PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA

Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data

Daniel Nunes Almeida e Silva

Análise dos acidentes de trânsito ocorridos nos últimos 10 anos nas rodovias federais brasileiras.

Belo Horizonte

2020

Daniel Nunes Almeida e Silva

Análise dos acidentes de trânsito ocorridos nos últimos 10 anos nas rodovias federais brasileiras.

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.

Belo Horizonte
2020

SUMÁRIO

1. Introdução

1.1. Contextualização

Os acidentes de trânsito são uma preocupação constante ao redor do mundo, a cada ano causam a morte de mais de 1,35 milhão de pessoas em todo o mundo, além de 50 milhões de feridos - e, segundo a Organização Mundial de Saúde (OMS), os acidentes são a principal causa de morte entre crianças e jovens com idade entre 5 e 29 anos.

Mais da metade de todas as mortes no trânsito ocorre entre usuários vulneráveis das vias: pedestres, ciclistas e motociclistas. Ainda, segundo dados da OMS, o Brasil é o quarto país com mais mortes no trânsito, atrás de China, Rússia, Índia e Estados Unidos.

Na abertura da última Conferência Global de Alto Nível da ONU sobre Segurança no Trânsito, a Declaração de Estocolmo afirmou que a previsão de um total de até 500 milhões de mortes em todo o mundo entre 2020 e 2030 “é uma crise evitável, que para ser evitada vai exigir um maior e mais significativo empenho político, liderança e ações abrangentes em todos os níveis na próxima década”.

Dessa forma, este trabalho tem como intuito fazer uma análise sobre os acidentes ocorridos nos últimos 10 anos (2011 a 2020) nas rodovias federais brasileiras, analisando as condições em que ocorreram esses acidentes.

1.2. O problema proposto

A Organização Pan Americana de Saúde (OPAS) releva que 93% das mortes no trânsito ocorrem em países de baixa e média renda, embora estes concentrem aproximadamente 60% dos veículos do mundo e custam à maioria dos países 3% de seu produto interno.

Dessa forma, por ser um tema bastante atual e de grande importância para os países envolvidos, principalmente o Brasil, que é o quarto país com mais mortes no trânsito, irei fazer uma análise exploratória dos dados referentes aos acidentes ocorridos nas rodovias federais brasileiras nos últimos 10 anos, verificando os dados relativos às ocorrências registradas e às pessoas envolvidas nos acidentes. Esses dados estão disponíveis no site da PRF, agrupados por **ocorrência** e por **pessoa.**

Com essa análise, pretendo verificar as condições em que ocorreram os acidentes nas estradas brasileiras no período de janeiro de 2011 a setembro de

2020, respondendo perguntas como: número de acidentes por ano (de 2011 a 2020), quais são os meses com maiores ocorrências de acidentes, as condições climáticas do momento, os horários em que ocorreram com maiores frequências, condições da estrada, idades dos acidentados, veículos envolvidos, etc...

Na próxima etapa será construído um modelo capaz de classificar se em um determinado acidente haverá vítimas fatais, pessoas feridas ou ilesas.

2. Coleta de Dados

A fonte de dados utilizada foi adquirida do site da Polícia Rodoviária Federal, no link: https://portal.prf.gov.br/dados-abertos-acidentes. Em novembro de 2020 foram baixados os dados referentes aos acidentes agrupados por ocorrência e por pessoa, dos anos de 2011 a 2020. Dessa forma, o período utilizado é relativo a 01/2011 a 09/2020 (último mês disponível no *dataset* de 2020).

Os dados disponíveis no site da PRF possuem algumas divergências em seu formato de acordo com os períodos utilizados. Eles estão divididos da seguinte forma, conforme justificativa abaixo:

- acidentes agrupados por **ocorrência** (até 2016)
- acidentes agrupados por **ocorrência** (2017 em diante)
- acidentes agrupados por **pessoa** (até 2016)
- acidentes agrupados por **pessoa** (2017 em diante)

"O registro de acidentes é realizado através do sistema BAT, que coleta informações referentes aos envolvidos (identificação, estado físico, se era passageiro, condutor, etc.), ao local, aos veículos, à dinâmica do acidente, etc. Os dados disponíveis têm origem nos sistemas BR-Brasil e BAT. O sistema BR-Brasil foi utilizado em nível nacional entre 2007 e 2016. O sistema BAT é utilizado desde 2017.

