a relação das variáveis explicativas com a variável dependente. Usando a linha de melhor ajuste, é possível prever um valor para uma variável dependente dada uma nova entrada (LONG *et al.*, 1993). Na Equação 2.9 é representada a linha de melhor ajuste,

$$Y_i = \alpha + \beta X_i + \epsilon_i \tag{2.9}$$

onde Y_i é a variável dependente; α é uma constante, que representa a interceptação da linha com o eixo vertical; β é outra constante, que representa a inclinação da linha (coeficiente angular); X_i é a variável explicativa (independente) que representa o fator explicativo; e, finalmente, ϵ_i é uma variável aleatória que representa os possíveis erros de medição.

Árvore de Decisão

Este é um algoritmo usado na classificação e regressão, baseado na obtenção de informações usando técnicas de categorização (e.g., Gini, Qui-quadrado e Entropia).

O ganho de informação por meio da Entropia pode ser expresso pela Equação 2.10,

$$G(A) = I(p,n)\sum_{v}^{i=1}\frac{p_i + n_i}{p+n}I(p_i;n_i) \tag{2.10}$$

onde

$$I(p,n) = -\frac{p}{p+n}log_2\frac{p}{p+n} - \frac{n}{p+n}log_2\frac{n}{p+n} \tag{2.11}$$

sendo p e n, respectivamente, instâncias positivas e negativas de uma variável de bipartição (LONG *et al.*, 1993).

Floresta Aleatória

É um algoritmo usado tanto na classificação (*Random Forest Classifier* - RFC) quanto na regressão (*Random Forest Regressor* - RFR) que gera a predição por um fator aleatório, onde a combinação de modelos de aprendizagem aumenta o resultado geral (HAN; PEI; KAMBER, 2011). Ele é baseado em uma floresta de árvores, utilizando uma abordagem de *Bootstrap* (descrito mais adiante neste capítulo), combinando árvores aleatoriamente para produzir uma saída com menor variância (BREIMAN, 2001). Além disso, as florestas aleatórias usam um mecanismo para criar as árvores com menos correlação entre si, já que cada árvore utiliza informações diferentes para tomar suas escolhas de nós. Isso torna o processo de tomada de decisão mais robusto para bases de dados com muitos dados ausentes.

Máquina de Vetor de Suporte

Este é um algoritmo usado tanto na classificação quanto na regressão que realiza a separação de um conjunto de objetos utilizando o conceito de vetores de suporte, que são coordenadas da observação individual. O algoritmo calcula esses vetores de suporte e escolhe a fronteira (linha ou hiperplano) que melhor separa os dados, diferenciando as classes, no caso da classificação. Na separação dos registros, a maneira que o SVM escolhe um hiperplano é parametrizável. Por exemplo, é possível definir a forma que o hiperplano é escolhido, baseando-se na margem que esse hiperplano tem dos registros. Assim, quanto maior a margem que é definida, mais o SVM será tolerante a erros (HEARST *et al.*, 1998).

Naive Bayes

Este é um algoritmo de classificação que, baseado-se no teorema de Bayes (STIGLER, 1982), desconsidera a correlação entre as variáveis, tratando cada uma de maneira

independente. Assim, este algoritmo baseia-se em distribuições de probabilidade, uma função que descreve a probabilidade de uma variável aleatória assumir certos valores, isto é $P(\mathbf{X} \mid C) = \prod_{i=1}^{n} P(X_i \mid C)$, onde $\mathbf{X} = (X_1, \cdots, X_n)$ é um vetor de variáveis e C é a classe (RISH *et al.*, 2001). Sendo assim, para classificar um registro ($\hat{y}$) de um novo evento X, o algoritmo escolhe o y cuja probabilidade é maior (Equação 2.12).

$$\hat{y} = \arg\max_{y} P(y) \prod_{i=1}^{n} P(x_i \mid y) \tag{2.12}$$

Pedregosa *et al.* (2011) implementaram o Naive Bayes (NB) de três maneiras: Gaussiana, que usa os dados em uma distribuição normal; multinomial, usando os dados em uma distribuição multinomial, uma generalização da distribuição binomial; e Bernoulli, que utiliza uma distribuição multivariada de Bernoulli (MURPHY *et al.*, 2006). Nesta pesquisa foi utilizada a forma Gaussiana, na qual o cálculo da probabilidade é igual à densidade de probabilidade da distribuição normal (Equação 2.13), onde a média (μ_y) é o valor médio de x_i, considerando as observações da classe y e o desvio padrão (σ_y) do atributo x_i, considerando as observações da classe y.

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \tag{2.13}$$

K-ésimo Vizinho mais Próximo

O algoritmo k-vizinhos mais próximos, pode ser tanto um classificador quanto um regressor. Em seu funcionamento, para uma instância de teste x_q, o algoritmo encontra os k vizinhos mais próximos de x_q no conjunto de treinamento. Em seguida, x_q é rotulado de acordo com a classe que é maioria entre os k-vizinhos. Quando esse algoritmo é aplicado na tarefa de regressão, o valor atribuído a x_q é calculado com base na média dos valores de seus vizinhos mais próximos (PETERSON, 2009).

A escolha ideal para o número de vizinhos k depende da disposição dos dados, já que um número muito alto, pode suprimir ruídos e construir um limite de classificação impreciso. O classificador disponibilizado por Pedregosa *et al.* (2011) usa uma medida uniforme para encontrar os vizinhos mais próximos. No entanto, para esse cálculo também pode ser usado outras medidas de distâncias (por exemplo, euclidiana, *hamming* e Manhattan).

Perceptron Multicamadas

O *Multi Layer Perceptron* (MLP), um dos tipos de redes neurais artificiais, foi inspirado no funcionamento dos neurônios biológicos e pode ser usado tanto para a tarefa de classificação, quanto para a tarefa de regressão. Na representação de neurônios, os sinais de entrada são representados por um conjunto $x = [x_1, x_2, x_3, \ldots, x_n]$. Esses sinais são

multiplicados por um conjunto de pesos $p = [p_1, p_2, p_3, \ldots, p_n]$, gerando um potencial de ativação (a), ao qual é acrescido um valor b, que corresponde ao viés (Equação 2.14).

$$a = \sum_{i=1}^{N} x_i p_i + b \tag{2.14}$$

O potencial de ativação é passado para uma função de ativação σ que produz o valor de saída do neurônio. Diversas funções de ativação podem ser utilizadas, como sigmoide, tangente hiperbólica, softmax e ReLU (GOODFELLOW; BENGIO; COURVILLE, 2016). A última delas, foi utilizada nos experimentos realizados nesta pesquisa, já que redes com essa função de ativação são fáceis de otimizar. Para a criação do MLP, diversas camadas de múltiplos neurônios são criadas. Assim, os valores de entrada são passados para a camada inicial, e são processados de uma camada para outra até a rede fornecer como resultado os valores de saída da última camada.

2.2.4 Comitês Homogêneos e Heterogêneos

Os comitês, também chamados de aprendizado por agrupamento ou *Ensemble Learning*, são combinações de diversos modelos preditivos que são treinados e geram um modelo preditivo mais robusto. Eles são chamados de homogêneos quando são compostos pelo mesmo tipo de modelo e de heterogêneos quando são compostos por modelos diferentes. De acordo com Han, Pei e Kamber (2011), um comitê combina uma série de k modelos, M_1, M_2, ..., M_k, com o objetivo de criar um modelo preditivo aprimorado, $M*$. Um determinado conjunto de dados, D, é usado para criar k conjuntos de treinamento, D_1, D_2, ..., D_k, onde D_i $(1 \leq i \leq k-1)$ é usado para gerar o comitê M_i. Assim, dado um novo registro para classificar, o comitê retorna uma previsão de acordo com a votação feita pelos modelos individuais, onde cada um teve sua própria previsão isolada.

Por exemplo, considerando um comitê de votação por maioria para classificação, dado um registro X que precisa ser classificado, o comitê coleta e compara as previsões retornadas por cada um dos modelos preditivos isoladamente e fornece a classe para o registro de acordo com a maioria. Assim, os classificadores isoladamente podem cometer erros, mas o comitê somente classifica erroneamente X quando mais da metade dos classificadores iniciais estiverem errados. Quando se trata de regressão, o comitê por votação calcula a média entre as previsões individuais.

Os comitês são mais eficientes quando há diversidade significativa entre as previsões nos modelos individuais e, geralmente, possuem desempenho melhor do que os modelos preditivos aplicados isoladamente (SAGI; ROKACH, 2018). De acordo com Han, Pei e Kamber (2011), na criação de comitês, os modelos preditivos individuais podem ser combinados de três maneiras:

- ***Bagging***: faz com que modelos homogêneos aprendam independentemente e em paralelo, combinando-os a partir de algum tipo de resultado médio. Em outras palavras, essa técnica consiste em subdividir o conjunto de dados em grupos menores, aplicar o mesmo modelo de predição nesses subgrupos e computar a média dos resultados como a predição do modelo final;

- ***Boosting***: faz com que modelos homogêneos aprendam de maneira adaptativa, sequencialmente, a partir de uma análise dos modelos treinados anteriormente. As RF são exemplos deste tipo de abordagem. O *Boosting* treina iterativamente novos modelos sempre com um enfoque nas observações que os modelos anteriores tiveram mais dificuldade, e, assim, o comitê é melhorado priorizando as predições com maior acurácia nas observações da base de teste. A maneira que essa melhoria é aplicada origina diferentes algoritmos para essa técnica, como Adaboost (SCHAPIRE, 2013), *Gradient Boosting* (NATEKIN; KNOLL, 2013) e XGBoost (CHEN *et al.*, 2015); e

- ***Stacking***: faz com que modelos heterogêneos aprendam em paralelo e sejam combinados treinando um metamodelo para gerar uma previsão com base nas previsões de diferentes modelos individuais. Por exemplo, para a tarefa de classificação, esta técnica combina vários modelos por meio de um meta-classificador. Os modelos de classificação individuais são treinados com base no conjunto de treinamento completo, então, o meta-classificador é ajustado com base nas saídas dos modelos de classificação individuais no conjunto (TANG; ALELYANI; LIU, 2015).

De maneira geral, o *Bagging* se concentra principalmente em obter um comitê com menos variância do que seus modelos individuais, enquanto que o *Boosting* e *Stacking* tentam produzir modelos menos tendenciosos do que seus componentes.

2.3 Aprendizagem Não-supervisionada

Diferentemente da aprendizagem supervisionada, a aprendizagem não-supervisionada analisa os registros sem consultar variáveis dependentes. Em muitos casos, isso acontece porque não há conhecimento prévio de uma variável dependente. Dessa forma, espera-se criar uma representação mais informativa e condensada dos dados (agrupamentos, regras de associação, etc), evidenciando semelhanças entre os objetos analisados para detectar similaridades e anomalias. Nesta pesquisa, a aprendizagem não-supervisionada foi utilizada somente para gerar agrupamentos, técnica que também é chamada de clusterização. Para esta finalidade o algoritmo *k-means* foi utilizado.

O *k-Means* tem a capacidade de agrupar uma série de dados com tempo de computação relativamente rápido e eficiente (LóPEZ-RUBIO; PALOMO; ORTEGA-ZAMORANO, 2018) e se tornou um dos algoritmos mais importantes na MD (WU

et al., 2007). O *k-means* agrupa os dados para separar amostras em n grupos, nos quais os atributos possuem alguma semelhança (VANDERPLAS, 2016). Este algoritmo requer que o número de agrupamentos (*clusters*) seja especificado. Esse algoritmo divide um conjunto de N amostras X em K *clusters* C separados, cada um com um ponto central (chamado de centróide) que é calculado pela média aritmética de todos os pontos pertencentes ao agrupamento.

Assim, o *k-means* escolhe k amostras de X para definir os centróides iniciais. Após a inicialização, o algoritmo é executado iterativamente atribuindo cada amostra ao seu centróide mais próximo e ajustando novos centróides tomando a média de todas as amostras atribuídas a cada centróide anterior. A diferença entre o antigo e o novo centróide é calculada para minimizar o critério de inércia até que os centróides não se movam significativamente.

Um método que ajuda a determinar a quantidade ideal de *clusters* para o *k-means* é o Método Cotovelo (BHOLOWALIA; KUMAR, 2014) que adiciona *clusters*, calcula o erro quadrático de soma (*Sum Squared Error* - SSE) por *cluster* até o número máximo de *clusters* que foram determinados, então, compara a diferença SSE de cada cluster e a diferença mais extrema forma o ângulo do cotovelo, mostrando o melhor número do cluster (UMARGONO; SUSENO; GUNAWAN, 2020).

2.4 Considerações Finais

A MD é parte de um processo iterativo de descoberta do conhecimento e faz uso da aprendizagem de máquina para reconhecer padrões nos dados. Sob essa perspectiva, a MD pode ser utilizada para explorar grandes quantidades de registros a procura de padrões em situações em que as técnicas tradicionais de exploração e análise de dados não sejam suficientes. Neste capítulo, foram apresentadas duas formas de aprendizagem de máquina, uma voltada para variáveis dependentes já presentes nos dados, outra procurando representações que também podem revelar padrões consistentes. Para isso, utiliza-se uma diversidade de algoritmos aplicados nos mais diversos contextos.

3 Trabalhos Relacionados

Este trabalho foi inspirado em Fanaee-T e Gama (2013) que propõem uma alternativa para rotular acontecimentos que ocorrem em Washington DC usando conhecimento prévio e informações de contexto. Os autores enriqueceram os registros de 2011 de SBC com informações meteorológicas e sazonais. Na presente pesquisa, esta abordagem foi replicada com um conjunto de dados mais recente para outro propósito, com o objetivo de gerar modelos preditivos para a quantidade de aluguéis de bicicletas. Nos sistemas de transporte atuais, informações precisas sobre a demanda de aluguéis são necessárias para os usuários, setores comerciais e, principalmente, para as empresas responsáveis pelo gerenciamento desses sistemas, com o potencial de melhorar a eficiência na gestão de recursos dos SBC.

O estado da arte atual na predição de demanda nos SBC utiliza registros históricos aliados às variáveis ambientais (por exemplo, temperatura) para prever a demanda futura (YANG *et al.*, 2020). A previsão de demanda tem ganhado cada vez mais atenção com o rápido desenvolvimento e implantação dos SBC e estudos com esse propósito foram aplicados em diversos SBC, como *Caltrans Performance Measurement System* - PeMS (Califórnia); *Seoul Citywide Bike Rental System* (Seoul); *Capital Bikeshare* (Washington); *Citi Bike* (Nova York), *Divvy* (Chicago); *London Shared Bike* (Londres); *Hangzhou Public Bicycle* (Hangzhou); *Bay Area* (São Francisco); e *Vélib* (Paris).

Na previsão de demanda de aluguéis, especificamente nos SBC, o objetivo não é meramente obter valores específicos de predição, mas otimizar o sistema para agendar o rebalanceamento de bicicletas compartilhadas (ZHANG *et al.*, 2019). Nesse sentido, os trabalhos relacionados para esse propósito são aplicados em três níveis: cidade, regiões e estação. No Quadro 2 são sumarizados os trabalhos discutidos neste capítulo, respondendo as seguintes questões:

- Q1) Qual a base de dados utilizada?
- Q2) A base de dados utilizada é enriquecida com informações meteorológicas?
- Q3) A base de dados utilizada é enriquecida com informações sazonais (ex. feriados)?
- Q4) Qual/quais o modelo(s) preditivo(s) utilizado(s)?
- Q5) O modelo preditivo é voltado para regressão (R) e/ou classificação (C)?
- Q6) A predição é em nível de cidade (C), região (R) ou estação (E)?
- Q7) Em que escala de tempo a predição acontece?

Quadro 2 – Fichamento dos trabalhos relacionados.

Autoria	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**
Lv *et al.* (2015)	*PeMS*	Não	Não	SAE; Regressão Logística	R	C	a cada 15, 30 e 45 minutos, e a cada hora
Sathishkumar, Jangwoo e Yongyun (2020)	*Seoul Citywide Bike Rental System*	Sim	Sim	LR; GMB; SVM; BT; XGBTree	R	C	a cada hora
Li e Zheng (2020)	*Capital Bikeshare*; *Citi Bike*	Sim	Não	SGPR	R	C; R	a cada hora
Li *et al.* (2015)	*Capital Bikeshare*; *Citi Bike*	Sim	Não	GBRT	R	C; R	a cada hora
Chen *et al.* (2016)	*Capital Bikeshare*; *Citi Bike*	Sim	Sim	WCN-MC	C	R	a cada hora
Yang *et al.* (2020)	*Citi Bike* ; Divvy	Sim	Não	XGBoost	R	R	a cada hora
Xu *et al.* (2020)	*Capital Bikeshare*; *London Shared Bike*	Sim	Sim	SOM-RT	R	E	a cada hora
Zhang *et al.* (2019)	Não especificado	Não	Não	XGBoost	R	R; E	a cada hora
Yang *et al.* (2016)	*Hangzhou Public Bicycle*; *Citi Bike*	Sim	Não	RF	R	E	a cada hora
Liu *et al.* (2016)	*Citi Bike*	Sim	Sim	MSWK	R	E	a cada hora
Ashqar, Elhenawy e Rakha (2019)	*Bay Area*	Sim	Não	NBRM; RF	R	E	a cada minuto
Gast *et al.* (2015)	*Vélib*	Não	Não	Probabilidade	R	E	a cada minuto
Presente trabalho	*Capital Bikeshare*	Sim	Sim	RF; DT; Comitês	C; R	C; R	a cada hora

Os trabalhos sumarizados no Quadro 2 são discutidos nas próximas seções.

3.1 Previsão por Cidade

Lv *et al.* (2015) usaram aprendizagem profunda (*deep learning*) para prever o fluxo de tráfego usando *Stacked Autoencoder* (SAE), um tipo de rede neural (BENGIO *et al.*, 2006). As correlações espaciais e temporais foram consideradas na predição. Um modelo SAE foi usado para extrair recursos de fluxo de tráfego e uma camada de regressão logística no topo da rede foi aplicada para a previsão de fato. Para treinamento do algoritmo foi utilizada a base de dados do PeMS (*Performance Measurement System*), com registros de veículos coletados entre Janeiro a Março de 2013 nas rodovias da Califórnia. No entanto, o modelo proposto não funciona bem em condições de baixo fluxo. A razão para isto é que pequenas diferenças entre o fluxo observado e o fluxo previsto podem causar um erro relativo maior quando a taxa de fluxo do tráfego é pequena. Os autores sugeriram investigar outros algoritmos de *deep learning* para previsão de fluxo de tráfego e aplicar esses algoritmos em diferentes conjuntos de dados abertos para examinar sua eficácia.

Sathishkumar, Jangwoo e Yongyun (2020) analisaram a relação entre o serviço de aluguéis de bicicleta e os diferentes parâmetros temporais e meteorológicos. Os autores utilizaram a base de dados do período de um ano (dezembro de 2017 a novembro de 2018) de um sistema de compartilhamento de bicicletas localizado em Seoul (Coréia do Sul) para construir e comparar a eficiência de vários modelos de regressão como *Linear Regression* (LR), *Gradient Boosting Machine* (GBM), *Support Vector Machine* (SVM), *Boosted Trees* (BT) e *Extreme Gradient Boosting Tree* (XGBTree). Os autores identificaram que o atributo *hour* é fundamental para a previsão da demanda e também apresentaram as informações climáticas como mecanismos que podem melhorar o desempenho dos modelos apresentados. Os autores propõem como trabalho futuro uma predição em nível de regiões, uma das abordagens de agrupamento que são mostradas no Capítulo 6 deste trabalho.

Li e Zheng (2020) aplicaram o algoritmo de agrupamento *Adaptive Transition Constraint* (AdaTC) nas estações, de modo que as estações fossem agrupadas com base em suas localizações geográficas e padrões de uso. Após isso, utilizaram um regressor de processo gaussiano (*Gaussian Process Regressor* - SGPR) baseado em similaridade para prever a demanda de aluguéis em um nível de cidade e agrupamento. A abordagem foi validada nos SBC de Washington DC e Nova York, com dados de abril a outubro de 2016. Os autores também utilizaram informações meteorológicas e comparam o modelo preditivo gerado com outros modelos presentes na literatura. Esse estudo também analisou os dados em níveis de regiões. Na próxima seção são apresentados estudos voltados para a análise em nível de regiões.

3.2 Previsão por Regiões

Li *et al.* (2015) utilizaram dados de abril a setembro de 2014 dos SBC localizados em Nova York e Washington. O número total de bicicletas alugadas nas cidades foi previsto pelo *Gradient Boosting Regression Tree* (GBRT) (FRIEDMAN, 2001), que é uma técnica de aprendizagem estatística para regressão. Em seguida, um modelo de inferência foi utilizado para quantificar a proporção de bicicletas alugadas e devolvidas em agrupamentos construídos pelo k-means. Indentificaram-se semelhanças entre os hábitos dos usuários de Washington DC e de Nova York, principalmente no que diz respeito a horários de pico de uso e padrões de utilização nos dias da semana. Embora os autores usem algumas informações meteorológicas para enriquecer a base, é necessário considerar outros fatores que também podem influenciar a quantidade de aluguéis (a presença de feriados) e que auxiliam na compreensão da predição em períodos anômalos. Esses fatores foram considerados na presente pesquisa, bem como a utilização de outros parâmetros climáticos (umidade, sensação térmica, pressão atmosférica) que não foram usados na pesquisa citada.

Yang *et al.* (2020) modelaram grafos a partir de registros de novembro de 2016 a setembro de 2017 dos SBC de Nova York e Chicago. Um método de agrupamento hierárquico foi aplicado para agrupar estações em 120 grupos para Nova York e 80 grupos para Chicago. Cada grupo de estações foi projetado como um nó, e o volume de viagens de bicicleta por hora entre os nós foi usado para gerar vértices que representam os fluxos de origem-destino dos aluguéis. Isso resultou em uma série de estruturas de grafo temporalmente ponderadas e direcionadas, a partir das quais várias propriedades de grafo foram calculadas, descrevendo o estado de cada nó em momentos diferentes. Os autores defendem que as relações entre as estações (nesse caso representado por grafos) são mais importantes na previsão de demanda do que as informações meteorológicas que são mais comumente usadas, já que informações estruturais do grafo capturam propriedades espaciais importantes. Assim, os autores utilizaram o *Extreme Gradient Boosting* (XGBoost) para avaliar e quantificar a relevância desses grafos na predição de demanda.

Chen *et al.* (2016) utilizaram as bases de dados de 2014 e 2015 dos SBC de Nova York e Washington DC e modelaram as relações entre as estações considerando diversas informações de contexto. O peso dessas relações foi calculado com base na correlação entre os registros dos SBC e os tipos de eventos que ocorreram nas proximidades das estações. Assim, os autores propuseram um método de agrupamento que utiliza eventos sociais e condições de tráfego, classificando a disponibilidade de bicicletas nos agrupamentos como "Cheio", "Normal" ou "Vazio" nas proximidades desses eventos. A quantidade de aluguéis e devoluções em cada agrupamento foi estimada com o *Weighted Correlation Network and Monte Carlo simulation* (WCN-MC), que quantifica o impacto dos eventos sociais e de tráfego próximos aquele agrupamento por meio de uma taxa de inflação (ZHENG *et al.*,

2015) e adota o método Monte Carlo (MOONEY, 1997) para simular o funcionamento do sistema, representando aluguéis e devoluções de bicicletas nos agrupamentos. No presente trabalho, a técnica de classificação foi utilizada diferentemente para ser uma alternativa à técnica de regressão, indicando a intensidade da demanda naquela hora.

No entanto, realizar predições por regiões pode ser um desafio, uma vez que um pequeno erro na previsão em nível de agrupamento seria multiplicado várias vezes, ocasionando em um resultado final insatisfatório (ZHANG *et al.*, 2019). Na presente pesquisa, os agrupamentos foram gerados considerando divisões geográficas já existentes na cidade e a proximidade entre estações.

3.3 Previsão por Estação

A previsão de demanda em cada estação de aluguel é desafiadora e atrai considerável interesse de pesquisadores. Gast *et al.* (2015) aplicaram uma abordagem probabilística para predizer a disponibilidade de bicicletas em cada estação de aluguel. Para isso, utilizaram os registros históricos em dois formatos: um que descreve a disponibilidade de bicicletas em todas as estações de aluguéis por minuto e outro contendo as informações individuais dos aluguéis. Esses registros foram utilizados com a finalidade de estimar as taxas de aluguel e devolução a cada minuto. A predição apresentada pelos autores foi realizada por minuto, enquanto que no presente trabalho as predições foram realizadas por hora. Além disso, Gast *et al.* (2015) apresentaram uma abordagem de predição puramente estatística, ignorando a não-linearidade dos registros de aluguéis. Por isso, acredita-se ser esta uma das causas do baixo poder preditivo da abordagem apresentada pelos autores.

Xu *et al.* (2020) propuseram SBC baseados em *edge computing*, que remetem a um novo paradigma da computação distribuída que possibilita o processamento de dados no local ou nas proximidades de onde os dados são coletados (ROMAN; LOPEZ; MAMBO, 2018). Dessa maneira, seria possível prever a demanda de aluguel de cada estação na rede utilizando o processamento em nós próximos. Para isso, os autores utilizaram um modelo de aprendizado de máquina híbrido que combina uma rede de mapeamento auto-organizável com uma árvore de regressão (*Self-Organizing Mapping network with a Regression Tree* - SOM-RT) para prever a demanda por bicicletas de uma determinada estação. As redes neurais auto-organizáveis se adaptam aos parâmetros e estruturas da rede, encontrando automaticamente propriedades essenciais para descrever as amostras. A abordagem apresentada por Xu *et al.* (2020) pode ser vista como uma versão modificada de uma rede de mapeamento auto-organizável (*Self-Organizing Mapping* - SOM), com o adicional de construir uma árvore de regressão para cada neurônio após o processo de treinamento do algoritmo. Os autores utilizaram a base de dados de SBC de Washington (janeiro de 2011 a dezembro de 2012) e de Londres (janeiro de 2015 a dezembro de 2016).

Ashqar, Elhenawy e Rakha (2019) quantificaram o efeito de várias condições meteorológicas nos aluguéis de bicicletas e avaliaram a viabilidade de um modelo genérico que fosse aplicável nas 70 estações do sistema *Bay Area* (São Francisco) com registros de agosto de 2013 a agosto de 2015. Para tal, foram aplicados o *Poisson Regression Model* (PRM) e o *Negative Binomial Regression Model* (NBRM) para criar modelos preditivos para prever a quantidade de aluguéis em cada uma das estações e utilizou o RF para selecionar o modelo mais preciso. O critério bayesiano de informação (*Bayesian Information Criterion* - BIC) foi selecionado para comparar os modelos seguindo uma regressão progressiva guiada pelos resultados do RF.

A abordagem de gerar um modelo genérico, como no trabalho supracitado, pode ser eficaz quando se tem um *big data* e diversas variáveis, mas, pode comprometer a precisão em estações específicas, já que o modelo proposto apresentou erros de predição aumentados em estações localizadas no centro da cidade. Além disso, identificou-se que a umidade era a principal variável meteorológica que influenciava a quantidade de aluguéis naquela região. No entanto, as informações de clima são fortemente relacionadas às localizações geográficas e este pode ser um fenômeno exclusivo daquela região, que pode não representar o hábito dos usuários de SBC em outras localidades. No presente trabalho, por exemplo, identificou-se que a temperatura tem um papel mais significativo na predição da quantidade de aluguéis do que a umidade.

Zhang *et al.* (2019) usaram o algoritmo *XGBoost* para predição da demanda de aluguéis em cada hora nas estações. Além disso, utilizaram o *k-means* para auxiliar na classificação das estações de acordo com três níveis de atividades, considerando as similaridades de demanda em cada estação. Os autores realizaram predições e analisaram o uso de cada bicicleta, identificando bicicletas que são usadas demasiadamente e buscando minimizar os custos de manutenção. Os pesquisadores do trabalho consideraram que os fatores temporais são os que mais afetam o serviço de aluguel e, por essa razão, não utilizaram uma base de dados enriquecida com informações meteorológicas. Além disso, embora seja destacado o intervalo de tempo dos registros utilizados (maio a agosto de 2015), não é explicitado a quais SBC os dados pertencem. Além disso, quando a escala de tempo semanal e mensal é utilizada ou tenta-se caracterizar as estações de maneira genérica, o desempenho do modelo preditivo gerado pelo *XGBoost* é reduzido.

Yang *et al.* (2016) desenvolveram uma metodologia de previsão de demanda por estação, considerando as correlações espaço-temporais entre as estações e fatores de clima. Para isso, usaram um modelo probabilístico para descrever os movimentos das bicicletas dentro de SBC e estimaram a quantidade e duração dos aluguéis em diferentes estações, usando registros do sistema no intervalo de um ano. A pesquisa foi aplicada nos SBC de Hangzhou (China) e de *Citi Bike* (Nova York). Nas duas cidades onde os sistemas estão localizados, perceberam-se que a demanda de aluguéis é menor de segunda-feira à sexta-

feira e maior no fim de semana. Os padrões apresentados na Seção 5.1 da presente pesquisa indicam que o oposto acontece em Washington DC. Os autores também apresentaram alguns padrões que são semelhantes, como o fato dos usuários desse sistema se sentirem mais confortáveis para pedalar em temperaturas mais altas.

Liu *et al.* (2016) propuseram um modelo *Meteorological Similarity Weighted K-nearest neighbor* (MSWK) que usa dados meteorológicos em conjunto com dados históricos de julho de 2013 a maio de 2015 de um dos SBC de Nova York para prever a demanda de aluguéis que acontecerão por hora em cada estação. Além disso, os autores apresentaram um modelo de predição de transição das bicicletas entre estações (*Inter Station Bike Transition* - ISBT) para prever a demanda de entrega, considerando a duração e frequência de viagens entre ambas estações. Assim, a presente pesquisa está mais focada na identificação das relações origem-destino em cada estação e, para auxiliar na tarefa de rebalanceamento, apresentou uma abordagem para agrupar as estações de acordo com diferentes níveis de fluxos de aluguéis.

3.4 Considerações Finais

Como apresentado ao decorrer do capítulo, a predição da demanda de aluguéis é útil para entender o hábito dos ciclistas. O estado da arte nesta área utiliza dados históricos e variáveis de contexto (por exemplo, informações climáticas) para prever a demanda futura. Nesse sentido, os trabalhos apresentados destacaram aplicações úteis tanto de um ponto de vista do usuário do sistema, quanto de um ponto de vista gerencial, sobretudo no que diz respeito ao rebalanceamento de bicicletas nos SBC.

Como abordado, a predição de demanda em SBC também pode acontecer em diferentes níveis. Embora o presente trabalho não pretenda aprimorar os algoritmos apresentados pelos trabalhos relacionados, na presente pesquisa diversos algoritmos clássicos são comparados e uma demonstração experimental é apresentada. A análise dos registros em nível de cidade pode ser útil para entender o dinamismo do sistema, bem como para identificar alguns aspectos que, de maneira geral, influenciam no serviço de aluguéis.

Em relação aos trabalhos apresentados, um diferencial da abordagem proposta é a utilização dos modelos utilizando dados em tempo real, o que é possibilitado pela maneira que os dados são disponibilizados por alguns SBC (Capítulo 7). Além disso, no presente trabalho é utilizada tanto a regressão quanto a classificação, identificando o impacto das técnicas de pré-processamento nos modelos preditivos gerados por ambas as abordagens.

Com os trabalhos relacionados percebeu-se que a análise em nível de regiões resulta em certas limitações, uma vez que a forma que os agrupamentos são construídos pode influenciar no desempenho dos modelos preditivos. Semelhantemente a Yang *et al.* (2020), no presente trabalho, as estações de aluguéis foram agrupadas considerando a proximidade

entre elas, porém, nos trabalhos relacionados foram encontradas outras abordagens para o agrupamento das estações de aluguéis, como considerar a similaridade de uso das estações (LI *et al.*, 2015; CHEN *et al.*, 2016).

