PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA
Pós-graduação Lato Sensu em Ciência de Dados e Big Data

Carlos Eduardo Ribeiro Leite

PREVISÃO DE RISCO DE FRAUDES DE VALORAÇÃO EM IMPORTAÇÃO DE KITS DE TRANSMISSÃO DE MOTOCICLETAS

Teresina - PI
2021

Carlos Eduardo Ribeiro Leite

PREVISÃO DE RISCO DE FRAUDES DE VALORAÇÃO EM IMPORTAÇÃO DE KITS DE TRANSMISSÃO DE MOTOCICLETAS

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Ciência de Dados e Big Data como requisito parcial à obtenção do título de especialista.

Teresina

2021

SUMÁRIO

1. Introdução

1.1. Contextualização

A Globalização é um fenômeno expressado pela diminuição das barreiras e distâncias entre as nações do mundo moderno, seja no âmbito social, cultural, econômico, ambiental ou político, e uma de suas maiores consequências é a integração comercial entre os países.

Com essa conexão mundial acontecendo dentro do âmbito das relações comerciais e econômicas de forma avassaladora, a competitividade foi extremamente acirrada, criando uma disputa mundial, em especial pelos mercados de grandes volumes de negociações.

Dentro desse panorama é impossível uma análise que não cite o aumento de competitividade dos produtos chineses, tendo em vista que a china há 40 anos não exportava quase nada e nos últimos 25 anos saltou de cerca de 200 bilhões de dólares para mais de 3 trilhões de dólares em 2020[1], um crescimento de mais de 15 vezes.

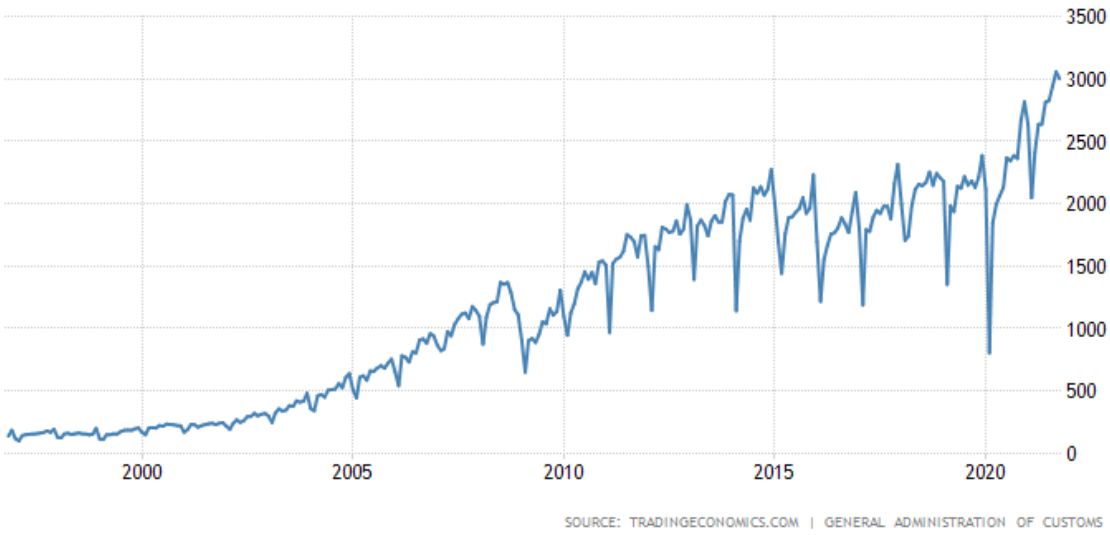


Figura 1 - Gráfico Exportações da China 25 anos

[1] Fonte: https://pt.tradingeconomics.com/china/exports. Acesso em 09/11/2021 às 16:30.

Especificamente para o mercado brasileiro, somente em 2019, a China exportou quase 35 bilhões de dólares, dos quais mais de 250 milhões de dólares foram somente de partes e acessórios de motocicletas e bicicletas[2].

Para que fosse possível a identificação das mercadorias em circulação no mundo foi implementado o Sistema Harmonizado de Descrição e Codificação de Mercadorias, geralmente denominado "Sistema Harmonizado" ou simplesmente "SH" (ou HS em inglês: *Harmonized System*).

O SH é uma nomenclatura internacional de produtos, desenvolvida pela Organização Mundial das Aduanas (OMA). O Brasil utiliza esse sistema como base para a classificação de mercadorias na Nomenclatura Comum do Mercosul (NCM).

O Brasil utiliza a NCM para classificar qualquer mercadoria que entre ou circule no seu território, sendo utilizada nas operações de comércio internacional, seja importação ou exportação, assim como no comércio doméstico, sendo exigência na emissão de documentos fiscais.

A classificação de uma mercadoria é possível de ser realizada com qualquer mercadoria existente ou que venha a existir, posto que há códigos residuais que são utilizados seguindo as regras de classificação estabelecidas dentro do previsto no Acordo Geral sobre Tarifas e Comércio 1994 (GATT 1994) em seu artigo 3º.

O acordo do GATT também prevê em seu artigo 6º as formas de valoração de uma mercadoria em circulação entre países, estabelecendo que preferencialmente o valor da transação seja utilizado. Entretanto, o disposto no acordo deixa margem para conluios entre os intervenientes de comércio exterior com vistas a reduzir o valor declarado de mercadorias, reduzindo do mesmo modo os tributos incidentes na transação; prática comumente chamada de subvaloração ou subfaturamento.

Para o mercado de peças de motocicletas há dois grandes problemas para a fiscalização aduaneira quanto à subvaloração. O primeiro consiste em que praticamente todas as peças de motocicletas, incluindo suas milhares de partes, marcas e modelos são classificadas em uma mesma posição da Tabela de Classificação da Nomenclatura Comum do

[2] Fonte: https://www.trademap.org/. Acesso em 09/11/2021 às 18:58.

Mercosul – NCM (87141000), enquanto o segundo consiste em determinar valores razoáveis e reais das transações, principalmente as oriundas do mercado chinês.

Esta questão de uma mesma classificação fiscal englobar diversos itens diferentes, impossibilitando uma automatização direta do trabalho de análise de risco não acontece somente no mercado de peças de motocicletas, existindo outros casos, como o de brinquedos, por exemplo.

A atuação da aduana, como controladora do comércio exterior nacional, é fundamental na proteção às indústrias brasileiras e na manutenção de empregos, em seus esforços para evitar as fraudes no comércio exterior, o contrabando e o descaminho, bem como a introdução, no mercado nacional, de mercadorias a preços subfaturados, com os quais nossas indústrias não podem competir.

Desse modo, seria extremamente útil uma forma de se analisar os muitos produtos diferentes todos os dias importados e que utilizam a mesma classificação fiscal de partes e peças de motocicletas, separando-os, classificando-os e estabelecendo se os valores declarados podem ser considerados dentro do espectro normalmente comercializado em condições normais no mercado internacional.

1.2. O problema proposto

A previsão adequada de um limite justificável de valor é importante para que se possa fazer o correto controle aduaneiro na entrada de mercadorias no país, especialmente em um mercado onde os itens importados são tão presentes como o de peças de motocicletas.

O grande problema existente é que praticamente todas as peças de motocicletas se concentram em uma única classificação fiscal, ou seja, estão agrupadas pelo código NCM 8714.10.00, encontrando-se a distinção somente no campo descrição preenchido pelo importador.

Essa grande quantidade de tipos de peças e itens completamente diferentes em uma mesma NCM causa um trabalho muito grande para os Auditores responsáveis pela verificação da regularidade de uma declaração de importação e dificulta sobremaneira a automatização do processo de trabalho, principalmente quanto à análise possíveis fraudes de valor declarado.

A análise aqui proposta baseia-se nos dados governamentais disponibilizados no Sistema Siscori. Para isso, extraíram-se as importações do Capítulo 87 da NCM (Veículos automóveis, tratores, ciclos e outros veículos terrestres, suas partes e acessórios).

Devido ao tamanho da base de dados obtida, limitou-se o escopo deste trabalho especificamente aos Kits de Transmissão presentes na subposição 8714.10.00.

Figura 2 - Kit de Transmissão (corrente, coroa e pinhão)

Deseja-se realizar atividades de ciência de dados para determinar estatisticamente as faixas de valores declarados para uma aplicação específica e comparar ao valor declarado em uma importação, observando sua dispersão quanto aos percentis do conjunto de declarações estudadas.

O objetivo é que se possa estabelecer um grau de risco quanto à existência de fraude de valor no item analisado de dada Declaração de Importação.

Para esse estudo utilizou-se os dados referentes às importações que compreendem o período de janeiro de 2020 a junho de 2021 (18 meses).

1.3. Objetivos

O objetivo é a partir dos dados de uma Declaração de Importação contendo kits de transmissão (corrente, coroa e pinhão), analisar sua descrição, determinar a classificação da aplicação do kit importado e indicar visualmente um grau de risco de fraude quanto ao valor declarado no item analisado da Declaração de Importação.

2. Coleta de Dados

Dentro do estudo proposto no escopo desse trabalho, utilizamos quatro *datasets* obtidos de distintas fontes e têm as características expostas a seguir.

Tabela NCM

A tabela NCM utilizada foi gerada através do gerador encontrado no sítio https://portalunico.siscomex.gov.br/classif/#/nomenclatura/tabela?perfil=publico mantido pelo Portal Único do Comércio Exterior Siscomex.

Nome da coluna	Descrição	Tipo
Código	Número de ordem (chave primária)	Inteiro
Descrição	Descrição do item da NCM	String (Texto)
Data Início	Início da vigência do item	Data
Data Fim	Início da vigência do item	Data
Ato Legal	Ato legal criador do item	String (Texto)
Número	Número do ato legal	Inteiro
Ano	Ano do ato legal	Inteiro

Tabela 1 - Estrutura de Dados da Tabela NCM

Tabela Marcas e Modelos de Motocicletas

A base de dados orginalmente disponível para download[3] é composta de vários arquivos do tipo CSV, dos quais foram utilizados dois especificamente.

