Algoritmos de Classificação Aplicados ao Controle Interno em Economicidade de Combustíveis

Classification algorithms applied to internal control of fuel economy

Ailton de Souza Leite [1]
orcid.org/0000-0001-7880-7942

André Luiz da S. Xavier [1]
orcid.org/0000-0001-2106-9720

George Fragoso Andrade[2]
orcid.org/0000-0003-1022-8853

Igor Vitor Teixeira [1]
orcid.org/0000-0002-5334-884X

Rubens Karman P. Silva[1]
orcid.org/0000-0001-9388-9889

[1]Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil. E-mail: asl@ecomp.poli.br

[2]Polícia Militar de Pernambuco, Recife, Brasil. E-mail: george.fragoso@sds.pe.gov.br

DOI: 10.25286/repa.v6i5.1751

Como citar este artigo pela NBR 6023/2018: LEITE, A. S.; XAVIER, A. L. S.; ANDRADE, G. F.; TEICEIRA, I. V.; SILVA, R. K. P. Algoritmos de Classificação Aplicados ao Controle Interno em Economicidade de Combustíveis. Revista de Engenharia e Pesquisa Aplicada, Recife, v.6, n. 5, p. 27-36, Novembro, 2021.

RESUMO

Este artigo apresenta técnicas de aprendizagem de máquina para identificar não conformidades no consumo de combustível dos veículos da Polícia Militar de Pernambuco - PMPE. A base de dados utilizada refere-se ao período de 2017 a 2020, com 18 atributos e mais de 900 mil registros de abastecimentos. Dos 124 modelos distintos, a pesquisa utilizou os modelos com maior frequência absoluta. Sendo estes: SPIN, SPACEFOX, XRE 300, HILUX e S10, distribuídos no interior do estado, Região Metropolitana do Recife - RMR e capital. Após as etapas de mineração dos dados, a base foi segmentada em 70% para treino (472.155 registros) e 30% para testes (202.353 registros). Em seguida, foram aplicadas técnicas de grid search, para avaliação dos classificadores Naive Bayes, KNN e Random Forest, considerando as métricas de acurácia, sensibilidade e precisão. O KNN foi o classificador que apresentou o melhor desempenho. Enquanto que, o Naive Bayes apresentou os piores resultados, obtendo uma acurácia inferior aos demais (94,30%).

PALAVRAS-CHAVE: Aprendizagem de Máquina, Algoritmos de Classificação; Mineração de Dados; Polícia Militar de Pernambuco - PMPE, Economicidade de Combustível.

ABSTACT

This article presents machine learning techniques to identify non-conformities in the fuel consumption of vehicles of the Military Police of Pernambuco - PMPE. In this sense, a database used refers to the period from 2017 to 2020, with 18 attributes and more than 900 thousand supply records. Of the 124 distinct models, a survey using the models with the highest absolute frequency. These being: SPIN, SPACEFOX, XRE 300, HILUX and S10, distributed in the interior of the state, Metropolitan Region of Recife - RMR and capital. After the data mining steps, the base was segmented into 70% for training (472,155 records) and 30% for tests (202,353 records). Then, the grid research techniques were applied to evaluate the Naive Bayes, KNN and Random Forest classifiers, considering the metrics of accuracy, sensitivity and precision. The KNN was the classifier that presents the best performance. While, Naive Bayes presents the worst results, obtaining a lower accuracy than the others (94.30%).

KEY-WORDS: Machine Learning, Classification Algorithms; Data Mining; Military Police of Pernambuco - PMPE, Fuel Economy;

1 INTRODUÇÃO

Gradualmente, a gestão pública vem modernizando suas rotinas internas, através do uso de recursos tecnológicos, a fim de oferecer serviços mais transparentes, eficientes e idôneos, tanto do ponto de vista econômico como operacional. Exemplos como os dados públicos abertos, portais de transparência e ferramentas que possibilitam o registro de solicitações por parte dos cidadãos, são medidas pertinentes para empoderar as ações de controle social, que consistem em uma maior fiscalização da sociedade sobre o Estado.

Em sintonia com esse processo de modernização, os mecanismos de controle interno do estado de Pernambuco não são diferentes, realizando a compreensão do planejamento e a orçamentação dos meios, a execução das atividades planejadas e a avaliação periódica da atuação.

A auditoria interna, desenvolvida por técnicos do quadro da organização, tem uma fundamental importância, abrangendo o estudo e a avaliação permanente do sistema de controle interno, sua adequação e desempenho [1]. O processo de auditoria é realizado através do monitoramento das folhas de pagamento, serviços terceirizados, locações e custeio para a manutenção dos serviços públicos. Não poderia ser diferente com os combustíveis, objeto de pesquisa deste trabalho, item de vital importância para a conservação e continuidade de vários serviços oferecidos ao cidadão. O presente trabalho visa identificar irregularidades no consumo de combustível dos veículos da PMPE (Polícia Militar de Pernambuco).

A PMPE, através do artigo 144 da constituição federal, lhe é atribuída a qualidade de polícia ostensiva e preservação da ordem pública do estado de Pernambuco. Em sintonia com a PMPE, a CPA (Comissão Permanente de Auditoria), instituída através do Decreto Estadual 24.629/2002 tem por objetivo avaliar o desempenho da gestão operacional analisando os indicadores de eficiência, eficácia e efetividade relacionadas às atividades operacionais, prezando pela qualidade da gestão financeira, patrimonial, combustíveis, materiais bélicos, almoxarifado e recursos humanos [2].