Na análise em nível de estação de aluguel, a quantidade reduzida de registros para cada estação dificulta a etapa de treinamento dos algoritmos e, além disso, fatores incertos (como eventos sociais nas proximidades e condições de trânsito) podem dificultar a predição (CHEN *et al.*, 2016). Por razão desses desafios, no presente trabalho, a predição de demanda em nível de estação não foi discutida, uma vez que a predição de demanda em nível de cidade e regiões já podem ser suficientes para proporcionar e apoiar decisões estratégicas nos SBC.

4 Metodologia

Neste capítulo é apresentada a metodologia utilizada no trabalho, detalhando-se a base de dados utilizada nesta pesquisa (Seção 4.1), os procedimentos de pré-processamento que foram utilizados (Seção 4.2) e descrevendo as novas bases geradas com cada um desses procedimentos (Seção 4.3). As bases de dados foram particionadas usando validação cruzada, de modo que os mesmos registros fossem usados para treinamento e teste.

A metodologia deste trabalho é inspirada no processo clássico de extração de conhecimento em banco de dados (*Knowledge Discovery in Databases* - KDD) (Seção 2.1), utilizando técnicas de aprendizagem de máquina para auxiliar na tomada de decisão no gerenciamento dos SBC (WICKHAM; GROLEMUND, 2017).

Com a metodologia utilizada (Figura 2), é possível utilizar bases de dados reais de SBC enriquecidas com informações de contexto, identificando atributos que influenciam no serviço de aluguel de bicicletas e gerando modelos preditivos e descritivos que auxiliem na gestão estratégica desses sistemas.

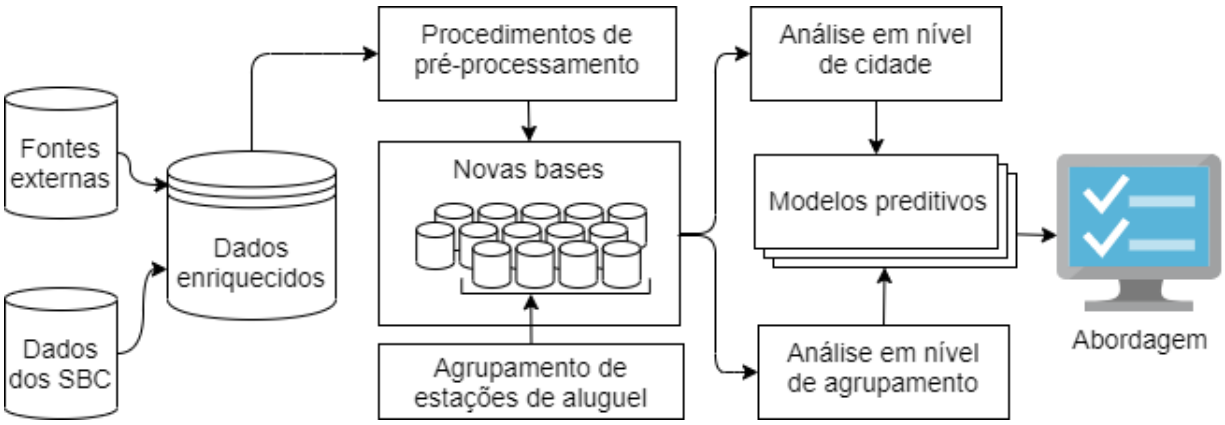


Figura 2 – Representação da metodologia da pesquisa.

As seções seguintes detalham a Figura 2, descrevendo a base de dados utilizada na pesquisa, os procedimentos de pré-processamento que foram aplicados, as novas bases criadas, as análises realizadas e a configuração dos modelos preditivos gerados.

4.1 Base de Dados

Nesse trabalho foi utilizada a base de dados do *Capital BikeShare*[1], um dos maiores SBC do mundo, localizado nos Estados Unidos, em Washington D.C. Esse sistema está em funcionamento desde 2010 e atualmente possui cerca de 500 estações e 4300 bicicletas.

[1] capitalbikeshare.com

A base de dados original do ano de 2017 do *Capital BikeShare* possui cerca de 3,75 milhões de registros. Os registros de aluguéis são realizados com dois tipos de usuários: *member*, que é o usuário com registro anual ou mensal, e *casual*, aquele que faz aluguéis eventuais ou faz uso do serviço por 1, 3 ou 5 dias. As estações de aluguéis dispõem de um sistema informatizado que armazena os registros de aluguéis contendo duração da viagem, horário de início, horário de término, informações sobre a estação inicial e final, identificação da bicicleta e tipo de usuário (CAPITAL BIKE SHARE SYSTEM, 2018). Os registros deste sistema em específico estão disponíveis publicamente, motivando a análise e exploração dessas informações.

A base de dados utilizada nesta pesquisa é referente aos aluguéis que aconteceram entre 1º de janeiro a 31 de dezembro de 2017, possuindo cerca de 3,75 milhões de registros, com aproximadamente 35,4% de aluguéis realizados por usuários casuais e 64,6% por usuários cadastrados. Além disso, também utilizaram-se bases de dados geradas a partir da base original, por exemplo, com o enriquecimento de informações meteorológicas, agrupamentos da quantidade dos aluguéis por hora, discretização, normalização e agrupamento das estações. Essas bases secundárias serão descritas na Seção 4.3.

Para enriquecer as bases de dados com informações de clima que não existiam na base original, foi desenvolvido um *script* de coleta de dados na Web (*Web Scraping*), recuperando informações meteorológicas do site Freemeteo[2]. Para o processo de integração dessas informações aos registros, foi considerada a medição meteorológica mais próxima do início e término da viagem.

Além dos dados meteorológicos, as informações sobre feriados em Washington D.C foram coletadas no site do departamento de recursos humanos (*Department of Human Resources* - DCHR)[3]. Para extrair o máximo de informações possíveis do atributo que indica quando o aluguel aconteceu (*start_date*), extraíram-se dele os atributos hora, dia, mês, estação do ano, dia da semana e dia útil (*hour*, *day*, *month*, *season*, *weekday* e *workday*, respectivamente).

Os atributos que identificam onde os aluguéis começam e terminam (*start_station*, *start_station_name*, *end_station* e *end_station_name*) contêm a identificação e o nome da estação onde a bicicleta foi alugada e devolvida. Mesmo existindo um atributo que indica a duração do aluguel (*duration*), ainda utilizou-se a interface de programação de aplicações (*Application Programming Interface* - API) do Google Maps[4] para calcular a distância mais curta entre *start_station* e *end_station* (*m_distance*, em metros) e a duração que o aluguel teria se esta rota fosse usada (*m_duration*, em minutos). Foram utilizados *scripts* em Python e em R para integrar e manipular essas informações adicionais[5].

[2] freemeteo.com
[3] dchr.dc.gov
[4] cloud.google.com/maps-platform
[5] https://github.com/johnattandouglas/Dissertacao/tree/main/Scripts

Para facilitar a identificação dos padrões e análise dos dados, foi realizado um agrupamento por horário dos registros. A base de dados com os registros agrupados por hora possui 8737 registros e 17 atributos, listados e descritos no Quadro 3.

Quadro 3 – Listagem dos atributos presentes na base de dados enriquecida com os registros agrupados por hora.

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

*Agrupamento dos registros de aluguéis em cada hora do dia.

A variável *qtd* mostrada no Quadro 3 foi a variável classe, pois representa a totalidade de aluguéis que aconteceram em cada hora. Para inserção das variáveis climáticas nos registros agrupados, considerou-se a média das medidas climáticas naquela hora.

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. No que diz respeito aos atributos de clima, foram identificados valores discrepantes nos atributos *wind* e *pressure*. Na Tabela 1 são apresentados os valores de mínimo, máximo, mediana e média das variáveis climáticas na base de dados.

Tabela 1 – Resumo estatístico das variáveis climáticas na base de dados.

Atributo	**Mínimo**	**Máximo**	**Mediana**	**Média**
temperature	-9	36	17	15,83
r_temperature	-17	43	17	15,24
wind	0	63	13	14,11
humidity	13	100	66	65,13
dew_point	-19	27	10	8,60
pressure	990	1042	1016	1017

Nos registros analisados perceberam-se discrepâncias em algumas variáveis climáticas. A temperatura, por exemplo, variou de -9ºC a 36ºC, enquanto que a sensação térmica oscilou entre -17ºC e 43ºC. Esse grande intervalo de valores para essas variáveis são explicáveis porque as estações do ano em Washington DC são bem definidas e variam de nevascas no inverno ao calor no verão. O atributo *pressure* tem relação com *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 (DUPUIS; FIELD, 2004), o que explica alguns valores discrepantes encontrados nos registros (como o valor máximo 63 encontrado nessa variável). Como os atributos *pressure* e *wind* praticamente não possuem correlação significativa com a quantidade de aluguéis (ver Seção 5.2), os valores discrepantes nessas duas variáveis não foram tratados.

4.2 Pré-processamento

Para melhorar o desempenho dos algoritmos a serem comparados neste trabalho foram aplicados seis procedimentos de pré-processamento na base de dados original e nas bases de dados enriquecidas:

- **procedimento A**: remoção de registros incompletos. Alguns atributos foram removidos da base, a saber: *id*, que não é útil para o propósito deste trabalho; *date*, uma coluna redundante, já que outras colunas (ex. *month*, *day*) foram derivadas dela; *holiday*, que possuía 8449 valores faltosos, indicando se o dia era 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*). 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, utilizaram-se os métodos *cut* e *qcut* da biblioteca Pandas (MCKINNEY, 2012, pp. 199). O *cut* segmenta e classifica valores de dados em partições usando equidistância e o *qcut* gera partições que possuem aproximadamente a mesma quantidade de registros. Na Tabela 2 são especificadas as partições geradas por cada uma dessas abordagens e é apresentada a quantidade de registros presentes em cada partição. Como exposto na Tabela 2, o método *qcut* gerou partições de tamanhos similares,

enquanto que o método *cut* resultou em partições de tamanhos bem diferentes, onde a maior partição foi a categoria "Muito Baixa" (com 4894 registros) e a menor foi a categoria "Muito Alta" (com apenas 108 registros);

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

Categorização	**Método *cut***		**Método *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*[6], que ajusta os valores no intervalo entre 0 (Min, menor valor) e 1 (Max, maior valor), de acordo com a Equação 4.1, onde x é o valor a ser normalizado;

$$x = \frac{x - \text{Min}}{\text{Max} - \text{Min}} \tag{4.1}$$

- **procedimento E**: aplicação da Análise de Componentes Principais (*Principal component analysis* - PCA) nas bases normalizadas. Esta é uma técnica de redução de dimensionalidade baseada na variância dos dados (SHLENS, 2014). Suponha que os dados a serem reduzidos consistam em vetores de dados descritos por n atributos. Com a aplicação da PCA procura-se por k vetores ortogonais n-dimensionais que podem representar os dados, onde $k \leq n$. Os dados originais são, portanto, projetados em um novo conjunto de dados com uma quantidade menor de atributos (HAN; PEI; KAMBER, 2011). Nos experimentos, a PCA foi aplicada para reduzir a dimensionalidade em 25% e 50%. Assim, a partir das bases que possuíam 12 atributos (excluindo a variável classe), utilizando a PCA, geraram-se bases que possuíam 9 e 6 atributos, respectivamente. Os atributos gerados com a PCA representam a variância dos registros anteriores, porém, são descorrelacionados. A redução da dimensionalidade nas outras bases geradas não se fez necessária, uma vez que essas já tinham o tamanho reduzido; e

- **procedimento F**: agrupamento das estações de aluguéis de bicicleta. Para definir os agrupamentos das estações foram utilizadas duas abordagens. A primeira considerou uma divisão geográfica já existente da cidade de Washington DC, por quadrantes: noroeste (*Northwest* - NW), nordeste (*North East* - NE), sudeste (*Southeast* - SE) e sudoeste (*South-west* - SW). A segunda abordagem utilizou centróides sugeridos por um algoritmo de clusterização. Ambas as abordagens são detalhadas no Capítulo 6.

[6] scikit-learn.org

4.3 Novas Bases Geradas

Os procedimentos descritos na Seção 4.2 foram usados individualmente e em conjunto para gerar quatorze novas bases. No Quadro 4 são mostradas as bases que foram geradas, especificando os procedimentos aplicados em cada uma delas. A Base 1 representa a base de dados original, somente com a remoção de registros incompletos, e foi utilizada para identificar padrões nos hábitos dos usuários (Seção 5.1). Na Base 2 a variável *season* foi ajustada e a partir dela as Bases de 3 a 10 foram geradas, combinando os procedimentos de pré-processamento para identificar aqueles que resultavam em modelos preditivos com melhor desempenho. A variável classe não é bem distribuída e a grande maioria dos registros estão nas categorias "Muito Baixa" e "Baixa". Nas Bases 11 a 14, nas quais os agrupamentos das estações de aluguéis foram aplicados, utilizou-se o método *cut* e dispensou-se o PCA. Detalhes sobre o agrupamento das estações de aluguéis são apresentados no Capítulo 6. Todas as novas bases de dados geradas estão disponíveis[7] para o desenvolvimento de estudos posteriores.

Quadro 4 – Procedimentos aplicados em cada base gerada.

	Procedimento							**Tipo de agrupamento**
	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	-
Base 11	X	X	X		X			por quadrante
Base 12	X	X	X		X			com 8 centroides
Base 13	X	X	X		X			com 13 centroides
Base 14	X	X	X		X			com 20 centroides

4.4 Análises em Nível de Cidade e Regiões

A análise em nível de cidade é detalhada no Capítulo 5, onde são identificados alguns padrões de utilização dos usuários do sistema, que foram úteis para auxiliar no entendimento da dinâmica de funcionamento desse sistema. Também são apresentadas análises de correlação entre os atributos das bases de dados utilizadas e é apresentada uma análise comparativa de modelos preditivos de regressão e classificação, bem como a aplicação de comitês homogêneos e heterogêneos.

[7] Disponível em: <https://github.com/johnattandouglas/Dissertacao>

A análise em nível de agrupamento é detalhada no Capítulo 6, onde são apresentados modelos descritivos utilizando *k-means* para agrupamento das estações de aluguéis por proximidade, que são aplicados em conjunto com os modelos voltados para a previsão de demanda. Avaliou-se a aplicação desses agrupamentos nas estações de aluguéis como abordagem para melhoria no desempenho dos modelos preditivos gerados. Após uma análise em nível de cidade e de regiões, é proposto um cenário de aplicação dos modelos preditivos gerados (Capítulo 7).

4.5 Modelos Preditivos

Para treinar e testar os algoritmos de regressão e classificação foi utilizada o método de validação cruzada em 10 partes, de modo que a mesma base foi usada para treinamento e teste dos algoritmos (JAMES *et al.*, 2013). A validação cruzada foi escolhida pelo fato de utilizar toda a base de dados, conforme explicado na Seção 2.2. Como cada mês e dia possuem demandas específicas, garantiu-se que registros de todo o ano fossem utilizados.

Foram gerados modelos preditivos utilizando Python, por meio do ambiente Jupyter[8]. Para a tarefa de regressão, seis regressores da biblioteca *Scikit Learn* (PEDREGOSA *et al.*, 2011) foram aplicados: *Linear Regression* (LR), *Support Vector Regression* (SVR), *K-Nearest Neighbors* (KNN), *MultiLayer Perceptron* (MLP), *Random Forest Regressor* (RFR) e *Decision Tree* (DT). Os parâmetros utilizados nos modelos de regressão são apresentados no Quadro 5.

Quadro 5 – Parâmetros utilizados nos modelos de regressão.

Algoritmo	**Parâmetros**
Linear Regression (LR)	fit_intercept = True, normalize = False, copy_X = True, n_jobs = None, positive = False
Support Vector Regression (SVR)	kernel = 'rbf', degree = 3, gamma = 'scale', coef0 = 0.0, tol = 0.001, C = 1.0, epsilon = 0.1, shrinking = True, cache$_size = 200, verbose = False, max_iter = -1$
K-Nearest Neighbors (KNN)	n_neighbors = 5, weights = 'uniform', algorithm = 'auto', leaf_size = 30, p = 2, metric = 'minkowski', metric_params = None, n_jobs = None,
MultiLayer Perceptron (MLP)	hidden_layer_sizes = 200, learning_rate = 'constant', activation = 'relu', solver = 'adam', alpha = 0.0001, batch_size = 'auto', max_iter = 200, power_t = 0.5, random_state = None
Random Forest Regression (RFR)	n_estimators = 100, criterion = 'mse', max_depth = None, min_samples_split = 2, min_samples_leaf = 1, min_weight_fraction_leaf = 0.0, max_features = 'auto', max_leaf_nodes = None
Decision Tree (DT)	criterion = 'mse', splitter = 'best', max_depth = None, min_samples_split = 2, min_samples_leaf = 1, min_weight_fraction_leaf = 0.0, max_features = None, random_state = None

[8] jupyter.org

Foram utilizados parâmetros padrões da biblioteca *Scikit Learn* para cada modelo de regressão. As métricas do R-quadrado (R^2), erro absoluto médio (*Mean Absolute Error*-MAE) e raiz quadrada do erro quadrático médio (*Root Mean Square Error* - RMSE) foram consideradas para avaliar o desempenho dos modelos preditivos de regressão (Capítulo 2).

Para a tarefa de classificação, também utilizou-se os algoritmos da biblioteca *Scikit Learn* (PEDREGOSA *et al.*, 2011): *Support Vector Classifier* (SVC), *Gaussian Naive Bayes* (NB), *k-Nearest Neighbors* (KNN), *Multi-layer Perceptron* (MLP), *Random Forest Classifier* (RFC) e *Decision Tree* (DT). Foram utilizados parâmetros padrões da biblioteca em cada algoritmo de classificação (Quadro 6).

Quadro 6 – Parâmetros utilizados nos modelos de classificação.

Algoritmo	**Parâmetros**
Support Vector Classifier (SVC)	C = 1.0, kernel = 'rbf', degree = 3, gamma =' scale'
Gaussian Naive Bayes (NB)	priors = None, var_smoothing = 1e-09
K-Nearest Neighbors (KNN)	n_neighbors = 5, weights = 'uniform', algorithm = 'auto', leaf_size = 30, p = 2, metric = 'minkowski', metric_params = None, n_jobs = None,
MultiLayer Perceptron (MLP)	hidden_layer_sizes = 200, activation = 'relu', solver = 'adam', alpha = 0.0001, batch_size = 'auto', learning$_r ate =' constant', power_t = 0.5, max_iter =$ $200, random_state = None$,
Random Forest Classifier (RFC)	n_estimators = 100, criterion = 'gini', max_depth = None, min_samples_split = 2, min_samples_leaf = 1, min_weight_fraction_leaf = 0.0, max_features = 'auto', max_leaf_nodes = None, bootstrap = True, n_jobs = None, random_state = None
Decision Tree (DT)	criterion = 'mse', splitter = 'best', max_depth = None, min_samples_split = 2, min_samples_leaf = 1, min_weight_fraction_leaf = 0.0, max_features = None, random$_s tate = None$

Como utilizou-se validação cruzada em 10 partes, considerou-se as 10 execuções dos experimentos para calcular a média da acurácia (%), do desvio padrão da acurácia, da precisão (%) e do *Recall* (%) para cada um dos modelos de classificação. Para a comparação dos múltiplos classificadores foram aplicados os teste de Friedman (1937) e Nemenyi (1963) (Capítulo 2).

A biblioteca *Scikit Learn* (PEDREGOSA *et al.*, 2011) ainda foi utilizada para gerar comitês homogêneos (por meio do *Bagging Classifier*), construindo um conjunto de modelos utilizando um algoritmo de aprendizagem simples por meio da combinação por votos para classificação (BREIMAN, 1996). Para geração de comitês heterogêneos, utilizou-se o *StackingClassifier* da biblioteca *Mlxtend*[9], treinando os modelos de classificação com base no conjunto de treinamento completo; então, um meta-classificador é ajustado com base nos resultados dos modelos de classificação individuais (TANG; ALELYANI; LIU, 2015).

[9] <rasbt.github.io/mlxtend/user_guide/classifier/StackingClassifier/>

Os *scripts*[10] criados para os experimentos foram executados em aproximadamente 187 horas de processamento em uma máquina com Windows 10 Home de 64 bits, com processador Intel core i5 e 8GB de RAM.

4.6 Considerações Finais

Como apresentado, a metodologia utilizada nesta pesquisa se assemelha ao processo clássico de KDD, focando em extrair conhecimento de uma base de dados do *Capital BikeShare* enriquecida com informações meteorológicas e sazonais, da qual foram geradas novas bases de dados com várias combinações de procedimentos de pré-processamento. A abordagem proposta é genérica e pode ser aplicada aos mais diversos tipos de sistemas de transporte compartilhados com o intuito de extrair padrões de mobilidade. A respeito dos SBC, com o conhecimento extraído por meio de modelos preditivos de demanda, tanto em nível de cidade como em nível de regiões, propõe-se uma abordagem para apoio a tomada de decisão dos gestores desses sistemas. As análises que foram realizadas são apresentadas nos próximos capítulos.

[10] https://github.com/johnattandouglas/Dissertacao/tree/main/Scripts

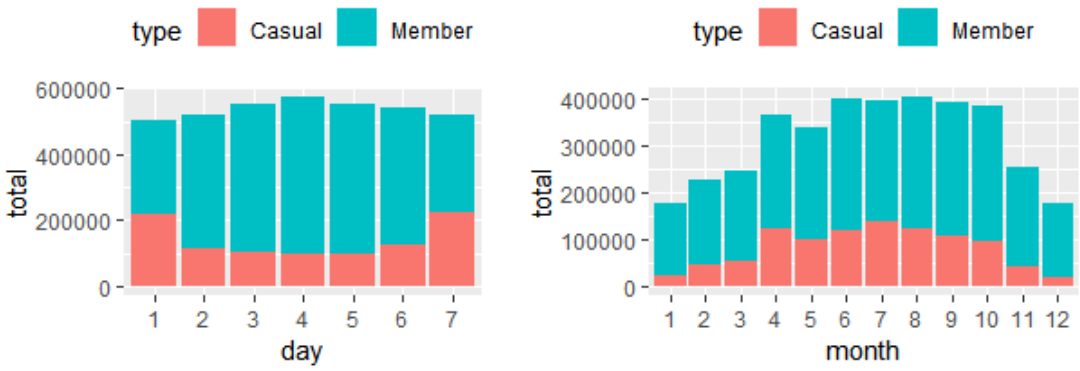
5 Análise em Nível de Cidade

Neste capítulo são apresentados padrões relacionados à utilização das bicicletas, em nível de cidade, que foram identificados na base de dados original (Seção 5.1). Além disso, é apresentada uma análise de correlação entre os atributos das bases de dados utilizadas (Seção 5.2). Com os aluguéis agrupados por hora, são também apresentadas análises comparativas de modelos preditivos de regressão (Seção 5.3) e de classificação (Seção 5.4) para os serviços de aluguel de bicicleta em nível de cidade.

5.1 Identificação de Padrões

Para as análises apresentadas nesta seção foi usada a base de dados original com informações adicionais, como clima e feriados (Seção 4.1). Analisando o atributo *type*, foi possível observar uma discrepância notável na quantidade de aluguéis de usuários casuais (sem registro no serviço) e registrados. No total, o número de aluguéis feitos por usuários registrados (2.775.979 aluguéis) representa quase o triplo do quantitativo realizado por usuários casuais (981.798 aluguéis). A duração média dos aluguéis de usuários registrados é de aproximadamente 12 minutos, enquanto que de usuários casuais é de aproximadamente 39 minutos. Assim, embora o número de aluguéis casuais seja inferior, eles tendem a ser mais longos, durando em média três vezes mais do que os aluguéis de usuários registrados.

Também identificou-se uma relação entre o atributo *type* e os dias da semana. Observou-se que os aluguéis casuais ocorrem em maior número aos sábados e domingos, enquanto que a quantidade de aluguéis dos usuários registrados é maior no meio da semana (especificamente às quartas-feiras). No gráfico da Figura 3(a) é representada a quantidade de aluguéis de cada tipo por dia da semana.



(a) Total em cada dia da semana.

(b) Total em cada mês.

Figura 3 – Quantitativo de aluguéis por tipo de usuário.

Analisando os valores de aluguéis por mês, observou-se que junho, julho, agosto, setembro e outubro são os meses com maior quantidade de aluguéis, conforme mostrado no gráfico da Figura 3(b). Janeiro e dezembro representam os meses com os menores quantitativos. Essa diminuição pode estar relacionada à temperatura média desses meses, que costumam ser os mais frios do ano na cidade de Washington DC.

Investigando a quantidade de aluguéis por estação do ano, verificou-se que no inverno ocorreram 572.440 aluguéis (15,2%), com uma média de 266 aluguéis por hora e cerca de 6.360 aluguéis por dia. Na primavera, aconteceram 1.072.608 aluguéis (28,5%), com uma média de 487 aluguéis por hora e cerca de 11.658 aluguéis por dia. No verão, foram registrados 1.197.255 (31,9%), com uma média de 542 aluguéis por hora e aproximadamente 13.013 aluguéis por dia. No outono ocorreram 915.474 (24,4%) aluguéis, com média de 419 por hora e cerca de 10.060 aluguéis por dia. Assim, o inverno é a estação com menos aluguéis e o verão é a estação com mais aluguéis, justificando os picos nos meses dessa estação, como mostrado anteriormente na Figura 3(b).

Existe uma relação proporcional entre *duration* e *season*, visto que a duração média dos aluguéis costuma ser maior nas estações do ano com mais aluguéis e menor nas estações do ano com menos aluguéis. Portanto, os aluguéis tendem a ser mais longos no verão. Na Figura 4 pode-se observar que na maioria dos meses a quantidade de aluguéis varia durante o mês, exceto nos meses do verão (junho a agosto), que o quantitativo permanece alto durante esse período.

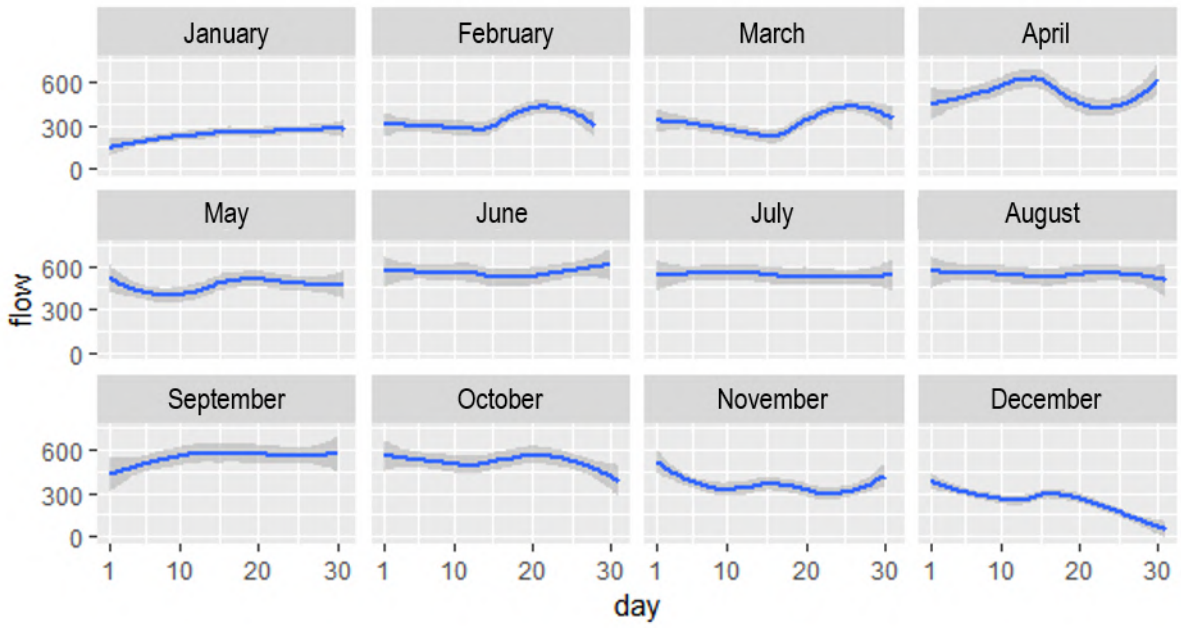
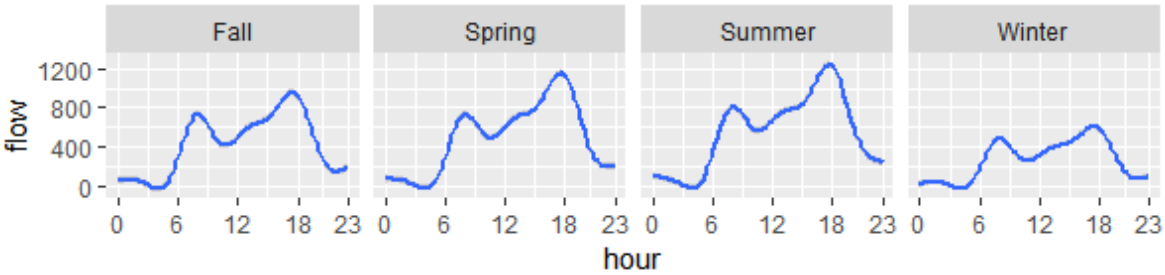


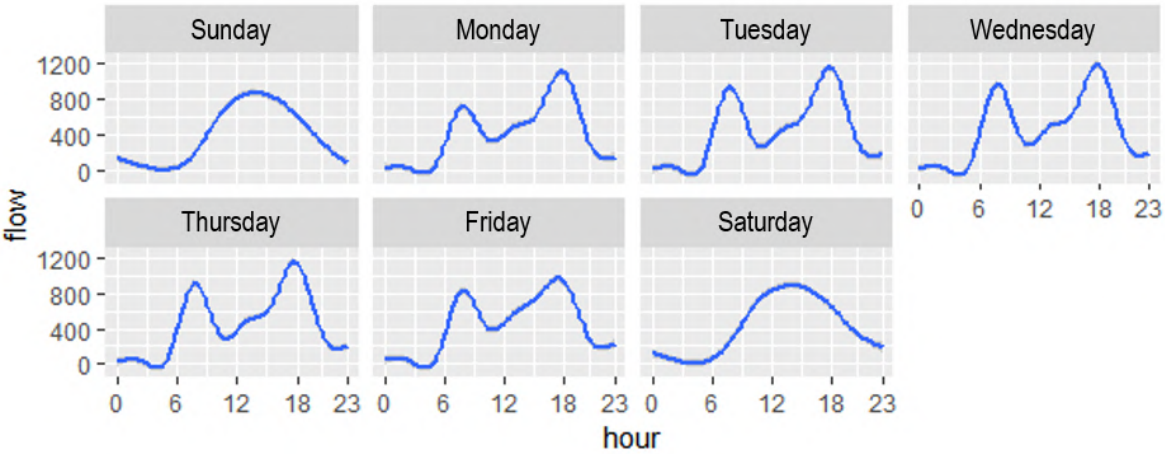
Figura 4 – Quantitativo de aluguéis por dia do mês.

Além disso, observou-se uma semelhança na quantidade de aluguéis por hora do dia, independente da estação do ano. Na Figura 5(a) são apresentados dois horários de pico em torno de 07:30 e 17:30. No inverno (janeiro a março), mesmo com a quantidade menor de aluguéis, os horários de pico ainda se assemelham às outras estações. Esses horários geralmente estão relacionados à quando as pessoas vão trabalhar e voltam para casa. Isso evidencia que, independente da estação do ano, a maioria dos usuários do *Capital BikeShare* o utiliza para o trajeto casa-trabalho.

Com relação ao atributo *hour*, a quantidade de aluguéis é maior nos momentos de maior movimentação na cidade. Como esperado, este quantitativo não é alto ao amanhecer e durante a madrugada. Um padrão de horário nos finais de semana e dias de semana foi identificado (Figura 5(b)). De segunda a sexta-feira, são evidentes dois picos de aluguel, que acontecem por volta das 07:30 e das 17:30. Aos sábados e domingos, verifica-se um aumento contínuo dos aluguéis a partir das 07:30 e uma diminuição a partir das 15:00. Nos feriados, os padrões de demanda em cada hora são semelhantes aos encontrados nos finais de semana. Os feriados também estão incluídos na Figura 5(a), porém, por serem menores em relação aos dias úteis, não foram suficientes para influenciar as curvas apresentadas nos gráficos, que são semelhantes às apresentas na Figura 5(b), de segunda à sexta.