O primeiro denominado *"marcas-motos.csv"* com a estrutura seguinte:

Nome da coluna	Descrição	Tipo
ID	Número de ordem (chave primária)	Inteiro
NOME	Nome da marca fabricante de motocicleta	String (Texto)

Tabela 2 - Estrutura de Dados da Tabela Marcas de Motocicletas

O outro arquivo se chama *"modelos-moto.csv"* e possui três colunas especificadas a seguir:

[3] Fonte: https://www.luiztools.com.br/post/base-de-dados-com-todas-as-marcas-e-modelos-de-veiculos/. Acesso em 13/12/2021 às 13:03.

Nome da coluna	Descrição	Tipo
ID	Número de ordem (chave primária)	Inteiro
IDMARCA	ID da marca que identifica o fabricante	Inteiro
NOME	Nome da marca fabricante de motocicleta	String (Texto)

Tabela 3 - Estrutura de Dados da Tabela Modelos de Motocicletas

A fim de auxiliar na classificação dos dados obtidos de forma que se determinem *features* para aplicar no aprendizado de máquina, consolidamos uma tabela de marcas e modelos de motocicletas existentes no mercado nacional a partir da base de dados com todas as marcas e modelos.

Utilizando-se desses dados, construiu-se uma tabela na qual foram identificados de forma separada os registros para cada modelo de motocicleta por marca e segregando os termos existentes na descrição do modelo para que se pudesse encontrar mais facilmente estes termos na descrição fornecida pelo importador.

Desse modo, para utilizar nesse trabalho, o *dataset* final construído tem a seguinte estrutura:

Nome da coluna	Descrição	Tipo
ID	Número de ordem (chave primária)	Inteiro
IDMARCA	Identificador da Marca	Inteiro
MARCA	Marca da motocicleta	String (Texto)
Modelo1	Característica 1 do Modelo da motocicleta	String (Texto)
Modelo2	Característica 2 do Modelo da motocicleta	String (Texto)
Modelo3	Característica 3 do Modelo da motocicleta	String (Texto)
NOME	Descrição completa do modelo da motocicleta	String (Texto)

Tabela 4 - Estrutura de Dados da Tabela Marcas e Modelos de Motocicletas

Dados de importações do capítulo 87 da NCM (Siscori)

Os dados referentes à importação foram obtidos através do sistema de apoio Siscori, disponível no sítio na internet localizado em https://siscori.receita.fazenda.gov.br, com acesso público aos dados.

O Sistema tem o objetivo de disponibilizar um conjunto de informações referentes às importações e exportações brasileiras, respeitando o sigilo fiscal, para apoio a outros sistemas e análises estatísticas em geral.

As extrações de importação e exportação são mensais e cada capítulo gera um arquivo separado, contendo somente os seus subitens que tenham tido operações promovidas por pelo menos quatro importadores no período de extração. Caso não exista nenhum subitem nessa situação, não será gerado o arquivo daquele capítulo.

A página de publicação automática dos arquivos estatísticos importação e exportação de apoio ao Siscori é amparada pela Lei de Acesso à Informação (Lei nº 12.527, de 18 de novembro de 2011) e por portaria normativa da RFB, publicada no Diário Oficial da União.

A tela a seguir mostra como foi realizada a extração dos dados utilizados. Desta forma foram obtidos os dados referentes às importações do capítulo 87 da NCM no período de janeiro de 2020 a junho de 2021, perfazendo um total de 18 arquivos do tipo CSV nomeados no padrão CAPXXAAMM.csv:

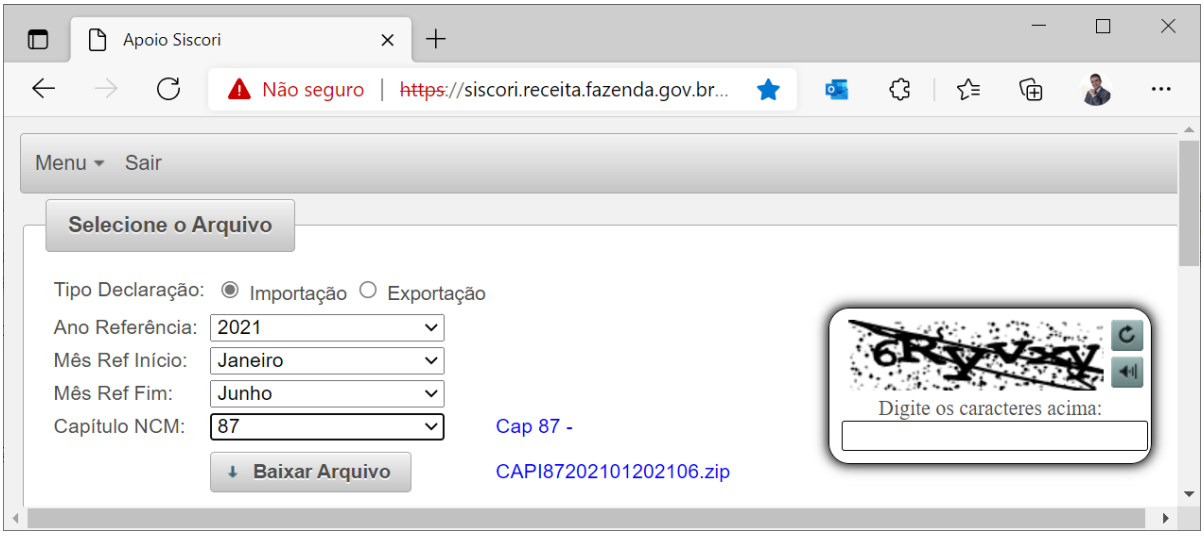


Figura 3 - Tela de importação de dados do Sistema Apoio Siscori

Cada arquivo CSV obtido engloba todas as importações realizadas no período especificado cujas classificações fiscais (NCM) estão contidas no capítulo 87 (Veículos automóveis, tratores, ciclos e outros veículos terrestres, suas partes e acessórios). Entretanto, é preciso salientar que o escopo deste trabalho se restringe aos itens classificados na subposição 8714.10.00 (Partes e acessórios de motocicletas, incluindo os ciclomotores)

Os dados obtidos no Siscori são fornecidos com a seguinte estrutura:

Nome da coluna	Descrição	Tipo
NUMERO DE ORDEM	Número de ordem (chave primária)	String (Texto)
ANOMES	Ano e mês do registro da importação	String (Texto)

Nome da coluna	Descrição	Tipo
COD.NCM	Classificação Fiscal (NCM) da mercadoria	String (Texto)
DESCRICAO DO CODIGO NCM	Descrição da mercadoria na tabela NCM	String (Texto)
PAIS	Código do país de origem	Inteiro
PAIS DE ORIGEM	Descrição do país de origem	String (Texto)
PAIS	Código do país de aquisição	Inteiro
PAIS DE AQUISICAO	Descrição do país de aquisição	String (Texto)
UND.ESTAT.	Código da unidade estatística da mercadoria	Inteiro
UNIDADE DE MEDIDA	Descriçaõ da unidade estatísica de medida	String (Texto)
UNIDADE COMERC.	Descrição da unidade de comercialização da mercadoria	String (Texto)
DESCRICAO DO PRODUTO	Descrição da mercadoria feita pelo importador	String (Texto)
QTDE ESTATISTICA	Quantidade na unidade de medida estatística	Float
PESO LIQUIDO	Peso líquido da mercadoria importada em quilos	Float
VMLE DOLAR	Valor da mercadoria no local de embarque em dólares	Float
VL FRETE DOLAR	Valor do frete da mercadoria em dólares	Float
VL SEGURO DOLAR	Valor do seguro da mercadoria em dólares	Float
VALOR UN.PROD.DOLAR	Valor unitário da mercadoria em dólares	Float
QTD COMERCIAL.	Quantidade na unidade de medida comercial	Float
TOT.UN.PROD.DOLAR	Total unitário da mercadoria em dólares	Float
UNIDADE DESEMBARQUE	Unidade administrativa de desembarque	String (Texto)
UNIDADE DESEMBARACO	Unidade administrativa de desembaraço	String (Texto)
INCOTERM	Incoterm - termos de negociação da mercadoria	String (Texto)
NAT.INFORMACAO	Natureza da informação prestada	String (Texto)
SITUACAO DO DESPACHO	Situação administrativa do despacho	String (Texto)

Tabela 5 - Estrutura de Dados das Tabelas Importadas do Siscori

Os dados fornecidos pelo Siscori, mesmo estando em um arquivo texto do tipo CSV, apresentam colunas de largura fixa iniciadas pelo caractere “@”. Os dados foram importados e trabalhados utilizando-se a biblioteca Pandas e a linguagem de programação Python dentro do ambiente do *Jupyter Notebook*.

Tabela referência ABIMOTO

Utilizaremos como dataset de comparação os dados fornecidos pela Associação Brasileira dos Importadores de Motopeças – ABIMOTO, especificamente a 13ª versão do Estudo de Valoração Aduaneira: partes e peças para motocicletas.

A tabela é composta dos seguintes dados:

Nome da coluna	Descrição	Tipo
ITEM	Número de ordem (chave primária)	Inteiro
FOTO	Fotografia de referência do item	Objeto
PARTES E PEÇAS	Descrição do item em português	String (Texto)
MOTO PARTS	Descrição do item em inglês	String (Texto)
NCM	Classificação fiscal do item	String (Texto)
UNI	Unidade de comercialização do item	String (Texto)
13ª Versão	Valor indicativo mínimo do item em dólar americano	Float

Tabela 6 - Estrutura de Dados da Tabela de Referência da ABIMOTO

Referência entre as tabelas

Os *datasets* da Tabela NCM, os dados obtidos no Sistema Siscori e a Tabela da ABIMOTO referenciam-se entre si incialmente pela classificação fiscal. Entretanto, como já explicitado, tal dado é insuficiente par uma relação precisa entre os *datasets*, pois uma mesma NCM contempla inúmeros itens diferentes, já que, no nosso caso, praticamente todas as peças de motos classificam-se na mesma posição da Tabela de Nomenclatura Comum do Mercosul.

