Clusterização de Falhas em Centrais Multimídias de Veículos Automotivos

Clusterization of media center's failures in automotive vehicles

Vinicius B. S. da Silva[1]
orcid.org/0000-0003-2947-4595

Carmelo J. A. Bastos Filho[2]
orcid.org/0000-0002-0924-5341

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

Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil. E-mail: carmelofilho@ecomp.poli.br

DOI: 10.25286/repa.v7i2.2221

Como citar este artigo pela NBR 6023/2018: Vinicius B. S. da Silva Carmelo J. A. Bastos Filho2. Clusterização de Falhas em Centrais Multimídias de Veículos Automotivos. Revista de Engenharia e Pesquisa Aplicada, v.7, n. 2, p. 96-107, 2022

RESUMO

A indústria automotiva mundial enfrenta por um momento crítico devido ao novo cenário econômico e social ocasionado pela pandemia do corona vírus. A paralisação de fábricas de vários setores ao redor do mundo fez com que uma escassez de matéria prima e produtos acabados surgisse. A ausência global de semicondutores, componentes com uso cada vez mais requisitados na indústria automotiva devido à maior quantidade de tecnologia embarcada nos veículos, fez com que várias montadoras parassem suas linhas produção aumentando o tempo de entrega de veículos novos e consequentemente, um aumento significativo no preço dos veículos. A falta de componentes também estabeleceu a carência de peças para troca em garantia no mercado, gerando insatisfação nos proprietários de veículos que apresentaram algum tipo de problema. Diante disto, este trabalho tem como intuito investigar relações entre dados de garantia de centrais multimídias e dados meteorológicos a fim de entender relações que possam ser utilizadas para melhorar a aquisição, distribuição e controle de estoque aprimorando o tempo de resolução do problema do consumidor final. Os resultados obtidos apontam que não existem relações significativas entre dados meteorológicos de temperatura e umidade com veículos que passaram por intervenção de garantia, mas é possível observar uma pré-disposição de veículos mais modernos apresentarem uma maior taxa de falha, visto que possuem centrais multimídias mais complexas.

PALAVRAS-CHAVE: Clusterização; Veículos Automotivos; Falhas; Centrais;

ABSTRACT

The global automotive industry is facing a critical moment due to the new economic and social scenario caused by the coronavirus pandemic. The shutdown of factories in various sectors worldwide caused a shortage of raw materials and finished products to arise. The global absence of semiconductors, components that are increasingly used in the automotive industry due to the increase amount of technology embedded in vehicles, caused several automakers to stop their production lines, thus increasing the delivery time of new vehicles and, therefore, a significant increase in the price of vehicles. The lack of components also established the absence of parts for warranty exchanges in the market, generating dissatisfaction among vehicle owners that presented some problems. Therefore, this work aims to investigate relationships between warranty data from media centers and meteorological data in order to understand relationships that can be used to improve the acquisition, distribution and inventory control, improving the resolution time of the final consumer's issue. The results obtained show that there are no significant relationships between meteorological data of temperature and humidity with vehicles that underwent warranty intervention. However, we observed a pre-disposition of more modern vehicles to have a higher failure rate, as they have more complex media centers.

REVISTA DE Engenharia e Pesquisa Aplicada

1 INTRODUÇÃO

Atualmente uma instabilidade na aquisição de semicondutores preocupa diversos setores da economia global, principalmente setores como os de automóveis e eletrônicos que dependem desses componentes para fabricação dos seus produtos, acarretando na paralisação de fábricas ao redor de todo o país e do mundo. Um dos principais fatores foi novo corona vírus ou COVID-19, que teve origem em Wuhan, China, em dezembro de 2019 e foi declarado pandemia em 11 de março de 2020 pela Organização Mundial da Saúde (OMS). O agravamento da pandemia causou interrupções econômicas e financeiras sem precedentes, visto que muitas famílias foram obrigadas a ficar em casa, ou limitar as interações sociais, para retardar a propagação do vírus **[1]**. No Brasil o cenário não foi diferente. De acordo com o Ministério da Saúde, até 12 de novembro de 2021, 21,94 milhões de casos foram registrados e 610.491 pessoas morreram devido a complicações com a doença **[2]**.

Os efeitos gerados pela pandemia não têm apenas consequências para a economia, mas inclui toda a sociedade afetada, levando assim a mudanças dramáticas na forma como as empresas agem e os consumidores se comportam **[3]**. De uma forma geral, o distanciamento social presenciado durante a pandemia fez com que smartfones e computadores se torneassem salas de aula online e conexões familiares, enquanto dispositivos de streaming de mídia registravam taxas sem precedentes de visualização de TV e filmes **[4]**. O uso de equipamentos eletrônicos aumentou tanto durante esse período que o mercado de produtos usados teve um aumento significativo, visto que os estoques dos varejistas não conseguiam acompanhar a demanda do consumidor **[5].** A paralisação global na indústria automotiva fez com que os produtores de semicondutores passassem a priorizar as empresas de tecnologia, pois a demanda por computadores, celulares e até videogames disparou durante este período.

Em contrapartida, nos últimos anos a utilização de semicondutores vem crescendo nos veículos automotivos. Segundo a montadora Volkswagem, um modelo SUV de médio porte, como o Volkswagen Taos, tem cerca de 300 chips **[6]**. Outro ponto importante é o crescimento da demanda de componentes eletrônicos com a popularização dos veículos elétricos e autônomos que, para permitir a direção em ambientes dinâmicos como os grandes centros urbanos, necessitam de sofisticados sistemas de direção que incluem uma série de sensores para percepção do ambiente, incluindo lidar, radar, câmeras e sensores ultrassônicos **[7]**.

Com a volta gradual das atividades do setor automotivo, montadoras começaram a encarar uma falta generalizada de semicondutores devido a realocação de recursos no mercado que desequilibraram a oferta e a demanda por chips automotivos que pode continuar gerando um risco de quebra da cadeia industrial **[8]**.

De acordo com a Associação Nacional dos Fabricantes de Veículos Automotores (ANFAVEA), em abril de 2021 cerca de 50% das fábricas de automóveis tiveram sua produção interrompida devido à falta de semicondutores **[9].** Como consequência da paralisação, um aumentando o tempo de entrega de veículos novos pode-se ser visualizado juntamente com um incremento significativo no preço dos veículos e peças de reposição.

1.1 OBJETIVO

O objetivo deste trabalho é realizar um estudo em dados de garantia de modelos de veículos automotores que apresentaram algum tipo de não conformidade em sua central multimídia, componente eletrônico de alto valor agregado que, devido à crise de semicondutores, apresenta baixo disponibilidade em pós-vendas e grande tempo para aquisição de novos componentes para reposição. Para este estudo, serão considerados apenas os casos de acionamento de garantia por falha nos componentes do veículo como sendo a causa principal da falha.