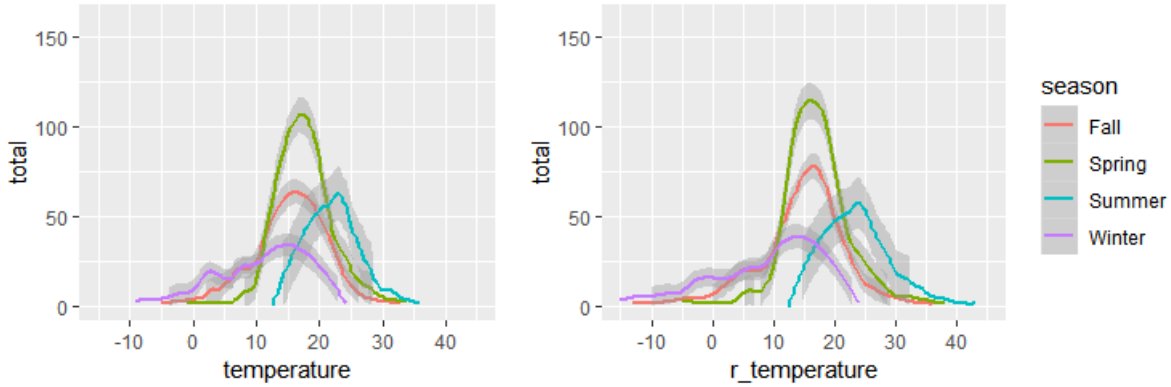
(a) Quantitativo de aluguéis por hora em cada estação do ano.



(b) Quantitativo de aluguéis por hora em cada dia da semana.

Figura 5 – Quantitativo de aluguéis por hora.

Também encontraram-se padrões no que diz respeito às variáveis climáticas. A análise da temperatura e da sensação de temperatura revelou uma relação desses dois atributos com a quantidade de aluguéis. Como mostrado na Figura 6, o maior fluxo de aluguéis ocorre na primavera e no verão, que são períodos de temperaturas amenas. A análise do quantitativo de aluguéis por estação do ano evidencia que conforme a intensidade desses dois atributos aumenta ou diminui, a quantidade de aluguéis se modifica. Assim, os usuários tendem a fazer menos aluguéis quando está extremamente frio ou muito quente. Na Figura 6 é demonstrado que, embora haja uma sutil variação entre os parâmetros *temperature* e *r_temperature*, a quantidade de aluguéis é maior entre 15°C e 20°C.



(a) Aluguéis por temperatura. (b) Aluguéis por sensação térmica.

Figura 6 – Relação do clima com o quantitativo de aluguéis em cada estação do ano.

Para entender a influência dos feriados na dinâmica dos aluguéis, foi feita uma comparação entre os tipos de usuários em cada uma nessas datas (Figura 7). Em alguns feriados, a quantidade de aluguéis de usuários casuais se aproxima ou supera a quantidade de aluguéis de usuários registrados.

Acredita-se que isso aconteça devido ao aumento do fluxo de visitantes na cidade nesses dias, considerando o potencial turístico dessas datas. Por exemplo, nos feriados do *Independence Day* e do *Memorial Day*, há um aumento significativo nos aluguéis, principalmente em estações próximas a monumentos históricos ou turísticos, como a Casa Branca e o Parque Nacional. No caso do *Independence Day*, o *National Park Service* (NPS) incentiva fortemente os turistas a usarem o transporte público para atividades na cidade, uma vez que o estacionamento público é extremamente limitado e os carros não são permitidos dentro ou ao redor do *National Mall* (NPS, 2020). Além disso, como várias estradas estão interditadas ao redor da *George Washington Memorial Parkway*, acredita-se que essa seja a principal razão para a grande quantidade de aluguéis neste dia. Por outro lado, em alguns outros feriados, a quantidade de aluguéis diminui, como é o caso do Natal e Ano Novo. Essa diminuição na demanda pode ser resultante das baixas temperaturas que ocorrem nesse período do ano.

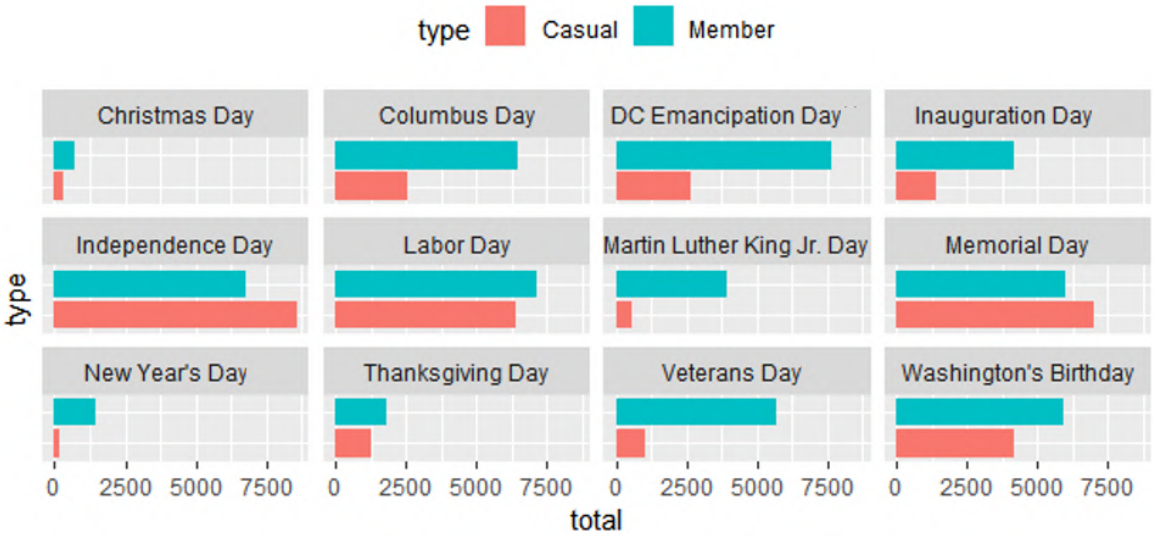


Figura 7 – Quantidade de aluguéis em cada feriado.

Destaca-se ainda a relação entre a duração do aluguel (*m_duration*) e a distância percorrida (*m_distance*). Nessa relação, apresentada em um gráfico de dispersão na Figura 8, foram identificados dois agrupamentos de perfis de aluguel, um que costuma ser mais rápido em que os usuários percorrem distâncias maiores, e outro que costuma demorar mais e as distâncias pecorridas são menores.

O usuário que aluga uma bicicleta para ir ao trabalho, por exemplo, faz o trajeto o mais rápido possível. Com base nas análises, acredita-se que outros usuários (como turistas) alugam a bicicleta por mais tempo, mas devolvem a bicicleta nas estações próximas, podendo mesmo devolvê-la no mesmo local do aluguel, resultando em distâncias mais curtas. No entanto, como é evidenciado na Figura 8, *type* não parece ser o atributo responsável por diferenciar esses dois grupos.

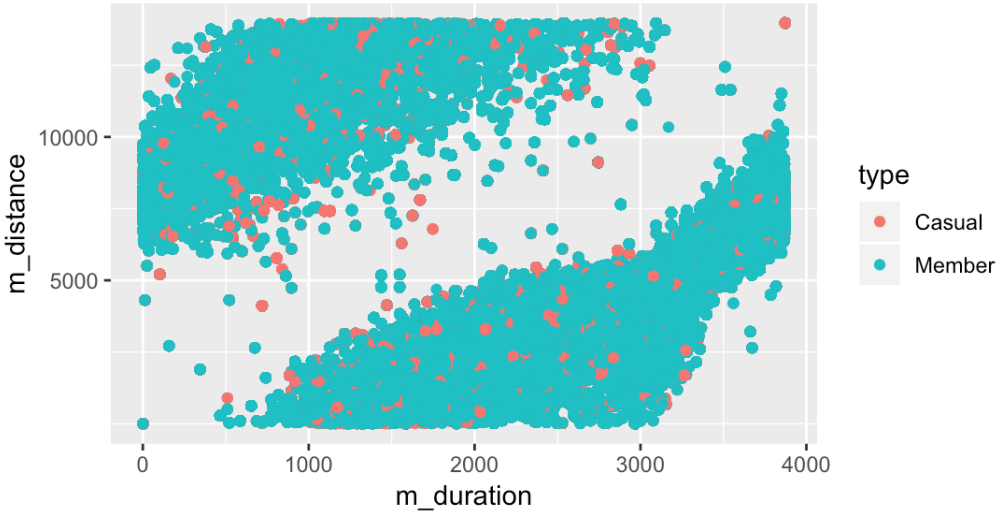


Figura 8 – Relação entre distância e duração do aluguel.

Nas análises, também foram testados os atributos *day*, *weekday* e *season*, no entanto, nenhuma delas explica efetivamente o agrupamento. Ao comparar o aluguel médio *m_duration* com *m_distance* baseado em *start_station* e *end_station* (calculado usando as coordenadas), percebeu-se que muitos aluguéis demoraram muito mais do que o esperado. Isso acontece principalmente nas estações próximas aos parques, durante feriados ou finais de semana. Assim, acredita-se que esses perfis de aluguel estejam relacionados à finalidade do aluguel, embora eventualmente os usuários cadastrados façam viagens de lazer e os ocasionais façam a trajetória casa-trabalho-casa.

5.2 Análise de Correlação

A análise de correlação foi realizada nas bases de dados para uma verificação da relação entre os atributos. 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 calculado por meio dos coeficientes de Pearson (Equação 5.1). Quanto mais próximo de 1 essa medida for, 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 é próximo de -1, existe uma correlação inversa (quanto maior uma, menor a outra). Contudo, quando o coeficiente de correlação é 0, não há correlação entre as variáveis (BUSSAB; MORETTIN, 2010).

$$\rho = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{5.1}$$

As matrizes de correlações geradas por esse método foram similares para as Bases 1 a 6 (Apêndice A). Para evitar redundância neste capítulo, na Figura 9 é mostrada somente a matriz de correlação da Base 6, que foi colorida conforme um intervalo de cores, no qual: a cor mais clara representa a correlação direta e a escura a correlação inversa. Valores próximos de zero estão coloridos em roxo, indicando a inexistência de correlação.

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. Nas Bases 7 a 10, geradas com a PCA com redução de 25% e 50%, não foi visualizada correlação entre os atributos (Apêndice C), o que era esperado, já que novos atributos (totalmente diferentes dos originais) foram gerados com a aplicação dessa técnica.

Na Figura 9 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

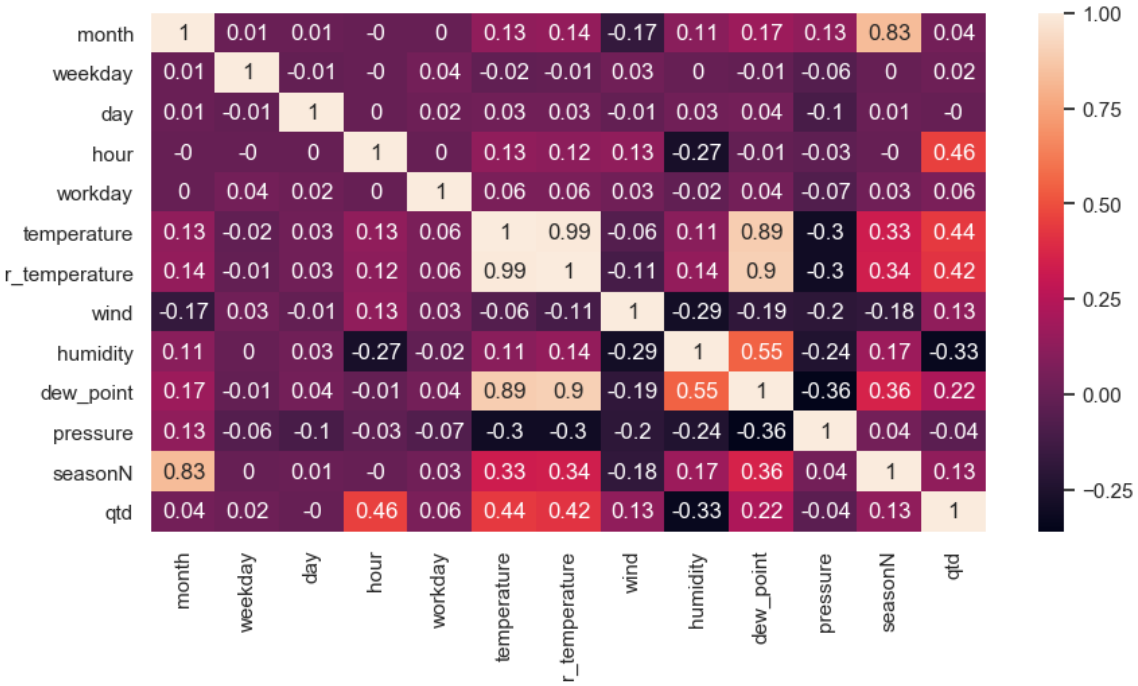


Figura 9 – Matriz de correlação (Base 6).

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, El-Assi e Habib (2015) no sistema de compartilhamento de bicicletas de 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).

5.3 Análise Comparativa de Modelos de Regressão

Como apresentado no Capítulo 4, algoritmos de regressão foram aplicados para gerar modelos preditivos da quantidade de aluguéis (atributo *qtd*) com o objetivo de estimar a demanda de aluguéis em cada hora. Esse tipo de estimativa é fundamental para os provedores desse serviço, uma vez que com essas informações eles conseguem organizar os recursos (bicicletas) para compartilhamento, identificando qual a demanda horária.

Na Tabela 3 é mostrado o desempenho de cada algoritmo de regressão na base de dados original em relação às métricas erro absoluto médio (*Mean Absolute Error*- MAE) e raiz quadrada do erro quadrático médio (*Root Mean Square Error* - RMSE).

De acordo com a Tabela 3, pode-se dizer que o modelo preditivo do SVR obteve menor desempenho em todas as métricas de avaliação. O LR também não obteve um bom

Tabela 3 – Comparação de desempenho entre regressores.

Regressor	**R^2**	**MAE**	**RMSE**
LR	0,390	244,102	320,275
SVR	0,240	249,340	357,577
KNN	0,592	182,405	262,098
MLP	0,595	189,649	261,126
RFR	**0,947**	**56,542**	**94,394**
DT	0,901	74,377	128,986

desempenho. Isso era esperado porque o fluxo de aluguel não obedece uma linearidade, conforme mostrado nas Figuras 5(a) e 5(b). Foi possível perceber a dinamicidade da quantidade de aluguéis em relação a diversos atributos, principalmente hora e dia da semana, que apresentam um padrão não-linear.

Os algoritmos baseados no ganho de informação, como RFR e DT, apresentaram melhores resultados, principalmente o RFR, que foi capaz de explicar aproximadamente 95% das demandas de aluguéis por hora por meio das variáveis independentes utilizadas nos respectivos modelos. A partir das árvores aleatórias geradas pelo algoritmo RFR, pôde-se calcular a relevância dos atributos (*feature importance*) pela porcentagem de aumento do erro quadrado médio (*Increased Mean Square Error* - IMSE), que descreve a capacidade preditiva do RMSE com atributos trocados aleatoriamente. Se essa permutação alterar drasticamente o valor previsto, o atributo é considerado crítico.

Também foi calculada a pureza do nó aumentada (*Increased Node Purity* - INP), que mede a função de perda quando os melhores nós da árvore são selecionados, revelando os atributos mais significativos para a predição (ECHEVERRY-GALVIS; PETERSON; CACERES, 2014). Na Figura 10 é representada a relevância dos atributos no modelo regressor construído com o algoritmo RFR.

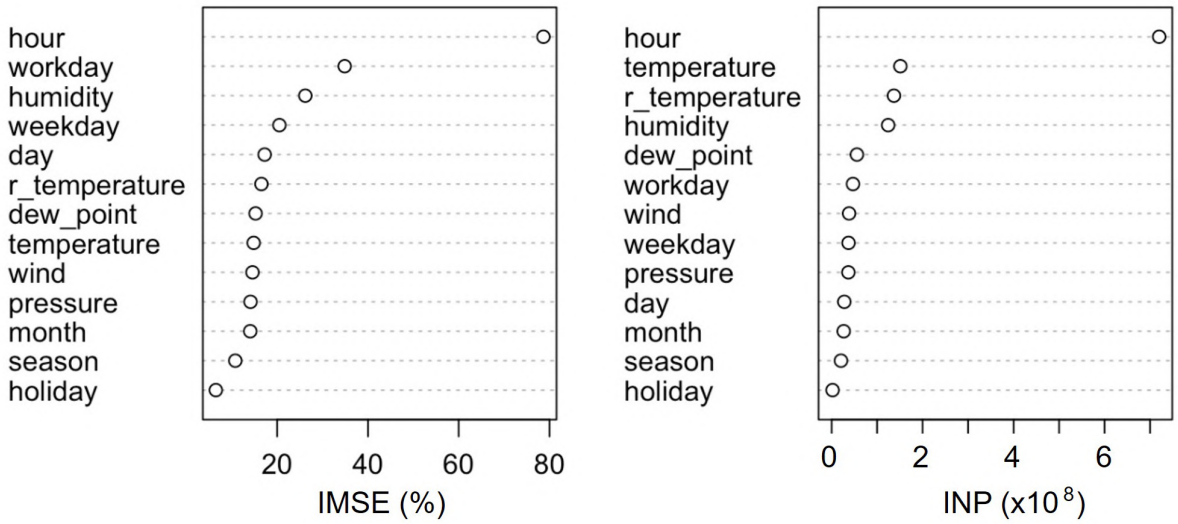


Figura 10 – Relevância dos atributos nos modelos preditivos gerados pelo algoritmo RFR.

Como exposto, os atributos identificados como relevantes na Seção 5.1 são os mais influentes nas árvores geradas pela floresta aleatória. Também pode ser visto que os valores de IMSE e INP diferem muito entre si, com exceção de *hour*, que é a raiz da árvore de acordo com ambas as métricas.

5.4 Análise Comparativa de Modelos de Classificação

Para os experimentos utilizou-se os seguintes algoritmos de classificação apresentados no Capítulo 4. Além disso, também foram aplicados comitês de algoritmos usando um conjunto dos mesmos classificadores (comitês homogêneos) e de diferentes classificadores (comitês heterogêneos) (KUNCHEVA, 2014).

O desempenho dos classificadores de acordo a acurácia (%), o desvio padrão da acurácia, a precisão (%) e o *Recall* (%) é mostrado na Tabela 4, com aproximação de duas casas decimais. Em negrito, está destacada a melhor acurácia obtida para cada uma das bases. A árvore de decisão (DT) foi o algoritmo com melhor desempenho nas bases onde a redução de dimensionalidade não foi aplicada (Bases 3 a 6). Nas outras (Bases 7 a 10), o MLP e o SVC se destacaram. De modo geral, os algoritmos não apresentaram bons resultados nas bases 9 e 10, mesmo a melhor acurácia obtida na Base 10 (47,23% com o MLP) foi inferior a 50%. Nas bases onde a redução de dimensionalidade foi aplicada (7 a 10) os maiores resultados alcançados para a precisão e o *recall* foram 67,91% e 65,68%, respectivamente.

A precisão e o *recall* apresentaram o maior resultado nas Bases 4 e 6 com o algoritmo DT. Porém, as melhores acurácias foram alcançadas nas Bases 3 e 5 com esse algoritmo, mesmo percebendo-se uma baixa nos resultados nas métricas de precisão e *recall*, que podem ser explicadas pelo método de discretização *cut* utilizado nessas bases (Seção 4.2), que gera quantidades desbalanceadas de registros em cada classe.

Na Figura 11 são representadas as acurácias dos algoritmos aplicados nas Bases 3 a 10. Os gráficos do lado esquerdo são das bases onde a discretização da variável classe (*qtd*) foi feita utilizando *cut*, e as do lado direito utilizando *qcut* (procedimento C da etapa de pré-processamento). A discretização por meio do método *cut* resultou em melhores resultados na acurácia dos algoritmos aplicados. Além disso, é possível perceber que a normalização (procedimento D da etapa de pré-processamento) diminuiu a variação da acurácia nos modelos gerados. É notável também que a PCA (procedimento E da etapa de pré-processamento) 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. Desta forma, observa-se que, neste caso, a técnica de redução da dimensionalidade aplicada com o PCA não melhorou o desempenho dos algoritmos de classificação aplicados.

Tabela 4 – Comparação de desempenho entre classificadores nas Bases 3 a 10.

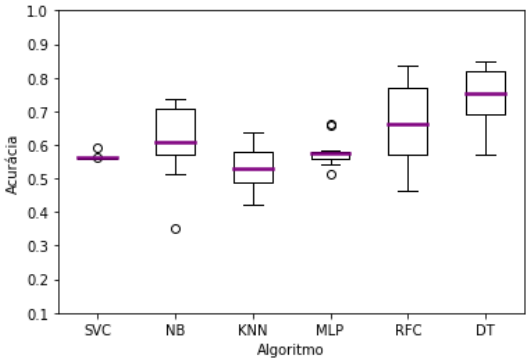
BASE	SVC			NB		
	Acurácia ± DP	Precisão	Recall	Acurácia ± DP	Precisão	Recall
3	56,38 ± 0,01	12,75	20,29	60,94 ± 0,11	39,29	46,49
4	36,47 ± 0,04	37,44	36,45	44,63 ± 0,07	43,58	44,61
5	66,32 ± 0,07	33,84	33,91	60,94 ± 0,11	39,29	46,49
6	50,55 ± 0,06	52,01	50,54	44,63 ± 0,07	43,58	44,61
7	**65,51** ± 0,07	33,96	33,57	63,31 ± 0,11	36,80	40,47
8	**63,35** ± 0,05	30,16	31,83	62,90 ± 0,04	33,01	35,40
9	49,68 ± 0,07	50,40	49,15	44,27 ± 0,07	42,65	44,26
10	45,63 ± 0,06	45,17	45,63	45,00 ± 0,04	43,41	45,00

BASE	KNN			MLP		
	Acurácia ± DP	Precisão	Recall	Acurácia ± DP	Precisão	Recall
3	52,91 ± 0,07	25,47	25,35	54,80 ± 0,06	28,11	31,98
4	35,07 ± 0,04	35,52	35,05	41,13 ± 0,05	36,13	40,94
5	46,48 ± 0,09	24,25	23,10	71,88 ± 0,09	51,47	44,09
6	34,87 ± 0,08	37,90	34,85	66,92 ± 0,10	69,43	66,55
7	45,95 ± 0,08	25,78	23,91	73,69 ± 0,08	61,53	57,81
8	47,87 ± 0,06	25,64	23,74	61,79 ± 0,04	34,48	32,58
9	34,42 ± 0,06	35,58	34,41	35,51 ± 0,09	67,91	65,68
10	31,79 ± 0,04	49,09	31,77	**47,23** ± 0,05	49,09	46,26

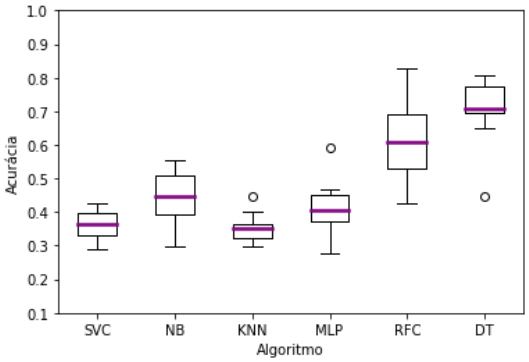
BASE	RFC			DT		
	Acurácia ± DP	Precisão	Recall	Acurácia ± DP	Precisão	*Recall*
3	65,44 ± 0,12	47,48	44,58	**75,08** ± 0,09	63,32	64,64
4	60,14 ± 0,13	63,60	60,92	**70,65** ± 0,09	71,95	70,34
5	65,90 ± 0,12	46,55	44,82	**74,63** ± 0,09	63,51	64,45
6	60,74 ± 0,12	63,86	59,73	**70,77** ± 0,09	71,83	70,52
7	61,55 ± 0,11	33,02	32,87	62,97 ± 0,07	43,49	42,03
8	50,79 ± 0,09	23,80	24,80	51,31 ± 0,06	29,58	29,24
9	51,37 ± 0,11	53,87	51,25	**53,49** ± 0,09	55,10	52,92
10	35,20 ± 0,08	35,38	34,78	34,79 ± 0,06	38,20	35,15

O teste de Friedman (abordado no Capítulo 2) foi aplicado para comparar o desempenho dos múltiplos classificadores apresentados na Tabela 4. No teste de Friedman, a hipótese nula (H_0) considera que todos os algoritmos são equivalentes e suas classificações médias são iguais, enquanto que a hipótese alternativa (H_1) considera que pelo menos um algoritmo tem desempenho superior aos demais. O teste de Friedman para os classificadores nas Bases 3 a 10 resultou em um valor-P = 0,0004, evidenciando que existe diferença significativa nas acurácias alcançadas nestas bases.

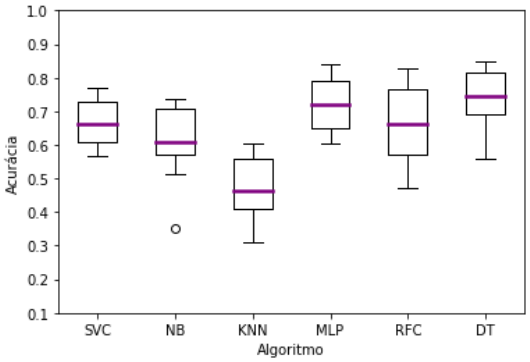
A diferença crítica entre os modelos de classificação foi verificada por meio do teste de Nemenyi. Para as acurácias mostradas na Tabela 4, as diferenças entre o SVC e o MLP se mostraram estatisticamente indistinguíveis, como ilustrado no Apêndice C. A análise também evidenciou a proximidade estatística entre os resultados alcançados com o NB e o RFC, bem como também explicitou o baixo desempenho do KNN.



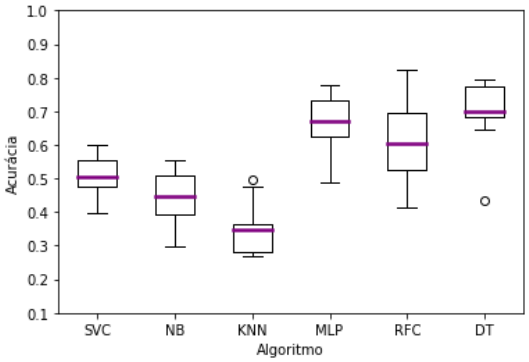
(a) Base 3 (*cut*).



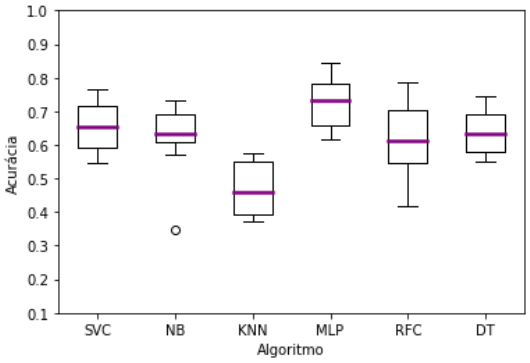
(b) Base 4 (*qcut*).



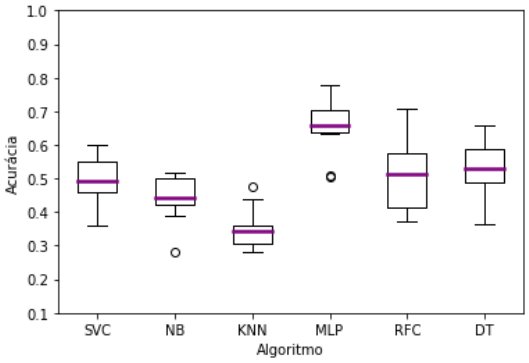
(c) Base 5 (*cut* e normalização).



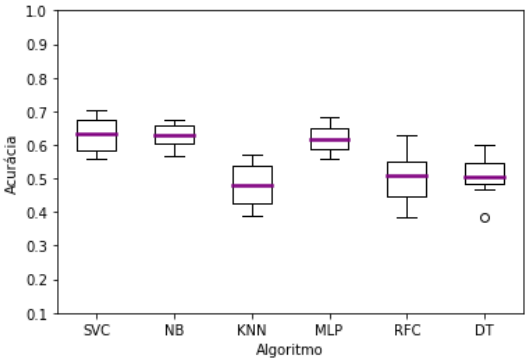
(d) Base 6 (*qcut* e normalização).



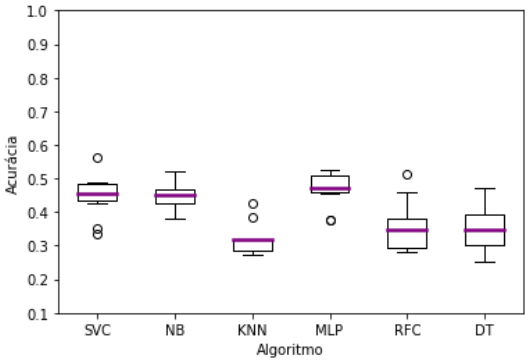
(e) Base 7 (*cut*, normalização e PCA 25%).



(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 11 – Acurácia dos classificadores aplicados nas Bases 3 a 10.

Para analisar a razão dos baixos resultados alcançados com as métricas de precisão e *recall* nos experimentos, foram gerados gráficos de curva ROC para ponderar quão bem os modelos podem distinguir as classes. Para isso, foi calculada a área sob a curva (AUC) de cada classe para cada algoritmo (Apêndice D). A curva ROC geralmente é utilizada para problemas de classificação binária. No entanto, neste caso, tratando-se de cinco classes que representam as quantidades de aluguéis ("Muito baixa", "Baixa", "Média", "Muito alta", "Alta") adaptou-se a abordagem realizando uma comparação pareada para a construção da matriz de confusão (LANDGREBE; DUIN, 2007).

Analisando os gráficos de curva ROC apresentados no Apêndice D, é possível perceber que nas bases discretizadas com o método *cut*, ou seja, nas classes desbalanceadas (Bases 3, 5, 7 e 8), os algoritmos SVC, NB, MLP e RFC possuem maiores AUC para a classe de quantidade "Muito alta" de aluguéis (classe 5) em relação às outras classes. Nesses mesmos algoritmos nessas bases desbalanceadas, as classes "Baixa" (classe 2) e "Média" (classe 3) tiveram as AUC menores, o que indica que os algoritmos distinguiram erroneamente essas duas classes. Analisando os gráficos de curva ROC do algoritmo KNN percebe-se que nas bases desbalanceadas, com exceção da classe "Muito baixa" e "Baixa", a classificação foi feita quase que aleatoriamente (AUC próximo de 0,5). Em relação ao algoritmo DT, notou-se que a aplicação da PCA diminuiu a AUC de todas as classes.

Em relação as bases discretizadas com o método *qcut*, ou seja, as bases balanceadas (Bases 4, 6, 9 e 10), percebe-se que em todos os algoritmos a classe "Muito baixa" (classe 1) tem AUC maior, o que indica que esta classe foi identificada corretamente pelos algoritmos. No entanto, notou-se um curva menor para as classes "Média" (classe 3) e "Alta" (classe 4), o que indica que os algoritmos SVC, KNN, RFC e DT distinguiram essas duas classes de maneira equivocada. Se tratando no KNN especificamente, essas duas classes foram classificadas quase aleatoriamente. Nas bases onde a PCA foi aplicada, destaca-se uma diminuição da AUC em todas as classes. Na Base 9, especificamente, a classe "Baixa" (classe 2) foi indevidamente classificada por todos os algoritmos com a aplicação da PCA. Além disso, nesta base, as classes "Média" (classe 3) e "Alta" (classe 4) passaram a ser classificadas erroneamente pelo algoritmo MLP, enquanto que as classes 3 e 4 foram classificadas quase aleatoriamente nos algoritmos SVC, KNN, RFC e DT.