Posteriormente, um dos objetivos desse trabalho é exatamente criar através do processamento de inteligência artificial uma subclassificação por aplicação (marca, modelo e demais características) para os kits de transmissão, utilizando-se a Tabela de Marcas e Modelos, de forma que se possa comparar os valores declarados.

Importante salientar que o *dataset* das marcas e modelos de motos servirá como base para classificação de cada descrição dentro desse universo de possíveis modelos onde são aplicados os kits de transmissão descritos na Declaração de Importação.

3. Processamento e Tratamento de Dados

A seguir se apresenta a forma de processamento e tratamento dos dados utilizados neste trabalho, mas primeiramente é importante esclarecer que ao longo da dissertação sobre o estudo proposto evita-se a apresentação de códigos de programação. Todo o código estará disponível em *notebooks Jupyter*, comentado em um nível suficiente para entendimento, como anexos desse trabalho.

Tabela NCM

Dataset de referência quanto à nomenclatura adotada quanto à classificação fiscal dos itens constantes dos arquivos CSVs com dados obtidos no Sistema de Apoio Siscori, dos quais foram extraídos especificamente a descrição do Capítulo 87, da posição 1, subposição 4, item 10, conforme segue:

Capítulo 87: Veículos automóveis, tratores, ciclos e outros veículos terrestres, suas partes e acessórios.

SubPosição 87.14: Partes e acessórios dos veículos das posições 87.11 a 87.13.

Item **8714.10.00**: - De motocicletas (incluindo os ciclomotores)

Tabela de referência ABIMOTO

Da tabela fornecida pela ABIMOTO foram filtrados os registros com o script **1b_trataABIMOTO13.ipynb** (Anexo II) utilizando os mesmos critérios dos dados de importações obtidos no Siscore, ou seja, foram filtrados somente os itens classificados na NCM 8714.10.00 e que se refiram a Kits de Transmissão de motocicletas, perfazendo 16 registros.

Importante salientar que a tabela inicial continha 129 registros, sendo que um dos campos era uma fotografia indicativa do item. Neste trabalho optamos por excluir este registro de imagem e mantivemos as demais colunas com informações pertinentes.

Posteriormente, utilizando-se do modelo de *machine learning* elaborado, foi realizada a classificação por aplicação para cada um dos registros, de forma que se pudesse comparar com os registros das importações também classificados.

Por fim, foi criada uma coluna para especificamente designar se a corrente do kit de transmissão importado é com retentor ou não. Esse dado é importante tendo em vista haver uma variação relevante do valor quando o kit contém ou não corrente com retentor.

Figura 4 - Corrente com retentor

Outro ponto a salientar é que a tabela foi fornecida em um arquivo PDF, que foi transformado em um arquivo no formato Excel para importação, sem haver nenhum registro nulo ou com valores ausentes.

Dados de importações do capítulo 87 da NCM

O tratamento desses dados foi um pouco mais complexo, para que se pudesse, após o processamento extrair somente os itens de interesse devidamente classificados.

Inicialmente o *dataset* obtido era composto de 7.693.634 de registros, todos com informações referentes a importações realizadas no período de janeiro de 2020 a junho de 2021 de produtos classificados no capítulo 87.

Para o processamento desses dados foi utilizado script em Python contido no *Jupyter Notebook* **1a_trataCSVsSiscori.ipynb**.

O script basicamente identificou todos os arquivos CSV a serem tratados e em seguida realizou, para cada um deles, as funções de importação dos dados contidos no arquivo CSV, estabelecendo os tipos de dados corretos e nomeando as colunas de maneira a facilitar o trabalho.

Em seguida optou-se por eliminar todos os registros cujo campo de descrição fosse maior que 131.072 caracteres, limite padrão. Esta eliminação não tem impacto sobre o trabalho tendo em vista que somente havia 27 registros nessa condição e nenhum deles era de itens de interesse (kits de transmissão). Da mesma forma foram ignoradas linhas onde havia erro de importação; igualmente insignificantes 19 registros.

Nos *datasets* importados não houve registros com campos nulos ou sem valor.

A seguir foram filtrados os registros com produtos classificados na NCM 8714.10.00 e cuja descrição continha os termos “kit”, “transm”, “corrente”, “coroa” e “pinhão” ou “pinhao”, independente se maiúsculas ou minúsculas. Por fim, todos os registros foram concatenados em um único *dataset* de trabalho contendo 19.092 registros, exportando-o para um arquivo a ser trabalhado chamado **dataframe.xlsx**.

Tabela Marcas e Modelos de Motocicletas

O processamento dos dados da tabela de marcas e modelos de motos foi realizado de forma manual com a mesclagem das duas tabelas obtidas utilizando o software de edição de planilhas Excel e em seguida o convertendo para o formato CSV.

Essa tabela é essencial, pois ela representa o universo de marcas e modelos que constituem as aplicações possíveis dos kits de transmissão para classificar cada item da base de dados de importação.

Na parte dedicada à classificação, o que se fará é exatamente vincular cada um dos registros a uma aplicação (marca e modelo) presente neste dataset.

4. Análise e Exploração dos Dados

Os dados brutos obtidos através do Siscori fornecem todas as importações realizadas no período solicitado referentes a determinado capítulo da Tabela de Nomenclatura Comum do Mercosul – NCM.

Esses dados foram observados e que foi percebido que havia inúmeros itens classificados na mesma NCM do objeto de estudo deste trabalho, qual seja o Kit de Transmissão para motocicletas composto de corrente, coroa e pinhão.

Dessa análise optou-se por filtrar os dados para contemplar somente os itens que estivessem dentro do escopo pretendido.

Em virtude do objeto de estudo ser determinado através de um campo de descrição, onde o importador ou seu representante faz a entrada livre de dados, sem qualquer exigência ou padronização, nossa análise exploratória foi convincente no sentido de que deveria ser feito um tratamento com a utilização de Programação de Linguagem Natural – PLN.

Dentro dessa análise exploratória de dados, observou-se a ocorrência de palavras que, embora frequentes, serão irrelevantes para a determinação da diferenciação entre os itens que estão classificados dentro do agrupamento representado pela NCM em estudo.

Comum no estudo de categorização de textos, o uso de *stopwords* é recomendado nesse caso, pois da mesma forma que artigos, pronomes e outras palavras tradicionalmente identificadas como não relevantes para a distinção, termos como kit, transmissão, corrente, coroa e pinhão também são igualmente irrelevantes para distinguir um item de outro.

Na biblioteca **Natural Language Processing Toolkit - nltk** já existe uma lista de *stopwords* para a língua portuguesa, constituída dos seguintes vocábulos:

> *“a”, “ao”, “aos”, “aquela”, “aquelas”, “aquele”, “aqueles”, “aquilo”, “as”, “até”, “com”, “como”, “da”, “das”, “de”, “dela”, “delas”, “dele”, “deles”, “depois”, “do”, “dos”, “e”, “ela”, “elas”, “ele”, “eles”, “em”, “entre”, “era”, “eram”, “essa”, “essas”, “esse”, “esses”, “esta”, “estamos”, “estas”, “estava”, “estavam”, “este”, “esteja”, “estejam”, “estejamos”, “estes”, “esteve”, “estive”, “estivemos”, “estiver”, “estivera”, “estiveram”, “estiverem”, “estivermos”, “estivesse”, “estivessem”, “estivéramos”, “estivéssemos”, “estou”, “está”, “estávamos”, “estão”, “eu”, “foi”, “fomos”, “for”, “fora”, “foram”, “forem”, “formos”, “fosse”, “fossem”, “fui”, “fôramos”, “fôssemos”, “haja”, “hajam”, “hajamos”, “havemos”, “hei”, “houve”, “houvemos”, “houver”, “houvera”, “houveram”, “houverei”, “houverem”,*

“houveremos”, “houveria”, “houveriam”, “houvermos”, “houverá”, “houverão”, “houveríamos”, “houvesse”, “houvessem”, “houvéramos”, “houvéssemos”, “há”, “hão”, “isso”, “isto”, “já”, “lhe”, “lhes”, “mais”, “mas”, “me”, “mesmo”, “meu”, “meus”, “minha”, “minhas”, “muito”, “na”, “nas”, “nem”, “no”, “nos”, “nossa”, “nossas”, “nosso”, “nossos”, “num”, “numa”, “não”, “nós”, “o”, “os”, “ou”, “para”, “pela”, “pelas”, “pelo”, “pelos”, “por”, “qual”, “quando”, “que”, “quem”, “se”, “seja”, “sejam”, “sejamos”, “sem”, “serei”, “seremos”, “seria”, “seriam”, “será”, “serão”, “seríamos”, “seu”, “seus”, “somos”, “sou”, “sua”, “suas”, “são”, “só”, “também”, “te”, “tem”, “temos”, “tenha”, “tenham”, “tenhamos”, “tenho”, “terei”, “teremos”, “teria”, “teriam”, “terá”, “terão”, “teríamos”, “teu”, “teus”, “teve”, “tinha”, “tinham”, “tive”, “tivemos”, “tiver”, “tivera”, “tiveram”, “tiverem”, “tivermos”, “tivesse”, “tivessem”, “tivéramos”, “tivéssemos”, “tu”, “tua”, “tuas”, “tem”, “tínhamos”, “um”, “uma”, “você”, “vocês”, “vos”, “à”, “às”, “é” e “éramos”.