1.1 DESCRIÇÃO DO PROBLEMA

Tendo em vista o dever de zelar pelos recursos públicos, a CPA tem a missão de aplicar as normas internas da legislação vigente e das diretrizes traçadas pela administração. A auditoria de combustível, que é uma das atribuições da CPA, é deveras complexa devido à sua pluralidade de possibilidades de alteração dos parâmetros básicos. Muitas vezes é exercida com base na experiência do auditor.

Nas unidades de gestão dos batalhões da PMPE, os colaboradores que são responsáveis pelas atividades internas, também são encarregados de extrair os dados referentes aos serviços realizados nos veículos do batalhão. Estes dados são extraídos de um sistema terceirizado, o Nutricash. Os dados extraídos são analisados previamente, a fim de detectar alguma anomalia ainda nessa etapa. Posteriormente, todas as unidades de gestão encaminham seus relatórios de serviços para a CPA, onde serão analisados pela comissão.

1.2 OBJETIVO

O objetivo deste projeto consiste em utilizar técnicas de aprendizagem de máquina para realizar a classificação dos dados, a fim de identificar, dentro do universo de registros existentes na base de dados, se determinado veículo está dentro da conformidade de economicidade ou não, baseados nos critérios adotados pela CPA. Ainda se espera que com esse trabalho a instituição tenha uma ferramenta que auxilie os auditores da CPA em seus processos diários.

1.3 JUSTIFICATIVA

Mensalmente os relatórios referentes aos serviços realizados nos veículos da PMPE são emitidos e encaminhados para o setor de auditoria da CPA. Oriundo de 97 unidades de gestão distribuídas ao longo do estado de Pernambuco, esses relatórios carregam dados de cerca de 4.444 veículos e mais de 900 mil registros de serviços automotivos de abastecimentos e outros.

Após o envio para a CPA, os relatórios passam por avaliação dos colaboradores lotados no referido órgão. Por ser um trabalho minucioso, envolvendo uma grande quantidade de informação e devido ao reduzido número de colaboradores, muitas vezes o tempo para análise dos relatórios não é apropriado.

O presente projeto não tem interesse em interferir nos critérios e percepções dos auditores em avaliar conformidade ou não conformidade dos veículos.

2 FUNDAMENTAÇÃO TEÓRICA

Atualmente, a PMPE é responsável por manter a ordem pública do estado, contando com um quadro de aproximadamente 17 mil efetivos para o funcionamento de suas atividades [2]. Em conjunto com o Corpo de Bombeiros Militar, Polícia Civil e Polícia Científica, compondo assim a SDS (Secretaria de Defesa Social), detém um orçamento anual de mais de 5 bilhões de reais para o ano de 2021 [3]. Devido aos volumosos recursos destinados à SDS, é necessário o acompanhamento e auditoria desses gastos a fim de garantir a idoneidade referente à gestão pública quanto à legalidade, legitimidade e economicidade.

A continuidade da função da PMPE depende, quase que integralmente, de sua mobilidade e acesso a veículos de trabalho. O presente trabalho visa analisar o consumo de combustível, que é algo indispensável para atividade policial e que, entre os anos de 2017 e 2020, custaram mais de 119 milhões de reais para os cofres públicos. Durante esse período foram registrados mais de 904 mil registros de abastecimento.

2.1 MINERAÇÃO DE DADOS

Impulsionado pela crescente utilização de recursos tecnológicos no dia a dia das pessoas e das organizações, considerando um aumento da capacidade de armazenamento e velocidade de leitura de dados, *hardware* e sistemas distribuídos escaláveis que oferecem poder de processamento cada vez maior convergindo assim para o aprofundamento das técnicas de mineração de dados, propiciando o aperfeiçoamento de métodos que possibilitam extrair, manipular e analisar grandes volumes de dados, levando à descoberta de conhecimento e posterior tomada de decisão.

A mineração de dados consiste no emprego de algoritmos capazes de extrair conhecimento a partir dos dados pré-processados [4]. Dentro de grandes quantidades de dados, a mineração de dados tem a finalidade de encontrar padrões ou similaridades que possibilitem tomada de decisão baseada em dados e percepções de eventos futuros.

As possíveis aplicações da mineração de dados, de acordo com [5], são: (i) classificação e estimativa probabilística que tem como objetivo prever, para cada indivíduo, qual classe ele pertence; (ii) regressão para estimar ou prever, para cada indivíduo, o valor numérico de alguma variável; (iii) agrupamento para reunir indivíduos de uma população por meio de suas similaridades; (iv) redução de dados para substituir um grande conjunto de dados por outro menor que contenham grande parte das informações importantes do conjunto maior; e entre outros exemplos citados pelo autor.

Neste trabalho foi utilizada uma abordagem utilizando algoritmos de classificação de dados.

2.1.1 Algoritmos de Classificação

Em diversos casos práticos é possível chegar à constatação de determinada característica ou evento, observando seus registros históricos, ou seja, é possível chegar a um determinado resultado ponderando as características e resultados anteriores.