A descrição dos dados de ambos os *datasets* estão disponibilizadas no site da PRF. Abaixo são mostradas as descrições das variáveis dos *datasets* relativos aos acidentes agrupados por **ocorrência:**

Nome da coluna/campo	**Descrição**	**Tipo (identificação automática)**
id	Variável com valores numéricos, representando o identificador do acidente	float64
data_inversa	Data da ocorrência no formato dd/mm/aaaa	Datetime64[ns] *importado dessa forma

dia_semana	Dia da semana da ocorrência. Ex.: Segunda, Terça, etc.	object
horário	Horário da ocorrência no formato hh:mm:ss.	Datetime64[ns] *importado dessa forma
uf	Unidade da Federação. Ex.: MG, PE, DF, etc.	object
br	Variável com valores numéricos, representando o identificador da BR do acidente.	Object
km	Identificação do quilômetro onde ocorreu o acidente, com valor mínimo de 0,1 km e com a casa decimal separada por ponto.	object
município	Nome do município de ocorrência do acidente	object
causa_acidente	Identificação da causa principal do acidente. Neste conjunto de dados são excluídos os acidentes com a variável causa principal igual a “Não”.	object
tipo_acidente	Identificação do tipo de acidente. Ex.: Colisão frontal, Saída de pista, etc. Neste conjunto de dados são excluídos os tipos de acidentes com ordem maior ou igual a dois. A ordem do acidente demonstra a sequência cronológica dos tipos presentes na mesma ocorrência	object
classificação_acidente	Classificação quanto à gravidade do acidente: Sem Vítimas, Com Vítimas Feridas, Com Vítimas Fatais e Ignorado.	object
fase_dia	Fase do dia no momento do acidente. Ex. Amanhecer, Pleno dia, etc	object
sentido_via	Sentido da via considerando o ponto de colisão: Crescente e decrescente.	object
condicao_metereologica	Condição meteorológica no momento do acidente: Céu claro, chuva, vento, etc.	object
tipo_pista	Tipo da pista considerando a quantidade de faixas: Dupla, simples ou múltipla.	object
tracado_via	Descrição do traçado da via	object
uso_solo	Descrição sobre as características do local do acidente: Urbano=Sim; Rural=Não.	object

ano	Ano do registro	float64
pessoas	Total de pessoas envolvidas na ocorrência.	int64
mortos	Total de pessoas mortas envolvidas na ocorrência.	int64
feridos_leves	Total de pessoas com ferimentos leves envolvidas na ocorrência.	int64
feridos_graves	Total de pessoas com ferimentos graves envolvidas na ocorrência.	int64
ilesos	Total de pessoas ilesas envolvidas na ocorrência.	int64
ignorados	Total de pessoas envolvidas na ocorrência e que não se soube o estado físico.	int64
feridos	Total de pessoas feridas envolvidas na ocorrência (é a soma dos feridos leves com os graves)	int64
veículos	Total de veículos envolvidos na ocorrência.	int64
latitude	Latitude do local do acidente em formato geodésico decimal.	float64
longitude	Longitude do local do acidente em formato geodésico decimal.	float64
regional	Dados internos da PRF	object
delegacia	Dados internos da PRF	object
uop	Dados internos da PRF	object

Observamos dessa análise inicial que a variável **ano** só estava prevista nos *datasets* referentes aos anos de 2011 a 2016, enquanto as variáveis **latitude**, **longitude**, **regional**, **delegacia** e **uop** só estavam previstas no período de 2017 a 2020.

Com relação às variáveis dos *datasets* relativos aos acidentes agrupados por **pessoa**, além de várias colunas similares, por exemplo, id, data_inversa, dia_semana, horario, etc, já descritas anteriormente, também constam as seguintes colunas:

Nome da coluna/campo	Descrição	Tipo (identificação automática)
pesid	Variável com valores numéricos, representando o identificador da pessoa envolvida	float64
id_veiculo	Variável com valores numéricos, representando o identificador do veículo envolvido.	object
tipo_veiculo	Tipo do veículo conforme Art. 96 do Código de Trânsito Brasileiro. Ex.: Automóvel, Caminhão, Motocicleta, etc.	object
marca	Descrição da marca do veículo.	Não foi importado
ano_fabricacao_veiculo	Ano de fabricação do veículo, formato aaaa	object
tipo_envolvido	Tipo de envolvido no acidente conforme sua participação no evento. Ex.: condutor, passageiro, pedestre, etc	object
estado_fisico	Condição do envolvido conforme a gravidade das lesões. Ex.: morto, ferido leve, etc.	object
idade	Idade do envolvido. O código “-1” indica que não foi possível coletar tal informação.	float64
sexo	Sexo do envolvido. O valor “inválido” indica que não foi possível coletar tal informação.	object

3. Processamento/Tratamento de Dados

Para processamento, tratamento e análise de dados utilizamos a linguagem Python, via Jupyter Notebooks para facilitar o acompanhamento e reprodução do que foi realizado.