Como essa biblioteca já possui a funcionalidade de complementação dessa lista, adicionou-se os termos seguintes, que pela sua característica não distinguem os itens dentro do *dataset* estudado:

“abaixo”, “acessorios”, “acessórios”, “aco”, “acondicionados”, “adaptavel”, “adaptável”, “almas”, “am”, “anel”, “ano”, “aplicacao”, “aplication”, “aplicavel”, “aplicação”, “aplicável”, “application”, “ate”, “atitanium”, “aç”, “aço”, “bravo”, “cada”, “caixa”, “caixas”, “cambio”, “certificado”, “cever”, “chain”, “chh”, “china”, “ciclomotores”, “cod”, “code”, “codigo”, “comando”, “combustão”, “comercial”, “comercialmente”, “commodity”, “compativel”, “compatível”, “compl”, “completo”, “completos”, “composto”, “composto”, “compostopor”, “compostpo”, “comum”, “condicao”, “condicoes”, “condição”, “condições”, “confeccionado”, “conhecido”, “conj”, “conjunto”, “conjuntos”, “constituido”, “constitutivo”, “constituído”, “contendo”, “coposto”, “coroa”, “corr”, “corrent”, “corrente”, “correntee”, “correntes”, “cx”, “câmbio”, “cãmbio”, “código”, “decreto”, “denominada”, “dente”, “dentes”, “descricao”, “descrição”, “descriçao”, “descrição”, “destaque”, “destaque”, “destaques”, “detransmissão”, “dimensao”, “dimensoes”, “dimensão”, “dimensões”, “diverso”, “diversos”, “dominado”, “durabilidade”, “elo”, “elos”, “embalagem”, “engrenagem”, “engrenagens”, “epinhao”, “epinhão”, “espessura”, “exclusivo”, “fabri”, “fabricada”, “fabricado”, “final”, “foiproduzida”, “funcao”, “funcao”, “funcão”, “funcão”, “funçao”, “função”, “gtin”, “hardt”, “ho”, “hp”, “imetro”, “in”, “incluso”, “indicado”, “ingles”, “inmetro”, “inv”, “invoice”, “iron”, “item”, “jc”, “kif”, “kit”, “kitr”, “kittr”, “kmc”, “ligacoes”, “ligaçoes”, “ligaçoes”, “ligações”, “marca”, “mark”, “match”, “material”, “maxx”, “medida”, “medidas”, “medindo”, “metal”, “mod”, “modelo”, “modelos”, “moto”, “motocicleta”, “motocicletas”, “motoneta”, “motonetas”, “motoparts”, “motor”, “motos”, “motos”, “movimento”, “nbsp”, “ncm”, “nome”, “normais”, “nova”, “novo”, “numero”, “número”, “onde”, “or”, “origem”, “oring”, “ox”, “papelao”, “papelão”, “part”, “partes”, “parts”, “pc”, “pcs”, “pec”, “pecas”, “perfil”, “peç”, “peças”, “pinhao”, “pinhão”, “posição”, “premium”, “procedencia”, “procedência”, “prodepe”, “produto”, “próprio”, “pç”, “qdes”, “qtd”, “qtds”, “qty”, “quantidade”, “ref”, “reforcada”, “reforçada”, “registro”, “rel”, “relacao”, “relação”, “reposicao”, “resp”, “respo”, “respons”, “responsa”, “responsav”, “responsave”, “responsavel”, “responsável”, “ret”, “retentor”, “riffel”, “ring”, “roda”, “sae”, “scud”, “semi”, “semi-”, “semi-kit”, “sendo”, “serve”, “shipping”, “sistema”, “sm”, “standard”, “standart”, “standartt”, “std”, “stdmodelo”, “steel”, “tambem”, “também”, “tec”, “temp”, “temperado”, “tipo”, “tipos”, “titanio”, “titaniu”, “titanium”, “titaniun”, “tr”, “tracao”, “tracão”, “trans”, “transmis”, “transmisao”, “transmiss”, “transmissa”, “transmissao”, “transmission”, “transmissão”, “transmitir”, “traseira”, “traçao”, “tração”, “und”, “unds”, “unid”, “unidade”, “unidade”, “unidades”, “unifort”, “uo”, “uso”, “utilizada”, “utilizadas”, “utilizado”, “utilizados”, “utilização”, “vem”, “venda”, “with”, “xy” e “year”.

Para demonstrar tal conclusão através de imagens, a figura abaixo retrata a **nuvem de palavras** antes da inclusão das *stopwords* citadas, construída com a utilização da biblioteca *wordcloud* disponível para a linguagem de programação Python.

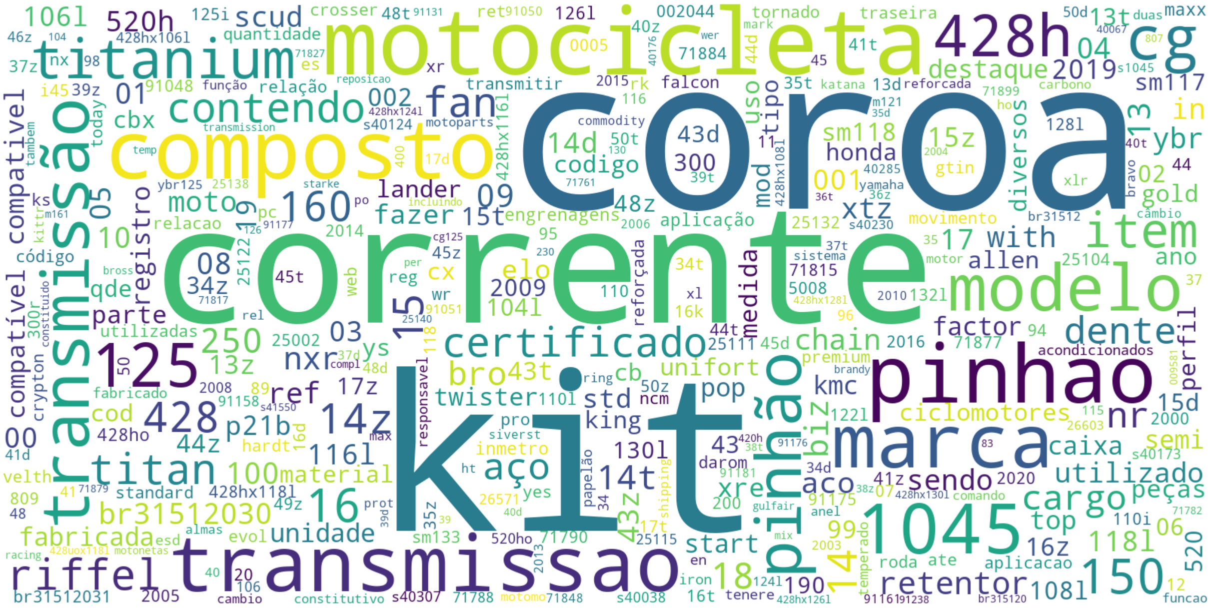


Figura 5 - *WordCloud* antes da remoção de stopwords

Do mesmo modo, a imagem a seguir mostra a **nuvem de palavras** após a inclusão das *stopwords* consideradas irrelevantes para o *dataset*.

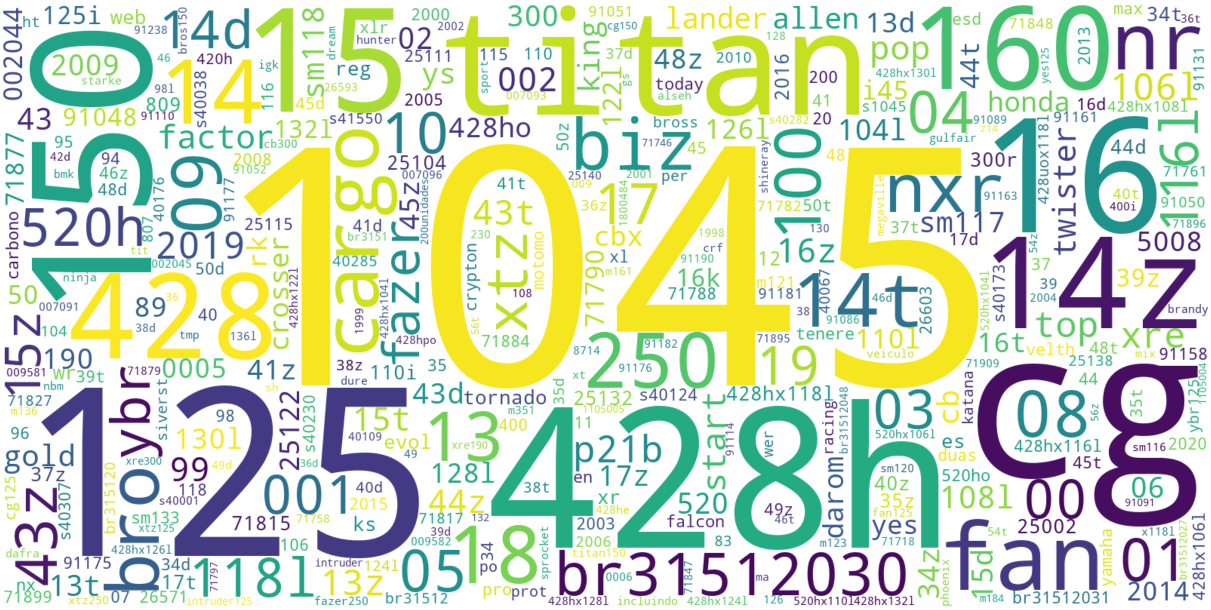


Figura 6 - *WordCloud* após da remoção de stopwords

Os códigos para criação das nuvens de palavras estão disponíveis no *notebook Jupyter* denominado **2_geraWordclouds.ipynb**.

5. Criação de Modelos de *Machine Learning*

A grande dificuldade na tarefa de análise de valores compatíveis na importação de peças de motocicletas, em especial dos kits de transmissão, se dá no fato de que milhares de importadores adquirem essas peças no exterior e informam sua descrição em um campo texto livre.

Nem mesmo a utilização da classificação fiscal normatizada no Mercosul, chamada de Nomenclatura Comum do Mercosul – NCM ajuda nesse caso específico, tendo em vista que grande parte das peças de motocicletas e todos os kits de transmissão são classificados em uma mesma posição na tabela da NCM.

Para que se possa tratar corretamente o *dataset* obtido na nossa etapa de processamento e tratamento de dados, assim como com a simplificação dos termos com a exclusão dos irrelevantes através dos filtros da nossa análise exploratória de dados, fez-se necessária uma classificação por aplicação para cada um dos itens importados e em seguida utilizou-se essa classificação para treinar o modelo de *machine learning* que fará a classificação dos futuros itens.

Somente através de uma classificação coerente é que se poderá agrupar itens semelhantes para que se possam efetuar os cálculos e obter as medidas estatísticas e probabilidades quanto ao valor declarado.