Dentre as abordagens de aprendizado para a classificação se encontra o aprendizado supervisionado, de acordo com [5], para essa proposta um alvo específico dentro da tabela de dados é exigido. Além disso, segundo o autor, deve haver outros dados sobre o alvo, chamados de rótulo individual, pois sua ausência implicará na ausência do alvo.

Considerando os algoritmos de classificação encontrados na literatura, para o presente trabalho, serão utilizados os algoritmos *Naive Bayes, K-Nearest Neighbors* (KNN) e *Random Forest*.

O algoritmo *Naive Bayes*, é um classificador que utiliza técnicas probabilísticas, utilizando como base os eventos mais prováveis de ocorrer, de acordo com as suas variáveis independentes [6]. Este algoritmo possui diferentes tipos de implementação, como Bernoulli, Gaussiano, Multinomial, entre outros. Para este trabalho, foi selecionado o tipo Gaussiano.

Já o algoritmo KNN determina uma quantidade mínima de vizinhos, representado pela variável *k*, onde é calculado a distância entre cada atributo com relação aos *k* vizinhos. Com isso, a classe de saída que aparece com mais frequência dentre os *k* vizinhos é selecionada [7].

O algoritmo *Random Forest,* de acordo com [4], é uma estrutura em representação de árvore na qual cada *nó interno* corresponde a um teste de um atributo, cada *ramo* representa um resultado do teste e os *nós* folhas representam as classes ou as distribuições de classes. O nó folha mais elevado da árvore é conhecido como nó raiz, e cada caminho da raiz até um nó folha corresponde a uma regra de classificação.

DOI: 10.25286/repa.v6i5.1751

2.2 TRABALHOS RELACIONADOS

Os autores [8] exploraram o consumo de combustível das viaturas da PMPE entre o período de 2018 e 2020, a fim de auxiliar o controle interno, propondo a implementação de um modelo computacional para realizar a previsão do consumo de combustíveis da frota de veículos da instituição. Identificando a sazonalidade existente nos dados, a ferramenta apresentou uma acurácia de 84% em prever o consumo de combustíveis das viaturas para momentos futuros, o que, segundo os autores, venha a facilitar o planejamento financeiro da corporação. Além disso, caso a previsão seja aplicada à um momento presente, os resultados podem vir a auxiliar os auditores em distinguir uma possível não conformidade nos registros devido a discrepância entre o que foi previsto e o que foi realizado.

No artigo, os autores [9], apontam que os valores padrões da FTP (do inglês, *Federal Test Procedure*) e o WLTC (do inglês, *Worldwide Harmonized Light Vehicles Test Cycle*) não podem ser usados para estimar o consumo real de combustível, nem as emissões dos veículos em uma região. Os autores afirmam que, além do ciclo de condução, ou seja, o valor padrão de emissões e consumo médio de combustível em determinada região, há de se considerar o padrão de direção que é descrito como um conjunto de parâmetros característicos. Os conjuntos de parâmetros podem ser descritos por um conjunto de velocidade média, energia cinética positiva e porcentagem de tempo de marcha lenta. O referido artigo é relevante por afirmar que existem outros fatores que contribuem para o aumento ou diminuição da eficiência de consumo de combustível por veículos de passeio. Corroborando com o método de auditoria utilizado pela CPA em, não apenas se basear nos parâmetros do Inmetro, mas levando em consideração seu histórico, região onde a viatura atua e componentes agregados ao veículo.

Os autores [10], em seu trabalho, expõem um estudo de caso de abordagens de técnicas de aprendizagem de máquina baseada em classificação que podem apoiar o planejamento estratégico de auditorias, tendo em vista o objetivo de maximizar os benefícios da auditoria e minimizar seus custos. A base de dados do referido trabalho consistiu em declarações fiscais integradas com outras fontes de dados. E como resultado, é apresentada uma metodologia para a construção de perfis de contribuintes fraudulentos, com o objetivo de apoiar a auditoria, onde foi adotada a abordagem KDD (do inglês, *Knowledge Discovery in Databases*) e a técnica de classificação por árvore de decisão. Diferentemente do trabalho citado, o presente artigo foca em aspectos relacionados a valores discrepantes de abastecimento e a metodologia adotada é a CRISP-DM (do inglês, *Cross Industry Standard Process for Data Mining*). No entanto, deixa explícito o potencial das abordagens de classificação para apoiar ações de auditoria.

3 MATERIAIS E MÉTODOS

3.1 DESCRIÇÃO DA BASE DE DADOS

Os dados foram obtidos a partir de relatórios extraídos mensalmente do sistema Nutricash, com dados referentes aos gastos de cada batalhão da PMPE. A base de dados é composta pelos serviços realizados para a manutenção da frota de veículos, como abastecimento, troca de óleo, troca de lubrificante e serviços de borracharia. Também são encontrados dados relacionados ao consumo de combustível dos veículos e a quilometragem percorrida desde o último abastecimento. Esses dados são discriminados para cada veículo do batalhão, sendo informado a marca, o ano de fabricação, o ano do modelo do veículo, como também o tipo do veículo.