Como objetivo adicional, busca-se identificar uma possível correlação entre o número de casos de acionamento da garantia com a existência de condições meteorológicas específicas em determinada região, visto produtos que possuem semicondutores podem sofrer alterações em suas performances e redução de vida útil quando submetidos a diferentes níveis de temperatura e umidade. Para isto os dados de garantia foram cruzados com os dados meteorológicos disponíveis pelo INMEP (Instituto Nacional de Meteorologia).

De posse das duas bases de dados devidamente combinadas, modelos de clusterização serão desenvolvidos para verificação de padrões que

DOI: 10.25286/repa.v7i2.2222

possam vir a ajudar na identificação de alguma similaridade nos grupos estudados.

1.2 JUSTIFICATIVA

Como a demanda pare reposição de peças não é previsível, existem várias dificuldades de gerenciamento e planejamento tornando muito difícil entregar os produtos no prazo, levando a margens de lucro muito baixas e insatisfação dos clientes, uma vez que nesta fase os clientes tendem a estarem chateados com a marca **[10]**.

A identificação de padrões nos dados é necessária para melhor priorização na aquisição, alocação e distribuição de componentes para substituição de um determinado grupo. Tal condição torna o processo mais ágil, aumentando margem e satisfação do consumidor final.

1.3 ESCOPO NEGATIVO

A produção de um automóvel, mesmo de modelos mais simples, envolve uma longa e complexa sequência de etapas e processos envolvendo centenas de fornecedores. A grande quantidade de agentes envolvidos acarreta o surgimento de problemas diversos, com características e complexidade diferentes, demandando maior ou menor esforço para sua solução. Cada variável adicionada na análise dos dados, se não tratada adequadamente, pode ganhar um aspecto negativo, deixando a análise cada vez mais complexa.

Neste trabalho como forma de simplificação do problema, alguns pontos deixaram de ser analisados, dada a complexidade que a inclusão desses itens traria à análise final. As informações climáticas agregadas a cada indivíduo analisado foram obtidas utilizando como base os dados da estação meteorológica da capital do estado ao qual o veículo passou por intervenção. Logo, tais informações não traduzem fidedignamente as reais condições a que os veículos foram expostos ao durante o espaço de tempo contido entre compra e reparação. No entanto, mesmo que os dados utilizados nas análises tivessem sido os obtidos nas estações mais próximas disponíveis, ainda não seria possível dizer que esta era a exposição real do veículo, devido a situações tais como, casos onde o veículo esteve abrigado em garagem, deslocamentos para outras regiões, etc.

2 FUNDAMENTAÇÃO TEÓRICA

Nesta seção será tratada de forma objetiva as principais tecnologias e abordagens utilizadas para a realização deste projeto, com ênfase na indústria automotiva e métodos de mineração de dados.

2.1 ÁREA DO NEGÓCIO

O objetivo principal de uma empresa automotiva é projetar e fabricar veículos que atendam às todas as necessidades e expectativas de seus clientes. O histórico de falhas em campo que causaram acionamento da garantia do veículo é o principal modo de entendimento de desempenho em relação à capacidade do produto de realizar sua função pretendida nas mãos do cliente **[11]**.

De acordo com SHOKOUHYAR **[12]**, a garantia desempenha um papel importante na retenção da fidelidade dos consumidores, aumentando a vantagem competitiva e o lucro das empresas de uma forma geral. Para chegar em um nível de maior excelência, esforços cada vez maiores estão sendo tomados para identificação cada vez mais precisa e veloz das lamentações dos consumidores, como por exemplo, em modelos que monitoram diariamente dados de redes sociais. Melhorar a precisão das previsões permite que as empresas minimizem com eficácia os custos relacionados à garantia, os níveis de estoque, o desperdício e a insatisfação do cliente, maximizando o retorno sobre o investimento, lucro, eficiência e satisfação do cliente se tornando cada vez mais competitivas dentro do mercado.

2.2 CLUSTERIZAÇÃO

Algoritmos de clusterização são ferramentas adequadas para identificar perfis em base de dados extensas. Diferente da classificação que é comumente utilizada como um método de aprendizagem supervisionada, a clusterização ou agrupamento tem como principal aplicação a aprendizagem não supervisionada (alguns modelos de agrupamento podem ser utilizados para ambos) **[13]**.

O objetivo primário dos agrupamentos é descobrir um novo conjunto de categorias que tenham instâncias ou objetos semelhantes. Para realização desta atividade a utilização, o uso de algum tipo de métrica se faz necessária para determinação de similaridade de dois objetos. Um dos principais tipos de medidas utilizadas para estimar tais relações são medidas de distâncias. Um exemplo bastante utilizado na mineração de dados é o k-means, que utiliza como critério de agrupamento soma da distância Euclidiana entre cada elemento e seu centro de agrupamento mais próximo (centroide) **[14]**.

Um problema encontrado na utilização de métodos que utilizam distâncias é que geralmente os dados utilizados contêm valores numéricos e categóricos. A maneira tradicional de tratar atributos categóricos como numéricos nem sempre produz resultados significativos porque muitos domínios categóricos não são ordenados. Visto isso,

dois métodos foram utilizados neste trabalho para realização dos estudos na base de dados

Um problema encontrado na utilização de métodos que utilizam distâncias em dados que contêm valores numéricos e categóricos (base de dados deste trabalho) é que, a maneira tradicional de tratar atributos categóricos como numéricos nem sempre produz resultados significativos porque muitos domínios categóricos não são ordenados **[15]**. Visto isso, dois métodos foram escolhidos para realização dos agrupamentos da base de dados: k-modes e clusterização hierárquica.

2.1.1 K-MODES

Proposto em 1998 por HUANG **[16]**, o método k-modes tem como objetivo resolver o paradigma do k-means para agrupar dados categóricos utilizando uma medida de dissimilaridade de correspondência simples para objetos categóricos, modas em vez de meios para agrupamentos e um método baseado em frequência para atualizar os modos no processo de agrupamento à moda k-means para minimizar a função de custo de agrupamento.

De forma a simplificar o entendimento, um o pseudo-código para a implementação do k-modes pode ser observado à seguir:

1. Escolha K observações aleatoriamente e utilize como líderes / grupos;
2. Calcule as diferenças e atribua cada observação ao seu agrupamento mais próximo;
3. Defina novas modas para os clusters;
4. Repita as etapas 2-3 até que não haja necessidade de reatribuição.