5.1. Criação das Aplicações para o aprendizado supervisionado

Em um primeiro momento, devido à inexistência de uma classificação para que se pudesse utilizar o aprendizado supervisionado, pensou-se na utilização de técnicas de aprendizado não supervisionado na determinação da motocicleta de aplicação do kit de transmissão analisado em cada registro.

No entanto, ao se concluir a análise exploratória dos dados, excluindo-se da descrição as *stopwords* da língua portuguesa e, em seguida, as palavras irrelevantes para a determinação da aplicação, decidiu-se pela criação de uma classificação que embasa o aprendizado supervisionado.

Para melhor entendimento, a ideia foi utilizar técnicas de processamento de linguagem natural e extrai para uma nova coluna chamada Modelo no *dataset* as informações relevantes que pudessem classificar a aplicação contida na descrição.

Tal procedimento foi realizado através dos códigos contidos nos seguintes *notebooks* Jupyter:

3_classificarAplicação.ipynb: código responsável pela criação da coluna DESCRICAO, que, a partir da coluna original DESCRICAO DO PRODUTO, fará a limpeza dos termos irrelevantes, de toda a sujeira que não contribui com a determinação da aplicação, buscará as palavras chaves existentes no *dataset* Aplicações e, por fim, acrescentará as marcas de acordo com os modelos dos fabricantes, já que a grande maioria dos modelos são de nomes patenteados pelas marcas.

Visualmente podemos observar como essa limpeza afetou a WordCloud:

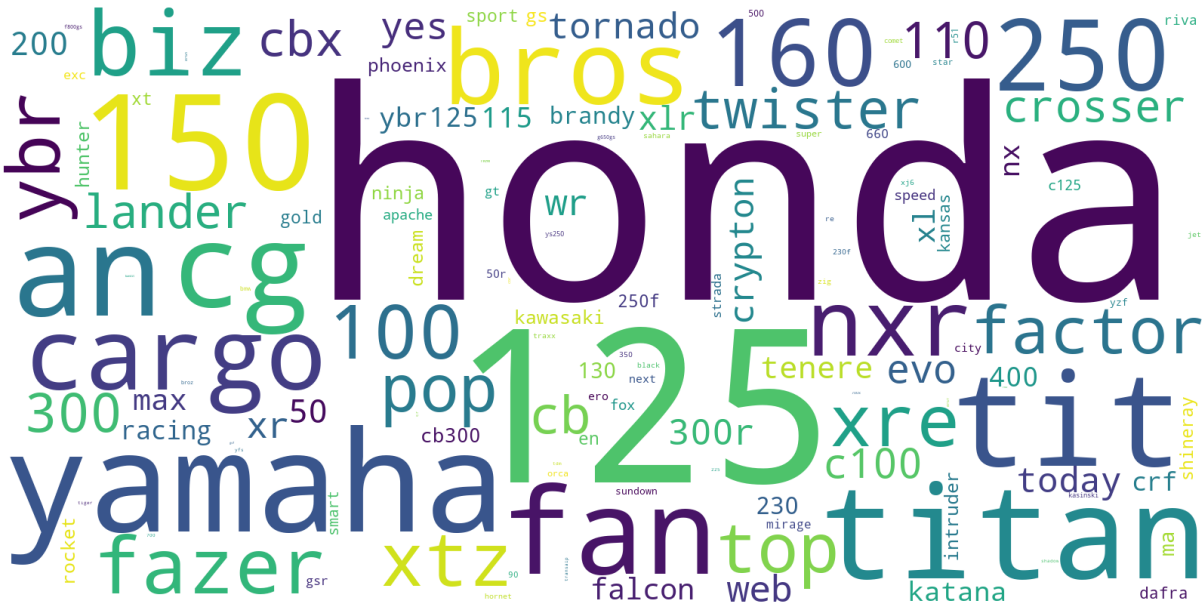


Figura 7 - *WordCloud* após limpeza

4_classificarDESCRICAO.ipynb: Diante da necessidade de se ter uma série de registros classificados para o aprendizado supervisionado houve a necessidade da classificação prévia de toda a base de dados para a aplicação do aprendizado de máquina que fará predições futuras dessa classificação.

O código contido nesse *notebook* importa a base de dados de Aplicações e através de técnicas de processamento de linguagem natural – PLN busca os termos singulares nas descrições para determinar a classificação.

Em seguida busca a comparação entre os termos remanescentes limpos na coluna Modelo e o conteúdo das possíveis Aplicações, classificando para uma aplicação quando todos os termos estão contidos dentro da descrição da Aplicação.

Caso haja mais de uma possibilidade nas análises acima, a opção foi pela criação de uma lista contendo cada uma das aplicações possíveis dentro dos critérios mencionados.

Paralelamente, utilizando técnicas de *machine learning*, especificamente os algoritmos ***Linear SVC***, ***Naive Bayes Multinomial*** e ***Regressão Logística***, treinou-se os modelos utilizando a descrição do *dataset* de Aplicações criado a partir das marcas e modelos de motos no *notebook* a seguir:

5_NLP_modeloClassificador.ipynb: As funções de aprendizado foram a base da classificação para o aprendizado inicial e, em seguida, aplicou-se à coluna Modelo de todo o *dataset*.

Analisando-se as classificações para os três modelos de algoritmos de inteligência artificial aplicados, em virtude do tempo de processamento excessivamente maior e também da enormidade de divergências, decidiu-se por abandonar o uso da modelo de regressão logística.

A partir de então se obteve três classificações para análise do *dataset* que inicialmente não era classificado, nas seguintes colunas:

- APLICACAO: classificação por análise dos termos contidos na descrição e que foram extraídos para a coluna Modelo.
- APLICACAOSVC: classificação obtida através das predições realizadas pela função criada com o algoritmo de *machine learning* Linear SVC.
- APLICACAOMNB: classificação obtida através das predições realizadas pela função criada com o algoritmo de *machine learning* Multinomial Naive Bayes.

Através da comparação entre eles decidiu-se pela seguinte lógica para escolha da classificação final a ser utilizada:

1) Se APLICACAOMNB e APLICACAOSVC forem iguais, APLICACAOFIM será APLICACAOMNB.
2) Se APLICACAOMNB e APLICACAOSVC forem diferentes:
   a) Se APLICACAO não tiver valor, então a APLICACAOFIM será uma lista com a combinação dos valores de APLICACAOSVC e APLICACAOMNB;

b) Se APLICACAO for igual a APLICACAOSVC, então a APLICACAOFIM será igual ao valor da APLICACAOSVC;
c) Se APLICACAO for igual a APLICACAOMNB, então a APLICACAOFIM será igual ao valor da APLICACAOMNB;
d) Se APLICACAO for uma lista:
    i) Se APLICACAOSVC estiver na lista, então APLICACAOFIM será APLICACAOSVC;
    ii) Se APLICACAOMNB estiver na lista, então APLICACAOFIM será APLICACAOSMNB;
    iii) Se não estiverem na lista, então APLICAFIM será a lista inicial acrescida de APLICACAOSVC e APLICACAOMNB.
e) Se APLICACAO, APLICACAOSVC e APLICACAOMNB forem diferentes, então APLICACAOFIM será uma lista contendo APLICACAO, APLICACAOSVC e APLICACAOMNB.

Fazendo uma análise do *dataframe* obtido, constatou-se que somente 163 registros permaneciam com classificação dúbia, ou seja, APLICACAOFIM continha uma lista. Desse modo, optou-se por exportar o *dataset* e tratar essas exceções no seguinte *notebook*:

6_NLP_modeloClassificador_parteManual.ipynb: Observando-se os registros há opções bem repetidas, então a solução foi criar um dicionário com o texto do modelo e sua correspondente classificação de Aplicação. Para os demais que eventualmente restarem, o código mostra as opções e pede que o usuário faça a classificação manualmente.

Ainda neste *notebook* faz-se a criação da coluna que identifica se a corrente do kit de transmissão possui ou não retentor, posto que esta questão influencia no preço do produto, fazendo-se necessário criar uma coluna do tipo *boolean* para indicar ou não a presença de retentor na corrente.

5.2. Modelos de *Machine Learning* para Classificação

Concluída a classificação de todos os itens presentes no *dataset*, finalmente foi possível dar início à criação do modelo de predição para que fosse possível classificar a aplicação para novos itens.

Para realizar essa tarefa, treinou-se e testou-se a classificação das aplicações utilizando três algoritmos de *machine learning* (Linear SVC, Naive Bayes Multinomial e Regressão Logística), para se ter certeza qual a melhor escolha para o caso em análise.

O código presente no *Jupyter Notebook* **7_treinamentoClassificador.ipynb** faz a leitura da base de dados de importações para, em seguida, separar a base de dados em treinamento e teste, treinar o modelo e avaliar os resultados.

Para cada um dos modelos de algoritmos de *machine learning* o procedimento foi basicamente o mesmo: leitura da base de dados, vetorização do *dataset*, transformação TF/IDF, separação da base em treinamento e teste (30%), treinamento do modelo, criação da função de predição dado um modelo, predição do modelo, avaliação através das métricas (acurácia, precisão, recall e score F1), construção da matriz d confusão, construção da matriz de confusão para um item considerando o paradigma de um *versus* outros (OvR, do inglês *one versus rest*) e, por fim, o teste de classificação de uma descrição da base de dados em sua respectiva aplicação, utilizando-se, *a priori*, a função de limpeza na descrição.

5.3. Módulo funcoesTCC

Durante o trabalho de limpeza dos dados, algumas funções e variáveis foram criadas e elas precisam ser reutilizadas em outros *notebooks* que compõem o trabalho.