A base de dados foi obtida segmentada em duas partes distintas, a primeira com registros referentes ao ano de 2017 e a segunda com registros entre 2018 e 2020. Após a integração das bases, foi verificada a quantidade de 904.863 registros, com 18 atributos.

A fim de realizar a integração dos dados, foi necessária a adequação dos atributos de ambas as bases, padronização e preservação dos atributos que possuíam em comum.

Para melhor análise dos dados e entendimento do problema de negócio, foram realizadas uma série de procedimentos que, neste trabalho, iremos considerar como etapa de integração e adequação das bases para as etapas seguintes.

Seguem os passos realizados neste primeiro momento: como a base foi recebida fragmentada em duas partes, um arquivo com registros de 2017 e outro com 2018, 2019 e 2020, foi necessário padronizar seus cabeçalhos e remover colunas que não estavam contempladas em ambas; Para melhorar a visualização dos dados, os veículos foram separados em 5 classes e distribuídos por

região onde atuam, esse procedimento será descrito melhor abaixo; Foram removidos os registros que não possuíam informações referentes ao centro de custo, como também foram removidos os registros com consumo menor ou igual a 0 (zero); Para extrair as métricas de consumo de combustível, os dados da coluna CONSUMO foram convertidos para o tipo de dado ponto flutuante. Os atributos preservados estão descritos no dicionário de dados e estão descritos no Quadro 1.

Quadro 1: Dicionário de Dados

ATRIBUTO	DESCRIÇÃO
Autorização	Identificação do registro de lançamento
Nome Fantasia	Nome fantasia do local que foi realizado o serviço
Cidade	Cidade em que foi realizado o serviço
Uf	Estado em que foi realizado o serviço
Serviço	Descrição do serviço que foi realizado
Hodômetro	Registro da quilometragem do veículo no momento da execução do serviço
Deslocamento	Quilometragem que o veículo irá realizar quando o serviço é abastecimento
Consumo	Consumo médio de combustível do veículo que executou o serviço
Quantidade	Quantidade de itens para o serviço executado
Unitário	Valor unitário do serviço executado
Valor	Valor total do serviço executado
Datahora Trans	Data e hora do momento que foi registrado a execução do serviço
Veículo	Identificação do veículo em que foi realizado o serviço
Tipo Veiculo	Tipo do veículo que foi realizado o serviço
Modelo Veiculo	Modelo do veículo que foi realizado o serviço
Condutor	Condutor do veículo que foi realizado o serviço
Centro De Custo	Centro de custo do batalhão que irá receber a despesa da execução do serviço
Região	Região onde o veículo trafega (Capital, Região metropolitana do Recife ou Interior).

Fonte: Os Autores.

3.2 ANÁLISE DESCRITIVA DOS DADOS

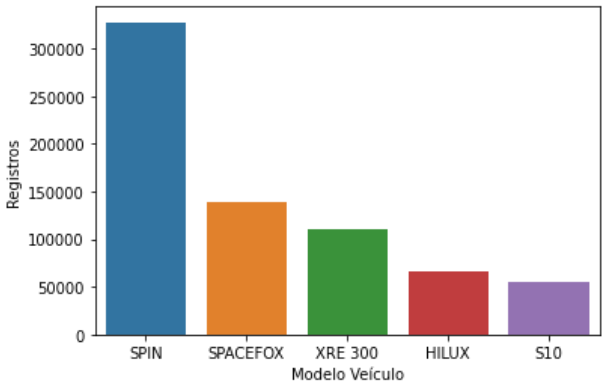
Inicialmente a base de dados contém registros de abastecimento de 7.432 veículos utilizados pelos batalhões da PMPE, os quais são categorizados em 124 modelos de veículos diferentes.

Para realizar a análise descritiva dos dados, foram selecionados os atributos CONSUMO e MODELO VEICULO, os quais representam o consumo do veículo por quilômetro por litro (km/L) e o modelo do veículo, respectivamente. O primeiro foi considerado como um atributo numérico, e o segundo, nominal. O consumo do veículo é calculado a partir do atributo de DESLOCAMENTO, que é medido em quilômetros, e é dividido pelo valor do atributo QUANTIDADE, que é referente a quantidade de litros que foi inserido no veículo.

Como existem 124 diferentes modelos de veículos na base, um agrupamento foi realizado, mantendo os modelos que possuíam uma maior frequência absoluta, sendo eles: SPIN, SPACEFOX, XRE 300, HILUX e S10. Os demais modelos não foram considerados na análise pois representam, aproximadamente, 23% da base, e possui uma grande variedade de modelos de veículos.

A Figura 1 mostra o histograma referente aos modelos de veículos, possibilitando assim uma visualização da frequência absoluta da distribuição dos dados entre os modelos.

Figura 1: Histograma dos modelos de veículo.



Fonte: Os Autores.

O Quadro 2 contém a distribuição de frequência para os modelos de veículos analisados. É perceptível no Quadro 2, através de observação das frequências absolutas e relativas, que a maior parte dos registros de abastecimentos ocorrem no interior do estado, seguido da RMR (Região Metropolitana do Recife) e da capital. Isso se deve a grande extensão territorial que corresponde a cada área e a quantidade de centrais de custos nessas regiões. As informações de limite inferior, limite superior, média e mediana fazem referência ao consumo de seus respectivos modelos de veículos.