Em 2009, CAO, Liang e Bai **[17]** propuseram uma nova forma de inicialização para dados categóricos foi proposta, no qual a distância entre os objetos e a densidade do objeto é considerada. Os resultados experimentais encontrados mostram que o método de inicialização proposto é superior ao método de inicialização aleatória.

2.1.1 CLUSTERIZAÇÃO HIERÁRQUICA

A Clusterização hierárquica é uma família geral de algoritmos de agrupamento que constroem clusters aninhados mesclando-os ou dividindo-os sucessivamente. Esta hierarquia de clusters é representada como uma árvore (também chamada de dendrograma). A raiz da árvore é o único cluster que reúne todas as amostras, sendo as folhas os aglomerados com apenas uma amostra **[18]**.

A mescla ou divisão de clusters por hierarquia é realizada de acordo com alguma medida de similaridade, escolhida de forma a otimizar algum critério pré-estabelecido. Os métodos de agrupamento hierárquico podem ser posteriormente divididos de acordo com a maneira como a medida de similaridade é calculada **[19]**.

• Agrupamento de link único (também encontrado como método mínimo ou o método do vizinho mais próximo) - Métodos que consideram a distância entre dois clusters igual à distância mais curta de qualquer membro de um cluster a qualquer membro do outro cluster;

• Agrupamento de link completo (também chamado de diâmetro, método máximo ou método do vizinho mais distante) - métodos que consideram a distância entre dois clusters igual à distância mais longa de qualquer membro de um cluster para qualquer membro do outro cluster;

• Agrupamento de link médio (ou método de variância mínima) - métodos que consideram a distância entre dois clusters igual à distância média de qualquer membro de um cluster a qualquer membro do outro cluster.

3 MATERIAIS E MÉTODOS

Esta seção tem como objetivo apresentação dos materiais, ferramentas e métodos utilizados no decorrer deste projeto além das aplicações dos mesmos.

3.1 DESCRIÇÃO DA BASE DE DADOS

Foram utilizados dados reais de garantia focados em um conjunto específico de peças (central multimídia) produzidos em uma fábrica de automóveis, compreendendo o período de 2017 a julho de 2021 totalizando 14.597 lamentações. Também serão utilizados dados disponibilizados pelo INMET (Instituto Nacional de Meteorologia) em **[20]** referentes a 589 estações de meteorologia espalhadas por todo território nacional. Os dados de garantia foram extraídos de um Data Warehouse de falhas contendo dados de veículos produzidos na América do Sul e armazenados em um arquivo no formato XLSX. Já os dados do INMET foram armazenados em pastas específicas separadas por ano. Os nomes destes arquivos em formato CSV permitem identificar a quais cidades ou estações se referem.

Os Quadros 1 e 2 apresentam os dicionários simplificados de dados da base de garantias e das bases INMET com as principais informações que serão utilizadas para construção do modelo.

DOI: 10.25286/repa.v7i2.2222

Quadro 1 - Dicionário de dados simplificado base garantia

CAMPO	DESCRIÇÃO	TIPO	TAM	VAL
Chassi	Numeração única do veículo	Char	8	ABC1234
Market	Mercado ao qual o veículo foi comercializado	Num	4	1234
Modelo	Modelo do Veículo	Num	3	123
Versão	Versão do veículo	Char	3	-
Km	Quilometragem no momento da intervenção	Num	-	-
Dealer	Concessionário que realizou a intervenção	Num	5	12345
Estado	Estado onde houve a intervenção	Char	2	AB
Desenho	Código da peça substituída	Char	13	-
Data_Prod	Data de produção do veículo	Char	10	DD/MM/AAAA
Data_Venda	Data de venda do veículo	Char	10	DD/MM/AAAA
Data_Int	Data da intervenção no veículo	Char	10	DD/MM/AAAA

Fonte: Os Autores

Quadro 2 - Dicionário de dados simplificado base INMET

CAMPO	DESCRIÇÃO	TIPO	TAM	VAL
Data	Data de aquisição da medição	Char	10	DD/MM/AAAA
Hora UTC	Hora da aquisição dos dados	Char	8	HHHH UTC
TEMP. MÁX.(◦C)	Temp. de orvalho na medição	Num	-	-
UMIDADE REL. MÁX.(%)	Umidade máxima na medição	Num	-	-
RADIAÇÃO (Kj/mˆ2)	Radiação na medição	Num	-	-

Fonte: Os Autores

3.2 PRÉ-PROCESSAMENTO DOS DADOS

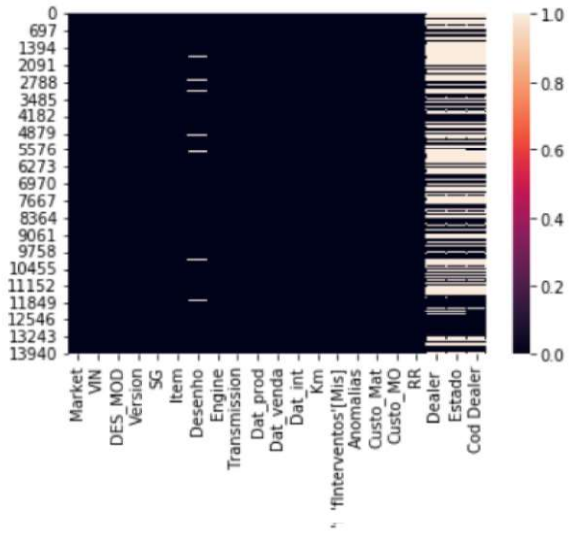
A etapa de pré-processamento dos dados foi iniciada com uma análise para identificar a distribuição de dados ausentes ou nulos na base de dados da garantia. De forma geral, foi observado um grande número de informações faltantes, como mostrado na Figura 1. É nítido no mapa de calor que várias amostras, isto é, garantias realizadas possuem dados que estão incompletas com informações que são se suma importância para o seguimento do estudo que diz respeito à concessionária que realizou a reparação fazendo com que não seja possível entender qual o estado brasileiro que a reparação ocorre uma vez que isso impediria de associar os dados climáticos adequadamente. Também se vê algumas amostras que não possuem o Part Number da peça, o que impede da identificação de qual peça foi substituída no veículo.

De posse destas informações utilizou-se a técnica de limpeza dos dados por meio da exclusão das linhas que continham dados NaN em atributos que seriam necessários para as análises futuras. Também foram excluídas colunas que possuíam dados duplicados como Chassis e VIN e Código da concessionária e Nome Fantasia da mesma. Com isso, o banco de dados passou de 14.597 amostras para 7.780 amostras.