Desse modo, foi necessária a criação e um módulo em python denominado ***funcoesTCC.py*** contendo essas funções, que podem ser facilmente importadas e utilizadas em qualquer *notebook Jupyter*, conforme podemos ver na imagem a seguir.

```
[2]: # importa funções criadas no módulo 3_classificarAplicação.ipynb (criarModelo.py)
     from funcoesTCC import *
     # Funções: criaModelo, limpaDescricao, achaPalavraChave, pegaChave, acrescentaMarca, retentorAux
     # Variávis: stopwords, palavrasChave, Marcas
     # Datasets: dftemp (Aplicações)

     ( . . . )

In [8]: descricao = df.iloc[13454]['DESCRICAO DO PRODUTO']
        # verifica a descrição do produto
        descricao

Out[8]: 'KIT DE TRANSMISSAO , MARCA RIFFEL, TITANIUM (1045) PARA MOTOCICLETAS CG 160 TITAN (16-19)/ CG 160 FA
        N (16-19) / CG 160 START (16-19)/ CG 160 CARGO (16-19), COMPOSTO DE CORRENTE 428 X SM118 + COROA S403
        07 44Z + PINHAO 25111 15Z (CERTIFICADO NR.BR31512'

In [9]: classificaAplicacaoSVC(criaModelo(descricao))

Out[9]: 'HONDA CG TIT TITAN 125 150 160'
```

Figura 8 - Módulo funcoesTCC

A utilização do módulo contendo as funções em vez de colocar as funções em todos os códigos facilita sobremaneira a manutenção do código, visto que quaisquer alterações só precisam ser feitas no módulo e já se refletem em todos os locais.

5.4. Predição do Risco dada uma Importação

O objetivo agora é recebendo os dados de uma importação contendo **kit de transmissão de motocicletas** fazer a sua classificação e em seguida apresentar uma análise de risco visual baseada na estatística existente de outras importações.

Este processamento é realizado dentro do *Jupyter notebook* intitulado **8_predição Risco.ipynb**.

É importante salientar que esta análise não substitui a acurada análise a ser realizada por um Auditor Fiscal, tendo em vista que o resultado será uma lista de parâmetros objetivos que indicarão graus de observação para o gerenciamento de risco.

Desse modo, tendo como entrada o registro de importação contendo todos os campos, extrai-se a descrição do produto, procede-se à limpeza dessa descrição.

A seguir, a função de previsão do risco determinará a classificação da aplicação e com base nos valores declarados de outras importações da mesma classificação, fará a comparação analítica com o valor da importação em análise, gerando as informações e os gráficos finais.

6. Interpretação dos Resultados

Desde a análise e exploração dos dados, observou-se a importância crucial da classificação das aplicações para que o *dataset* pudesse realmente ser utilizado com seu propósito de fazer uma previsão de risco na importação de kits de transmissão de motocicletas.

Como primeira grande tarefa abordada, a classificação exigiu esforço grande, posto que a base de dados original não contém qualquer tipo de classificação que pudesse ser utilizada.

Assim, foi necessária a obtenção de uma lista de aplicações para servisse ao classificador como opções de resultado, mas para isso foi preciso antes uma limpeza grande na descrição para remover todos os termos desnecessários.

A classificação realizada com o estudo dos termos importantes da descrição, adicionando-se as marcas para modelos conhecidos de motocicletas permitiu que fosse criada uma coluna identificadora da aplicação através da aplicação dos modelos de *machine learning*.

Com a base de dados classificada foi necessário partir para a criação de um modelo genérico de classificador, que pudesse pegar qualquer descrição do produto apresentada pelo contribuinte e este pudesse classificar corretamente a aplicação.

Nessa fase foram testados três modelos de classificador – Linear SVC, Multinomial Naive Bayes e Regressão Logística. Todos performaram muito bem, com alta acurácia e precisão, conforme se pode observar na tabela abaixo:

Modelo	Acurácia	Precisão	Recall	F1_Score
Linear SVC	1.00	1.00	1.00	1.00
Multinomial Naive Bayes	0.93	0.93	0.93	0.93
Regressão Logística	0.99	0.99	0.99	0.99

Tabela 7 - Métricas dos Modelos dos Classificadores

Importante observar que todos os modelos performaram muito bem e cumprem muito bem o papel de determinar a aplicação dada uma descrição. Pode-se confirmar esta afirmação observando-se, na figura a seguir, a matriz de confusão gerada pelos modelos com a mesma aparência.

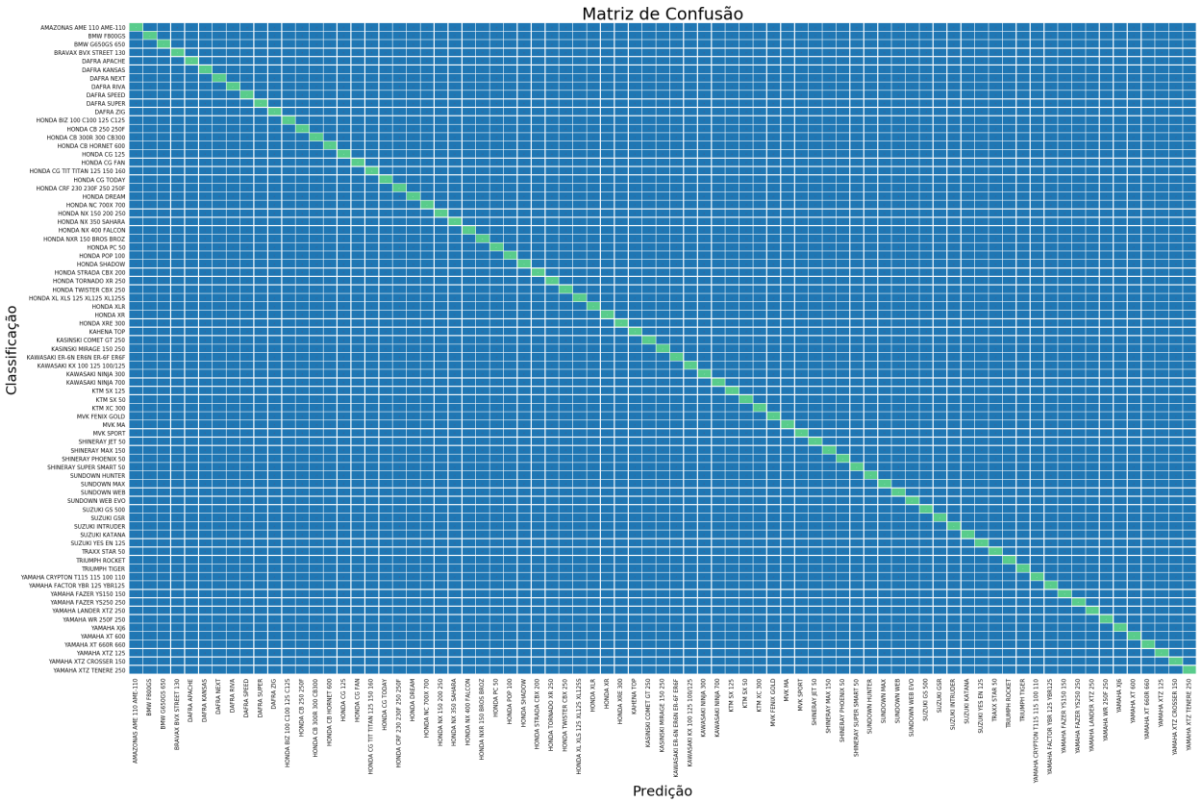


Figura 9 - Matriz de Confusão do Modelo

Apesar do grande número de classes, pode-se observar a linha formada pelos acertos na diagonal da imagem, demonstrando o alto índice de acerto do modelo.

Para melhorar a visão, esboçaremos a matriz de confusão para o item "**HONDA BIZ 100 C100 125 C125**" versus o resto (OvR), que permitirá ver a classificação como se uma classificação binária do tipo ou é a classificação demonstrada ou é o resto.

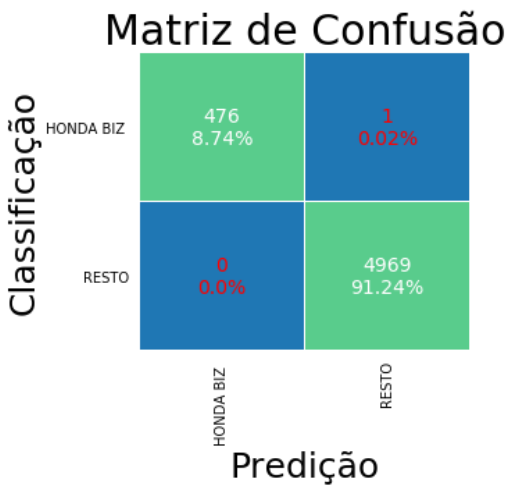


Figura 10 - Matriz de Confusão (HONDA BIZ 100 C100 125 C125)

Reafirma-se que todos os modelos tiveram performance em acertos semelhantes e qualquer um dos escolhidos desempenharia bem o papel de classificar as aplicações.

A **escolha** tomou por base a performance em tempo de execução, sendo o modelo **Linear SVC** o mais rápido dos três analisados, além de ter obtido as melhores métricas na classificação.

Finalizada a tarefa de classificar todos os itens, chegou a hora de utilizar essa classificação para dada uma importação, classificar a aplicação do item pela sua descrição de produto.

Com o item classificado foi possível separar a base somente dos itens com a mesma aplicação e filtrar para itens cuja corrente fosse com ou sem retentor, de acordo com o item em análise.

Uma observação muito importante a ser feita, foi quanto aos valores *outliers*, isto é, valores atípicos, posto que alguns itens eram discrepantes, fosse por ser uma sub ou sobrevaloração do item ou fosse porque alguns importadores insistem em, erroneamente, declarar caixas com vários kits em vez de unidades.

A identificação de *outliers*, distantes dos demais pontos da série, pode ser feita utilizando o intervalo interquartílico (IQR). Entende-se como outlier, valores menores que Q1-1,5*IQR ou valores maiores que Q3+1,5*IQR.