5.4.1 Aplicação de Comitês

Na Base 5, na qual obtiveram-se as melhores acurácias com os algoritmos isolados, foram aplicados os comitês homogêneos utilizando o *BaggingClassifier* (BREIMAN, 1996). Foram aplicados comitês homogêneos de tamanho $T = \{5, 10, 15, 20\}$ para os algoritmos SVC, NB, KNN, MLP e DT. Esses valores de tamanho de comitês foram escolhidos por uma análise empírica, buscando identificar tamanhos ideais. Por meio dos experimentos, percebeu-se que os comitês gerados com tamanho maiores que vinte não apresentavam

melhoria significativa na acurácia.

Na Tabela 5 são sumarizados os resultados dos experimentos na Base 5. Os comitês compostos somente com o algoritmo DT alcançou acurácias superiores a 77,30%. Além do mais, dentre os comitês mostrados, somente o DT apresentou resultados satisfatórios para as métricas de Precisão (70,11%) e *Recall* (66,39%). De maneira geral, percebe-se que o tamanho dos comitês não aumentou significativamente a acurácia dos modelos. A melhoria é apenas de um décimo, sendo estatisticamente irrelevante. Os comitês de MLP, mesmo com acurácias superiores a 71% possuem precisão e *recall* baixos que impossibilitariam este modelo preditivo na prática. Com a aplicação do teste de Friedman, obteve-se um valor-P = 0,56, e assim, conclui-se que existe uma diferença estatisticamente considerável entre as acurácias alcançadas com os comitês homogêneos na Base 5. O gráfico com as diferenças críticas calculadas nesta análise são mostrados no Apêndice C.

Tabela 5 – Comparação de desempenho de comitês homogêneos na Base 5.

T	SVC			NB		
	Acurácia ± DP	Precisão	Recall	Acurácia ± DP	Precisão	Recall
5	66,64 ± 0,06	34,01	34,12	61,19 ± 0,11	41,76	47,41
10	66,30 ± 0,07	33,30	33,81	61,00 ± 0,11	41,14	47,20
15	66,46 ± 0,07	33,33	33,94	60,93 ± 0,11	41,23	47,58
20	66,48 ± 0,33	33,43	33,93	61,03 ± 0,11	39,45	47,54

T	KNN			MLP		
	Acurácia ± DP	Precisão	Recall	Acurácia ± DP	Precisão	Recall
5	44,40 ± 0,11	24,77	23,41	71,88 ± 0,08	54,64	43,54
10	44,71 ± 0,25	24,72	23,50	71,88 ± 0,58	57,82	44,03
15	44,51 ± 0,10	24,56	23,09	71,88 ± 0,09	56,40	44,44
20	44,37 ± 0,10	24,47	22,85	71,83 ± 0,09	55,11	43,87

T	RFC			DT		
	Acurácia ± DP	Precisão	Recall	Acurácia ± DP	Precisão	*Recall*
5	67,37 ± 0,12	46,41	42,52	**77,30** ± 0,10	70,11	66,39
10	67,56 ± 0,12	47,14	43,11	**78,48** ± 0,09	71,62	67,85
15	67,33 ± 0,12	46,45	42,94	**78,47** ± 0,09	71,61	69,33
20	67,29 ± 0,12	46,17	42,75	**78,30** ± 0,09	71,34	68,93

Para compor os comitês heterogêneos foram usados os três algoritmos com melhor desempenho na aplicação dos comitês homogêneos (MLP, RFC e DT). Esses algoritmos foram combinados para compor os seguintes comitês heterogêneos: MLP e RFC; MLP e DT; RFC e DT; MLP, RFC 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).

Na Tabela 6 são detalhados os resultados obtidos para cada um desses comitês (nos tamanhos 5, 10, 15 e 20) com cada meta-classificador usado. Nota-se que a aplicação dos comitês heterogêneos não foi tão proveitosa quanto a dos homogêneos.

O comitê de MLP e DT alcançou as melhores métricas e permaneceu com a mesma acurácia, independente do meta-classificador utilizado. Este comitê alcançou acurácia de 75,14%, Precisão de 63,66% e *Recall* de 64,40%. Aplicando o teste de Friedman nas acurácias obtidas na Tabela 6, obteve-se um valor-$P = 0{,}102$, indicando que realmente a acurácia do comitê heterogêneo MLP e DT tem superioridade estatística sobre os demais. O gráfico de diferenças críticas desta análise é mostrado no Apêndice C. No entanto, a acurácia de aproximadamente 75% nos comitês de MLP e DT foi inferior aos resultados obtidos anteriormente com o comitê homogêneo de DT. Portanto, a utilização dessa técnica nessa situação não trouxe vantagem no que diz respeito à melhoria do desempenho do agrupamento dos classificadores, uma vez que as métricas não melhoraram após a aplicação dos comitês heterogêneos.

5.5 Considerações Finais

Como apresentado neste capítulo, o algoritmo RFR obteve um desempenho notável na regressão e o DT na classificação, destacando a estratégia de ganho de informação adotada por algoritmos baseados em árvore como promissora para a previsão da quantidade de aluguéis em SBC. Além disso, na construção dos modelos preditivos de classificação, as técnicas de PCA e a categorização da variável classe com partições do mesmo tamanho diminuíram o desempenho dessa abordagem e favoreceram o MLP. No entanto, houve uma melhoria nos comitês homogêneos de árvore de decisão. Mesmo assim, percebeu-se que a técnica de discretização utilizada no pré-processamento influencia diretamente nos resultados alcançados com a classificação.

Nesse sentido, a tarefa de regressão parece se adequar melhor para o propósito apresentado, uma vez que os resultados das métricas avaliativas foram mais satisfatórios, evidenciando modelos preditivos que poderiam ser utilizados na prática. Por fim, a análise de correlação se mostrou efetiva para relacionar o acontecimento de aluguéis às condições de clima, como temperatura e umidade, além de evidenciar que as horas do dia também influenciam no serviço de aluguel.

Os resultados apresentados na Seção 5.1 foram publicados no artigo intitulado "*Analyzing Patterns of a Bicycle Sharing System for Generating Rental Flow Predictive Models*" (VIANA *et al.*, 2019), que recebeu menção honrosa.

Os resultados apresentados nas Seções 5.2, 5.3 e 5.4 foram publicados no artigo intitulado "Análise Comparativa de Modelos Preditivos na Gestão Estratégica de Bicicletas Compartilhadas: Um Estudo de Caso" (VIANA *et al.*, 2020).

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

Meta-classificador	**T**	**MLP e RFC**			**MLP e DT**		
		Acurácia ± DP	Precisão	*Recall*	Acurácia ± DP	Precisão	*Recall*
MLP	**5**	66,12 ± 0,12	46,01	43,92	**75,14** ± 0,09	63,66	64,40
MLP	**10**	66,12 ± 0,11	45,22	42,68	**75,14** ± 0,09	63,66	64,40
MLP	**15**	66,93 ± 0,11	47,44	44,07	**75,14** ± 0,09	63,67	64,40
MLP	**20**	66,27 ± 0,12	44,59	43,34	**75,14** ± 0,09	63,67	64,40
DT	**5**	66,05 ± 0,12	46,39	43,88	**75,14** ± 0,09	63,67	64,40
DT	**10**	66,06 ± 0,12	44,81	43,27	**75,14** ± 0,09	63,66	64,40
DT	**15**	64,41 ± 0,12	47,20	42,52	**75,14** ± 0,09	63,67	64,40
DT	**20**	65,66 ± 0,12	44,34	43,28	**75,14** ± 0,09	63,67	64,40

Meta-classificador	**T**	**RFC e DT**			**MLP, RFC e DT**		
		Acurácia ± DP	Precisão	*Recall*	Acurácia ± DP	Precisão	*Recall*
MLP	**5**	71,64 ± 0,09	63,66	64,40	72,00 ± 0,10	59,42	51,38
MLP	**10**	72,92 ± 0,11	63,33	58,19	72,19 ± 0,10	61,75	51,19
MLP	**15**	73,41 ± 0,10	62,10	51,72	72,01 ± 0,10	59,93	51,31
MLP	**20**	73,22 ± 0,01	59,92	52,26	72,46 ± 0,10	62,83	51,41
DT	**5**	75,14 ± 0,09	63,67	64,40	72,83 ± 0,10	62,01	55,70
DT	**10**	75,01 ± 0,09	58,97	56,68	72,35 ± 0,09	55,28	51,79
DT	**15**	71,29 ± 0,09	55,32	47,13	68,36 ± 0,10	48,83	46,21
DT	**20**	71,26 ± 0,09	58,66	47,33	72,2 ± 0,10	58,00	50,62

6 Análise em Nível de Regiões

Neste capítulo, o *Capital BikeShare* foi analisado em nível de regiões geográficas definidas por agrupamentos das estações de aluguéis de bicicleta. Para definir os agrupamentos foram consideradas duas abordagens: a divisão por regiões geográficas já existente na estrutura de Washington D.C. (Seção 6.1) e a análise de agrupamentos por meio de um algoritmo de clusterização (Seção 6.2). Cerca de 0,1% da base original (3.836 registros) foi desconsiderada por possuir valores nulos para as posições geográficas de início/fim do aluguel, impossibilitando a identificação da origem do aluguel. Os resultados das aplicações dos modelos preditivos nesses agrupamentos são discutidos na Seção 6.3. Por fim, na Seção 6.4 são apresentadas as considerações finais do capítulo resumindo as conclusões das análises.

6.1 Agrupamento por Quadrante

Com essa abordagem agrupou-se as estações de aluguéis utilizando uma organização geográfica já existente na cidade. No caso de Washington DC, como mostrado na Figura 12, a cidade é dividida por ruas que partem do Capitólio de Washington, gerando quatro quadrantes: noroeste (*Northwest* - NW), nordeste (*North East* - NE), sudeste (*Southeast* - SE) e sudoeste (*South-west* - SW).

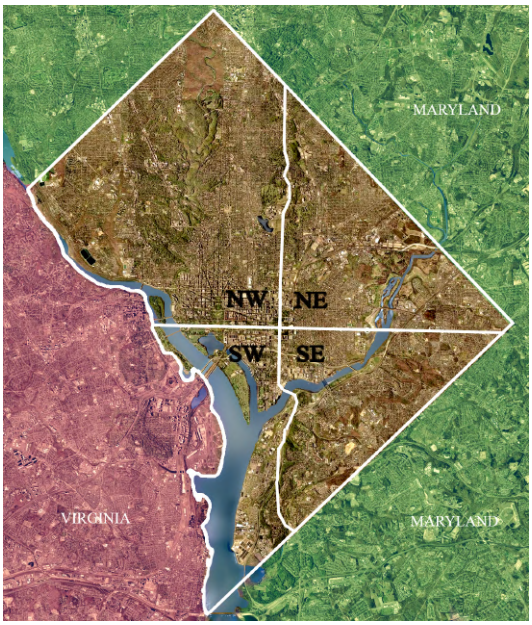


Figura 12 – Divisão em quadrantes e limite territorial de Washington D.C.

Fonte: Satélite do *United States Geological Survey* (USGS), 2002.

O quadrante NW é o maior e sedia o centro governamental dos Estados Unidos, bem como também é o principal centro econômico, comercial e universitário da cidade. O quadrante NE é predominantemente residencial (classe média e alta). O quadrante SW é o menor e possui várias estruturas governamentais, uma base aérea e um porto militar. Por fim, o quadrante SE é um centro comercial e residencial (classe baixa e média) (FOGLE, 2020). Esses quatro quadrantes foram usados para agrupar as 484 estações de bicicletas por meio da posição geográfica na qual o aluguel acontece, como segue:

- **quadrante NW**: 273 estações, com duração média dos aluguéis de 18,2 minutos;
- **quadrante NE**: 53 estações, com duração média dos aluguéis de 14,7 minutos;
- **quadrante SW**: 121 estações, com duração média dos aluguéis de 26,5 minutos; e
- **quadrante SE**: 37 estações, com duração média dos aluguéis de 15,7 minutos.

A Base 11 foi gerada com a quantidade de aluguéis agrupada por hora em cada quadrante, contendo 32.993 registros. Percebe-se que, embora o quadrante SW não contenha o maior número de estações, ele possui aluguéis que duram, em média, mais tempo, resultado de aluguéis que começam em alguma estação desse quadrante e terminam em alguma estação localizada em outro quadrante. No Quadro 7 são apresentadas as características de cada região geográfica.

Quadro 7 – Detalhamento de aluguéis em cada região geográfica (quadrante) na Base 11.

Quadrante	**Tipo de aluguel**	**Duração Média (minutos)**
NW	Casual (13,8%)	38,35
	Registrado (86,2%)	11,99
NE	Casual (23,6%)	33,00
	Registrado (76,4%)	11,82
SW	Casual (16,8%)	42,32
	Registrado (83,2%)	13,41
SE	Casual (45,2%)	34,88
	Registrado (54,8%)	11,79

Em todos os quadrantes a quantidade de aluguéis realizados por usuários registrados é superior à quantidade de aluguéis realizados por usuários casuais. Além disso, evidencia-se que os aluguéis no quadrante SW são cerca de 80% mais demorados em relação a aluguéis que acontecem no quadrante NE.

6.2 Agrupamento por Proximidade

Nesta abordagem de agrupamento por proximidade, foi utilizado o algoritmo de clusterização *k-means*[1], a fim de gerar grupos de estações de aluguéis utilizando a

[1] https://github.com/johnattandouglas/Dissertacao/tree/main/Scripts/NivelAgrupamento

localização geográfica das mesmas. Para tal, aplicou-se o método cotovelo (Figura 13), com o intuito de identificar o número ideal de agrupamentos (Capítulo 2).

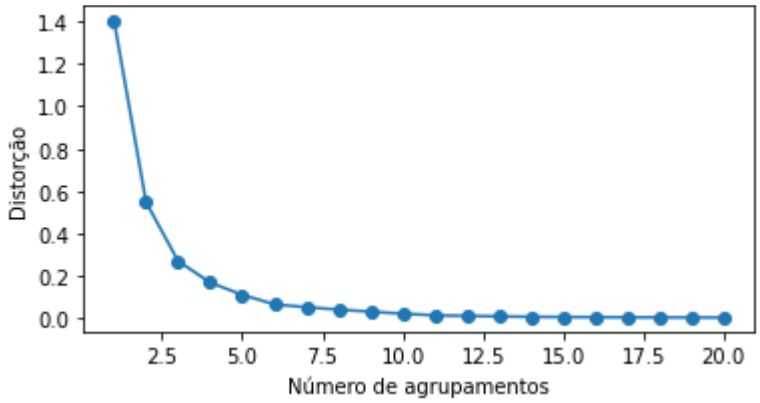


Figura 13 – Aplicação do método cotovelo.

Por meio do método cotovelo foi possível identificar que o número ideal de agrupamentos seria 3. No entanto, percebeu-se que esse número não seria viável para situação reais de aplicação, uma vez que um número maior de agrupamentos facilitaria no rebalanceamento na prática, já que seria possíel considerar as necessidades de cada agrupamento individualmente. Assim, optou-se por utilizar uma granularidade maior e definiram-se 8, 13 e 20 agrupamentos (Figuras 14, 15 e 16, respectivamente).

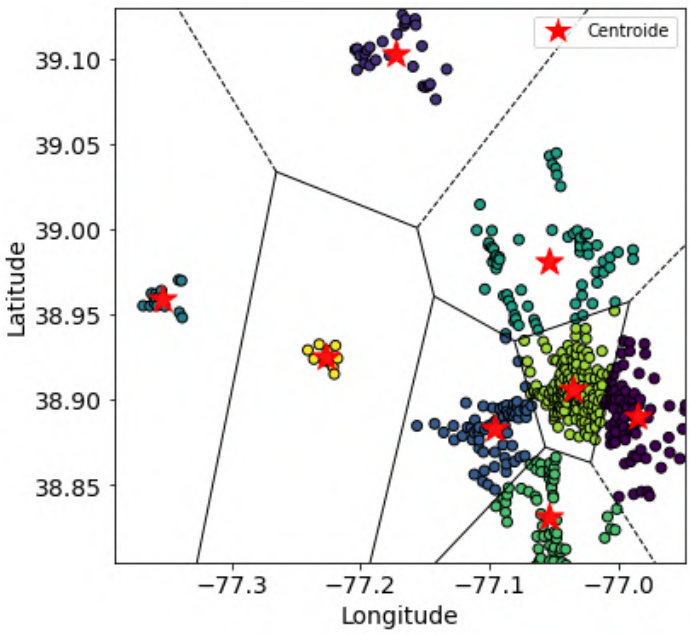


Figura 14 – Agrupamento das estações por proximidade com 8 centróides (Base 12).

A quantidade de aluguéis por hora resultou em bases de dados com 14 colunas (*month, day, hour, season_y, workday_y, temperature, r_temperature, wind, humidity, dew_point, pressure, weekday, cluster* e *qnt*), totalizando, respectivamente, 49.511, 79.586 e 111.501 registros nas bases com 8, 13 e 21 agrupamentos.

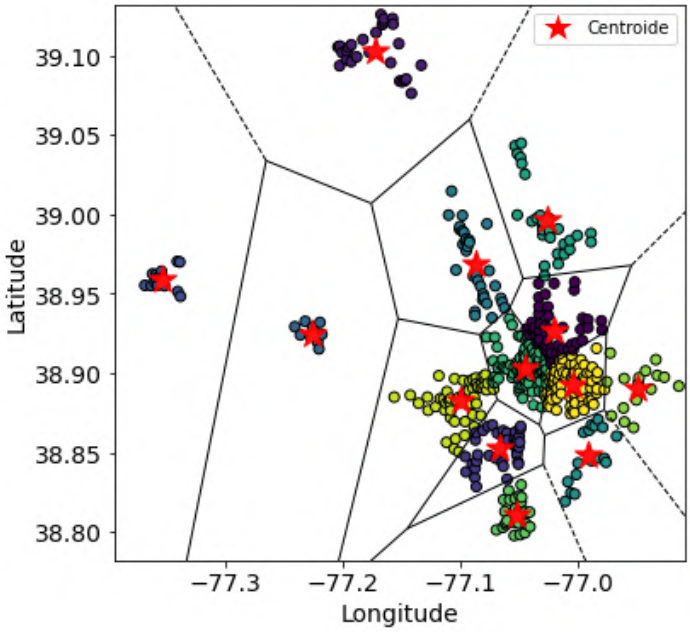


Figura 15 – Agrupamento das estações por proximidade com 13 centróides (Base 13).

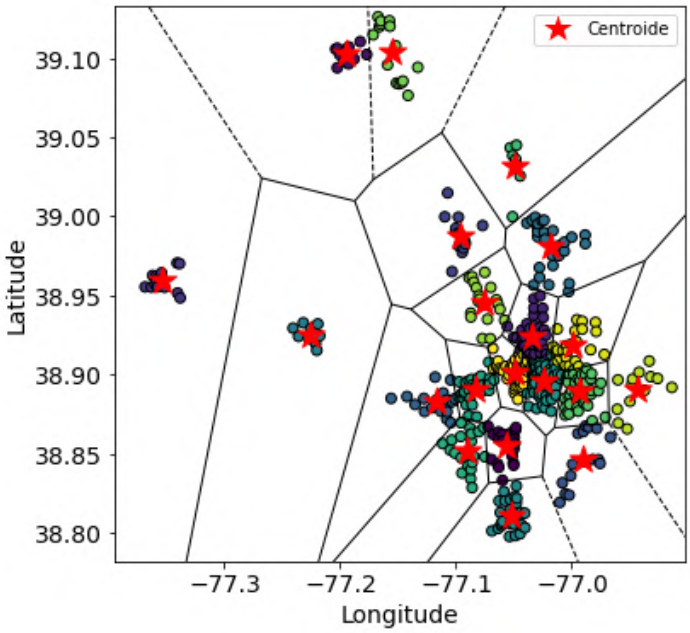


Figura 16 – Agrupamento das estações por proximidade com 20 centróides (Base 14).

Como mostrado nas Figuras 14, 15 e 16, o agrupamento das estações gerado pelo algoritmo *k-means* pode ser útil para auxiliar os gestores dos SBC na gestão estratégica dos recursos desses sistemas, direcionando bicicletas para regiões que possuem uma demanda maior e, dessa forma, tratando de maneira direcionada o problema de rebalanceamento.

6.3 Aplicação dos Algoritmos

Os regressores (Seção 6.3.1) e os classificadores (Seção 6.3.2) foram aplicados nas bases de dados 11 a 14, com as estações de aluguéis agrupadas por proximidade.

6.3.1 Regressão

Na Tabela 7(a) é demonstrada a aplicação dos regressores individualmente, onde os melhores resultados para o R^2 (coeficiente de determinação) estão em negrito. Percebe-se que o regressor floresta aleatória (RFR) obteve melhor desempenho, tanto com os melhores valores da métrica R^2, quanto com os do MAE e RMSE. No agrupamento com 13 centróides, o MLP, mesmo com um R^2 baixo, obteve a melhor resultado para o RMSE. O valor do R^2 médio para os regressores LR, SVR, KNN e MLP foi reduzido com os agrupamentos. No entanto, os agrupamentos ajudaram a melhorar o desempenho do RFR e DT. No agrupamento com 20 centróides, o R^2 diminuiu, devido à quantidade menor de aluguéis em cada agrupamento, diminuindo a quantidade de registros tanto para treinar quanto para testar os modelos preditivos.

O gráfico Quantil-Quantil (Q-Q) mostrado na Figura 17 apresenta a distribuição dos dados da Base 13, na qual os melhores resultados com o RFR individual foram alcançados. Os pontos seguem um padrão não-linear, sugerindo que os dados não são distribuídos em conformidade com a curva de normalidade (Gaussiana). A assimetria no Q-Q aparece com pontos dispostos em forma de arco. Uma distribuição perfeita seria indicada por uma linha de pontos em um ângulo de 45 graus da parte inferior esquerda do gráfico para a parte superior direita. Uma linha em vermelho foi traçada no gráfico para ajudar a tornar essa expectativa clara. Os desvios dos pontos da linha mostram um desvio da distribuição esperada. No caso do gráfico mostrado, o comportamento da curva cinza, é decorrente da distribuição assimétrica dos dados.

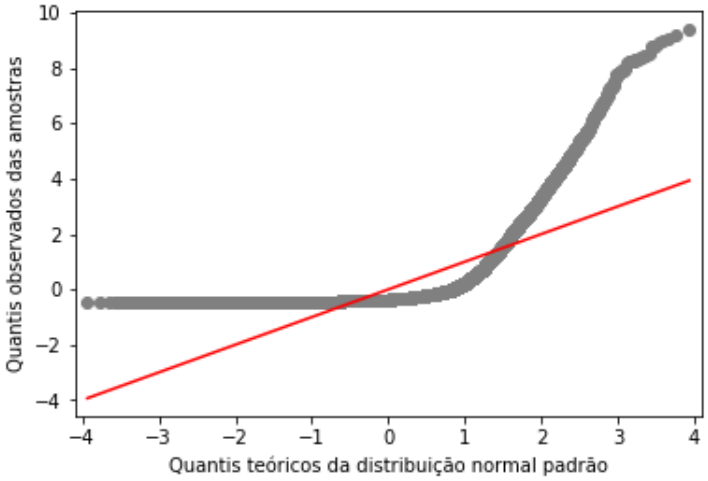


Figura 17 – Gráfico Quantil-Quantil da distribuição normal do RFR isolado na base 13.

Tabela 7 – Desempenho dos regressores nas bases de dados com estações agrupadas: (a) individualmente e (b) em comitês.

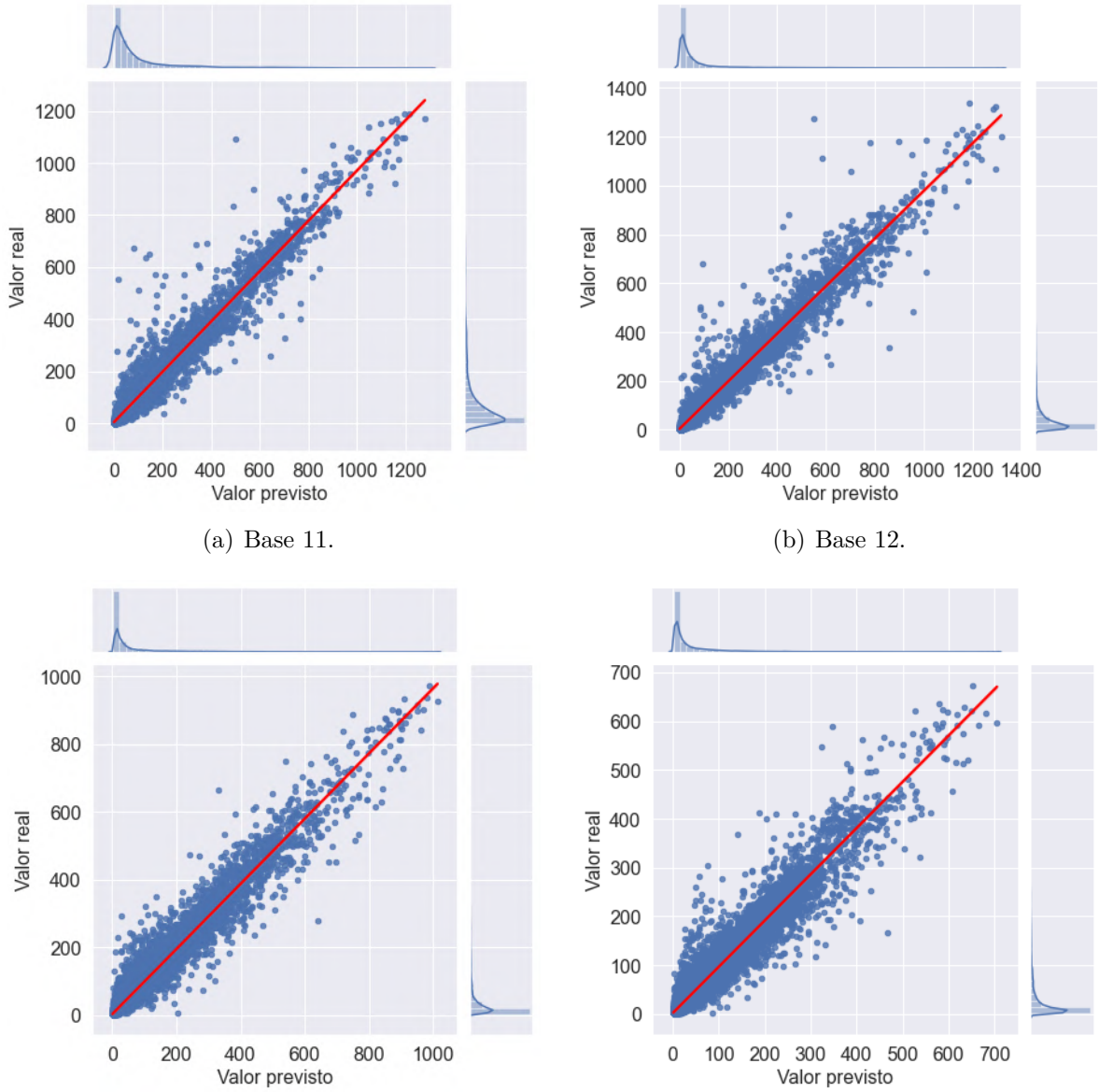
(a) Aplicação dos regressores individualmente.

Regressores	**Agrupamento por quadrantes**			**Agrupamentos com 8 centróides**			**Agrupamentos com 13 centróides**			**Agrupamentos com 20 centróides**		
	R^2	**MAE**	**RMSE**	**R^2**	**MAE**	**RMSE**	**R^2**	**MAE**	**RMSE**	**R^2**	**MAE**	**RMSE**
LR	0,353	93,702	135,881	0,078	89,393	147,661	0,035	58,060	100,882	0,074	37,319	62,521
SVR	0,313	72,264	140,072	0,082	65,914	160,003	0,108	42,908	108,126	0,046	29,762	66,438
KNN	0,665	54,210	97,816	0,194	84,157	138,056	0,107	55,010	97,073	0,019	37,826	64,345
MLP	0,728	54,862	88,057	0,681	55,426	86,926	0,393	48,790	20,279	0,233	35,685	56,897
RFR	**0,954**	17,514	36,079	**0,960**	12,047	30,698	**0,963**	8,385	30,698	**0,946**	6,853	15,0789
DT	0,914	23,631	49,642	0,934	15,716	40,671	0,930	10,784	27,233	0,910	8,731	19,532

(b) Aplicação dos regressores em comitês com 2, 3 e 4 algoritmos.

Regressores	**Agrupamento por quadrantes**			**Agrupamentos com 8 centróides**			**Agrupamentos com 13 centróides**			**Agrupamentos com 20 centróides**		
	R^2	**MAE**	**RMSE**	**R^2**	**MAE**	**RMSE**	**R^2**	**MAE**	**RMSE**	**R^2**	**MAE**	**RMSE**
KNN e MLP	0,740	49,710	86,085	0,522	65,277	106,311	0,299	49,738	85,981	0,156	36,066	59,673
KNN e RFR	0,893	31,143	55,318	0,771	44,561	73,603	0,751	28,557	51,235	0,718	19,983	34,523
KNN e DT	0,882	32,188	58,039	0,766	44,945	74,416	0,741	29,083	52,295	0,713	20,215	34,826
MLP e RFR	0,902	32,211	52,840	0,898	29,898	49,169	0,822	25,757	43,371	0,768	19,099	31,288
MLP e DT	0,893	33,441	55,362	0,884	32,515	52,325	0,832	24,567	42,158	0,763	18,895	31,628
RFR e DT	**0,947**	18,767	39,004	**0,954**	12,887	32,933	**0,957**	8,796	21,234	**0,938**	7,260	16,154
KNN, MLP e RFR	0,865	35,987	62,086	0,767	45,738	74,302	0,677	33,414	58,391	0,601	24,442	41,017
KNN, MLP e DT	0,861	36,227	63,039	0,774	44,717	73,112	0,667	33,892	59,247	0,598	24,525	41,172
KNN, RFR e DT	0,928	25,037	45,212	0,879	31,683	53,445	0,866	20,509	37,563	0,846	14,506	25,457
MLP, RFR e DT	0,888	33,551	56,661	0,931	23,316	40,494	0,897	18,858	32,906	0,869	13,782	23,500
KNN, MLP, RFR e DT	0,906	29,648	51,915	0,859	35,021	57,675	0,798	26,224	46,187	0,754	18,928	32,194

A distribuição dos dados de cada base utilizada no experimento pode ser visualizada na Figura 18, na qual apresentam-se relações entre os valores previstos e os valores reais da quantidade de aluguéis. Na parte superior e lateral, uma linha representa um histograma, a qual explicita essa distribuição assimétrica dos dados, indicando que, nas Bases 11, 12, 13 e 14, existem mais registros nos quais a quantidade de aluguéis está entre 0 e 50.



(a) Base 11.

(b) Base 12.

(c) Base 13.

(d) Base 14.

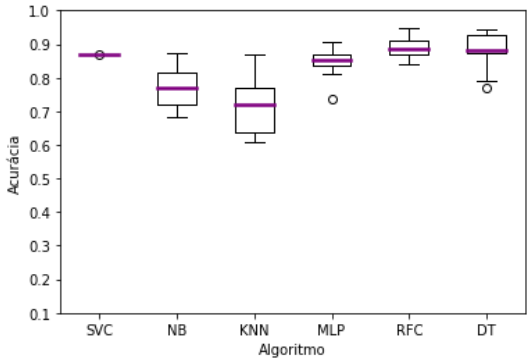
Figura 18 – Relação entre valores reais e previstos pelo RFR em cada base com estações agrupadas.

Para aplicação dos comitês, foi utilizado um comitê de votação voltado para regressão, o *Voting Regressor* da biblioteca *Scikit-Learn* (PEDREGOSA *et al.*, 2011). Para tal, os quatro algoritmos que obtiveram melhor desempenho individualmente foram considerados nesse experimento (KNN, MLP, RFR e DT), gerando combinações de tamanhos 2, 3 e 4.