A próxima etapa consiste na integração dos dados meteorológicos. Inicialmente cria-se as colunas para estes dados, tais como temperatura máxima e precipitação média à base das garantias, inicializando-as com o valor zero.

Haja visto que os arquivos meteorológicos estão segregados por ano e por estação, assim como o período de exposição varia em cada caso conforme as datas de fabricação e acionamento da garantia do veículo, não era viável realizar um pré-processamento geral destas bases, sendo este realizado linha por linha no momento de carregar os dados da base meteorológica para a base do projeto que inicialmente contava apenas com os dados das garantias.

Figura 1 – Mapa de calor dos dados faltantes da base de garantia



Fonte: Os Autores

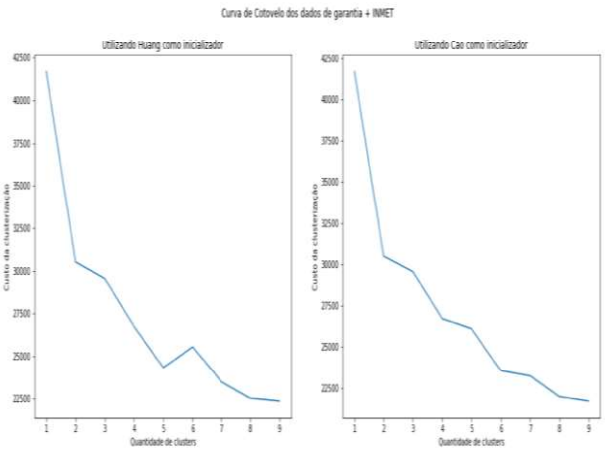
Foram implementadas funções para localizar os arquivos referentes à capital do estado onde o veículo foi comercializado, e para os anos desde a data de fabricação até o acionamento da garantia.

Como na base do INMET os dados estão registrados de hora em hora, foi necessário aplicar transformações de modo a se obter dados como temperatura máxima, temperatura média, precipitação total e precipitação média referente ao período de interesse. Dados ausentes ou não numéricos foram desconsiderados, assim como também foi necessário excluir as leituras de valor "-9999" que correspondem a falhas tais como falha de comunicação, sensor desligado ou inoperante, etc. e que ficariam influenciando o cálculo das médias ou somatório caso mantidas. Ao final desta etapa novamente foi realizada uma verificação se houveram dados NaN ou incongruentes como por exemplo, precipitação negativa ou temperatura máxima com valores acima de 200 ◦C. O total de amostras do dataset resultante foi 6.872.

3.2 METODOLOGIA EXPERIMENTAL

Como início do processo de análise dos dados, foi necessária definição dos parâmetros iniciais de partida do algorítimo. Logo, uma varredura para os números de clusters, parâmetro K, sendo o variando de 1 até 9 foi realizado para Huang e Cao, ambos inicializadores do método k-means. Como parâmetro de comparação, foi utilizada o custo para cada K e plotado, conforme a Figura 2, que mostra a curva de cotovelo obtida para ambos os métodos.

Figura 2 – Curva de Cotovelo obtida com variações de K



entre 1 e 9 na base Garantia + INMET para inicializadores HUANG e Cao

Fonte: Os Autores

É possível verificar nos gráficos que as diferenças mais significativas do custo que ocorrem nos valores de K mais baixos e passam a ter menos relevância para altos valores de K, apesar do aumento significativo do custo computacional. Isto se dá devido à complexidade e quantidade de instâncias que estão sendo estudados. Para uma escolha mais assertiva de qual K será utilizado como o número total de clusters para as análises do k-modes, o coeficiente de silhueta também foi calculado para cada valor de K. Os valores da média do coeficiente de silhueta para os valores de K podem ser encontrados nos Quadros 3 e 4.

Quadro 3 - Valores da média do coeficiente de silhueta para o inicializador Huang

Número de Clusters (K)	Avarege Silhouette Score
2	0,1188
3	0,1325
4	0,0914
5	0,0955
6	-0,0100

Fonte: Os Autores

DOI: 10.25286/repa.v7i2.2222

Quadro 4 - Valores da média do coeficiente de silhueta para o inicializador Cao

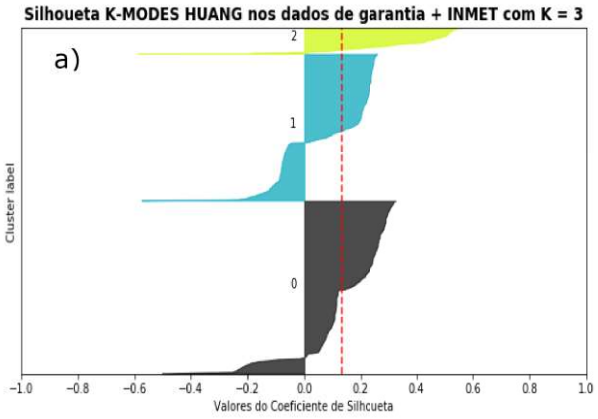
Número de Clusters (K)	Avarege Silhouette Score
2	0,1188
3	0,1038
4	0,0496
5	0,0461
6	0,0752

Fonte: Os Autores

Como é possível observar, os valores encontrados através da silhueta são pequenos, ou seja, os clusters encontrados pelo método k-modes apresentam valores muito baixos, fazendo com que o modelo em questão não apresente bons resultados. Isso pode ser melhor entendido através dos gráficos que serão apresentados a seguir nas Figuras 3a e 3b onde foram utilizadas as melhores configurações para os modelos com inicialização por Cao e o modelo com inicialização por Huang com um total de K=3 clusters.

Portanto, devido aos resultados abaixo do esperado utilizando o método k-modes faz-se necessário a utilização de outros algorítimos para um resultado mais preciso, logo foi escolhido um método que tenha mais variáveis que possam ser manipuladas.