Observando os dados de limite inferior e limite superior no Quadro 2, é possível identificar valores exorbitantes, tanto apontando para um alto quanto um baixo consumo. Claramente são valores *outliers*

DOI: 10.25286/repa.v6i5.1751

e não representam a realidade, provavelmente valores inseridos de maneira errônea. Na seção de pré-processamento será descrita a forma adotada para minimizar esses valores discrepantes.

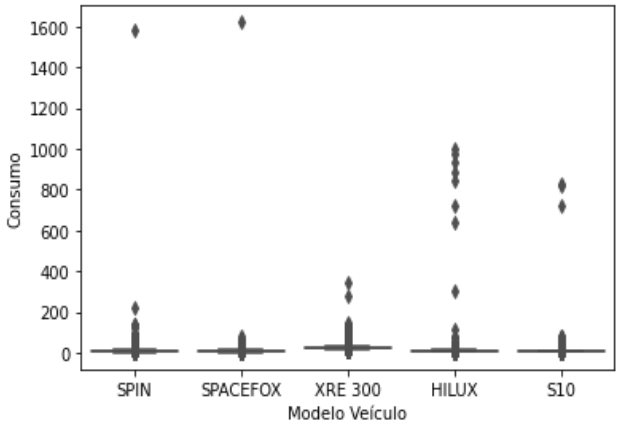
Quadro 2: Distribuição dos registros de abastecimento por modelo de veículo (MV) e área: frequência absoluta (FA), frequência relativa (FR), limite inferior (LI), limite superior (LS), média e mediana.

MV	AREA	FA	FR	LI	LS	MEDIA	MEDIANA
Hilux	Capital	45	0,11	6,68	11,29	9,27	9,33
	Interior	60496	0,11	6,20	11,50	8,84	8,85
	RMR	5082	0,11	6,22	11,54	8,87	8,88
S10	Capital	317	0,13	5,41	10,00	7,82	7,72
	Interior	42956	0,12	5,87	10,89	8,39	8,38
	RMR	12145	0,12	5,65	10,47	8,06	8,06
Space Fox	Capital	17671	0,16	4,49	8,33	6,40	6,41
	Interior	56249	0,13	5,54	10,28	7,86	7,91
	RMR	64998	0,14	5,15	9,55	7,28	7,35
Spin	Capital	47101	0,16	4,40	8,16	6,26	6,28
	Interior	148168	0,12	5,63	10,43	8,00	8,03
	RMR	132578	0,14	5,09	9,45	7,21	7,27
XRE 300	Capital	11974	0,04	15,97	29,61	22,80	22,80
	Interior	44563	0,04	16,03	29,77	22,93	2,90
	RMR	54335	0,04	16,90	31,36	24,15	24,13

Fonte: Os Autores.

Na Figura 2 é exibido o *boxplot* do consumo dos veículos para cada modelo, neste momento não levamos em consideração a área que pertence cada modelo de veículo. Para todos os modelos de veículos foi possível identificar que existe uma grande quantidade de *outliers*. Para quase todos os modelos, com exceção do XRE 300, a mediana de consumo foi semelhante, porém todos apresentaram *outliers* tanto acima do quartil superior, como também abaixo do quartil inferior. O modelo de veículo XRE 300 apresentou uma mediana de consumo mais elevado que os demais modelos, pois é do tipo de veículo motocicleta, sendo diferente dos demais, que são automóveis.

Figura 3: *Boxplot* do consumo por modelos de veículos.



Fonte: Os Autores.

3.3 PRÉ-PROCESSAMENTO DOS DADOS

Durante o processo de entendimento da base de dados, foi constatado que a mesma apresenta uma significativa quantidade de registros e uma quantidade pouca expressiva de dados ausentes. Todavia, foi necessário aplicação de algumas etapas de pré-processamento, a fim de melhorar a análise dos dados e ganhar performance nas avaliações.

Primeiramente foi realizada uma limpeza nos dados onde, a priori, foram selecionadas as colunas mais relevantes, sendo elas: SERVIÇO, UF, MODELO VEICULO, CONSUMO, CENTRO DE CUSTO e AREA. Quanto às demais, uma vez que essas colunas não são de interesse para continuidade da análise, priorizou-se a sua remoção.

Em seguida, foram aplicadas uma série de reduções na base de dados. Primeiramente foram selecionados os registros do estado de Pernambuco, pois os mesmos representam 99,97% dos dados da base (904.580 registros). Em seguida, foram selecionados os registros referentes a serviços de abastecimento, desconsiderando assim os demais tipos de serviços, como ARLA32, borracharia, troca de óleo, lubrificante e NOX.

Foram realizadas algumas transformações com o propósito de agrupar as classes trabalhadas e simplificar o modelo. Primeiro foram agrupados os serviços que apresentaram similaridade entre si, como os de tipo Diesel (Diesel, Diesel S10/S50 e S10/S50); segundo, os modelos de veículos foram agrupados baseando-se em sua frequência de ocorrência, por exemplo, os modelos que mais se repetiram foram Spin, Spacefox, XRE 300, Hilux e S10, onde juntos representam 77% da base original, mas que após os pré-processamentos totalizaram 701.882 registros restantes. Os demais

modelos de veículos foram desconsiderados para a análise.