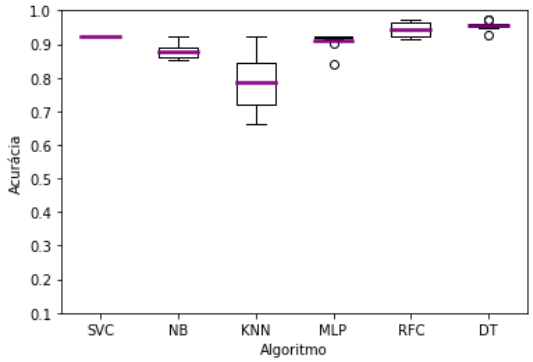
Como mostrado na Tabela 7(b), de modo geral, o R^2 teve uma pequena redução se comparado aos resultados alcançados sem os agrupamentos. Porém, em alguns casos, as métricas de erro diminuíram consideravelmente, como é o caso dos comitês com dois algoritmos (RFR e DT) aplicados na Base 17, na qual as estações estavam agrupadas em 20 centróides. Essa combinação de algoritmos destacou-se justamente porque ambos utilizam a abordagem de ganho de informação, o que se mostrou vantajosa de acordo com os experimentos anteriores.

6.3.2 Classificação

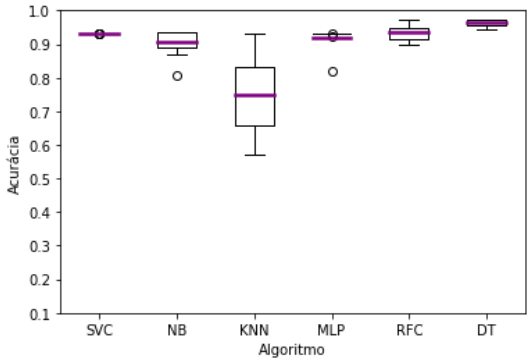
Uma vez que a aplicação do método *cut* resultou em melhores acurácias nas bases de dados nas quais ele foi aplicado, aplicou-se esse método para discretização da quantidade de aluguéis na base com os agrupamentos por quadrantes, utilizando as seguintes partições, definidas automaticamente pelo método de discretização: 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 mostrou resultados similares aos apresentados anteriormente (Seção 5.2). Na Figura 19 é apresentada a acurácia dos algoritmos de classificação aplicados nas bases de dados com agrupamentos das estações (Bases 11 a 14).



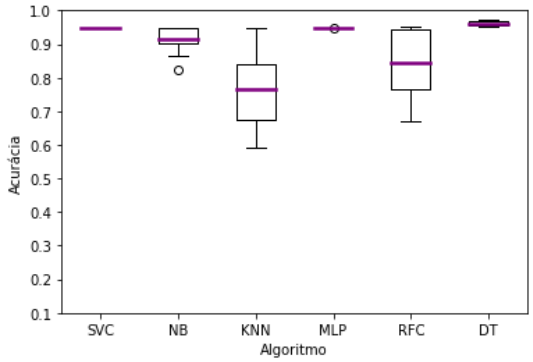
(a) Agrupamento por quadrantes (Base 11).



(b) Agrupamento com 8 centróides (Base 12).



(c) Agrupamento com 13 centróides (Base 13).



(d) Agrupamento com 20 centróides (Base 14).

Figura 19 – Aplicação dos algoritmos de classificação nas bases com agrupamentos.

Na Tabela 8 são detalhados os desempenhos dos classificadores nas Bases 11 a 14. Os algoritmos de classificação obtiveram melhor acurácia nestas bases com as estações de aluguéis agrupadas, evidenciando os agrupamentos como abordagem que pode aumentar a eficácia na predição. No entanto, as métricas de precisão e *recall* foram baixas com todos os classificadores, com exceção do DT, que mais uma vez se destacou. Na Base 11, com agrupamentos por quadrante, o RFC se destacou, com a acurácia de 88,60%. Na Bases de 12 a 14, o algoritmo DT se destacou com 95,57%, 96,21%, 96,13%, respectivamente.

Tabela 8 – Comparação de desempenho entre classificadores nas Bases 11 a 14.

BASE	**SVC**			**NB**		
	Acurácia ± DP	Precisão	Recall	Acurácia ± DP	Precisão	*Recall*
11	86,75 ± 0,00	17,35	20,00	76,90 ± 0,06	36,05	55,51
12	92,11 ± 0,00	18,42	20,00	87,78 ± 0,02	33,31	52,04
13	93,27 ± 0,00	18,65	20,00	90,49 ± 0,04	26,53	46,98
14	94,57 ± 0,00	18,91	20,00	91,53 ± 0,04	25,45	41,21

BASE	KNN			MLP		
	Acurácia ± DP	Precisão	Recall	Acurácia ± DP	Precisão	*Recall*
11	71,81 ± 0,09	17,66	17,27	87,12 ± 0,10	27,53	27,28
12	78,65 ± 0,08	18,56	17,53	91,69 ± 0,02	18,42	22,95
13	74,99 ± 0,12	18,54	16,49	93,08 ± 0,01	18,65	22,34
14	76,97 ± 0,12	19,78	16,41	94,59 ± 0,00	23,72	20,08

BASE	RFC			DT		
	Acurácia ± DP	Precisão	Recall	Acurácia ± DP	Precisão	*Recall*
11	**88,60** ± 0,03	44,26	40,49	88,06 ± 0,05	64,68	66,34
12	94,16 ± 0,02	46,51	41,43	**95,57** ± 0,01	67,40	66,13
13	93,73 ± 0,03	37,91	35,91	**96,21** ± 0,01	66,41	65,53
14	84,17 ± 0,11	26,38	26,56	**96,13** ± 0,01	61,73	61,23

Com a aplicação do teste de Friedman (valor-P = 0,029) comprovou-se que a acurácia alcançada pelo algoritmo DT é estatisticamente superior e diferente das alcançadas pelos outros algoritmos nestas bases. Além disso, com a aplicação do teste de Nemenyi, evidenciou-se que os resultados alcançados pelo MLP e SVC são estatisticamente indistinguíveis, como apresentado no Apêndice C.

Gráficos de curva ROC para os algoritmos individuais aplicados nas Bases 11 a 14 são apresentados no Apêndice D. Os algoritmos MLP e RFC tiveram classes com AUC altas, em especial quando nos referimos à classe "Muito alta" (classe 5), que alcançou AUC próximo à 1,0. Na Base 11 e 12, o NB distinguiu erroneamente a classe "Baixa" (classe 2) das outras. Na Base 13, este algoritmo também passou a classificar incorretamente a classe "Média" (classe 3). Nas Bases 13 e 14, o algoritmo SVC somente conseguiu distinguir corretamente as classes "Alta" (classe 4) e "Muito alta" (classe 5). Nestas bases, o algoritmo KNN continuou sem distinguir as classes e na maioria dos experimentos ele realizou a classificação quase que aleatoriamente (AUC próximo de 0,5).

Os comitês homogêneos foram aplicados na Base 14, onde as estações de aluguéis são agrupadas em 20 centróides. Os desempenhos dos comitês homogêneos e heterogêneos são apresentados nas Tabelas 9 e 10, respectivamente. A aplicação dos comitês homogêneos não resultou em melhora significativa nas métricas avaliativas (Tabela 9). Evidencia-se ainda a baixa precisão e *recall* mesmo com a utilização dos comitês homogêneos. Com a aplicação do teste de Friedman (valor-P = 0,8) identificou-se que os resultados alcançados com o MLP e SVC são estatisticamente indistinguíveis e realmente inferiores aos alcançados pelo algoritmo DT. O teste de Nemenyi evidenciou que o algoritmo DT realmente apresentou desempenho superior quando comparado aos outros algoritmos, como apresentado no gráfico de diferença crítica no Apêndice C.

Tabela 9 – Comparação de desempenho de comitês homogêneos na Base 14.

T	**SVC**			**NB**		
	Acurácia ± DP	Precisão	*Recall*	Acurácia ± DP	Precisão	*Recall*
5	94,57 ± 0,00	18,91	20,00	91,50 ± 0,04	25,20	41,19
10	94,57 ± 0,00	18,91	20,00	91,57 ± 0,04	25,94	42,15
15	94,57 ± 0,00	18,91	20,00	91,55 ± 0,04	25,68	42,00
20	94,57 ± 0,00	18,91	20,00	91,52 ± 0,04	25,75	42,03

T	**KNN**			**MLP**		
	Acurácia ± DP	Precisão	*Recall*	Acurácia ± DP	Precisão	*Recall*
5	76,06 ± 0,12	18,90	16,59	94,57 ± 0,00	18,91	20,00
10	75,93 ± 0,12	18,80	16,36	94,57 ± 0,00	18,91	20,00
15	75,86 ± 0,12	18,79	16,34	94,57 ± 0,00	18,91	20,00
20	75,86 ± 0,19	12,17	16,36	94,57 ± 0,00	18,91	20,00

T	**RFC**			**DT**		
	Acurácia ± DP	Precisão	*Recall*	Acurácia ± DP	Precisão	*Recall*
5	87,60 ± 0,08	25,56	25,90	**96,44** ± 0,01	67,13	59,71
10	87,23 ± 0,08	26,58	25,76	**96,69** ± 0,01	67,14	58,53
15	87,32 ± 0.08	26,33	25,69	**96,66** ± 0,01	66,41	59,57
20	87,21 ± 0,08	26,36	25,64	**96,76** ± 0,01	67,43	59,18

Adicionalmente, como exposto na Tabela 10, foram construídos comitês heterogêneos com a combinação dos algoritmos que obtiveram melhores acurácias individualmente na Base 14, usando cada um deles como meta-classificadores (SVC, MLP e DT), porém, não obtiveram acurácias maiores do que as já alcançadas. O melhor resultado alcançado com os comitês heterogêneos foi de 96,17%, obtido com o comitê composto pelo MLP, SVC e DT, usando o MLP como meta-classificador. Os resultados dos testes estatísticos (valor-P = 0,428) evidenciaram que os três comitês heterogêneos com acurácia invariada de 96,17% possuem semelhança estatística.

Tabela 10 – Acurácia e desvio padrão na Base 14 dos comitês heterogêneos para cada meta-classificador*.

Meta-classificador	**T**	**MLP e SVC**			**MLP e DT**		
		Acurácia ± DP	Precisão	Recall	Acurácia ± DP	Precisão	*Recall*
MLP	**5**	94,56 ± 0,00	19,07	20,02	96,17 ± 0,00	61,52	61,12
MLP	**10**	94,56 ± 0,00	23,54	21,27	96,17 ± 0,00	61,52	61,12
MLP	**15**	94,57 ± 0,00	26,35	22,15	96,17 ± 0,00	61,52	61,12
MLP	**20**	94,57 ± 0,00	24,85	22,66	96,17 ± 0,00	61,52	61,12
DT	**5**	94,57 ± 0,00	26,86	21,11	96,17 ± 0,00	61,52	61,12
DT	**10**	94,42 ± 0,01	28,27	22,56	96,17 ± 0,00	61,52	61,12
DT	**15**	94,43 ± 0,01	29,53	22,90	96,17 ± 0,00	61,52	61,12
DT	**20**	94,45 ± 0,01	31,22	23,54	96,17 ± 0,00	61,52	61,12

Meta-classificador	**T**	**SVC e DT**			**MLP, SVC e DT**		
		Acurácia ± DP	Precisão	Recall	Acurácia ± DP	Precisão	*Recall*
MLP	**5**	96,17 ± 0,00	61,52	61,12	96,17 ± 0,00	61,52	61,12
MLP	**10**	96,17 ± 0,00	61,52	61,12	96,17 ± 0,00	61,52	61,12
MLP	**15**	96,17 ± 0,00	61,52	61,12	96,17 ± 0,00	61,52	61,12
MLP	**20**	96,17 ± 0,00	61,52	61,12	96,17 ± 0,00	61,52	61,12
DT	**5**	96,17 ± 0,00	61,52	61,12	96,17 ± 0,00	61,52	61,12
DT	**10**	96,17 ± 0,00	61,52	61,12	96,17 ± 0,00	61,52	61,12
DT	**15**	96,17 ± 0,00	61,52	61,12	96,17 ± 0,00	61,52	61,12
DT	**20**	96,17 ± 0,00	61,52	61,12	96,17 ± 0,00	61,52	61,12

* Os resultados alcançados são estatisticamente indistinguíveis (ver Apêndice C).

6.4 Considerações Finais

De um ponto de vista operacional para os SBC, pode ser mais útil predizer a quantidade de aluguéis em determinadas regiões da cidade, de modo a alocar recursos conforme as necessidades de cada região. Na classificação da quantidade de aluguéis utilizando esta abordagem, a melhor acurácia obtida foi de 96,21%, alcançada com o algoritmo de árvore de decisão (DT) na base de dados com as estações agrupadas com 13 centróides. Assim, evidencia-se que os agrupamentos, além de melhorar a acurácia, também são mais úteis na solução prática do problema de desbalanceamento. No entanto, a utilização da técnica de comitês homogêneos e heterogêneos não foi vantajosa nos experimentos apresentados.

Parte dos resultados apresentados utilizando os agrupamentos por quadrantes (Seção 6.1) foram publicados no artigo intitulado "Modelos de aprendizagem de máquina para a gestão estratégica de bicicletas compartilhadas" (VIANA *et al.*, 2021).

7 Cenário de Aplicação dos Modelos Preditivos

O cenário de aplicação apresentado neste capítulo visa demonstrar como integrar os modelos preditivos gerados para estimar as demandas de aluguéis em SBC a uma aplicação que utiliza dados em tempo real como entrada de tais modelos para auxiliar não só os usuários, mas também os gestores desses serviços no processo de tomada de decisão. Na Figura 20 esta abordagem é ilustrada, identificando suas etapas e respectivos componentes envolvidos em cada uma delas, como os dados de entrada, os modelos preditivos utilizados na análise desses dados e as saídas geradas. Cada uma dessas etapas é descrita, respectivamente, nas Seções 7.1, 7.2 e 7.3. Na Seção 7.4 são apresentadas discussões a respeito da abordagem.

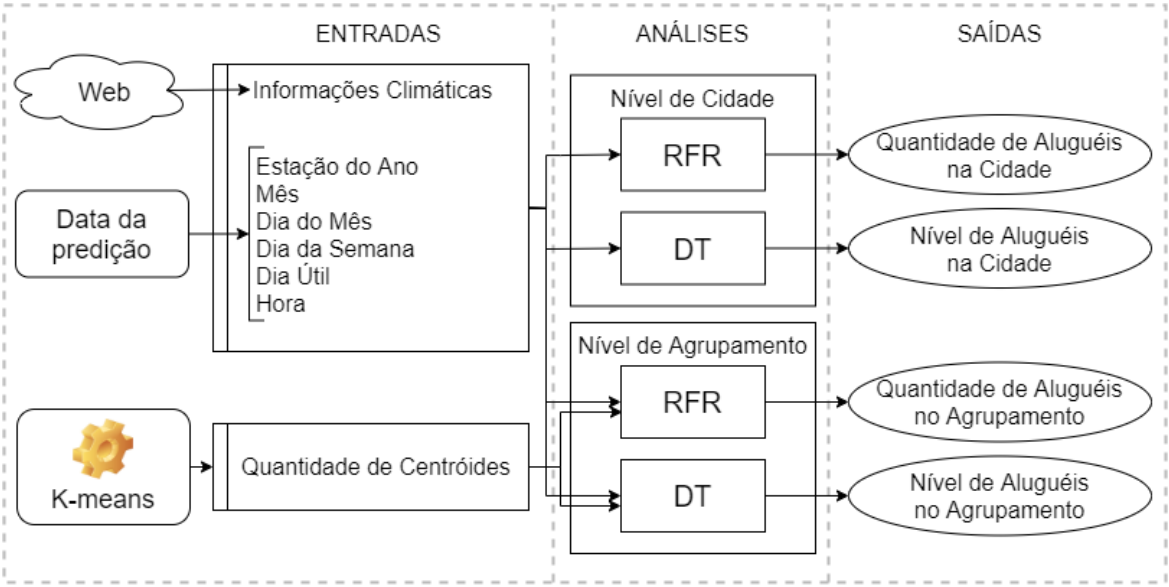


Figura 20 – Abordagem para utilização dos modelos preditivos gerados.

7.1 Entradas

Para a aplicação dos modelos preditivos apresentados anteriormente, diversos dados precisam ser fornecidos como valores de entrada. Para a predição de demanda de uma determinada hora/dia, são necessárias as informações climáticas desse momento, que podem ser obtidas por meio de coleta de dados em *websites* de previsão de tempo. A partir da data de previsão, derivam-se variáveis como estação do ano, mês, dia do mês, dia da semana, dia útil e hora.

Para utilização dos modelos preditivos em nível de agrupamento, além dos dados já citados é necessário especificar a quantidade de centróides que serão utilizados na previsão. Nos experimentos mostrados, foram gerados modelos preditivos para 8, 13 e 20 centróides gerados por meio do algoritmo *K-means*. Na prática, a quantidade de centróides poderia ser dinâmica, sendo ajustada conforme o desejo de granularidade do gestor. Dessa forma, seria configurável o tamanho da região onde os modelos preditivos seriam aplicados para prever a demanda de aluguel em cada região.

7.2 Análises

As análises possíveis com o cenário de aplicação apresentado são em níveis de cidade e de agrupamentos. Os modelos de predição em nível de cidade são úteis quando deseja-se analisar a demanda em todo o sistema, considerando fenômenos que aumentarão a demanda de maneira geral, não somente em uma região específica.

Os modelos de predição em nível de agrupamento fornecem a demanda considerando os centróides fornecidos como entrada e por isso são utilizados para uma análise mais focada em determinadas regiões. Neste caso, os modelos preditivos seriam usados justamente para calcular a quantidade de bicicletas disponíveis em determinado grupo de estações. Dessa forma, em agrupamentos onde a demanda costuma ser alta, os gestores conseguiriam elaborar um planejamento estratégico de reposição de bicicletas para balanceamento da quantidade de bicicletas em cada estação. Além disso, seria possível configurar alertas que indicariam situações de escassez extrema, como, por exemplo, um conjunto de estações bastante utilizada possuindo poucas ou nenhuma bicicleta disponível.

Na Figura 21 são apresentados, como exemplo, os registros em tempo real de aluguéis de um dos SBC de Washington D.C. (Capital BikeShare) por meio de uma aplicação móvel, que também executariam os modelos preditivos gerados. Em nível de cidade, a predição de demanda seria útil para gerar indicadores gerais à respeito da demanda, enquanto que as análises em nível de agrupamento seriam voltados para uma análise mais direcionada à uma região, ambas as análises auxiliando no apoio estratégico do serviço de aluguéis.

Como mostrado na Figura 21, os dados em tempo real do sistema poderiam ser utilizados em um sistema de gerenciamento dos serviços de SBC, onde os dados das estações de aluguéis, bem como a disponibilidade de bicicletas seriam apresentados. Esse sistema possuiria funcionalidades adicionais associadas ao emprego dos modelos preditivos gerados, tanto em nível de cidade quanto em nível de agrupamentos para, por exemplo, realizar o acompanhamento de utilização das bicicletas tanto em tempo real quanto como predição, e também monitorar a demanda em determinadas regiões.

Figura 21 – Proposta de integração dos modelos preditivos por meio de um aplicativo móvel.

7.3 Saídas

As saídas geradas pelos modelos preditivos são de dois tipos: a quantidade de aluguéis prevista e o nível de demanda (muito baixa, baixa, média, alta e muito alta). Os dois tipos apresentam informações similares, mas podem ser utilizadas para propósitos diferentes. As saídas que são predições da quantidade de aluguel são usadas quando espera-se o valor exato da demanda, enquanto que o nível é mais adequado para aplicações onde apenas o nível de demanda é suficiente.

De um ponto de vista dos gestores de SBC, as saídas podem ser utilizadas para auxiliar na tarefa de rebalanceamento, identificando regiões geográficas com estações vazias (ou cheias) em um determinado momento. Os gestores podem utilizar as informações geradas como saídas para adotar ações preventivas, como realizar um rebalanceamento para colocar mais bicicletas nas regiões que provavelmente serão esvaziadas dentro de um intervalo de tempo, o que também pode ser utilizado como parâmetro para a implantação de novas estações em regiões onde a demanda costuma ser mais alta.

De um ponto de vista do usuário de um serviço de aluguel de bicicletas compartilhadas, a experiência no uso do serviço de aluguel pode ser bastante prejudicada se eles se depararem com uma estação que não possui mais bicicletas disponíveis, pois precisam encontrar outra estação nas proximidades para realizar o aluguel. O mesmo pode acontecer quando for necessário finalizar um trajeto, tentando devolver a bicicleta em uma estação que não tem mais lugares vazios. As saídas podem ser utilizadas para oferecer serviços voltados para os usuários que indiquem a disponibilidade de bicicletas em determinadas horas e regiões específicas.

Como mostrado na Figura 22, as saídas em quantidade também podem ser utilizadas para geração de mapas de calor, indicando regiões de alta demanda (em vermelho). Na figura, os mapas de calor foram gerados com a biblioteca Folium[1] que permite a visualização de mapas da biblioteca Leaflet.js[2] usando Python. Por meio dessa visualização, também é possível analisar cada agrupamento de estações individualmente, identificando áreas para implantação de novas estações de aluguéis.

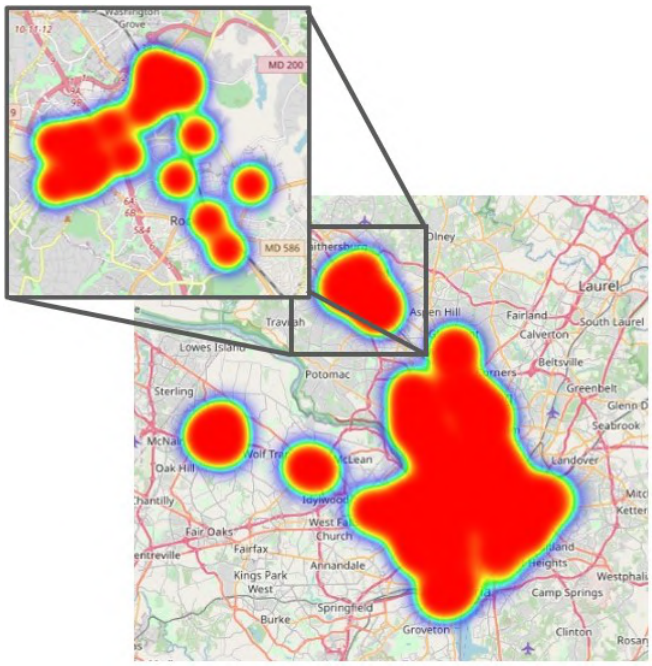


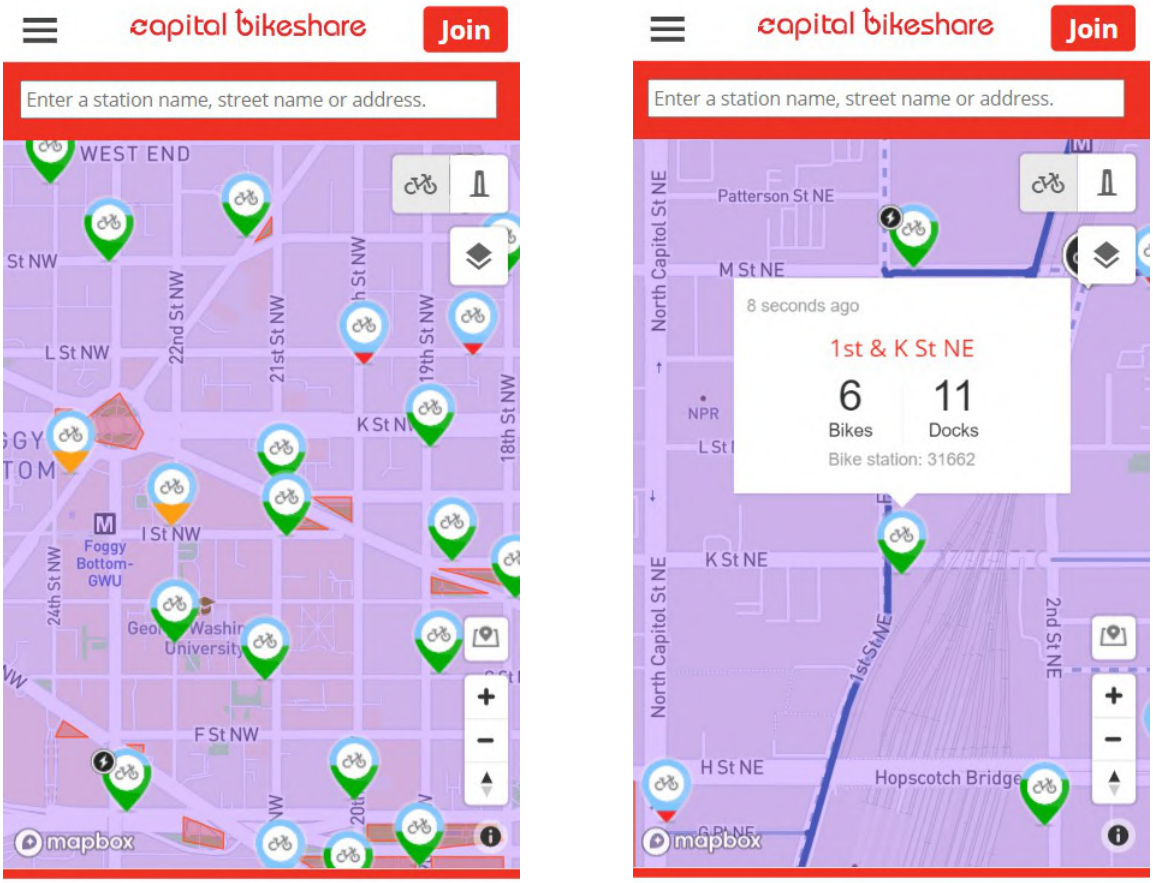
Figura 22 – Visualização em mapa de calor com ênfase em um agrupamento.

7.4 Dados em Tempo Real

O *Capital BikeShare* possui um aplicativo móvel (disponível para Android[3] e iPhone[4]), que utiliza dados em tempo real e possibilita que o usuário confira pelo celular as estações de aluguéis na proximidade (Figura 23(a)), bem como quantas bicicletas estão livres em cada estação (Figura 23(b)).

Atualmente, duzentos e nove SBC publicam as bases de dados seguindo uma especificação própria para os registros de sistemas de transporte, denominada de *General Bikeshare Feed Specification* (GBFS)(NABSA, 2021). O *Capital BikeShare* utiliza essa especificação e disponibiliza dados sobre o seu serviço em tempo real. Esses dados são disponibilizados por meio de *Uniform Resource Locator* (URL[5]) com diversos arquivos no formato *JavaScript Object Notation* (JSON) e representam diferentes aspectos do sistema.

[1] https://python-visualization.github.io/folium/
[2] https://leafletjs.com/
[3] https://play.google.com/store/apps/details?id=com.motivateco.capitalbikeshare
[4] https://apps.apple.com/us/app/capital-bikeshare/id1233403073
[5] https://gbfs.capitalbikeshare.com/gbfs/gbfs.json



(a) Disponibilidade de bicicletas nas estações.

(b) Situação individual da estação.

Figura 23 – Aplicativo do *Capital BikeShare* com informações em tempo real.
Fonte: Capital Bike Share System (2018)

Dentre os arquivos JSON disponibilizados pelo *Capital BikeShare*, um arquivo[6] em específico é referente à situação das estações de aluguéis, e possui dados relacionados a quantidade de bicicletas disponíveis, a quantidade de bicicletas que foram alugadas, a identificação da estação e a sua respectiva localização geográfica. Para utilização dessa fonte de dados em uma aplicação, poderia-se desenvolver uma API que consuma os dados de forma dinâmica, e dessa forma, associar os dados em tempo real com as saídas geradas pela abordagem proposta neste trabalho.

No aplicativo do *Capital BikeShare*, as estações são coloridas com verde, amarelo e vermelho de acordo com a quantidade de bicicletas disponíveis, onde verde indica que todas as bicicletas ainda estão disponíveis e vermelho indica onde não existem mais bicicletas disponíveis. O aplicativo também exibe as opções de ônibus e metrôs nas proximidades do usuário. Dessa forma, o usuário consegue conciliar o seu translado com o transporte público. A abordagem proposta neste trabalho poderia ser integrada a esse aplicativo para prover funcionalidades adicionais relacionadas à disponibilidade de bicicletas em períodos

[6] https://gbfs.capitalbikeshare.com/gbfs/en/station_status.json

futuros, utilizando os modelos preditivos apresentados. Dessa forma, o usuário conseguiria planejar passeios com antecedência tendo como parâmetro a disponibilidade estimada de bicicletas em cada horário e região da cidade. Nas análises e previsões discutidas nos capítulos anteriores utilizou-se registros do *Capital BikeShare* de 2017. Porém, é possível reutilizar a abordagem apresentada para gerar modelos preditivos para outros SBC e em outros sistemas de transporte compartilhado que possuam dados abertos.

Os modelos preditivos elaborados neste trabalho podem ser utilizados para disponibilizar funções em um aplicativo que, por exemplo, permita que o usuário pesquise a disponibilidade de bicicletas em determinadas regiões. Ressalta-se que a disponibilização dos dados em tempo real possibilita a criação de outros aplicativos que não sejam necessariamente disponibilizados pela empresa responsável pelo *Capital BikeShare*.

7.5 Considerações Finais

O cenário de aplicação dos modelos preditivos de demanda de aluguéis de bicicletas concebidos neste trabalho pode trazer vantagens tanto para os usuários quanto para os gestores desses serviços. Além disso, a disponibilização de dados dos SBC em tempo real possibilita o desenvolvimento de novas aplicações que não precisam, necessariamente, pertencer as empresas responsáveis por esses sistemas. Na prática, tal cenário pode servir como processo estratégico para melhorar a qualidade do serviço de aluguel, independente do tipo de transporte compartilhado.

8 Conclusões

Com este trabalho evidenciou-se a possibilidade de gerar modelos preditivos e descritivos usando informações de contexto em SBC. Nesse trabalho foi apresentada uma abordage genérica que pode ser aplicada em STC com o intuito descrever padrões de mobilidade utilizando modelos descritivos e preditivos.

Assim, foi fornecido aos gestores dessa modalidade de serviços informações precisas acerca do serviço de aluguel de bicicletas para auxiliá-los no entendimento do hábito dos usuários. Com isso, foram produzidos modelos de predição de demanda de aluguéis tanto usando algoritmos de regressão quanto de classificação.

Para uma predição em nível de cidade, o melhor modelo regressor gerado foi do algoritmo *Random Forest Regressor* (RFR) que alcançou coeficiente de determinação de 0,947. Na tarefa de classificação, a acurácia foi mais alta com o algoritmo *Decision Tree* (DT) aplicado em comitê homogêneo (78,48%) do que quando aplicado individualmente (75,08%). Os testes de Friedman e Nemenyi evidenciaram que esses resultaram alcançados são realmente estatisticamente superiores aos outros algoritmos utilizados. Porém, com a análise dos gráficos de curvas ROC percebeu-se dificuldade dos algoritmos em distinguir corretamente, em geral, as classes “Média” e “Alta”.

Na análise utilizando os agrupamentos de estações de bicicletas percebeu-se que um número muito grande de agrupamentos pode dificultar o desempenho dos algoritmos, uma vez que à medida que a quantidade de agrupamentos cresce, a quantidade de registros para treinamento e teste desses algoritmos diminui. Mesmo assim, os melhores resultados foram encontrados com o agrupamento que utilizou 13 centroides, no qual o RFR se destacou na regressão com coeficiente de determinação de 0,963 e o DT se destacou na classificação com acurácia de 96,21%.

Percebeu-se que a tarefa de regressão parece mais adequada para auxiliar o problema apresentado, uma vez que as métricas de precisão e *recall* apresentaram valores baixos na maioria dos resultados alcançados com a tarefa de classificação. A abordagem proposta neste trabalho poderia ser integrada a aplicativos utilizados pelas empresas responsáveis pelos SBC para prover funcionalidades adicionais relacionadas à disponibilidade de bicicletas em períodos futuros, utilizando os modelos preditivos apresentados. Dessa forma, o usuário conseguiria planejar passeios com antecedência tendo como parâmetro a disponibilidade estimada de bicicletas em cada horário e região da cidade.