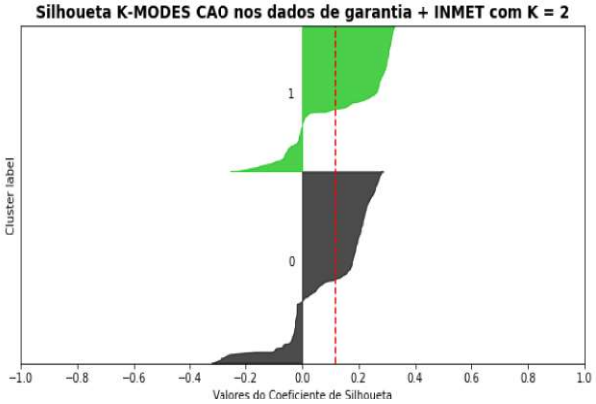
Figura 3 – Gráfico de Silhueta (a) Inicializador Huang com K = 3 e (b) Inicializador Cao com K = 2



Fonte: Os Autores

Utilizando Agrupamento Hierárquico

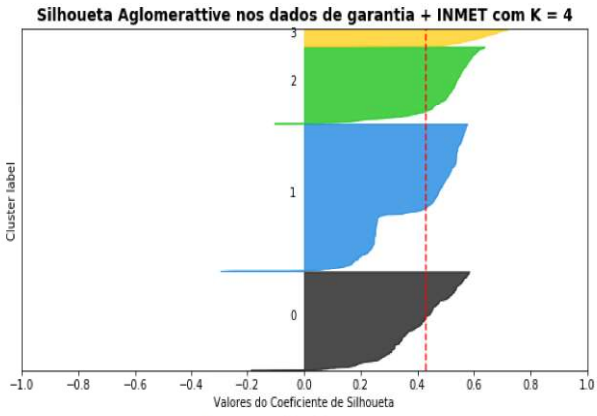
Como demonstrado anteriormente, os valores não foram satisfatórios então foi necessário o uso



de outros métodos para realização de análises mais robustas. Com ajuda da biblioteca ScikitLearn, foi utilizado o algorítimo de agrupamento hierárquico para identificação dos clusters.

Assim como realizado no k-modes, será necessário achar os melhores parâmetros que entregue os melhores indicadores. Para o caso do aglomerativo, além da métrica da silhueta também foi utilizado o índice Calinski-Harabasz (CH Index) para medição de performance do modelo. Para o modelo de agrupamento hierárquico, além do número de clusters, também foram variados os tipos de afinidades dos clusters (Euclidiana, Manhattan, L1, L2 e Cosine) e da forma de ligação (linkage) entre os clusters posteriormente calculados (completa, média e simples). Logo, para a variação de clusters escolhido entre 2 e 6, foram estudados ao total 90 diferentes configurações. No Quadro 5, são apresentados os resultados encontrados para K=4.

Figura 4 – Curva de Silhueta para K=4



Fonte: Os Autores

Quadro 5 - Valores de silhueta e CH index encontrados para K = 4 variando *Affinity* e *linkage*

K	Affinity	Linkeage	Avg. Silhouette	CH Index
4	Euclidian	Complete	0,2271	1815,59
4	Euclidian	Average	0,4297	4692,27
4	Euclidian	Single	0,0913	11,99
4	L1	Complete	0,4117	4331,01
4	L1	Average	0,3436	2686,85
4	L1	Single	0,2478	6,83
4	L2	Complete	0,2271	1815,59
4	**L2**	**Average**	**0,4297**	**4692,27**
4	L2	Single	0,0913	11,99
4	Manhattan	Complete	0,4117	4341,01
4	Manhattan	Average	0,3436	2686,85
4	Manhattan	Single	0,2478	6,83
4	Cosine	Complete	0,1092	936,62
4	Cosine	Average	0,2736	898,00
4	Cosine	Single	0,1992	5,17

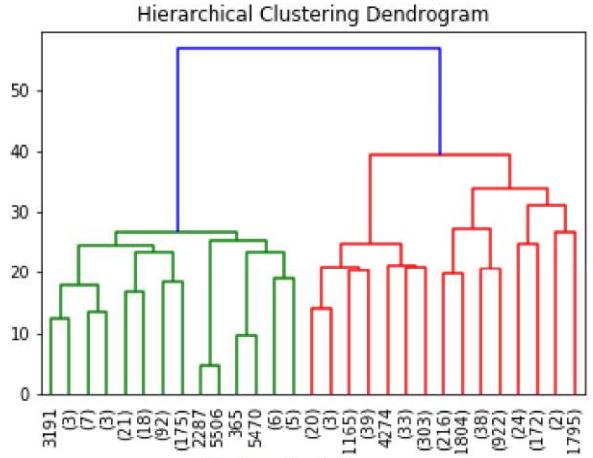
Fonte: Os Autores

Utilizando como ponto de partida os parâmetros onde a melhor configuração encontrada (em destaque no Quadro 5), onde K=4, Affinity = L2 e Linkeage = Avarege, o gráfico de silhueta do modelo em questão foi plotado e pode ser verificado na Figura 4.

Na Figura 4 é possível observar que os 4 clusters que foram encontrados através do algorítimo possuem uma boa separação, onde poucas amostras encontram-se em zonas de misturas entre os clusters.

Para compreensão e comparação com o melhor resultado obtido através da variação dos parâmetros, um Dendograma foi elaborado. Explicando de uma forma simples, o processo aglomerativo ou Clustering Bottom-up começa essencialmente a partir de um cluster individual (cada ponto de dados é considerado um cluster individual, também chamado de folha) e, em seguida, cada cluster calcula sua distância entre eles. Os dois clusters com a distância mais curta um do outro se unem, criando o que chamamos de nó. Clusters recém-formados, mais uma vez, têm sua distância calculada do membro de seu cluster com outro cluster fora de seu cluster. O processo é repetido até que todos os pontos de dados sejam atribuídos a um cluster denominado raiz. A Figura 5 mostra o Dendograma obtido utilizando os dados de garantia mais meteorológicos.

Figura 5 – Dendograma para obtido para os dados de Garantia + INMET

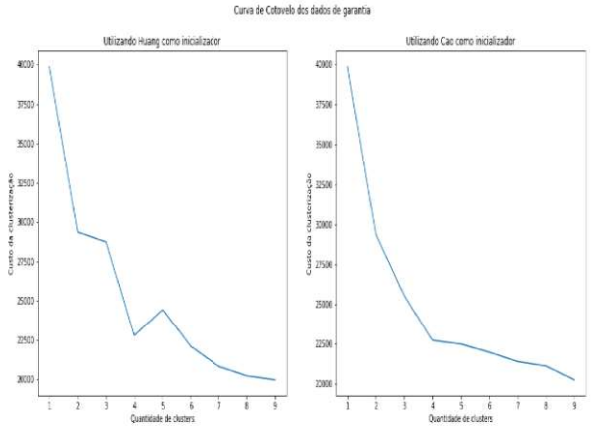


Fonte: Os Autores

Para efeito de comparação e entendimento do aumento da complexidade do modelo após a junção das bases de dados (garantia e meteorologia), os mesmos passos demonstrados até o presente momento foram realizados apenas com os dados da garantia.

A Figura 6 mostra os gráficos de cotovelo para os dados de garantia realizados no modelo k-modes com Huang e Cao como inicializadores, respectivamente.