3.1 Importação dos dados

Os dados relativos aos acidentes agrupados por ocorrência foram importados todos de uma mesma pasta, já que as únicas divergências entre eles são algumas colunas, o que não impede o tratamento em conjunto. Após a importação das 10 planilhas, os arquivos foram concatenados em um único dataframe, gerando um *dataset* com 1.223.039 linhas e 30 colunas.

Já os dados relativos aos acidentes agrupados por pessoa, os separadores utilizados nos arquivos csv são diferentes. Até o ano de 2015, o separador utilizado foi a vírgula (,), enquanto nos anos seguintes o separador utilizado foi o ponto e vírgula (;). Dessa forma, os arquivos foram distribuídos em pastas diferentes, para que a importação fosse realizada de uma única vez.

Como muitas das variáveis são similares, nesse caso só foram importadas 8 colunas: **id**, **pesid**, **tipo_veiculo**, **ano_fabricacao_veiculo**, **tipo_envolvido**, **sexo**, **idade** e **estado_fisico.** Essas variáveis enriquecerão o *dataset* anterior com informações das vítimas e dos veículos acidentados. Após a importação das 10 planilhas, os arquivos foram concatenados em um único dataframe, gerando um *dataset* com 2.704.697 linhas e 8 colunas.

Em ambas as importações (ocorrências e pessoas) foram utilizados o módulo interno **glob**, conforme abaixo (os comandos são similares, mudando apenas a pasta dos *datasets*):

```
# Importando todos os dados relativos a acidentes agrupados por ocorrência de 2012 a 2020

path = r'C:\TCC\DATASETS'

all_files = glob.glob(path + "/*.csv")

li = []

for filename in all_files:

    df = pd.read_csv(filename, sep = ';', decimal=",", encoding = "ISO-8859-1",

                index_col=False, parse_dates=[["data_inversa", "horario"]])

    nRow, nCol = df.shape

    li.append(df)

df = pd.concat(li, axis=0, ignore_index=True)

nRow, nCol = df.shape
```

Dessa forma, foram gerados 02 dataframes, **df** (acidentes agrupados por ocorrência) e **dfpessoas** (acidentes agrupados por pessoas). Esses dataframes foram unidos com base na variável **id,** conforme abaixo:

```
#Unindo os dois dataframes
df = df.merge(dfpessoas, how = "left", on = "id")
```

Após essa união, foram excluídos do dataframe gerado as seguintes colunas: **ano**, **latitude**, **longitude**, **regional**, **delegacia**, **uop**, **km**, **municipio**, **ignorados**, **ilesos**, **mortos**, **feridos_leves**, **feridos_graves**, **feridos, classificação_acidente**. As informações apresentadas nessas variáveis não são úteis para a análise proposta. Algumas, como ano, latitude, longitude, regional, delegacia e uop, só estavam previstas em anos específicos (2011 a 2016), o que acabaria trazendo uma grande quantidade de dados ausentes. Além disso, as informações dessas variáveis, com exceção da coluna **ano** são irrelevantes para o propósito desse trabalho.

Com a junção do dataframe e com a exclusão dessas variáveis, o novo dataframe foi gerado com 2.704.815 linhas e 22 colunas.

3.2 Dados duplicados

Como no dataframe gerado foi incluída a variável **pesid**, representando o identificador da pessoa envolvida no acidente, irão existir várias linhas com o mesmo **id**, mas como cada linha será relativa a uma pessoa acidentada, não poderemos analisar dados duplicados com base apenas nessa variável. Dessa forma, foram analisados dados iguais em todas as variáveis – 79 registros duplicados, que foram removidos com o comando df.drop_duplicates(), deixando o arquivo com 2.704.736 linhas.

3.3 Dados nulos

As variáveis abaixo possuem dados vazios:

```
br                          1087
causa_acidente                 0
classificacao_acidente         4
condicao_metereologica         7
data_inversa_horario           0
dia_semana                     0
fase_dia                       1
id                             0
pessoas                        0
sentido_via                    0
tipo_acidente                  0
tipo_pista                     0
tracado_via                    0
uf                             0
uso_solo                       0
veiculos                       0
pesid                          4
tipo_veiculo                4002
ano_fabricacao_veiculo     51777
tipo_envolvido                 0
estado_fisico                  3
idade                      98651
sexo                         574
dtype: int64
```

Como a variável **pesid** é a que diferencia as pessoas envolvidas nos acidentes, foram analisadas as linhas em que ela estava vazia. Dessa análise, verificou-se que todas as variáveis importantes também estavam ausentes, por isso, decidi pela remoção dessas 4 linhas. Após isso, também foram removidas as variáveis **pesid** e **id**, que já não seriam mais importantes para esse projeto. Isso resultou em um dataframe com 2.704.732 linhas e 20 colunas. Os outros registros vazios serão detalhados no item abaixo.