Os modelos preditivos apresentados, sobretudo os regressores, seriam úteis na predição da quantidade de aluguéis, tanto em nível de cidade quanto em nível de regiões específicas da cidade, definidas pelos modelos descritivos de agrupamentos. Além disso, os resultados apresentados podem ser utilizados também para auxiliar na atividade de rebalanceamento de bicicletas nas estações.

Como trabalhos futuros, espera-se implementar uma ferramenta que utilize os modelos de previsão elaborados neste trabalho consumindo dados em tempo real de SBC, auxiliando na tomada de decisão e gestão estratégica desses sistemas. Além disso, destacam-se também como novas perspectivas de pesquisa novas definições de centróides para os agrupamentos, utilizando fatores relacionados à outros sistemas de transporte (como estações de metrô ou ônibus nas proximidades), características da região (como densidade populacional) ou pontos de interesse (como quantidade de lojas, pontos turísticos, etc.).

Referências

AIT-AMIR, B.; POUGNET, P.; HAMI, A. E. 6 - meta-model development. In: HAMI, A. E.; POUGNET, P. (Ed.). *Embedded Mechatronic Systems 2*. Elsevier, 2015. p. 151 – 179. ISBN 978-1-78548-014-0. Disponível em: <http://www.sciencedirect.com/science/article/pii/B9781785480140500062>. Citado na página 26.

ALVAREZ-VALDES, R. *et al.* Optimizing the level of service quality of a bike-sharing system. *Omega*, v. 62, p. 163 – 175, 2016. ISSN 0305-0483. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0305048315002005>. Citado na página 18.

ANSERE, J. A.; HAN, G.; WANG, H. A novel reliable adaptive beacon time synchronization algorithm for large-scale vehicular ad hoc networks. *IEEE Transactions on Vehicular Technology*, v. 68, n. 12, p. 11565–11576, 2019. Citado na página 19.

ASHQAR, H. I.; ELHENAWY, M.; RAKHA, H. A. Modeling bike counts in a bike-sharing system considering the effect of weather conditions. *Case Studies on Transport Policy*, v. 7, n. 2, p. 261 – 268, 2019. ISSN 2213-624X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S2213624X16301018>. Citado 2 vezes nas páginas 36 e 40.

BEARD, R. E. Biometrika tables for statisticians. *Journal of the Institute of Actuaries*, Cambridge University Press, v. 93, n. 1, p. 154–155, 1967. Citado na página 28.

BENGIO, Y. *et al.* Greedy layer-wise training of deep networks. In: *Proceedings of the 19th International Conference on Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2006. (NIPS'06), p. 153–160. Citado na página 37.

BHOLOWALIA, P.; KUMAR, A. Ebk-means: A clustering technique based on elbow method and k-means in wsn. *International Journal of Computer Applications*, Citeseer, v. 105, n. 9, 2014. Citado na página 34.

BREIMAN, L. Bagging predictors. *Machine learning*, Springer, v. 24, n. 2, p. 123–140, 1996. Citado 2 vezes nas páginas 50 e 63.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Citado na página 30.

BUSSAB, W. O.; MORETTIN, P. A. *Estatística Básica*. São Paulo: Saraiva, 2010. Citado na página 57.

CAPITAL BIKE SHARE SYSTEM. *Capital bike sharing trip history data*. 2018. Disponível em: <http://www.capitalbikeshare.com/trip-history-data>. Acesso em: 2018. Citado 2 vezes nas páginas 44 e 83.

CHEN, L. *et al.* Dynamic cluster-based over-demand prediction in bike sharing systems. In: *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. New York, NY, USA: Association for Computing

Machinery, 2016. (UbiComp '16), p. 841–852. ISBN 9781450344616. Disponível em: <https://doi.org/10.1145/2971648.2971652>. Citado 3 vezes nas páginas 36, 38 e 42.

CHEN, T. *et al.* Xgboost: extreme gradient boosting. *R package version 0.4-2*, p. 1–4, 2015. Citado na página 33.

CHIAVENATO, I. *Introdução à teoria geral da administração*. Rio de Janeiro: Editora Manole., 2014. Citado na página 17.

DELL'AMICO, M. *et al.* The bike sharing rebalancing problem: Mathematical formulations and benchmark instances. *Omega*, v. 45, p. 7 – 19, 2014. ISSN 0305-0483. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0305048313001187>. Citado na página 19.

DEMSAR, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, JMLR.org, v. 7, p. 1–30, dez. 2006. ISSN 1532-4435. Citado na página 28.

DUPUIS, D. J.; FIELD, C. A. Large wind speeds: Modeling and outlier detection. *Journal of Agricultural, Biological, and Environmental Statistics*, [International Biometric Society, Springer], v. 9, n. 1, p. 105–121, 2004. ISSN 10857117. Disponível em: <http://www.jstor.org/stable/1400709>. Citado na página 46.

DÜSING, R. Knowledge discovery in databases. *Wirtschaftsinformatik*, v. 42, n. 1, p. 74–75, Feb 2000. ISSN 1861-8936. Disponível em: <https://doi.org/10.1007/BF03250720>. Citado 2 vezes nas páginas 9 e 23.

ECHEVERRY-GALVIS, M.; PETERSON, J.; CACERES, R. The social nestwork: Tree structure determines nest placement in kenyan weaverbird colonies. *PloS one*, v. 9, p. e88761, 02 2014. Citado na página 59.

FANAEE-T, H.; GAMA, J. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, Springer Berlin Heidelberg, p. 1–15, 2013. ISSN 2192-6352. Disponível em: <https://link.springer.com/article/10.1007/s13748-013-0040-3>. Citado na página 35.

FISHMAN, E.; WASHINGTON, S.; HAWORTH, N. Bike share: A synthesis of the literature. *Transport Reviews*, Routledge, v. 33, n. 2, p. 148–165, 2013. Disponível em: <https://doi.org/10.1080/01441647.2013.775612>. Citado na página 17.

FOGLE, J. M. *Washington, D.C.* 2020. Online: <https://www.britannica.com/place/Washington-DC>. Acessado em 30/10/2020. Citado na página 68.

FRIEDMAN, A. Urban design and shared transport. In: _____. *Fundamentals of Sustainable Urban Design*. Cham, Switzerland: Springer International Publishing, 2021. p. 197–203. ISBN 978-3-030-60865-1. Citado na página 17.

FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, The Institute of Mathematical Statistics, v. 29, n. 5, p. 1189–1232, 10 2001. Disponível em: <https://doi.org/10.1214/aos/1013203451>. Citado na página 38.

FRIEDMAN, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, v. 32:200, p. 675–701, 1937. Citado 2 vezes nas páginas 28 e 50.

GAST, N. *et al.* Probabilistic forecasts of bike-sharing systems for journey planning. In: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. New York, NY, USA: Association for Computing Machinery, 2015. (CIKM '15), p. 703–712. ISBN 9781450337946. Disponível em: <https://doi.org/10.1145/2806416.2806569>. Citado 3 vezes nas páginas 19, 36 e 39.

GÉRON, A. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. Sebastopol, Califórnia: O'Reilly Media, 2019. Citado na página 28.

GHOSH, S. *et al.* Dynamic repositioning to reduce lost demand in bike sharing systems. *J. Artif. Int. Res.*, AI Access Foundation, El Segundo, CA, USA, v. 58, n. 1, p. 387–430, jan. 2017. ISSN 1076-9757. Citado na página 18.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. Cambridge, Massachusetts: MIT Press, 2016. <http://www.deeplearningbook.org>. Citado na página 32.

HAMILTON, T. L.; WICHMAN, C. J. Bicycle infrastructure and traffic congestion: Evidence from dc's capital bikeshare. *Journal of Environmental Economics and Management*, v. 87, p. 72 – 93, 2018. ISSN 0095-0696. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0095069616300420>. Citado na página 17.

HAN, J.; PEI, J.; KAMBER, M. *Data Mining: Concepts and Techniques*. 3. ed. Waltham, Massachusetts: Elsevier, 2011. Citado 5 vezes nas páginas 23, 25, 30, 32 e 47.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer Science & Business Media, 2009. Citado na página 29.

HAWKINS, D. M. The problem of overfitting. *Journal of chemical information and computer sciences*, ACS Publications, v. 44, n. 1, p. 1–12, 2004. Citado na página 26.

HEARST, M. A. *et al.* Support vector machines. *IEEE Intelligent Systems and their applications*, IEEE, v. 13, n. 4, p. 18–28, 1998. Citado na página 30.

HOO, Z. H.; CANDLISH, J.; TEARE, D. What is an roc curve? *Emergency Medicine Journal*, British Association for Accident and Emergency Medicine, v. 34, n. 6, p. 357–359, 2017. ISSN 1472-0205. Disponível em: <https://emj.bmj.com/content/34/6/357>. Citado na página 29.

JAMES, G. *et al. An Introduction to Statistical Learning with Applications in R*. New York: Springer Texts in Statistics. Springer., 2013. Citado na página 49.

JAPPINEN, S.; TOIVONEN, T.; SALONEN, M. Modelling the potential effect of shared bicycles on public transport travel times in greater helsinki: An open data approach. *Applied Geography*, v. 43, p. 13 – 24, 2013. ISSN 0143-6228. Disponível em: <http://www.sciencedirect.com/science/article/pii/S014362281300132X>. Citado na página 17.

KASSAMBARA, A. *Machine Learning Essentials: Practical Guide in R*. [S.l.]: STHDA, 2018. Citado 3 vezes nas páginas 25, 26 e 27.

KOTSIANTIS, S. B. *et al.* Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, IOS Press, Amsterdam, v. 160, n. 1, p. 3–24, 2007. Citado na página 25.

KUNCHEVA, L. I. *Combining Pattern Classifiers: Methods and Algorithms*. 2nd. ed. New Jersey: Wiley Publishing, 2014. ISBN 1118315235. Citado na página 60.

LANDGREBE, T. C.; DUIN, R. P. Approximating the multiclass roc by pairwise analysis. *Pattern Recognition Letters*, v. 28, n. 13, p. 1747–1758, 2007. ISSN 0167-8655. Disponível em: <https://www.sciencedirect.com/science/article/pii/S016786550700150X>. Citado 2 vezes nas páginas 29 e 63.

LAYTON, R. *Learning Data Mining with Python*. Birmingham, England: Packt Publishing Ltd, 2017. Citado 2 vezes nas páginas 23 e 24.

LI, Y.; ZHENG, Y. Citywide bike usage prediction in a bike-sharing system. *IEEE Transactions on Knowledge and Data Engineering*, v. 32, n. 6, p. 1079–1091, 2020. Citado 2 vezes nas páginas 36 e 37.

LI, Y. *et al.* Traffic prediction in a bike-sharing system. In: *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*. New York, NY, USA: Association for Computing Machinery, 2015. (SIGSPATIAL '15). ISBN 9781450339674. Disponível em: <https://doi.org/10.1145/2820783.2820837>. Citado 3 vezes nas páginas 36, 38 e 42.

LIU, J. *et al.* 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*. New York, NY, USA: Association for Computing Machinery, 2016. (KDD '16), p. 1005–1014. ISBN 9781450342322. Disponível em: <https://doi.org/10.1145/2939672.2939776>. Citado 5 vezes nas páginas 18, 19, 20, 36 e 41.

LONG, W. J. *et al.* A comparison of logistic regression to decision-tree induction in a medical domain. *Computers and Biomedical Research*, Elsevier Science, v. 26, n. 1, p. 74–97, 1993. Citado 2 vezes nas páginas 29 e 30.

LV, Y. *et al.* Traffic flow prediction with big data: A deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, v. 16, n. 2, p. 865–873, 2015. Citado 2 vezes nas páginas 36 e 37.

LóPEZ-RUBIO, E.; PALOMO, E. J.; ORTEGA-ZAMORANO, F. Unsupervised learning by cluster quality optimization. *Information Sciences*, v. 436-437, p. 31–55, 2018. ISSN 0020-0255. Disponível em: <https://www.sciencedirect.com/science/article/pii/S002002551830015X>. Citado na página 33.

MAHMOUD, M.; EL-ASSI, W.; HABIB, K. N. Effects of built environment and weather on bike sharing demand: Station level analysis of commercial bike sharing in toronto. *Annual Meeting of Transportation Research Board*, v. 94, p. 589–613, 2015. Disponível em: <https://link.springer.com/article/10.1007/s11116-015-9669-z>. Citado na página 58.

MCKINNEY, W. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. Sebastopol, California: "O'Reilly Media, Inc.", 2012. Citado na página 46.

METZ, C. E. Basic principles of roc analysis. In: ELSEVIER. *Seminars in nuclear medicine*. [S.l.], 1978. v. 8, n. 4, p. 283–298. Citado na página 29.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. *Introduction to linear regression analysis*. New Jersey: John Wiley & Sons, 2012. v. 821. Citado na página 25.

MOONEY, C. Z. *Monte carlo simulation*. London: Sage, 1997. Citado na página 39.

MURPHY, K. P. *et al.* Naive bayes classifiers. *University of British Columbia*, v. 18, p. 60, 2006. Citado na página 31.

MÁTRAI, T.; TÓTH, J. Comparative assessment of public bike sharing systems. *Transportation Research Procedia*, v. 14, p. 2344 – 2351, 2016. ISSN 2352-1465. Transport Research Arena TRA2016. Disponível em: <http://www.sciencedirect.com/science/article/pii/S2352146516302678>. Citado na página 17.

NABSA. *Understanding the General Bikeshare Feed Specification (GBFS)*. 2021. Acessado em 15/01/2021. Citado na página 82.

NATEKIN, A.; KNOLL, A. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, Frontiers, v. 7, p. 21, 2013. Citado na página 33.

NEILL, S. P.; HASHEMI, M. R. Chapter 8 - ocean modelling for resource characterization. In: NEILL, S. P.; HASHEMI, M. R. (Ed.). *Fundamentals of Ocean Renewable Energy*. Academic Press, 2018, (E-Business Solutions). p. 193 – 235. ISBN 978-0-12-810448-4. Disponível em: <http://www.sciencedirect.com/science/article/pii/B9780128104484000082>. Citado na página 26.

NEMENYI, P. B. *Distribution-Free Multiple Comparisons*. Tese (Doutorado) — Princeton University, 1963. Citado 2 vezes nas páginas 28 e 50.

NPS, N. P. S. *Closures*. 2020. Acessado em 06/10/2020. Disponível em: <https://www.nps.gov/subjects/nationalmall4th/closures.htm>. Citado na página 55.

O'BRIEN, O. *et al.* *The Meddin Bike-sharing World Map*. 2021. Acessado em 14/02/2021. Disponível em: <https://bikesharingworldmap.com/>. Citado na página 17.

PASQUALE, P. *et al.* *Comunicação Integrada de Marketing: A Teoria na Prática*. [S.l.]: Elsevier, 2011. Citado na página 17.

PEDREGOSA, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado 4 vezes nas páginas 31, 49, 50 e 73.

PETERSON, L. E. K-nearest neighbor. *Scholarpedia*, v. 4, n. 2, p. 1883, 2009. Citado na página 31.

REN, Y. *et al.* Rebalancing bike sharing systems for minimizing depot inventory and traveling costs. *IEEE Transactions on Intelligent Transportation Systems*, v. 21, n. 9, p. 3871–3882, 2020. Citado na página 19.

RISH, I. *et al.* An empirical study of the naive bayes classifier. In: *IJCAI 2001 workshop on empirical methods in artificial intelligence*. [S.l.: s.n.], 2001. v. 3, n. 22, p. 41–46. Citado na página 31.

ROMAN, R.; LOPEZ, J.; MAMBO, M. Mobile edge computing, fog et al.: A survey and analysis of security threats and challenges. *Future Generation Computer Systems*, v. 78, p. 680 – 698, 2018. ISSN 0167-739X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167739X16305635>. Citado na página 39.

SAGI, O.; ROKACH, L. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley Online Library, v. 8, n. 4, p. e1249, 2018. Citado na página 32.

SATHISHKUMAR, V. E.; JANGWOO, P.; YONGYUN, C. Using data mining techniques for bike sharing demand prediction in metropolitan city. *Computer Communications*, v. 153, p. 353–366, 2020. ISSN 0140-3664. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0140366419318997>. Citado 2 vezes nas páginas 36 e 37.

SCHAPIRE, R. E. Explaining adaboost. In: *Empirical inference*. [S.l.]: Springer, 2013. p. 37–52. Citado na página 33.

SHLENS, J. A tutorial on principal component analysis. *CoRR*, abs/1404.1100, 2014. Disponível em: <http://arxiv.org/abs/1404.1100>. Citado na página 47.

SOUZA, L. C.; GOMES, E. T. A. 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*. Letra1, 2014. p. 384–395. ISSN 978-85-63800-17-6. Disponível em: <http://www.editoraletra1.com/anais-congeo/arquivos/978-85-63800-17-6-p384-395.pdf>. Citado na página 17.

STIGLER, S. M. Thomas bayes's bayesian inference. *Journal of the Royal Statistical Society: Series A (General)*, Wiley Online Library, v. 145, n. 2, p. 250–258, 1982. Citado na página 30.

TANG, J.; ALELYANI, S.; LIU, H. Data classification: Algorithms and applications. In: *Data Classification: Algorithms and Applications*. [S.l.]: Data Mining and Knowledge Discovery Series, CRC Press, 2015. p. 484–509. Citado 2 vezes nas páginas 33 e 50.

UMARGONO, E.; SUSENO, J. E.; GUNAWAN, S. V. K-means clustering optimization using the elbow method and early centroid determination based on mean and median formula. In: *Proceedings of the 2nd International Seminar on Science and Technology (ISSTEC 2019)*. Atlantis Press, 2020. p. 121–129. ISBN 978-94-6239-168-0. ISSN 2352-5398. Disponível em: <https://doi.org/10.2991/assehr.k.201010.019>. Citado na página 34.

VANDERPLAS, J. *Python data science handbook: Essential tools for working with data*. 1. ed. [S.l.]: O'Reilly Media, Inc., 2016. Citado 2 vezes nas páginas 24 e 34.

VIANA, J. D. F. *et al.* Analyzing patterns of a bicycle sharing system for generating rental flow predictive models. In: *Anais do III Workshop de Computação Urbana*. Porto Alegre, RS, Brasil: SBC, 2019. p. 57–70. ISSN 2595-2706. Disponível em: <https://sol.sbc.org.br/index.php/courb/article/view/7468>. Citado na página 65.

VIANA, J. D. F. *et al.* Análise comparativa de modelos preditivos na gestão estratégica de bicicletas compartilhadas: Um estudo de caso. In: *Anais do LII Simpósio Brasileiro de Pesquisa Operacional*. Campinas, Brasil: SBPO, 2020. Citado na página 65.

VIANA, J. D. F. *et al.* Modelos de aprendizagem de máquina para a gestão estratégica de bicicletas compartilhadas. *Brazilian Journal of Development*, v. 7, p. 2209–2227, 2021. ISSN 2525–8761. Disponível em: <https://www.brazilianjournals.com/index.php/BRJD/article/view/22793/18273>. Citado na página 78.

WICKHAM, H.; GROLEMUND, G. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. 1. ed. Sebastopol: O'Reilly Media, 2017. Citado 2 vezes nas páginas 24 e 43.

WU, X. *et al.* Top 10 algorithms in data mining. *Knowl. Inf. Syst.*, Springer-Verlag, Berlin, Heidelberg, v. 14, n. 1, p. 1–37, dez. 2007. ISSN 0219-1377. Disponível em: <https://doi.org/10.1007/s10115-007-0114-2>. Citado na página 34.

XU, T. *et al.* A hybrid machine learning model for demand prediction of edge-computing-based bike-sharing system using internet of things. *IEEE Internet of Things Journal*, v. 7, n. 8, p. 7345–7356, 2020. Citado 4 vezes nas páginas 18, 19, 36 e 39.

YANG, Y. *et al.* Using graph structural information about flows to enhance short-term demand prediction in bike-sharing systems. *Computers, Environment and Urban Systems*, v. 83, p. 101521, 2020. ISSN 0198-9715. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0198971520302544>. Citado 4 vezes nas páginas 35, 36, 38 e 41.

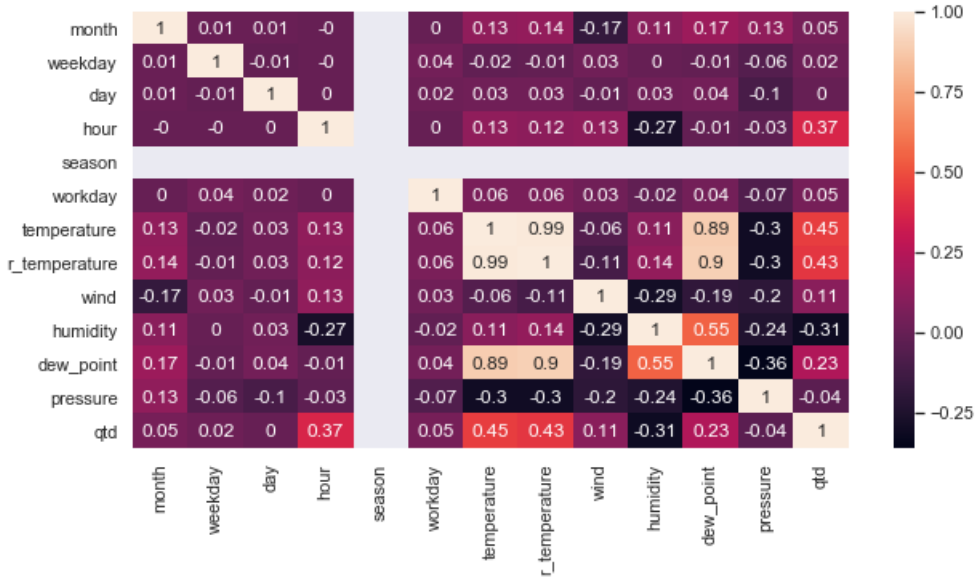
YANG, Z. *et al.* Mobility modeling and prediction in bike-sharing systems. In: . New York, NY, USA: Association for Computing Machinery, 2016. (MobiSys '16), p. 165–178. ISBN 9781450342698. Disponível em: <https://doi.org/10.1145/2906388.2906408>. Citado 2 vezes nas páginas 36 e 40.

ZHANG, Y. *et al.* ibike: Intelligent public bicycle services assisted by data analytics. *Future Generation Computer Systems*, v. 95, p. 187 – 197, 2019. ISSN 0167-739X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167739X18322787>. Citado 4 vezes nas páginas 35, 36, 39 e 40.

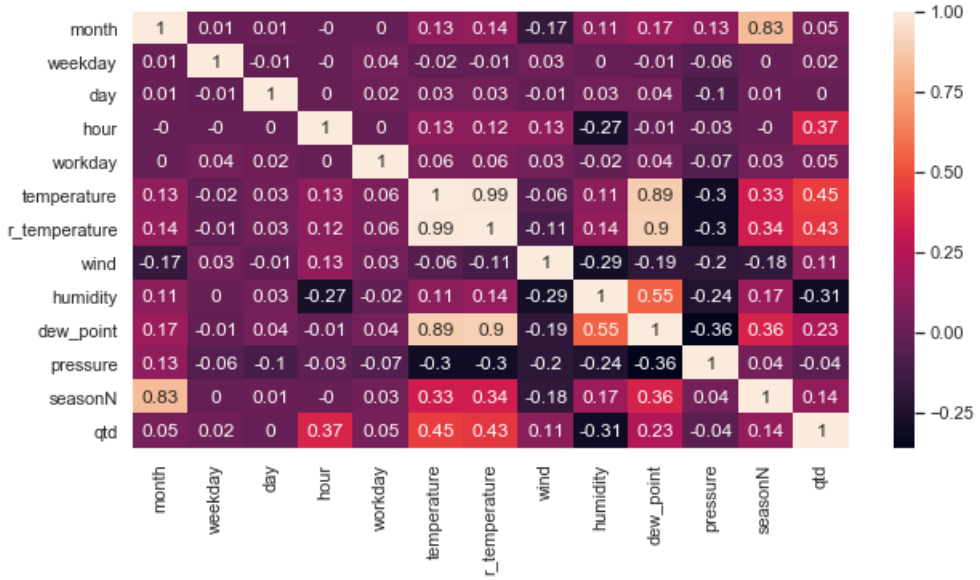
ZHENG, Y. *et al.* Forecasting fine-grained air quality based on big data. In: *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. [S.l.: s.n.], 2015. p. 2267–2276. Citado na página 39.

APÊNDICE A – Correlação nas Bases sem Redução de Dimensionalidade (Bases 1 a 6)

Legenda: A cor mais clara representa a correlação direta e a escura a correlação inversa. Valores próximos de zero estão colorido em roxo, indicando a inexistência de correlação.

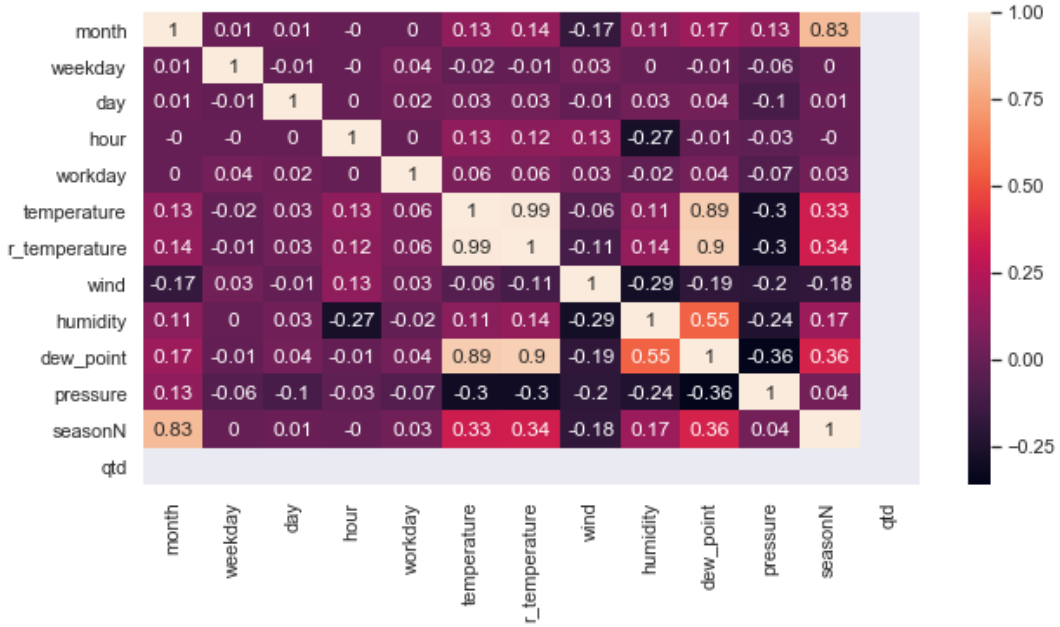


Matriz de correlação da Base 1, na qual os valores faltosos foram removidos.

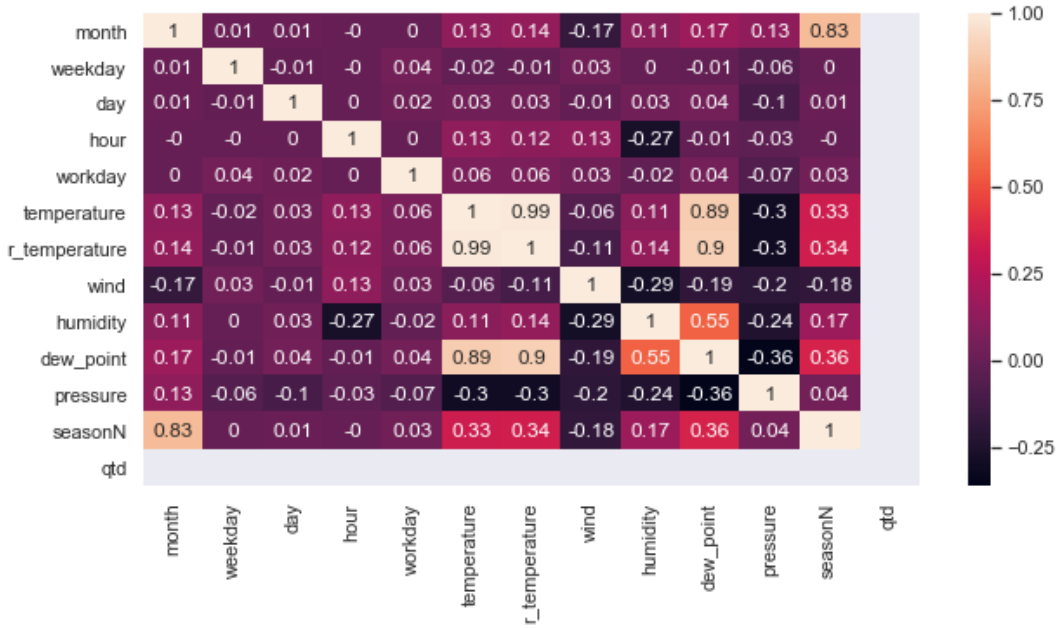


Matriz de correlação da Base 2, na qual a variável *season* foi convertida para um valor numérico.

Legenda: A cor mais clara representa a correlação direta e a escura a correlação inversa. Valores próximos de zero estão colorido em roxo, indicando a inexistência de correlação.

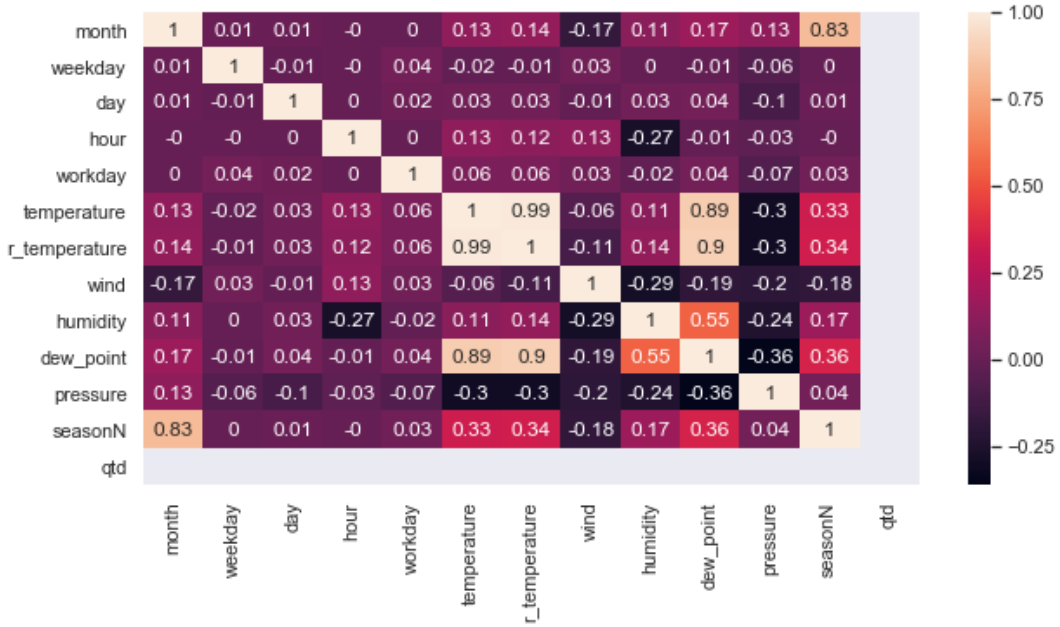


Matriz de correlação da Base 3, na qual a variável clase (*qtd*) foi discretizada utilizando o método *cut*.