Figura 6 – Curva de Cotovelo obtidas com variações de K entre 1 e 9 para inicializadores HUANG e Cao sem dados do INMET



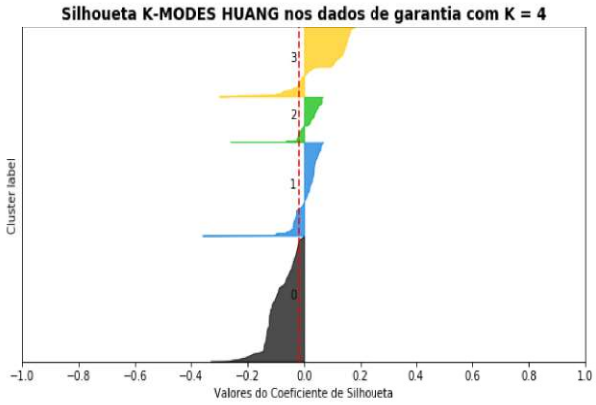
Fonte: Os Autores

DOI: 10.25286/repa.v7i2.2222

É possível verificar um comportamento mais uniforme das curvas, principalmente com o inicializador pode densidade Cao, onde é possível comprovar de forma mais clara a diminuição de forma exponencial em K=4, confirmando o que a necessidade de utilização de 4 clusters, assim como encontrado no algoritmo aglomerativo hierárquico. As Figuras 7 e 8, mostram o comportamento dos gráficos de silhueta para ambos os inicializadores do k-modes, utilizando K=4.

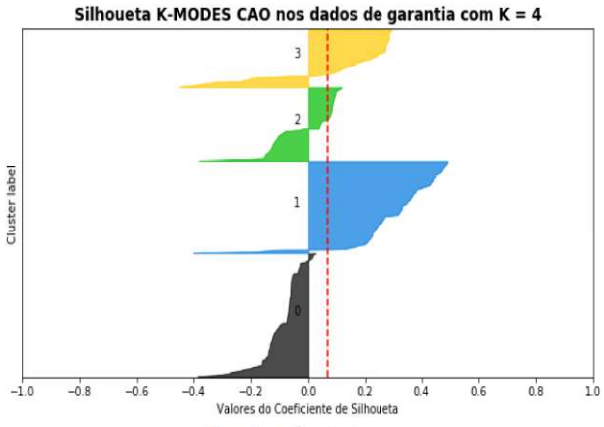
Figura 7 – Inicializador Huang com K = 4 para os dados Garantia

Fonte: Os Autores



Da mesma maneira, o modelo aglomerativo hierárquico foi realizado para os dados de garantia. As Figuras 9 e 10 mostram a melhor configuração encontrada e o dendograma, respectivamente.

Figura 8 – Inicializador Cao com K = 4 para os dados Garantia

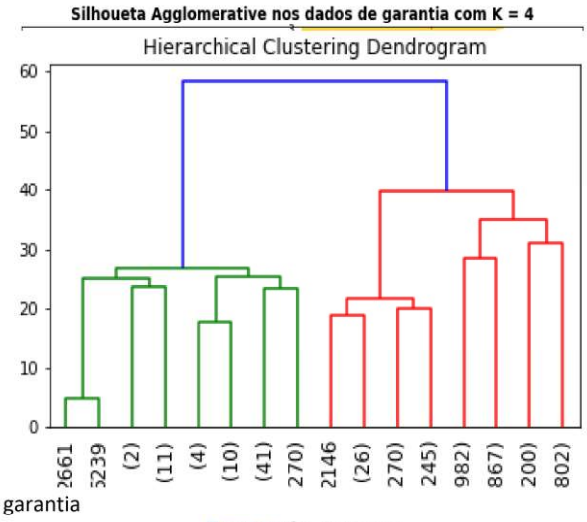


Fonte: Os Autores

Figura 9 – Curva de Silhueta para K=4 aplicado para os dados de garantia

Fonte: Os Autores

Figura 10 – Dendograma obtido apenas para dados de



Fonte: Os Autores

4 Análise e Discussão dos Resultados

Nesta seção, serão apresentados os resultados que foram obtidos através dos algoritmos desenvolvidos que foram aplicados nas bases de dados, além da comparação entre as análises realizadas entre os dados de garantia e dados meteorológicos quanto apenas garantia.

4.1 RESULTADOS

Como demostrado nos gráficos anteriores apresentados na subseção 3.3, o modelo aglomerativo hierárquico se mostrou mais eficiente para o estudo em questão, tanto para os dados singelos de garantia quanto com a adição das informações meteorológicas. O Quadro 6 mostra de forma resumida a comparação dos melhores resultados obtidos na última seção através do método de aglomeração hierárquica.

. **Quadro 6 -** Valores da média do coeficiente de silhueta para o inicializador Cao

Fonte: Os Autores

Como esperado, a utilização apenas dos dados de garantia gera melhores resultados nas distribuições devido à diminuição das variáveis e complexidade do modelo.

Seguindo com o estudo, foi realizada uma separação por bins nos dados meteorológicos para facilitar a visualização. Os gráficos presentes na Figura 11 mostra as distribuições dos clusters encontrados com ajuda do modelo aglomerativos.

Observando o primeiro gráfico referente a temperatura máxima, é possível verificar uma

grande concentração de amostras na região de 30-40 ºC. Este comportamento já era esperado, visto que os dados de garantia são de veículos da América do Sul, muito concentrado nos maiores mercados Brasil em primeiro e Argentina em segundo, países tropicais onde temperaturas altas são comuns na maior parte do ano. Partindo para temperatura média, é possível perceber um comportamento parecido onde a maior parte dos veículos falhados encontram-se em regiões de temperatura média entre 20-30 ºC. Pode-se perceber que a soma dos dois primeiros bins para temperatura média e temperatura máxima, apresentam basicamente a mesma quantidade de elementos.