Com intuito de eliminar os *outliers* existentes na coluna CONSUMO constatados na seção anterior, foram calculadas as médias do consumo considerando os modelos de veículos de acordo com a área de atuação dos mesmos. A partir dessa média calculada, após reuniões com os *stakeholders*, foi definido que todos os valores que estivessem abaixo ou acima de 50% do consumo médio do modelo de veículo que atua em determinada área, seriam desconsiderados. No Quadro 3 mostra o consumo médio e os limites.

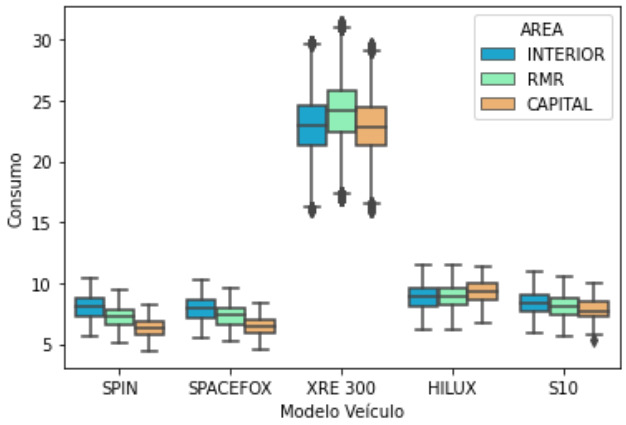
Quadro 3: Consumo médio e limites inferior e superior.

MV	AREA	CONSUMO	LI	LS
Hilux	Capital	9,33	6,53	12,13
	Interior	8,85	6,20	11,51
	RMR	8,88	6,22	11,54
S10	Capital	7,72	5,40	10,04
	Interior	8,38	5,87	10,89
	RMR	8,06	5,64	10,48
Space Fox	Capital	6,41	4,49	8,33
	Interior	7,91	5,54	10,28
	RMR	7,35	5,15	9,56
SPIN	Capital	6,28	4,40	8,16
	Interior	8,03	5,62	10,44
	RMR	7,27	5,09	9,45
XRE 300	Capital	22,80	15,96	29,64
	Interior	22,90	16,03	29,77
	RMR	24,13	16,89	31,37

Fonte: Os Autores.

Finalizado a etapa de pré-processamento e remoção de dados não relevantes, foi possível verificar que os dados ficaram mais coerentes com a realidade. Conforme é possível verificar na Figura 4, os *boxplots* apresentam quartis mais próximos da mediana, representando uma distribuição simétrica dos dados. Além disso, é possível constatar maior economicidade em determinados trechos que esses veículos trafegam, podendo ser devido às áreas com congestionamento menor, poucos aclives e declives ou regiões que possibilitam uma maior velocidade. O comportamento é semelhante na maioria das classes, porém destoam quando são observadas as classes XRE 300 e Hilux, as quais apresentam economia média maior na RMR e capital respectivamente.

Figura 4: *Boxplot* de consumo após o pré-processamento



Fonte: Os Autores.

3.4 METODOLOGIA EXPERIMENTAL

Após a etapa de pré-processamento, foi realizado um processo de rotulagem da base de dados, utilizando valores binários para representar quando um veículo pode ser considerado dentro da conformidade, sendo representado por 0, ou não, sendo representado por 1. Para determinar o rótulo de cada veículo, foi considerada a média do consumo (coluna CONSUMO) para o modelo do veículo (coluna MODELO VEICULO), considerando a área que o veículo trafega (coluna AREA). Na situação em que o consumo for menor ou maior que 30% da média do consumo do modelo do veículo, de acordo com a área que o mesmo trafega, o registro será rotulado como fora dos conformes (1). Caso o consumo esteja dentro do intervalo determinado, o mesmo será rotulado como dentro dos conformes (0).

Para realizar o treinamento dos modelos, a base de dados foi dividida em 70% para treino (472.155 registros) e 30% para testes (202.353 registros).

Em seguida, foram aplicadas técnicas de *grid search*, seguido dos treinamentos e das avaliações dos classificadores *Naive Bayes*, KNN e *Random Forest*, considerando as métricas de acurácia, sensibilidade e precisão. Os hiperparâmetros utilizados no *grid search* podem ser visualizados no Quadro 4.

DOI: 10.25286/repa.v6i5.1751

Quadro 4: Hiperparâmetros utilizados no *grid search*.

CLASSIFICADOR	HIPERPARÂMETROS
Random Forest	n_estimators: 100 e 150 criterion: gini e entropy
KNN	n_neighbors: 5 e 10 algorithm: auto e ball_tree
Naive Bayes	-

Fonte: Os Autores.

4 ANÁLISE E DISCUSSÃO DOS RESULTADOS

4.1 RESULTADOS

Nesta seção são apresentados os resultados obtidos a partir das execuções da técnica de *grid search* e das avaliações dos classificadores. Todos os classificadores obtiveram resultados elevados, porém ainda ocorrendo uma variação entre si.