Estes valores também são representados no *boxplot*, mas como pontos acima ou abaixo da linha conectada à caixa. Vejamos um desenho de um *boxplot* com todos estes conceitos:

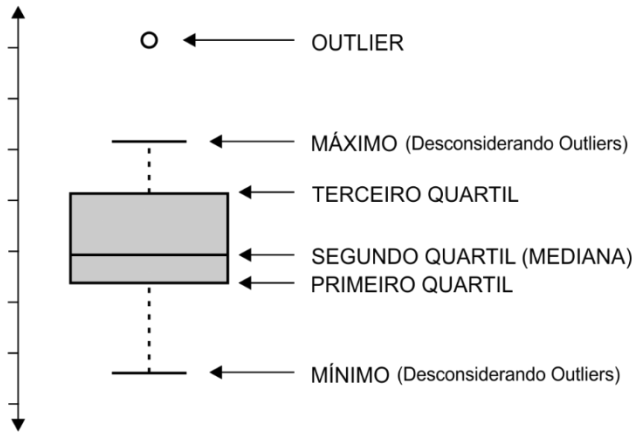


Figura 11 - BarPlot mostrando Outliers

Apesar da bibliografia indicar o uso de uma vez e meia o intervalo entre o primeiro e o terceiro quartil para definir os *outliers*, optou-se nesse trabalho, por uma questão de segurança, em utilizar o valor de três vezes.

Determinados os valores válidos que podem ser utilizados para uma aplicação selecionada, a base do gerenciamento de risco fica definida em como se comporta o valor declarado dentro do grupo de percentis da base histórica.

O que se quer observar, em outras palavras, é em que percentil se encaixa o valor declarado.

Para isso, cogitou-se várias possibilidades de apresentação desse resultado, e ao final uma apresentação gráfica foi escolhida como a melhor solução, tendo em vista que somente a apresentação dos valores ainda careceria de muita observação e tempo na análise dos números.

Abaixo vislumbra-se na figura como aparece um histograma da frequência dos valores declarados para uma aplicação classificada.

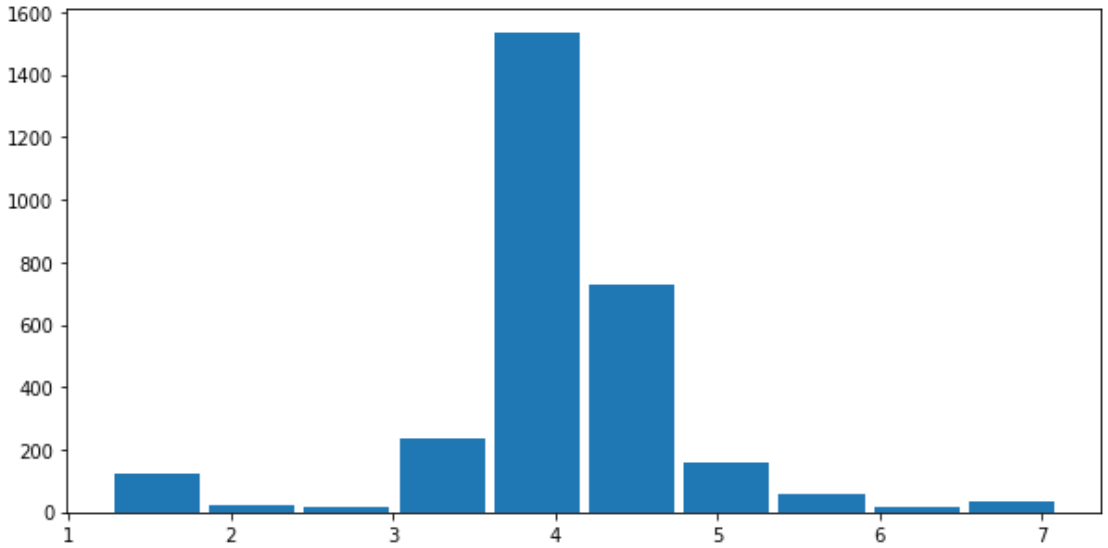


Figura 12 - Histograma por decil de uma aplicação aleatória

7. Apresentação dos Resultados

Este é o resultado:

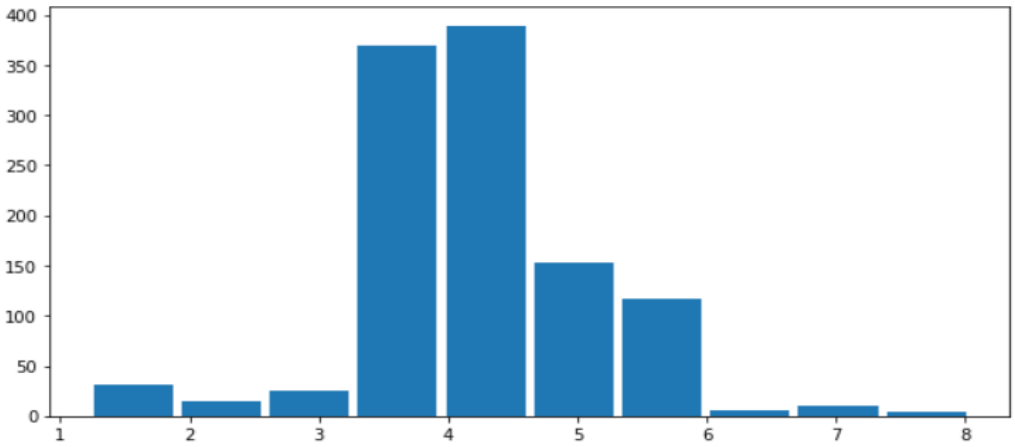
Dados da Declaração de Importação
Descrição: TM10205 - KIT DE TRANSMISSAO COMPOSTO DE COROA, PINHAO E CORRENT
E PARA MOTOCICLETA - UTILIZADAS NAS MOTOCICLETAS YBR125 FACTOR 03/08 - (45T
/ 14T 428H X 118L), SUA FUNCAO E TRANSMITIR O MOVIMENTO DA CAIXA DE CAMBIO
ATE A RODA TRASEIRA, TAMBEM E RESPO
Origem: CHINA, REPUBLICA POP
Retentor: sem retentor
Aplicação: YAMAHA FACTOR YBR 125 YBR125
Valor DI: USD 3.78
Tabela de Referência ABIMOTO
Valor: USD 4.20

Estatísticas:

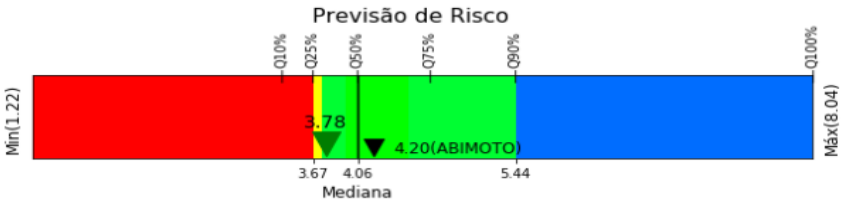
qtd de registros:	1123.000000
média simples:	4.152855
desvio padrão:	0.908013
valor mínimo:	1.216719
percentil 25%:	3.666567
percentil 50%:	4.060000
percentil 75%:	4.686000
valor máximo:	8.036000

Percentis:
'10=1.22', '20=3.40', '25=3.67', '30=3.58', '40=3.75', '50=3.95'
'60=4.06', '70=4.13', '75=4.69', '80=4.50', '90=4.83', '100=5.44'

Histograma por Decil:



Previsão de Risco:



ATENÇÃO: Valor declarado de USD3.78 menor que referência mínima ABIMOTO: USD4.20.

Figura 13 - Relatório de Risco

Para entendermos o resultado apresentado é preciso retomar o objetivo deste trabalho, qual seja, apresentar visualmente uma análise de risco nas importações de kits de transmissão de motocicletas.

A entrada para esta análise é a adição de uma declaração de importação contendo um item com kit de transmissão. Neste registro utilizaremos a descrição do produto e o valor fornecidos pelo contribuinte na declaração e apresentaremos os itens a seguir para a análise do Auditor Fiscal quanto ao risco de fraude de valor nesta importação.

```
Dados da Declaração de Importação
Descrição: TM10205 - KIT DE TRANSMISSAO COMPOSTO DE COROA, PINHAO E CORRENT
E PARA MOTOCICLETA - UTILIZADAS NAS MOTOCICLETAS YBR125 FACTOR 03/08 - (45T
/ 14T 428H X 118L), SUA FUNCAO E TRANSMITIR O MOVIMENTO DA CAIXA DE CAMBIO
ATE A RODA TRASEIRA, TAMBEM E RESPO
Origem:    CHINA, REPUBLICA POP
Retentor:  sem retentor
Aplicação: YAMAHA FACTOR YBR 125 YBR125
Valor DI:  USD 3.78
Tabela de Referência ABIMOTO
Valor:      USD 4.20
```

Figura 14 - Relatório de Risco: Dados da Declaração de Importação

Na imagem acima vemos o cabeçalho do Relatório de ricos, contendo os dados relevantes da declaração de importação, a classificação da aplicação do item e a informação se a corrente do kit possui ou não retentor.

```
Tabela de Referência ABIMOTO
Valor:      USD 4.20
```

Figura 15 - Relatório de Risco: Valor na Tabela ABIMOTO (se existir)

Para alguns itens, a ABIMOTO fornece em sua tabela de referência de valores mínimos esperados para importação de peças de motocicleta, valores para alguns kits de transmissão. Caso o item em análise seja um dos constantes da tabela, ele será apresentado.

A seguir o relatório apresenta as estatísticas para a base de dados histórica de importações para a aplicação em análise, conforme se observa na imagem abaixo:

```
Estatísticas:
qtd de registros:    1123.000000
média simples:          4.152855
desvio padrão:          0.908013
valor mínimo:           1.216719
percentil 25%:          3.666567
percentil 50%:          4.060000
percentil 75%:          4.686000
valor máximo:           8.036000
```

Figura 16 - Relatório de Risco: Estatísticas:

Também é fornecido ao Auditor Fiscal para análise os valores para todos os decis e ainda para o primeiro e terceiro quartis.

```
Percentis:
  '10=1.22', '20=3.40', '25=3.67', '30=3.58', '40=3.75', '50=3.95'
  '60=4.06', '70=4.13', '75=4.69', '80=4.50', '90=4.83', '100=5.44'
```

Figura 17 - Relatório de Risco: Percentis

O próximo item constante do relatório é o histograma por decil.