Figura 11 – Distribuição dos Clusters para dados Meteorológicos

Fonte: Os Autores

Figura 12 – Distribuição dos Clusters para principais dados de garantia

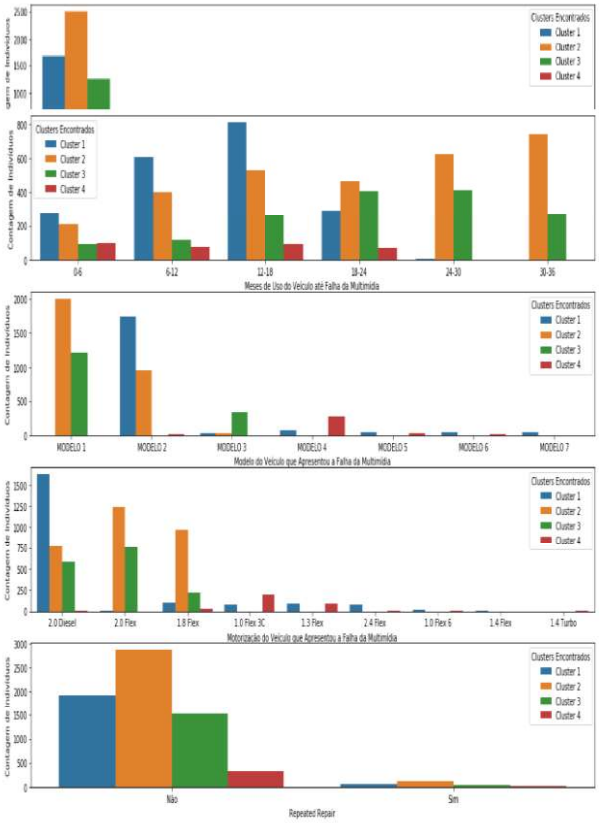
Fonte: Os Autores

Visualizando os resultados encontrados sobre precipitação, é possível perceber que a grande maioria dos veículos que falharam pertenciam a regiões na qual não há grandes volumes de chuvas visto a concentração das amostras em locais de baixa precipitação acumulada e média. Logo uma conclusão que pode ser observado é que a umidade pode não ter efeitos diretos para as falhas nas centrais multimídias em estudo, visto que quanto mais úmida a região, maior a quantidade de chuvas, aumentando a precipitação média e total.

Fazendo uma análise mais profunda na clusterização focando nos dados de garantia que sensíveis da empresa, os nomes dos modelos foram omitidos por confidencialidade. Os principais atributos foram plotados na Figura 12.

O primeiro gráfico da Figura 12 mostra o tempo em meses de uso do veículo até o momento da falha na multimídia. Assim como feito para os dados meteorológicos, os valores dos meses foram distribuídos em bins e plotados de forma cronológica para melhor entendimento. É possível verificar que existe uma distribuição de amostras ao longo dos 36 meses de garantia do veículo não apresentando concentrações ou picos em determinado tempo de uso. Em relação aos modelos dos veículos, é importante ter em mente que os modelos aqui chamados de MODELO 1, MODELO 2 e MODELO 3 são os veículos mais caros, chegando a valores superiores à R$ 180.000,00 e consequentemente mais tecnológicos. É nítida a predominância dos 3 modelos nos Clusters 1-Azul, 2-Laranja e 3-Verde, sendo responsáveis por cerca de 97% do total de amostras. Isto ocorre devido ao fato de veículos mais modernos possuírem mais funcionalidades que são geridas por interfaces presentes na central multimídia, fazendo com que as mesmas sejam mais robustas e complexas aumentando o risco de falha neste componente. Controle de temperatura, partida do motor, rádio, travamento de portas, seletor de terreno, verificação de calibração dos pneus e GPS, por exemplo, podem ser controlados pela central multimídia da maioria das versões dos modelos citados.

Olhando a motorização (terceiro plot), fica evidente que os motores mais robustos que estão presentes na maioria dos veículos mais modernos e tecnológicos, tenham a maior parte das amostras e concentração dos clusters. Outra informação relevante é que o motor 2.4 Flex, apesar de estar presente apenas nas versões topo de linha dos modelos, não apresentou um grande número de casos. Isto acontece devido ao fato da baixa



comercialização das versões com tal motorização atrelada principalmente pelo grande consumo de combustível. Por último, o gráfico de *Repeated Repair* ou Repetição de Reparo, mostra que a maioria dos casos e clusters estão concentrados em não, ou seja, após a primeira passagem na concessionária o problema foi resolvido. Para o Cluster 4-Vermelho, onde os modelos e versões são mais simples (conforme visto na distribuição por modelo) verifica-se que não houveram reparações repetidas o que faz total sentido, visto que as suas centrais são mais simples o que facilita o diagnóstico do problema.

DOI: 10.25286/repa.v7i2.2222

4.2 DISCUSSÃO

Conforme mostrado ao longo do trabalho, o fato das massas de dados não serem regulares fez com que o algoritmo aglomerativo hierárquico se saísse melhor nos estudos que foram apresentados. Quando os dados são utilizados sem adição dos dados meteorológicos do INMET, é possível verificar resultados ainda melhores nas distribuições, visto que o problema passa a ter menos variáveis, facilitando o processo de clusterização dos dados de garantia.

Nas análises iniciais onde foram utilizadas as variações do método k-modes para avaliar a junção dos dados de garantia mais meteorológico os resultados foram muito baixos sendo possível observar até valores negativos de silhueta para o inicializador aleatório (Huang) indicando que a boa parte das amostras foram atribuídas ao cluster errado. Tal comportamento também é visto quando se utiliza o inicializador por densidade (Cao) que é mais sofisticado, mas em menor quantidade. Tais resultados não geram confiança nas distribuições que foram encontradas.

Os resultados encontrados na aplicação do k-modes para os dados singelos de garantia também tiveram desempenhos relativamente baixos. Apesar da melhora no coeficiente de silhueta, quando plotado os gráficos para um total de 4 clusters, é possível verificar que a tendência de erro na escolha dos clusters.

Observando o método de aglomeração por hierarquia, os resultados chegaram a ser aproximadamente 75% melhores do que os encontrados pelo k-modes no que se diz respeito apenas ao coeficiente de silhueta. Ao observar os gráficos de silhueta com ou sem dados meteorológicos, é possível verificar que a massa de dados foi classificada em sua grande maioria de forma correta e que a separação dos clusters também estão melhores entre si.

O modelo de clusterização apresentou um coeficiente de silhueta aproximadamente 7% melhor quando os dados meteorológicos foram desconsiderados. Um dos motivos que justificam tal melhora é a forma com o qual os dados meteorológicos foram adicionados, visto que foram utilizados como ponto de partida da aquisição através da capital do estado da concessionária na qual o veículo foi reparado. Isso desconsidera alguns fatores importantes como por exemplo que o veículo foi reparado durante uma viagem e que cidades pertencentes ao mesmo estado, mas longe da capital podem ter climas totalmente diferentes, por exemplo.