O *Naive Bayes* por ser um classificador probabilístico, o mesmo não possui hiperparâmetros, impossibilitando a aplicação do *grid search*. O mesmo foi o classificador que apresentou os piores resultados, apresentando uma acurácia de 94,68%. O mesmo apresentou uma sensibilidade e precisão elevadas para a classe 0, obtendo os valores de 100% e 94,30% respectivamente. Já para a classe 1, os valores foram de 56,12% e 100%. A partir desses valores, constatou-se que esse classificador não conseguiu classificar corretamente os registros referentes a não conformidade (classe 1). Os resultados podem ser visualizados na matriz de confusão (Figura 5).

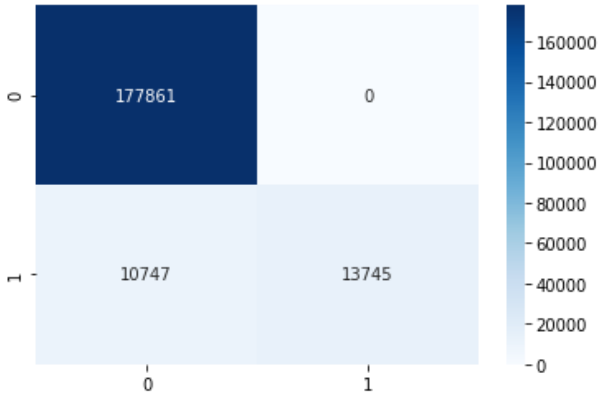
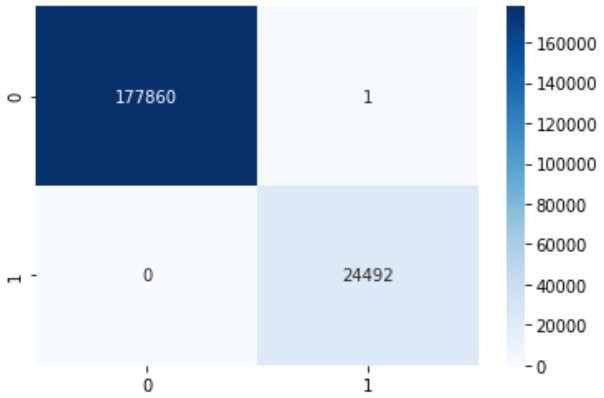


Figura 5: Matriz de confusão do algoritmo *Naive Bayes*

Fonte: Os Autores.

Para o KNN foram selecionados 5 vizinhos (*n_neighbors*), sendo utilizado o algoritmo automático (*auto*). O mesmo apresentou uma acurácia mais elevada ao ser comparado com o *Naive Bayes*, obtendo um valor de 99,99%. Além disso, também foi obtido para a classe 0 uma sensibilidade de 99,99% e uma precisão de 100%. Já para a classe 1, esses valores são de 100% e 99,99%, respectivamente. Os resultados podem ser observados na Figura 6.

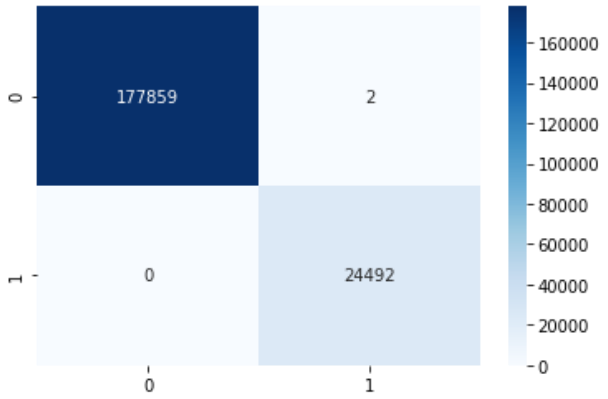
Figura 6: Matriz de confusão do algoritmo KNN.



Fonte: Os Autores.

Por fim, para o *Random Forest* foram selecionados 100 estimadores (*n_estimators*) e como critério de qualidade foi selecionado o coeficiente de Gini. O mesmo conseguiu chegar a valores semelhantes ao KNN, obtendo uma acurácia de 99,99%, a uma sensibilidade e uma precisão de 99,99% e 100%, respectivamente, para a classe 0. Já para a classe 1, o modelo obteve uma sensibilidade de 100% e uma precisão de 99,99%. Os resultados estão disponíveis na matriz de confusão da Figura 7.

Figura 7: Matriz de confusão do algoritmo *Random Forest.*



Fonte: Os Autores.

4.2 DISCUSSÃO

Nesta seção são descritas algumas discussões realizadas através dos resultados apresentados na seção anterior.

Todos os classificadores utilizados apresentaram resultados semelhantes, principalmente na acurácia. Porém, ocorreram algumas variações nos resultados da sensibilidade e da precisão. Os resultados de acurácia podem ser visualizados na Figura 8, os da sensibilidade na Figura 9 e os da precisão na Figura 10.

Como já citado anteriormente, dentre os classificadores utilizados neste trabalho, o *Naive Bayes* apresentou os piores resultados, obtendo uma acurácia inferior aos demais (94,30%), porém o mesmo apresentou uma maior dificuldade para acertar os registros fora dos conformes (classe 1), obtendo uma sensibilidade de 56,10% para esta classe.