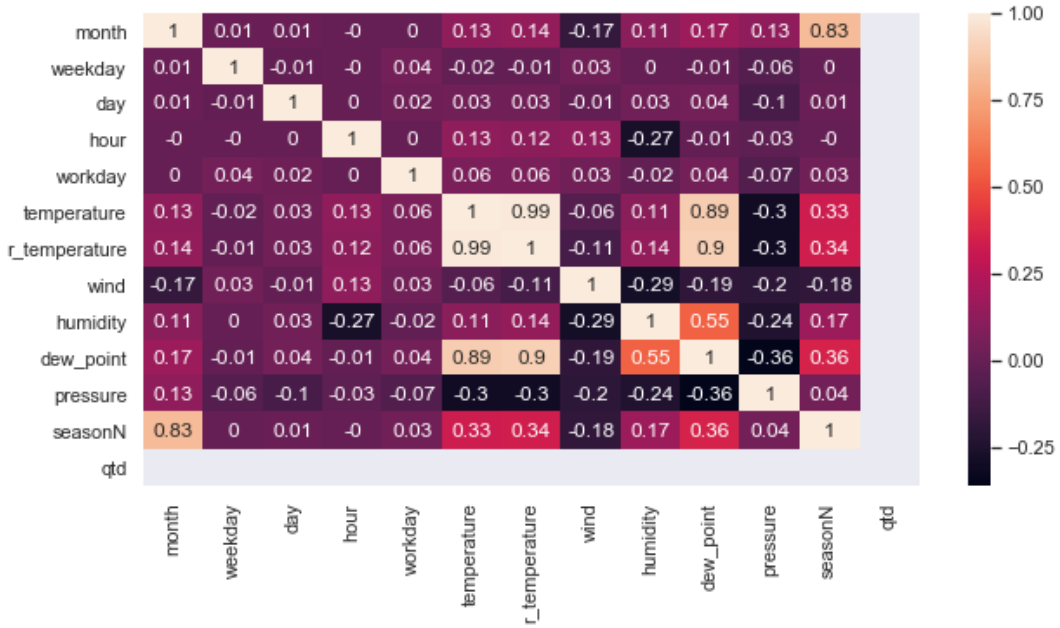


Matriz de correlação da Base 4, na qual a variável clase (*qtd*) foi discretizada utilizando o método *qcut*.

Legenda: A cor mais clara representa a correlação direta e a escura a correlação inversa. Valores próximos de zero estão colorido em roxo, indicando a inexistência de correlação.



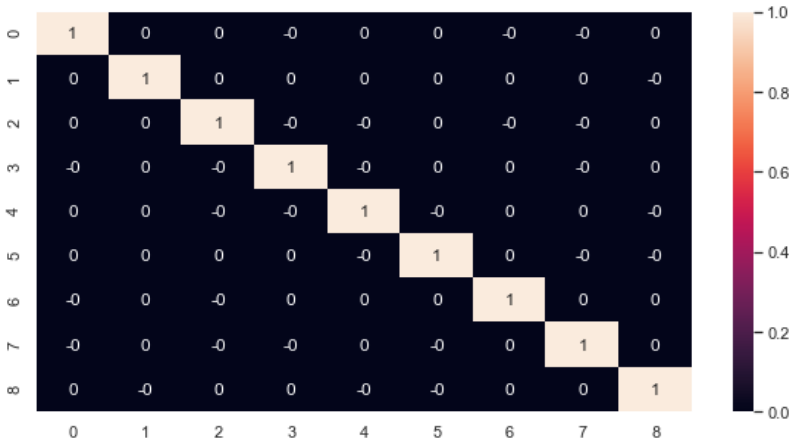
Matriz de correlação da Base 5, na qual a variável clase (*qtd*) foi discretizada utilizando o método *cut*, com normalização nas variáveis meteorológicas.



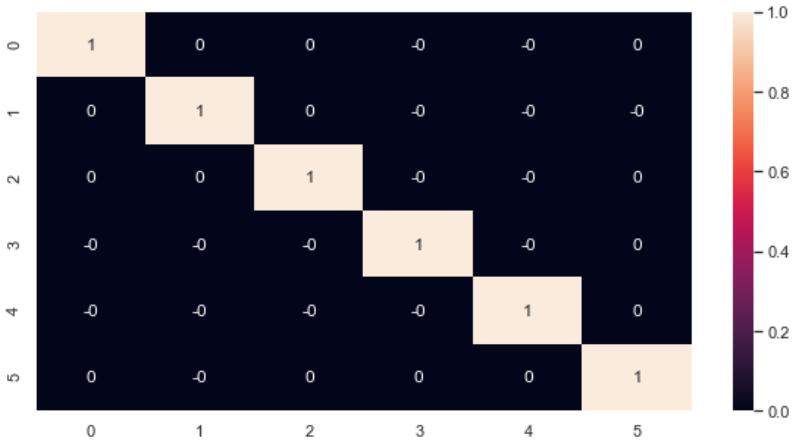
Matriz de correlação da Base 6, na qual a variável clase (*qtd*) foi discretizada utilizando o método *qcut*, com normalização nas variáveis meteorológicas.

APÊNDICE B – Correlação nas Bases com Redução de Dimensionalidade (Bases 7 a 10)

Legenda: A cor mais clara representa a correlação direta e a escura indica a inexistência de correlação.

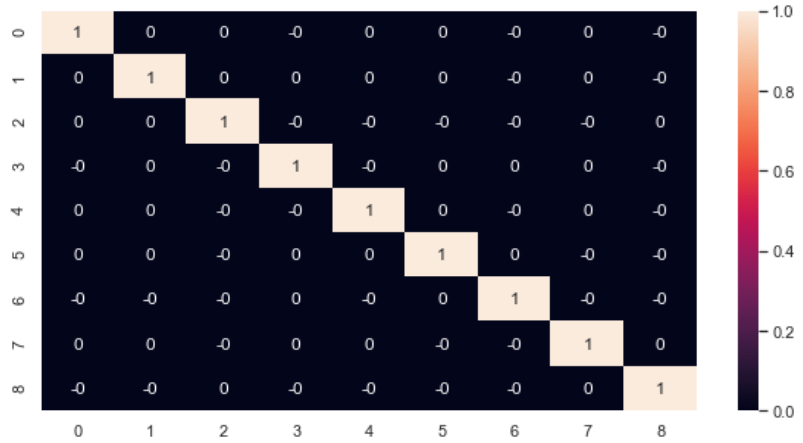


Matriz de correlação da Base 7, na qual a variável clase (*qtd*) foi discretizada utilizando o método *cut*, com normalização nas variáveis meteorológicas e redução da dimensionalidade em 25%.

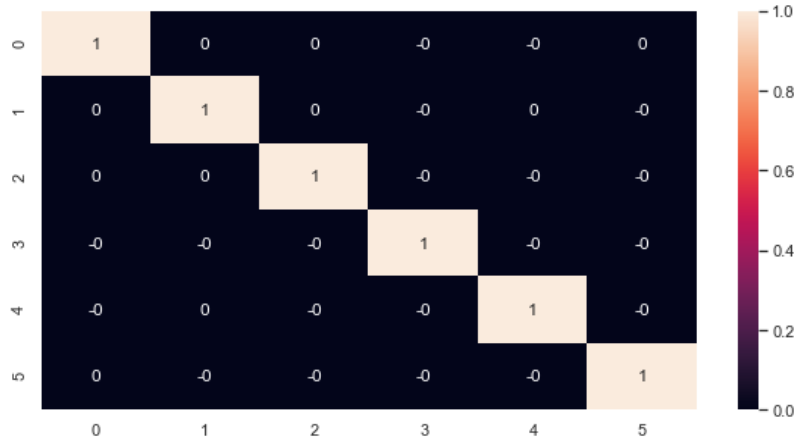


Matriz de correlação da Base 8, na qual a variável clase (*qtd*) foi discretizada utilizando o método *cut*, com normalização nas variáveis meteorológicas e redução da dimensionalidade em 50%.

Legenda: A cor mais clara representa a correlação direta e a escura indica a inexistência de correlação.



Matriz de correlação da Base 9, na qual a variável clase (*qtd*) foi discretizada utilizando o método *qcut*, com normalização nas variáveis meteorológicas e redução da dimensionalidade em 25%.



Matriz de correlação da Base 10, na qual a variável clase (*qtd*) foi discretizada utilizando o método *qcut*, com normalização nas variáveis meteorológicas e redução da dimensionalidade em 50%.

APÊNDICE C – Resultados dos Testes Estatísticos

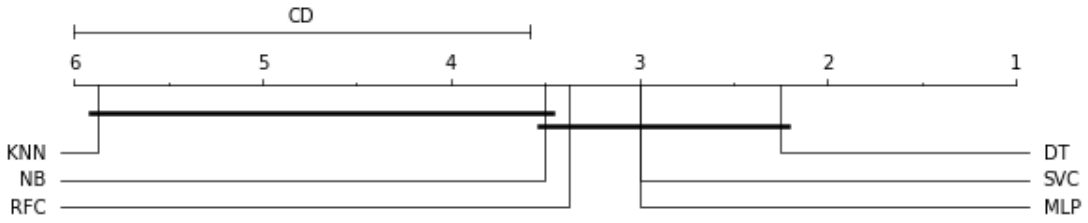


Diagrama de diferença crítica entre os modelos de classificação individuais nas Bases 3 a 10 usando o teste de Nemenyi com base nas acurácias mostradas na Tabela 4.

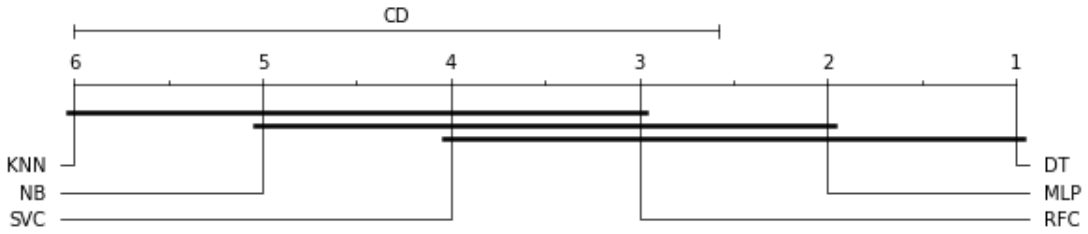


Diagrama de diferença crítica entre os comitês homogêneos aplicados na Base 5 usando o teste de Nemenyi com base nas acurácias mostradas na Tabela 5.

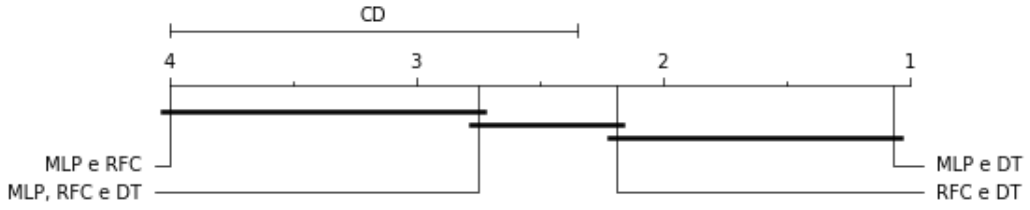


Diagrama de diferença crítica entre os comitês heterogêneos aplicados na Base 5 usando o teste de Nemenyi com base nas acurácias mostradas na Tabela 6.

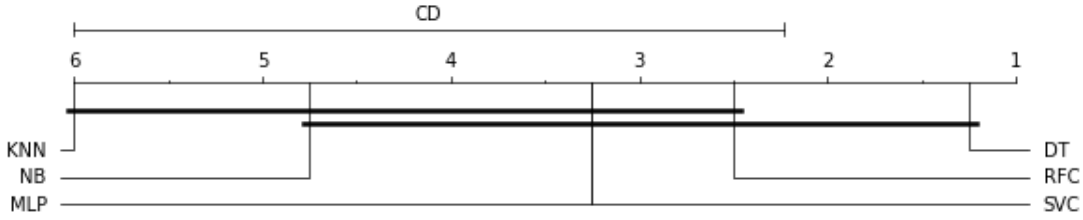


Diagrama de diferença crítica entre os modelos de classificação individuais nas Bases 11 a 14 usando o teste de Nemenyi com base nas acurácias mostradas na Tabela 8.

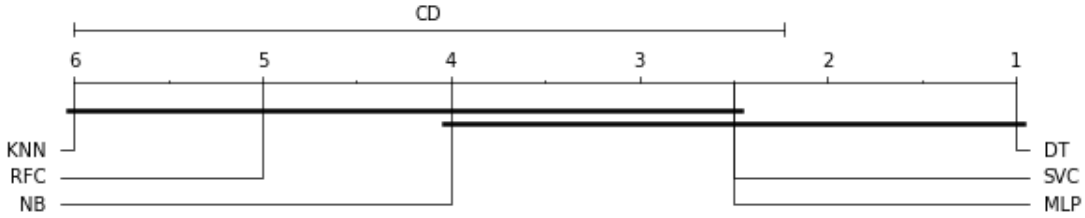


Diagrama de diferença crítica entre os comitês homogêneos aplicados na Base 14 usando o teste de Nemenyi com base nas acurácias mostradas na Tabela 9.

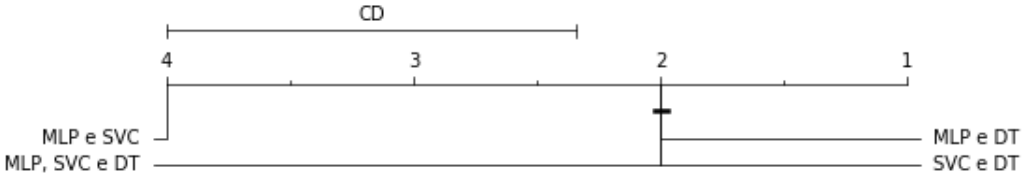
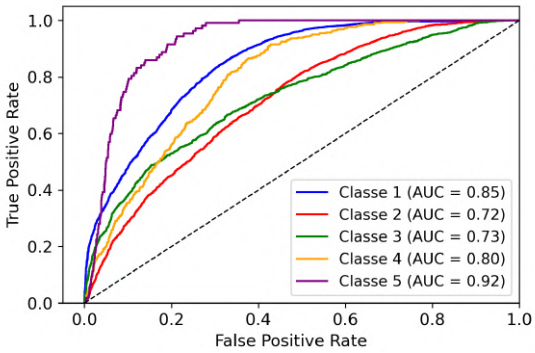
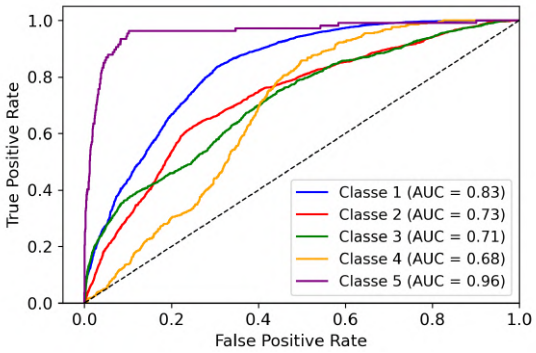


Diagrama de diferença crítica entre os comitês heterogêneos aplicados na Base 14 usando o teste de Nemenyi com base nas acurácias mostradas na Tabela 10.

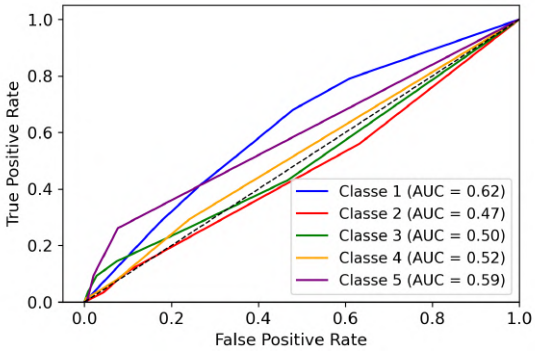
APÊNDICE D – Curva ROC e AUC dos Classificadores Individuais Aplicados nas Bases de Dados



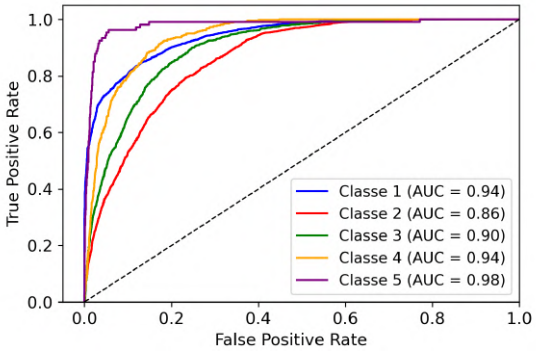
(a) SVC.



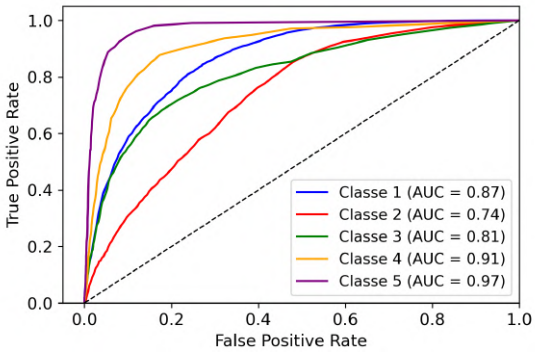
(b) NB.



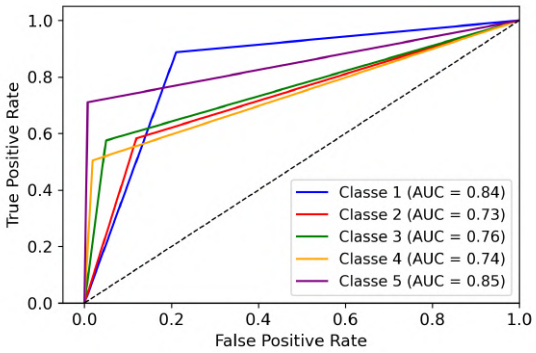
(c) KNN.



(d) MLP.

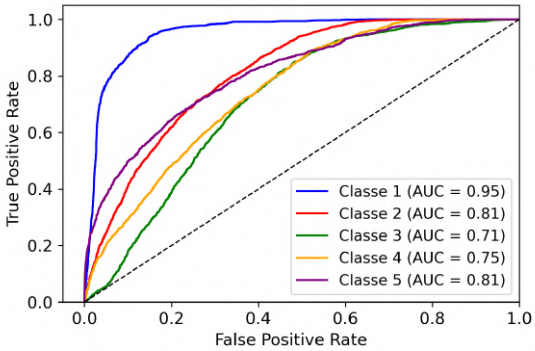


(e) RFC.

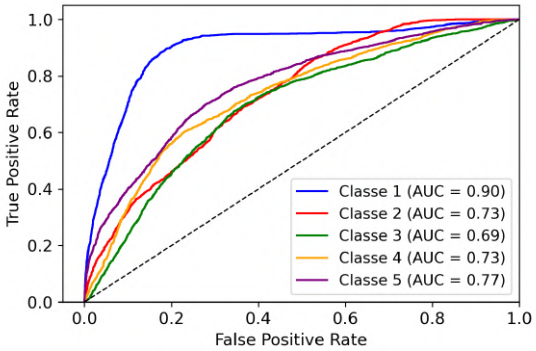


(f) DT.

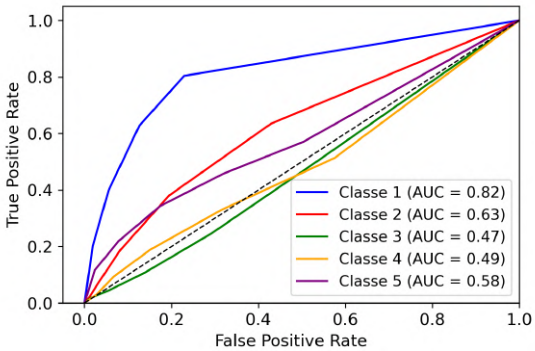
Curva ROC dos algoritmos individuais na Base 3 para cada classe. Legenda das classes que representam as quantidades de aluguéis: Classe 1: “Muito Baixa”; Classe 2: “Baixa”; Classe 3 : “Média”; Classe 4: “Alta”; Classe 5: “Multo Alta”.



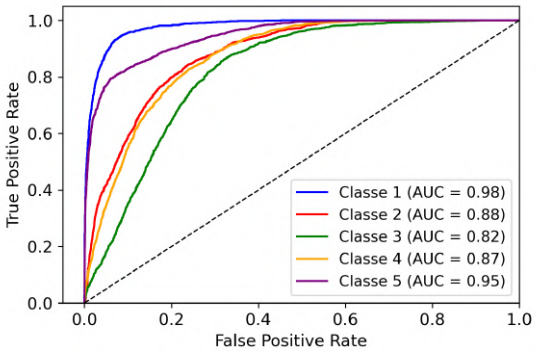
(a) SVC.



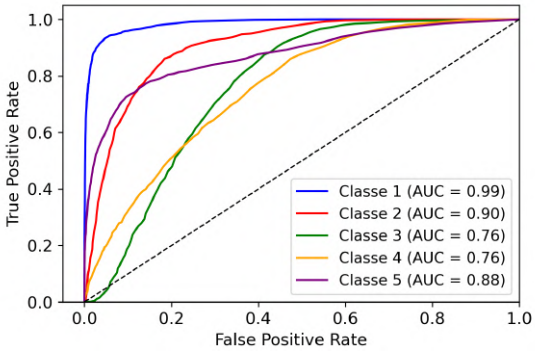
(b) NB.



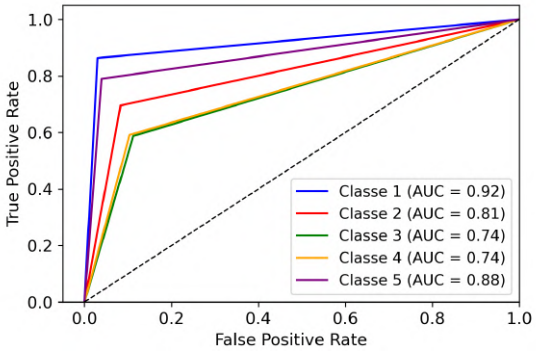
(c) KNN.



(d) MLP.

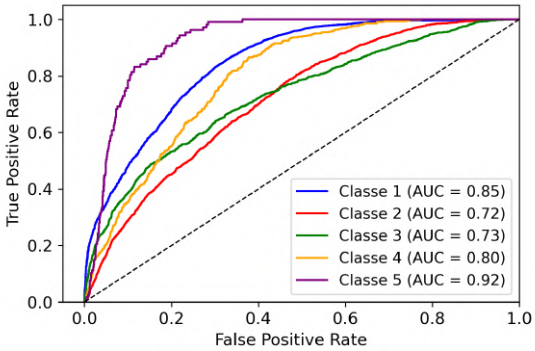


(e) RFC.

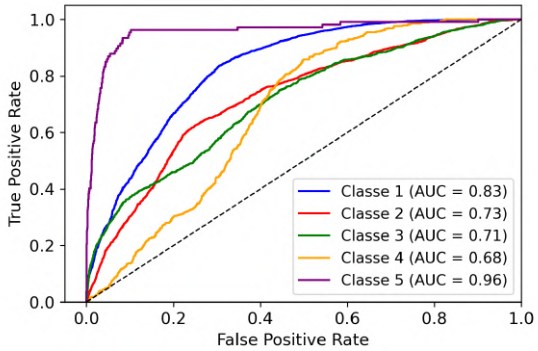


(f) DT.

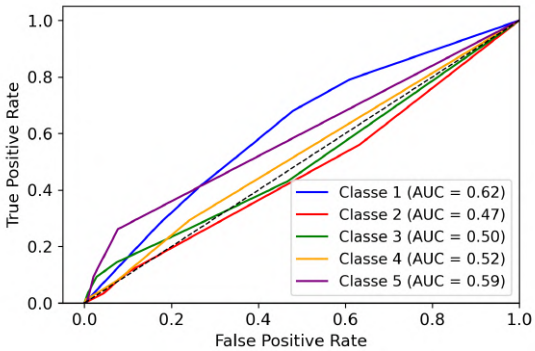
Curva ROC dos algoritmos individuais na Base 4 para cada classe. Legenda das classes que representam as quantidades de aluguéis: Classe 1 = “Muito Baixa”; Classe 2 = “Baixa”; Classe 3 = “Média”; Classe 4 = “Alta”; Classe 5 = “Multo Alta”.



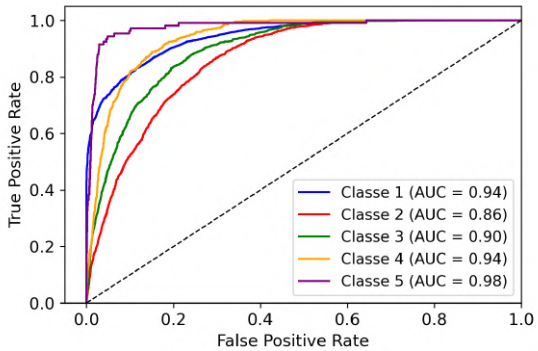
(a) SVC.



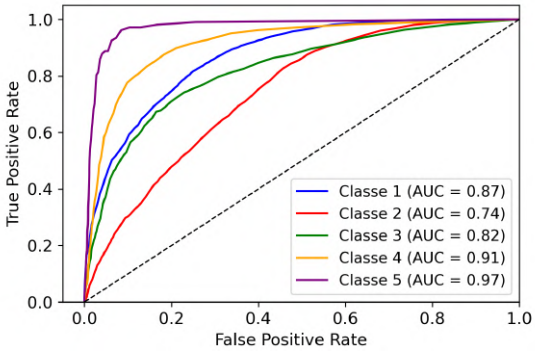
(b) NB.



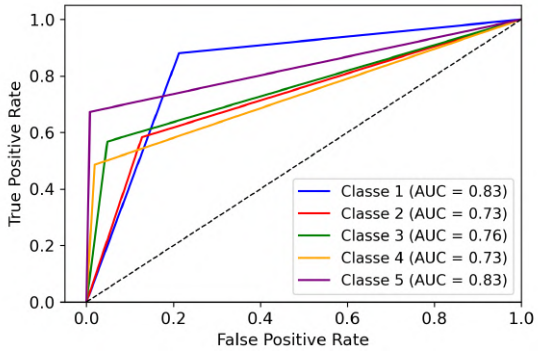
(c) KNN.



(d) MLP.

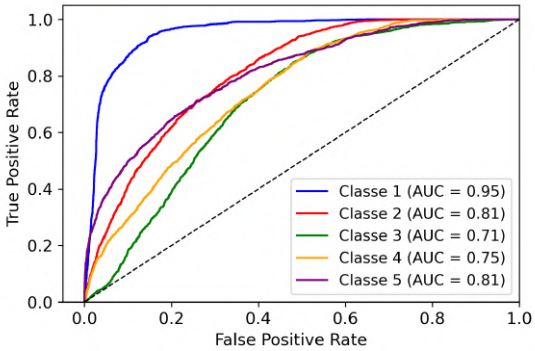


(e) RFC.

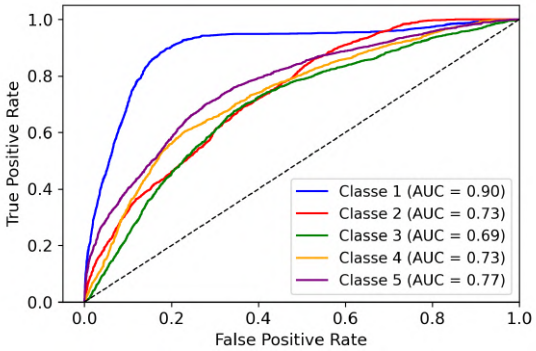


(f) DT.

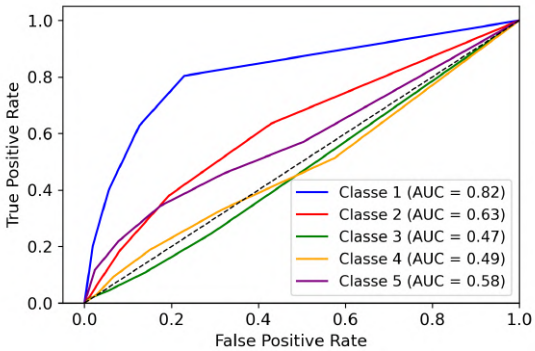
Curva ROC dos algoritmos individuais na Base 5 para cada classe. Legenda das classes que representam as quantidades de aluguéis: Classe 1 = “Muito Baixa”; Classe 2 = “Baixa”; Classe 3 = “Média”; Classe 4 = “Alta”; Classe 5 = “Multo Alta”.



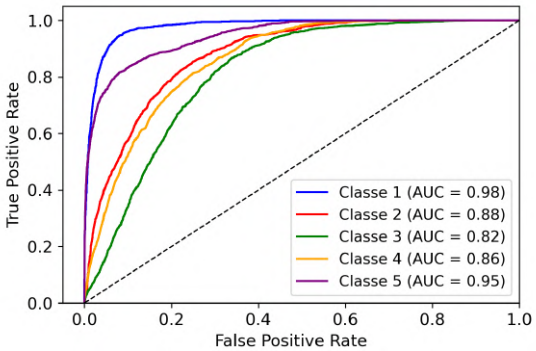
(a) SVC.



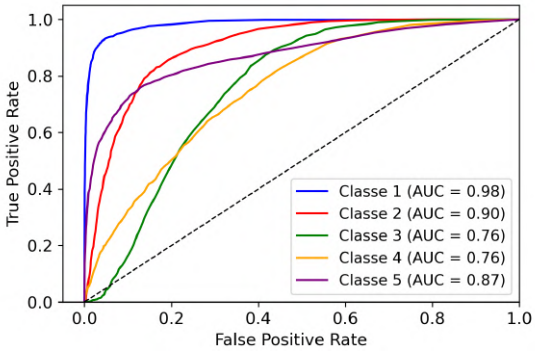
(b) NB.



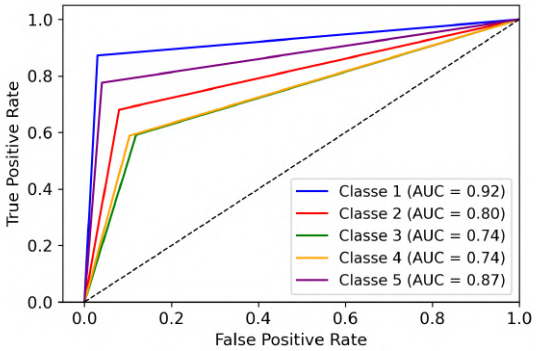
(c) KNN.



(d) MLP.

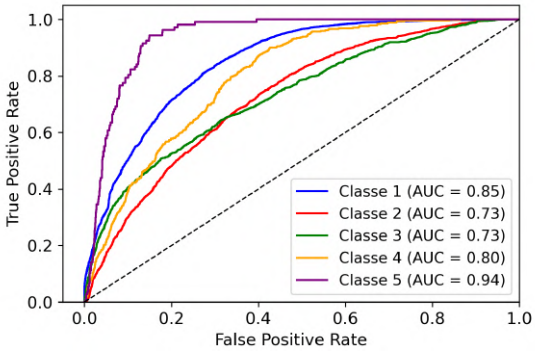


(e) RFC.

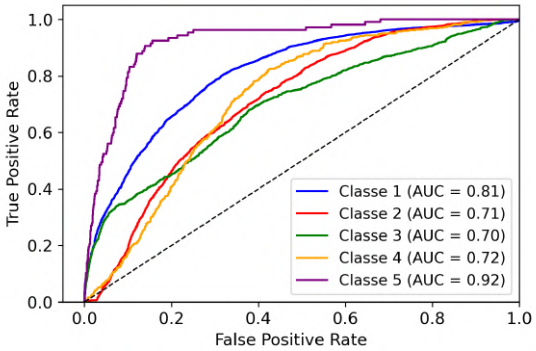


(f) DT.

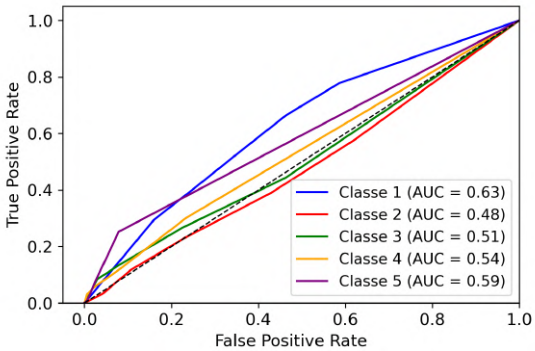
Curva ROC dos algoritmos individuais na Base 6 para cada classe. Legenda das classes que representam as quantidades de aluguéis: Classe 1 = “Muito Baixa”; Classe 2 = “Baixa”; Classe 3 = “Média”; Classe 4 = “Alta”; Classe 5 = “Multo Alta”.