As análises mais aprofundadas dos clusters indicam a grande concentração de anomalias e reclamações são dadas nos veículos mais completos, devido à maior robustez e maior quantidade de funcionalidades das centrais. Já na distribuição por tempo de uso, verifica-se que há uma distribuição em basicamente todos os meses ao qual o veículo está sob garantia, mas é possível ver uma elevação dos três principais clusters (modelos mais caros) até os 18 meses de garantia, seguidos de uma diminuição nos clusters Azul, um leve aumento do cluster Verde que volta a cair após o período de 24 meses e de um aumento gradativo no cluster Laranja após os 18 meses. O fato de não ser possível verificar uma curva ascendente, mostra que as falhas que estão ocorrendo não se devem à problemas de vida-útil das peças. Uma baixa concentração nos primeiros meses de utilização, também mostra que não foram problemas que saíram da montadora e não foram identificados. Tal distribuição de dados mostra que houveram problemas de qualidade nas multimídias que foram identificadas ao longo do tempo nos veículos.

Outro ponto interessante é que nos itens de Repeated Repair, os veículos mais simples não passam por mais de duas intervenções para o mesmo problema, o que afirma que quando um problema ocorre em multimídias mais simples, a diagnose é mais precisa e o problema é prontamente resolvido.

5 CONCLUSÕES E TRABALHOS FUTUROS

Mediante os resultados verificados durante o desenvolvimento do trabalho, pode se concluir que o foco principal da montadora em questão é priorizar as centrais multimídias dos veículos MODELO 1, MODELO 2 e MODELO 3 nos estoques, visto a grande concentração de casos ao longo de todo o período de garantia do veículo.

Os modelos criados utilizando os dados do INMET necessitam de uma melhor calibração. Como possibilidade de trabalho futuro, é possível o desenvolvimento de um modelo que utilize a cidade ou coordenada geográfica da concessionária que realizou a troca do componente para buscar qual a estação de coleta de dados meteorológica mais próxima, aumentando assim a acurácia dos dados utilizados. Sobre os dados de garantia, pretende-se realizar treinamento em modelos com um período maior de dados, mais modelos de veículos e também podem ser realizados trabalho no qual sejam utilizados outros componentes que possuam (ou não) componentes eletrônicos.

REFERÊNCIAS

[1] GORMSEN, Niels Joachim; KOIJEN, Ralph SJ. **Coronavirus: Impact on stock prices and growth expectations.** The Review of Asset Pricing Studies, Oxford University Press, v. 10, n. 4, p. 574–597, 2020.

[2] PAINEL Coronavírus - Ministério da Saúde - Governo Federal do Brasil. [S.l.: s.n.]. Acessado em 12 de novembro, 2021. Disponível em: https://covid.saude.gov.br

[3] DONTHU, Naveen; GUSTAFSSON, Anders. **Effects of COVID-19 on business and research.** [S.l.]: Elsevier, 2020.

[4] ALTHAF, Shahana; BABBITT, Callie W. **Disruption risks to material supply chains in the electronics sector**. Resources, Conservation and Recycling, Elsevier, v. 167, p. 105248, 2021.

[5] USED Tech and Gadget Repair Businesses Are Booming Right Now. [S.l.: s.n.]. Acessado em 10 de novembro, 2021. Disponível em: https://onezero.medium.com/used-tech-and-gadget-repair-businesses-%20are-booming-right-now-46531abf4bbc.

[6] ENTENDA o que são semicondutores e por que eles estão em falta no mundo todo. [S.l.: s.n.]. Acessado em 05 de novembro, 2021. Disponível em: https://epocanegocios.globo.com/Tecnologia/noticia/2021/03/entenda-o-que-sao-semicondutores-e-porque-eles-estao-em-falta-no-mundo-todo.html

[7] GUANETTI, Jacopo; KIM, Yeojun; BORRELLI, Francesco. **Control of connected and automated vehicles: State of the art and future challenges**. Annual reviews in control, Elsevier, v. 45, p. 18–40, 2018.

[8] WU, Xiling; ZHANG, Caihua; DU, Wei. **An analysis on the crisis of "chips shortage" in automobile industry**——Based on the double influence of COVID-19 and trade friction. In: IOP PUBLISHING, 1. JOURNAL of Physics: Conference Series. [S.l.: s.n.], 2021. v. 1971, p. 012100.

[9] BRASIL tem 29 fábricas de veículos paradas: 'Crise sem precedentes'. [S.l.: s.n.]. Acessado em 23 de novembro, 2021. Disponível em: https://economia.uol.com.br/noticias/bbc/2021/04/05/brasil-vive-crise-etem-29-fabricas-de-veiculos-paradas.htm

[10] REBELO, CGS; PEREIRA, MT; SILVA, JFG; FERREIRA, LP; SÁ, JC; MOTA, AM. **After sales service: key settings for improving profitability and customer satisfaction**. Procedia Manufacturing, Elsevier, v. 55, p. 463–470, 2021.

[11] LU, Ming-Wei. **Automotive reliability prediction based on early field failure warranty data.** Quality and Reliability Engineering International, Wiley Online Library, v. 14, n. 2, p. 103–108, 1998.

[12] SHOKOUHYAR, Sajjad; AHMADI, Sadra; ASHRAFZADEH, Mahdi. **Promoting a novel method for warranty claim prediction based on social network data**. Reliability Engineering & System Safety, Elsevier, v. 216, p. 108010, 2021.

[13] MAIMON, Oded; ROKACH, Lior. **Data mining and knowledge discovery handbook.** Springer, 2005.

[14] PENA, José M; LOZANO, Jose Antonio; LARRANAGA, Pedro. **An empirical comparison of four initialization methods for the k-means algorithm.** Pattern recognition letters, Elsevier, v. 20, n. 10, p. 1027–1040, 1999.

[15] HUANG, Zhexue. **Clustering large data sets with mixed numeric and categorical values**. In: CITESEER. PROCEEDINGS of the 1st pacific-asia conference on knowledge discovery and data mining,(PAKDD). [S.l.: s.n.], 1997. p. 21–34.

[16] HUANG, Zhexue. **Extensions to the k-means algorithm for clustering large data sets with categorical values**. Data mining and knowledge discovery, Springer, v. 2, n. 3, p. 283–304, 1998.

[17] CAO, Fuyuan; LIANG, Jiye; BAI, Liang. **A new initialization method for categorical data clustering**. Expert Systems with Applications, Elsevier, v. 36, n. 7, p. 10223–10228, 2009.

[18] HIERARCHICAL clustering. [S.l.: s.n.]. Acessado em 15 de novembro, 2021. Disponível em: https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering

[19] JAIN, Anil K; MURTY, M Narasimha; FLYNN, Patrick J. **Data clustering: a review**. ACM computing surveys (CSUR), Acm New York, NY, USA, v. 31, n. 3, p. 264–323, 1999.

[20] DADOS HISTÓRICOS ANUAIS. [S.l.: s.n.]. Acessado em 10 de setembro, 2021. Disponível em: https://portal.inmet.gov.br/dadoshistoricos

DOI: 10.25286/repa.v7i2.2222