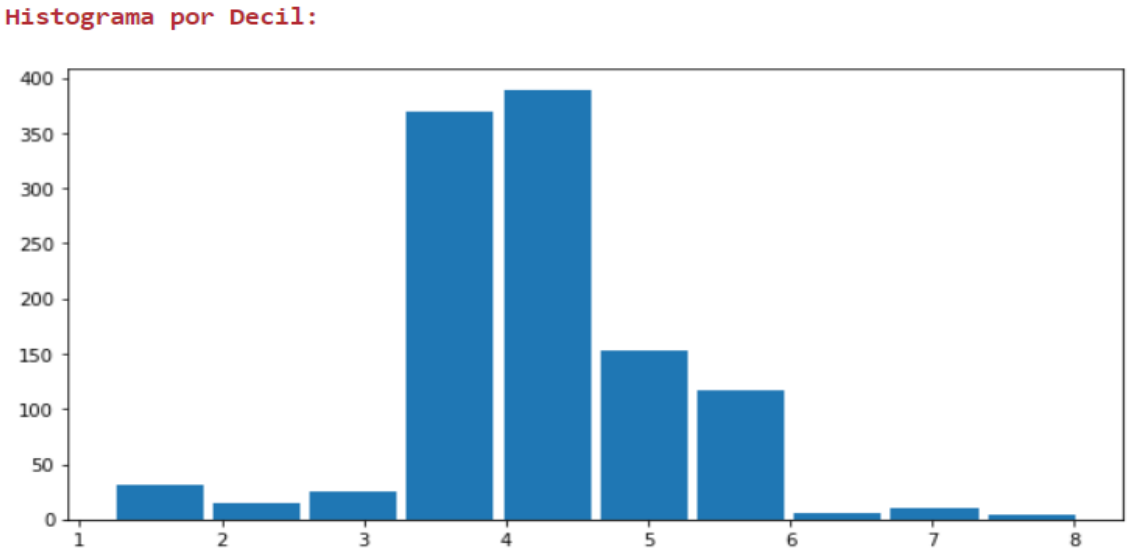


Figura 18 - Relatório de Risco: Histograma por decil

Por fim, vem o item mais importante do relatório, a imagem que posiciona o valor declarado indicando a sujeição de risco dentro dos valores históricos para a aplicação em análise.

Observa-se que a barra varia de acordo com os valores encontrados no dataset filtrado para a aplicação pretendida, contendo somente os valores que correspondam ao valor de retentor, variando do mínimo ao máximo valor encontrado, já excluídos todos os valores *outliers*.

As cores são bons indicativos do risco apresentado em relação ao valor da operação em análise. Na imagem podemos observar que o primeiro quartil é representado em vermelho; deste até o quarto decil é apresentado em amarelo; a partir daí é utilizado o verde até o nono decil; e daí em diante apresentamos em azul os valores considerados altos, pois, englobando somente os dez porcento mais altos dentre os valores históricos declarados.

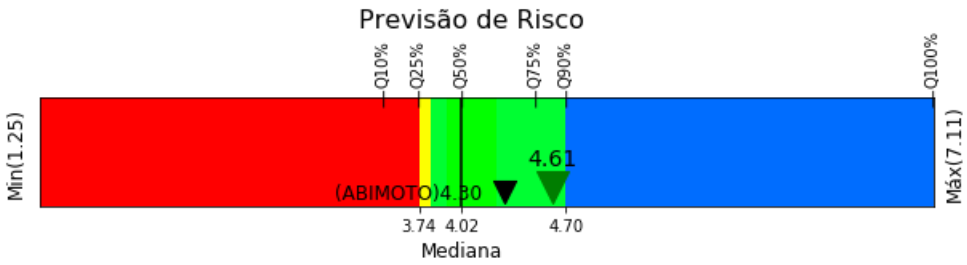


Figura 19 - Relatório de Risco: Previsão do Risco

Em virtude da importância dessa imagem e da quantidade de informações apresentadas, faz-se necessária uma análise mais minuciosa de cada item de informação fornecida

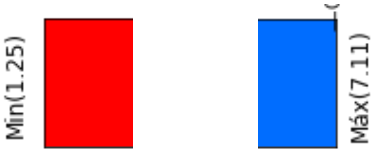


Figura 20 - Previsão de Risco: Valores mínimo e máximo

O gráfico da previsão de risco indica nas extremidades em posição vertical os valores mínimo e máximo, fora dos quais estão os valores considerados como *outliers*.

Na figura a seguir destaca-se a presença dos principais percentis graficamente indicados, com a presença em destaque na parte inferior da indicação da mediana com uma linha vertical cortando o gráfico.

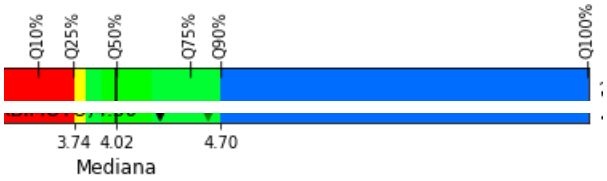


Figura 21 - Relatório de Risco: Indicação dos principais percentis com valores

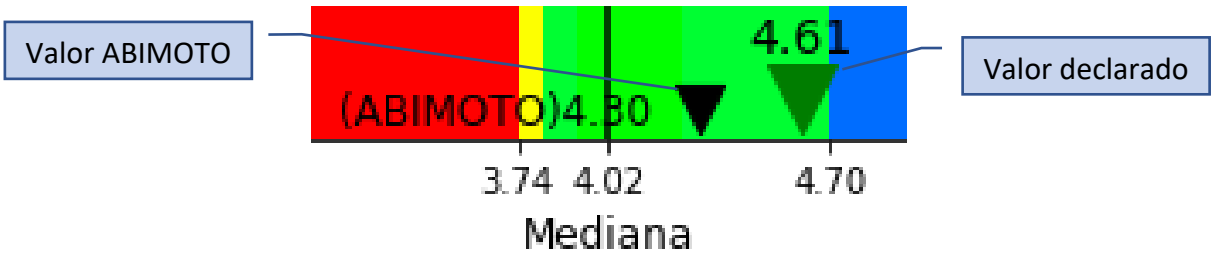


Figura 22 - Relatório de Risco: Indicação colorida para o valor declarado e ABIMOTO

As informações passam por alguns detalhes, como a seta indicativa que mostra o valor onde o preço do item está no conjunto de dados, com a cor da seta indicando o grau de risco, podendo assumir tons de vermelho, amarelo, verde ou azul, este último quando o valor declarado está no último decil.

Ainda é preciso ressaltar que há dois alertas que só aparecem caso sejam necessários.

O primeiro se refere a casos nos quais o valor declarado é abaixo do indicado na tabela ABIMOTO como o valor mínimo para importação daquele item, conforme se pode ver na figura a seguir.

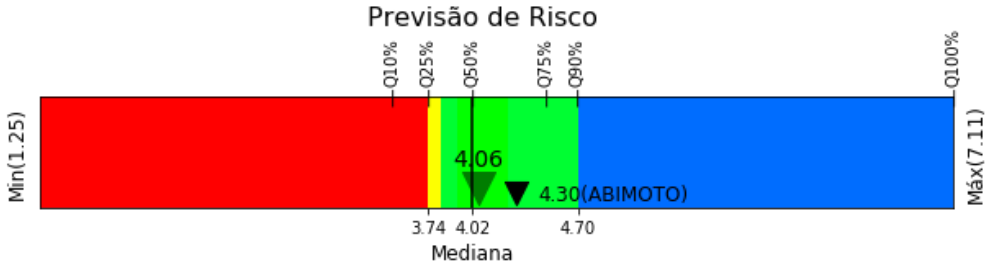


Figura 23 - Relatório de Risco: Alerta de valor abaixo do valor ABIMOTO

O segundo possível alerta se refere a quando o valor declarado se enquadra na definição de um *outlier* para o conjunto de dados. Observe na imagem abaixo que o valor declarado é maior que o limite superior, caracterizando-se como um valor discrepante da base de dados existentes.

Figura 24 - Relatório de Risco: Alerta de valor *outlier*

8. Links

Link para o vídeo: youtube.com/...

Link apresentação:

Link para o repositório: https://github.com/kaduleite/TCC_PUC_MG_2021

REFERÊNCIAS

ABIMOTO – Associação Brasileira dos Importadores de Autopeças. **Valoração Aduaneira: Partes e Peças para Motocicletas**. 13ª versão, 2021.

DUARTE, Luiz. **Base de dados com todas as marcas e modelos de veículos**. https://www.luiztools.com.br/post/base-de-dados-com-todas-as-marcas-e-modelos-de-veiculos/. Acesso em 13/12/2021.

ESTATISITE. **Localizando Outliers Através do Intervalo Interquartil**. Disponível em https://estatsite.com.br/2018/12/01/localizando-outliers-atraves-do-intervalo-interquartil-boxplot-codigo-sas/. Acesso em 10/01/2021

Governo Federal, Portal Único Siscomex. **Gerador de Tabela NCM**. Disponível em https://portalunico.siscomex.gov.br/classif/#/nomenclatura/tabela?perfil=publico. Acesso em 09/11/2021.

Governo Federal, Receita Federal do Brasil: **Sistema de Apoio Siscori**. Disponível em https://siscori.receita.fazenda.gov.br/. Acesso em julho a novembro de 2021.

MCKINNEY, William Wesley. **Python para análise de Dados.** São Paulo: Novatec Editora, 2018

Trademap. **Estatísticas de Comércio Internacional para Desenvolvimento Econômico e Comercial**. Disponível em https://www.trademap.org/. Acesso em 09/11/2021.

Trading Economics. **Estatísticas de Exportações da China**. Disponível em https://pt.tradingeconomics.com/china/exports. Acesso em 09/11/2021.

ANEXOS

Lista de Figuras

Lista de Tabelas

Notebooks Jupyter

1a_trataCSVsSiscori.ipynb

1b_trataABIMOTO13.ipynb

2_geraWordclouds.ipynb

3_classificarAplicação.ipynb

4_classificarDESCRICAO.ipynb

5_NLP_modeloClassificador.ipynb

6_NLP_modeloClassificador_parteManual.ipynb

7_treinamentoClassificador.ipynb

8_predição Risco.ipynb