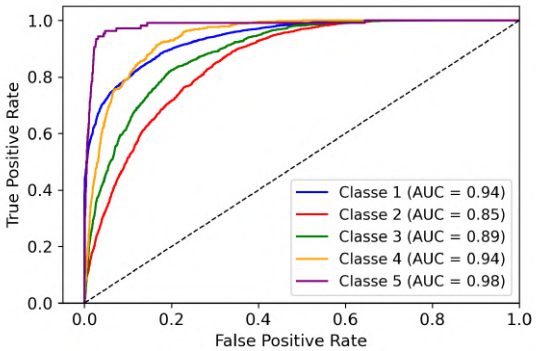
(a) SVC.



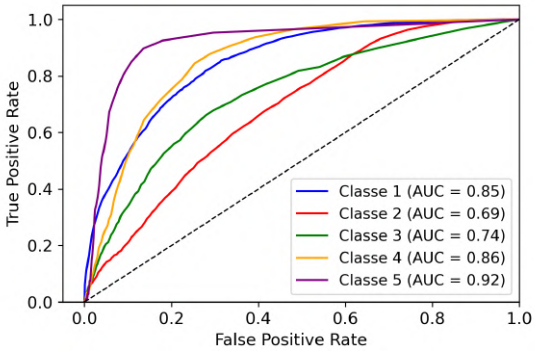
(b) NB.



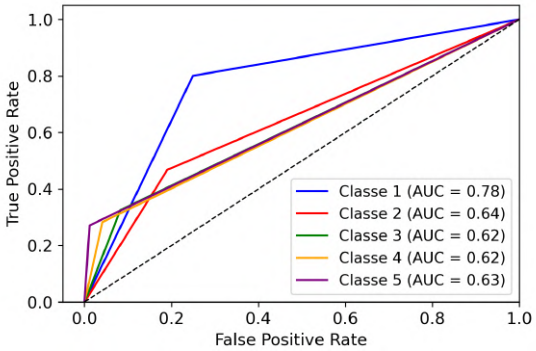
(c) KNN.



(d) MLP.

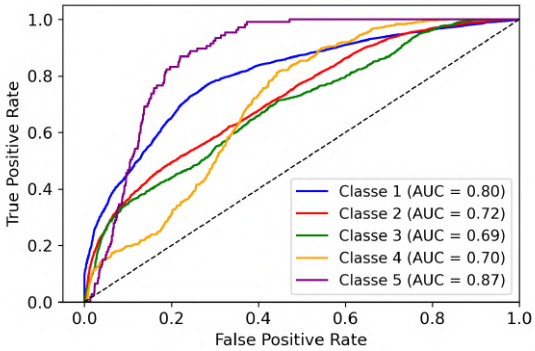


(e) RFC.

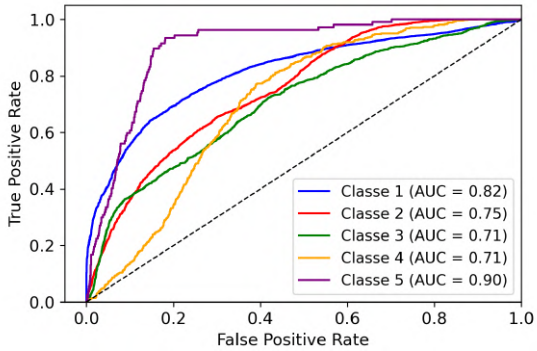


(f) DT.

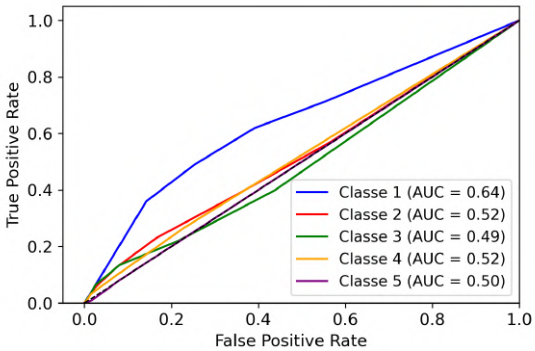
Curva ROC dos algoritmos individuais na Base 7 para cada classe. Legenda das classes que representam as quantidades de aluguéis: Classe 1 = “Muito Baixa”; Classe 2 = “Baixa”; Classe 3 = “Média”; Classe 4 = “Alta”; Classe 5 = “Multo Alta”.



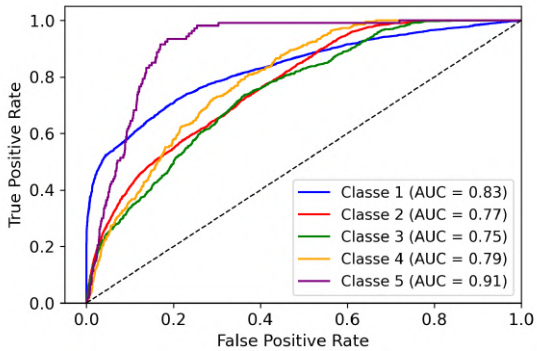
(a) SVC.



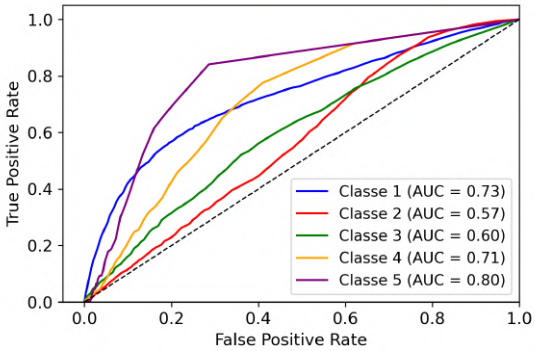
(b) NB.



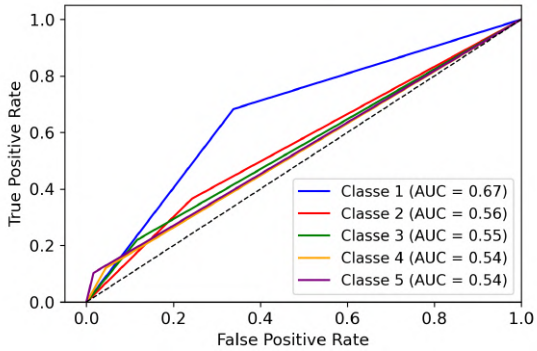
(c) KNN.



(d) MLP.

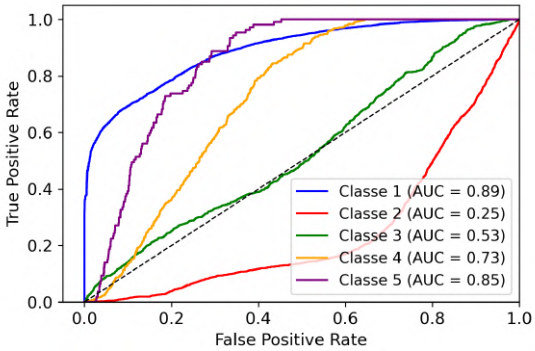


(e) RFC.

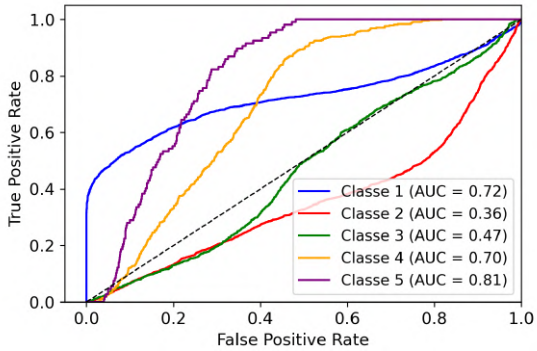


(f) DT.

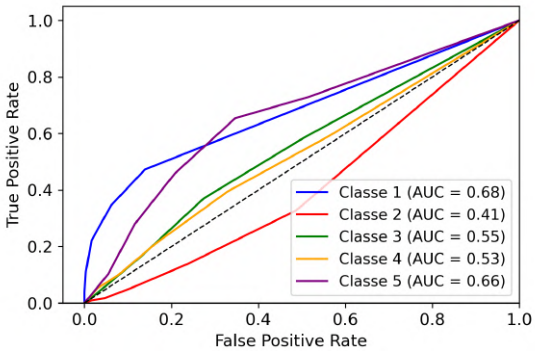
Curva ROC dos algoritmos individuais na Base 8 para cada classe. Legenda das classes que representam as quantidades de aluguéis: Classe 1 = “Muito Baixa”; Classe 2 = “Baixa”; Classe 3 = “Média”; Classe 4 = “Alta”; Classe 5 = “Multo Alta”.



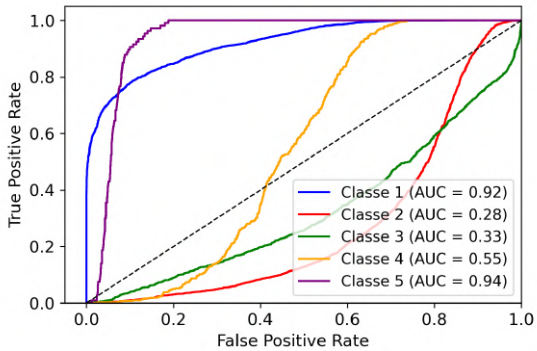
(a) SVC.



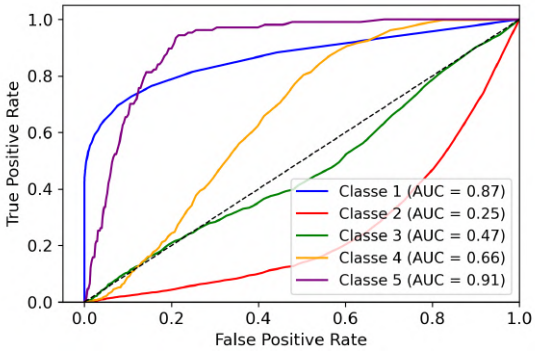
(b) NB.



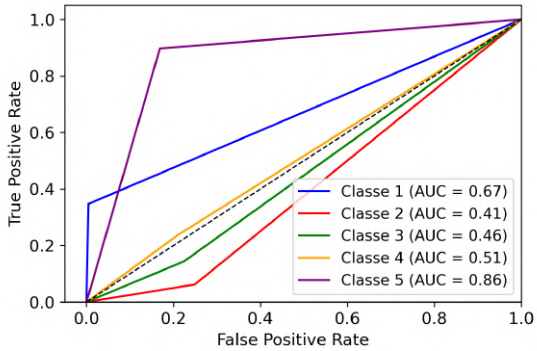
(c) KNN.



(d) MLP.

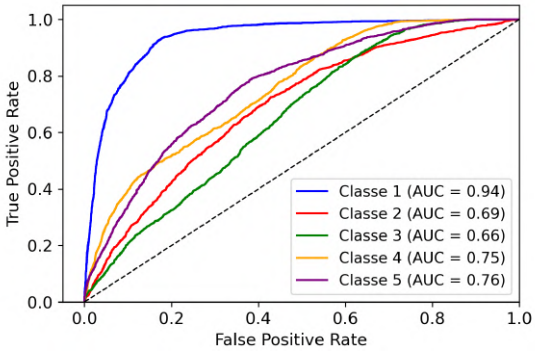


(e) RFC.

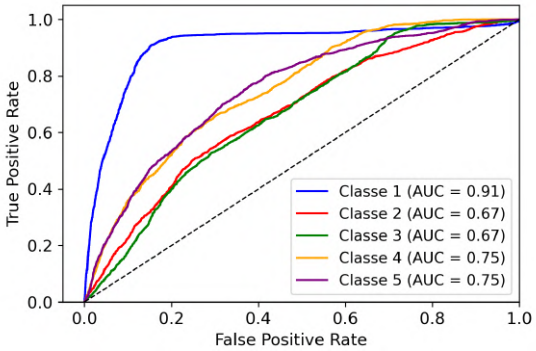


(f) DT.

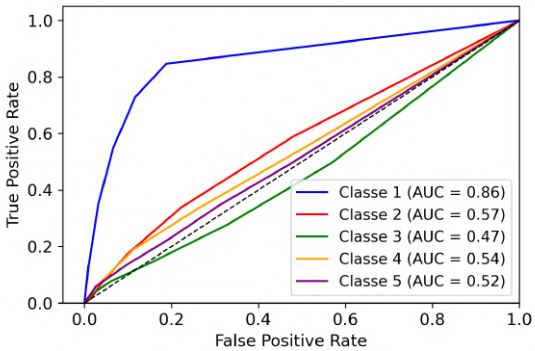
Curva ROC dos algoritmos individuais na Base 9 para cada classe. Legenda das classes que representam as quantidades de aluguéis: Classe 1 = “Muito Baixa”; Classe 2 = “Baixa”; Classe 3 = “Média”; Classe 4 = “Alta”; Classe 5 = “Multo Alta”.



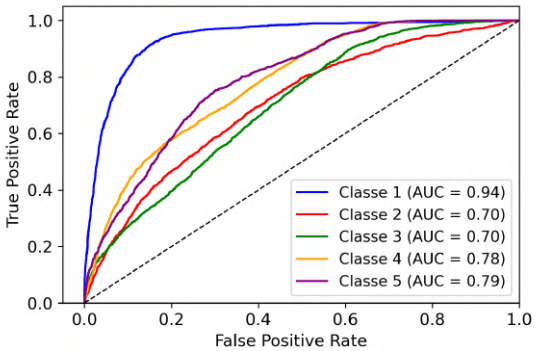
(a) SVC.



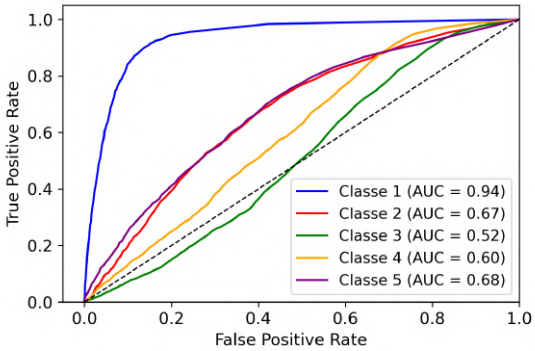
(b) NB.



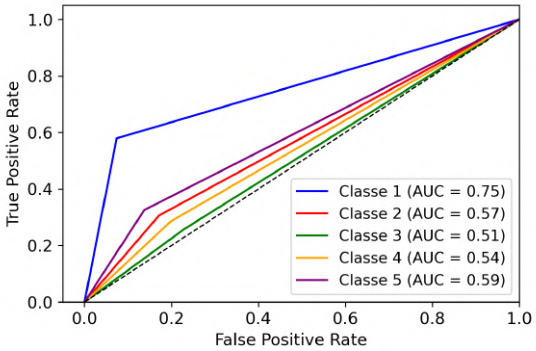
(c) KNN.



(d) MLP.

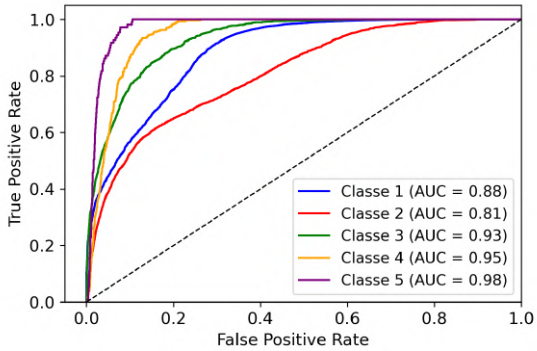


(e) RFC.

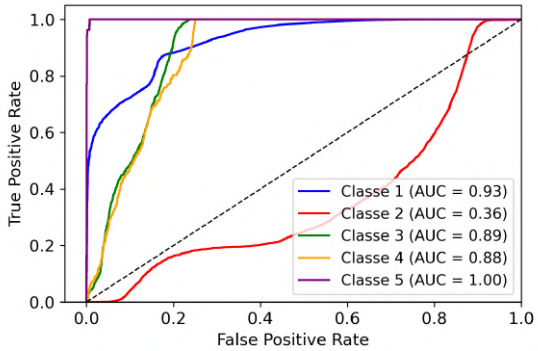


(f) DT.

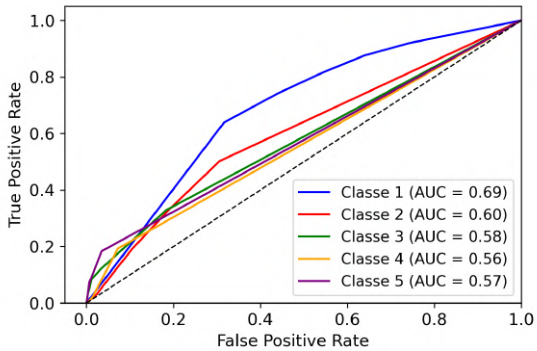
Curva ROC dos algoritmos individuais na Base 10 para cada classe. Legenda das classes que representam as quantidades de aluguéis: Classe 1 = “Muito Baixa”; Classe 2 = “Baixa”; Classe 3 = “Média”; Classe 4 = “Alta”; Classe 5 = “Multo Alta”.



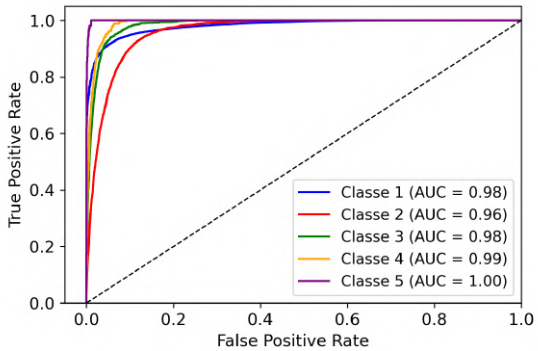
(a) SVC.



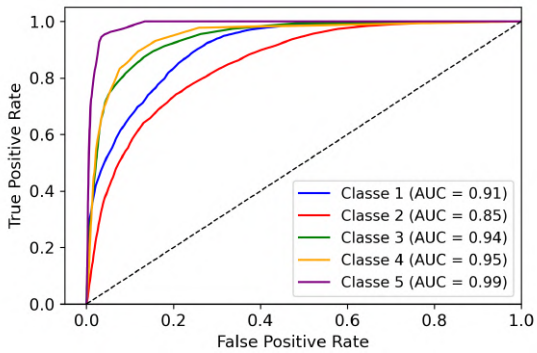
(b) NB.



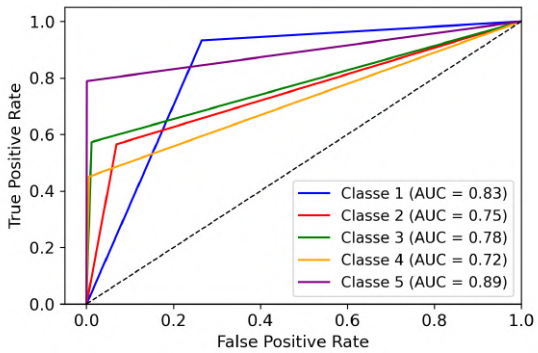
(c) KNN.



(d) MLP.

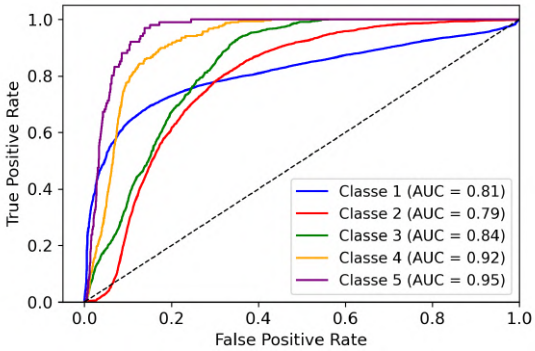


(e) RFC.

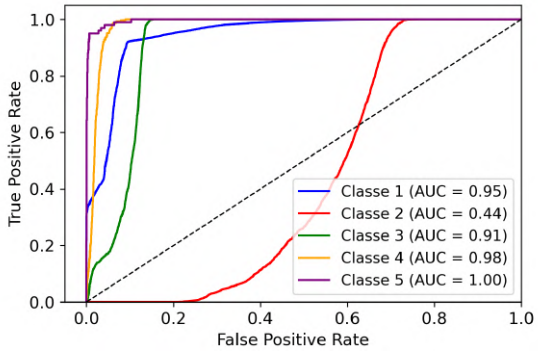


(f) DT.

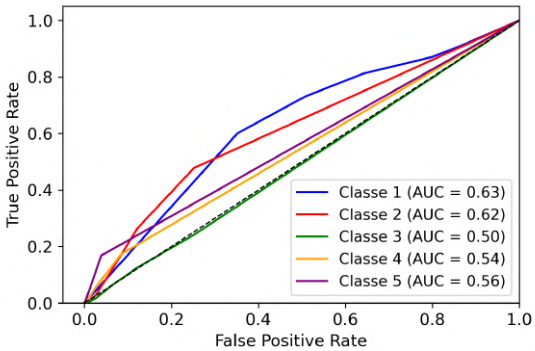
Curva ROC dos algoritmos individuais na Base 11 para cada classe. Legenda das classes que representam as quantidades de aluguéis: Classe 1 = “Muito Baixa”; Classe 2 = “Baixa”; Classe 3 = “Média”; Classe 4 = “Alta”; Classe 5 = “Multo Alta”.



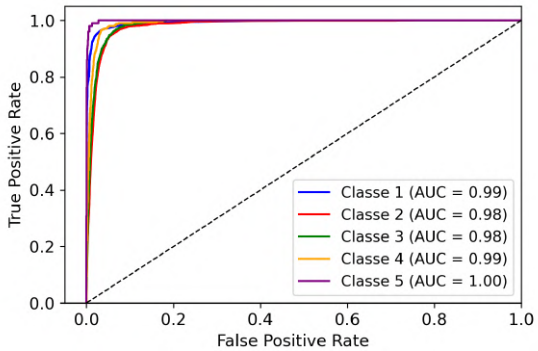
(a) SVC.



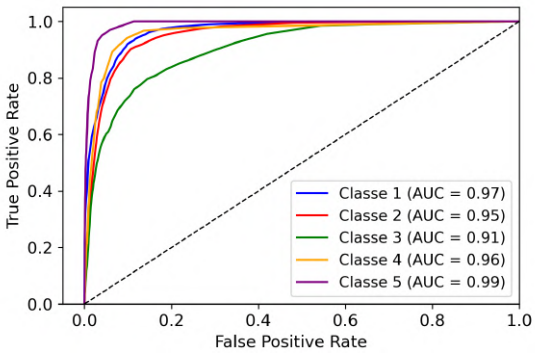
(b) NB.



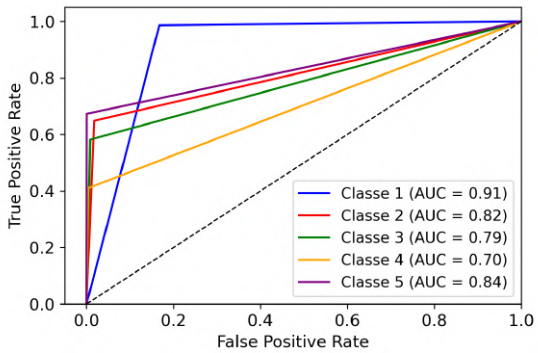
(c) KNN.



(d) MLP.

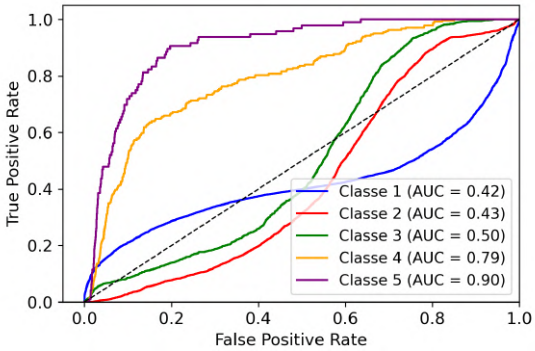


(e) RFC.

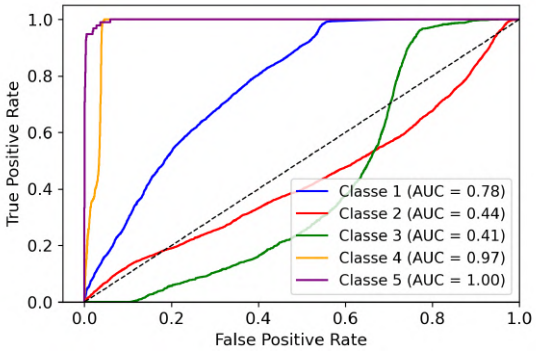


(f) DT.

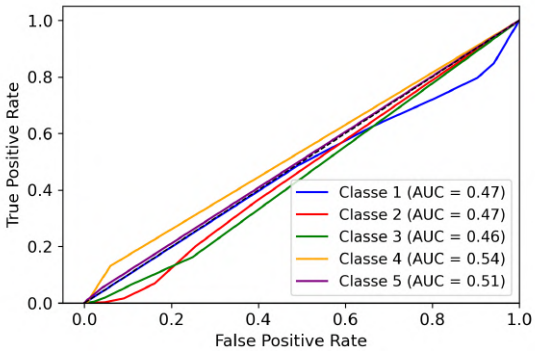
Curva ROC dos algoritmos individuais na Base 12 para cada classe. Legenda das classes que representam as quantidades de aluguéis: Classe 1 = “Muito Baixa”; Classe 2 = “Baixa”; Classe 3 = “Média”; Classe 4 = “Alta”; Classe 5 = “Multo Alta”.



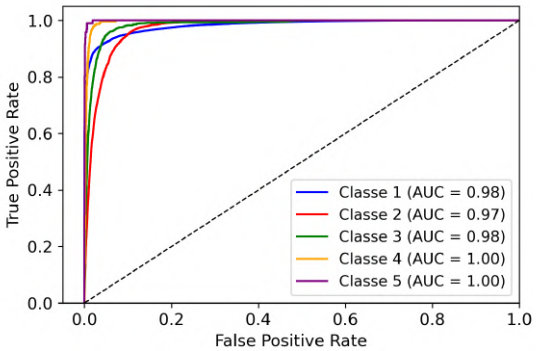
(a) SVC.



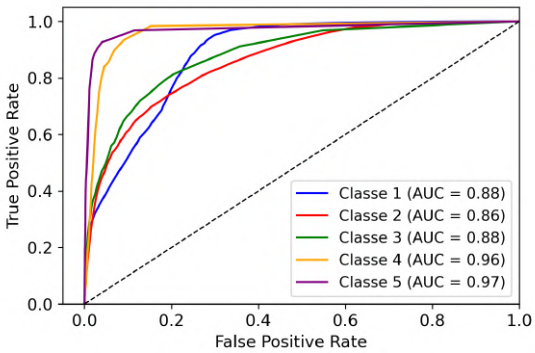
(b) NB.



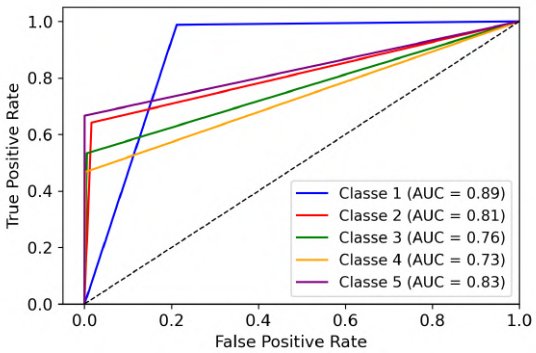
(c) KNN.



(d) MLP.

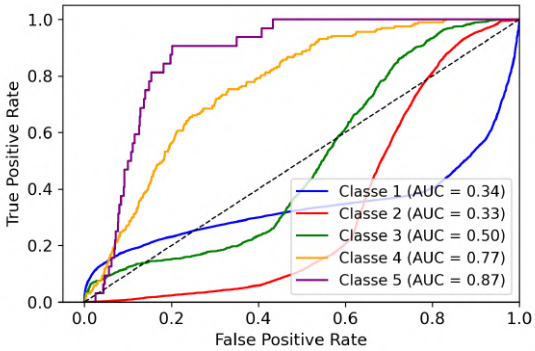


(e) RFC.

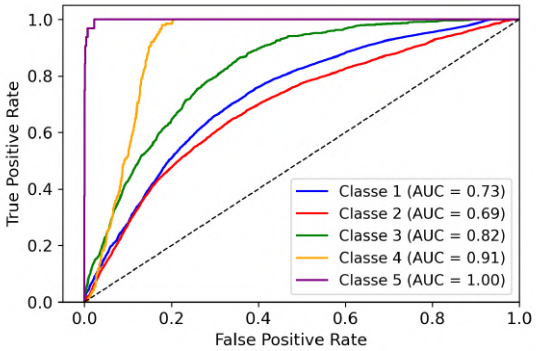


(f) DT.

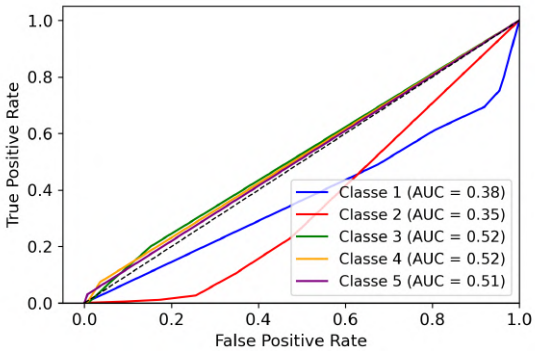
Curva ROC dos algoritmos individuais na Base 13 para cada classe. Legenda das classes que representam as quantidades de aluguéis: Classe 1 = “Muito Baixa”; Classe 2 = “Baixa”; Classe 3 = “Média”; Classe 4 = “Alta”; Classe 5 = “Multo Alta”.



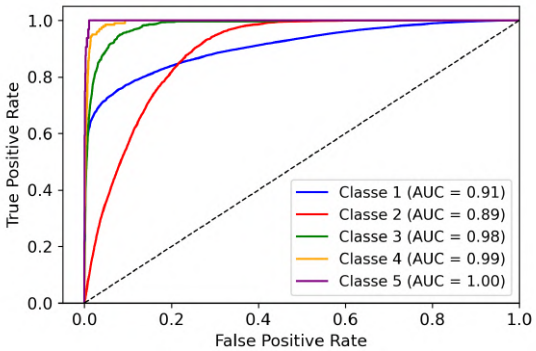
(a) SVC.



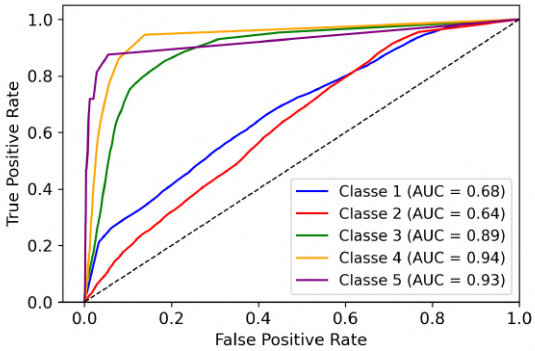
(b) NB.



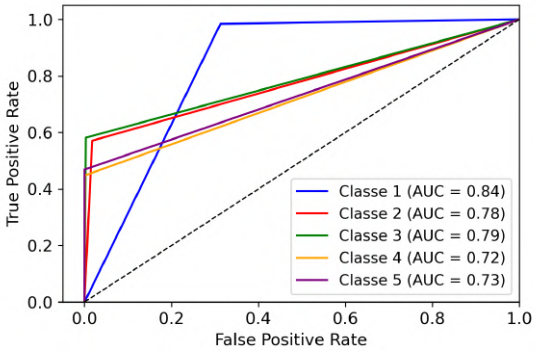
(c) KNN.



(d) MLP.



(e) RFC.



(f) DT.

Curva ROC dos algoritmos individuais na Base 14 para cada classe. Legenda das classes que representam as quantidades de aluguéis: Classe 1 = "Muito Baixa"; Classe 2 = "Baixa"; Classe 3 = "Média"; Classe 4 = "Alta"; Classe 5 = "Multo Alta".