Já o KNN e o *Random Forest* obtiveram uma acurácia 99,99%, e valores de sensibilidade e precisão iguais. A única diferença entre eles, é que o *Random Forest* previu erroneamente que 2 registros de abastecimento estavam dentro dos conformes (classe 0), mas na verdade estavam fora dos conformes (classe 1). Já o KNN, errou apenas 1 registro de abastecimento.

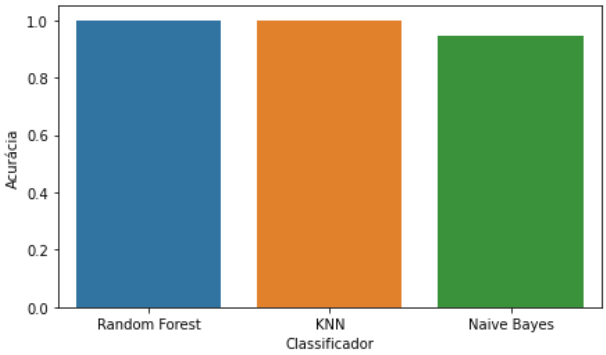
5 CONCLUSÕES E TRABALHOS FUTUROS

Nesta seção serão descritas as conclusões obtidas a partir dos resultados obtidos e também serão apresentados os trabalhos futuros.

Analisando os resultados, o KNN foi o classificador que apresentou os melhores valores, mas como foi citado na seção anterior, os resultados foram bem semelhantes, principalmente entre o KNN e o *Random Forest*.

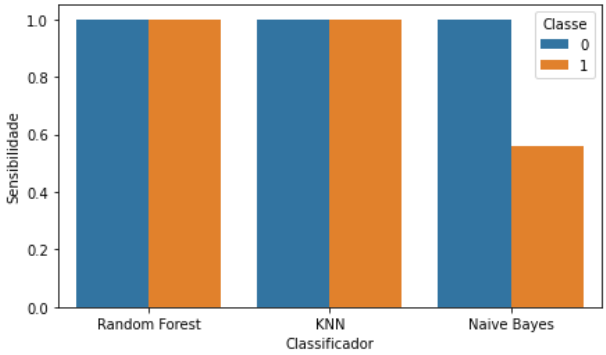
Como trabalhos futuros, é planejado realizar uma análise do balanceamento dos dados da base, aplicando técnicas como *undersampling* ou *oversampling* para evitar possíveis *overfitting* nos classificadores treinados. Por fim, é planejado realizar uma análise mais detalhada a respeito da técnica aplicada na rotulação dos dados, principalmente após a utilização dos classificadores propostos neste trabalho nos processos de auditoria da PMPE.

Figura 8: Resultados da acurácia.



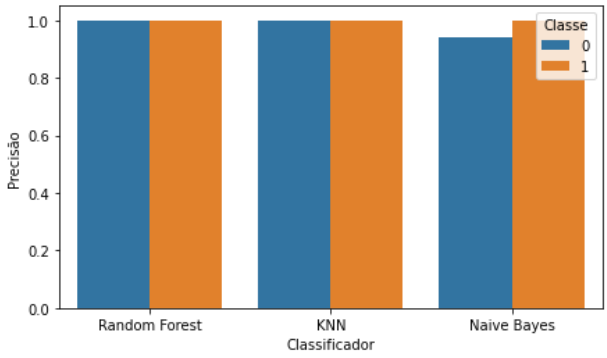
Fonte: Os Autores.

Figura 9: Resultados da sensibilidade.



Fonte: Os Autores.

Figura 10: Resultados da precisão.



Fonte: Os Autores.

DOI: 10.25286/repa.v6i5.1751

REFERÊNCIAS

[1] Silva, Pedro Gabril Kenne. **O papel do controle interno na administração pública.** Contexto, V.2, nº 2. 2002, UFRGS, ISSN 2175-8751

[2] PERNAMBUCO, **Portal de Lei de Acesso à Informação do Governo de Pernambuco**. https://www.lai.pe.gov.br/pmpe/. Acessado em: 15 de mar de 2021.

[3] _____, **Lei Orçamentária Anual**. Pernambuco, 2020, p. 459.

[4] Castro, Leandro Nunes. **Introdução à mineração de dados: conceitos básicos, algoritmos e aplicações.** São Paulo, Saraiva, 2016.

[5] Provost, F., Fawcett, T. **Data Science para Negócios.** Alta Books, 2016, Rio de Janeiro, RJ.

[6] Mitchell, TM. ***Machine Learning**, McGraw-Hill Higher Education*. New York, 1997.

[7] Cover, Thomas. Hart, Peter. ***Nearest neighbors pattern classificatio**n*. IEEE *transactions on information theory*. 1967.

[8] Pereira, A. C., Andrade, G., Silva, H., Amorim Neto, H., Lima, J. A., & Batista, M. **Análise do Consumo de Combustível da PM-PE com Foco no Controle**. Revista De Engenharia E Pesquisa Aplicada, 6(3), 39-48, 2021.

[9] Huertas J.I.; Giraldo, M.; Quirama, L.F.; Díaz, J. **Driving Cycles Based on Fuel Consumption**. Energies 2018, 3064.

[10] Bonchi, F., Giannotti F., Mainetto G., Pedreschi D., **A classification-based methodology for planning audit strategies in fraud detection**. 5º ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA, 1999.