3.4 Tratamentos dos dados

Aqui irei detalhar o tratamento realizado nos dados

Datas

As colunas data_inversa e horario foram importadas com o uso do comando parse_dates, da seguinte forma: parse_dates = [["data_inversa", "horario"]]. Isso gerou uma única variável no formato data: data_inversa_horario.

Também foram criadas 03 novas variáveis no dataframe, **ano**, **mes** e **hora**, levando-se em conta as informações disponíveis na variável data_inversa_horario.

Essa extração é fundamental para a análise dos dados, conforme informações apresentadas no item 4. Análise e Exploração dos Dados.

Idade

Os 98.647 registros vazios foram substituídos por -1, já que, conforme documentação da PRF, o código -1 indica que não foi possível coletar tal informação. Também foram encontrados registros com idade acima de 110, esses também foram substituídos pelo valor -1. Ainda, foi gerado um novo atributo (**faixa_etaria**) com base na idade, indicando as seguintes faixas etárias da vítima: ignorada (-1), 0-4 anos, 5-14 anos, 15-24 anos, 25-34 anos, 35-44 anos, 45-54 anos, 55-64 anos, 65-74 anos e +75 anos. Após a criação desse atributo, a variável **idade** foi excluída**.**

dia_semana

Todos os registros foram normalizados para letra minúscula e sem a palavra “feira”.

estado_fisico

Todos os registros foram normalizados para letra minúscula, limpeza de espaços em brancos a mais, substituição dos 03 registros vazios por “ileso”, além da normalização dos registros com “lesões graves” para “ferido grave”, “lesões leves” para “ferido leve”, “morto” para “óbito”, “não informado” para “ileso” e “(null)” para “ileso”.

Conforme o conteúdo do dicionário de variáveis da PRF, essa coluna é registrada da seguinte forma: “Condição do envolvido conforme a gravidade das lesões. Ex.: morto, ferido leve, etc.”. Dessa forma, podemos supor que os 46.781 registros, conforme abaixo, se referem ao estado físico “ileso”, já que as vítimas não foram registradas pela PRF.

- “ignorado”: 10883
- “não informado”: 35847
- “(null)”: 48; e
- vazios: 3

sexo

Todos os registros foram normalizados para letra minúscula, substituição dos registros com “m” para “masculino”, “f” para “feminino”, “i” para “inválido”, “não informado” para “inválido”, assim como a substituição dos 574 registros vazios para “inválido”.

Analisando esses dados, verificamos que os registros “ignorado” só passaram a existir a partir do ano de 2017 e uma quantidade bem pequena. Dessa forma, também substitui os registros “ignorado” por “inválido”, em respeito ao conteúdo do dicionário de variáveis da PRF (“O valor “inválido” indica que não foi possível coletar tal informação”).

ano	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
sexo										
feminino	65919.0	64666.0	68769.0	63788.0	51727.0	45842.0	44057.0	37212.0	37155.0	21587.0
ignorado	NaN	NaN	NaN	NaN	NaN	NaN	1126.0	1144.0	1104.0	484.0
inválido	15160.0	14284.0	15177.0	14100.0	10703.0	8850.0	12064.0	9146.0	9232.0	7080.0
masculino	331205.0	317962.0	321874.0	290642.0	206622.0	161571.0	147169.0	117299.0	114781.0	75231.0

tipo_veiculo

Todos os registros foram normalizados para letra minúscula, substituição dos 4002 registros vazios para “não informado”, assim como a normalização de vários registros, por exemplo: “(null)”, “micro-ônibus”, “charrete”, “motocicletas”, “carro-de-mão”, tratores, entre outros.

causa_acidente

Todos os registros foram normalizados para letra minúscula, assim como a normalização de vários registros, por exemplo: “(null)”, “condutor-dormindo”, “defeito mecânico no veículo”, entre outros.

fase_dia

Todos os registros foram normalizados para letra minúscula, substituição do registro vazio para “ignorado” e substituição do registro “(null)” para “ignorado”.

tipo_pista

Substituição dos registros “(null)” para “ignorado”.

tracado_via

Substituição dos registros “(null)” para “ignorado”.

br

Substituição dos 1087 registros vazios para 0, dos registros “(null)” para “0” e depois foi alterado o tipo de object para uint64, uniformizando registros similares.

uso_solo

Substituição dos registros “Sim” para “Urbano”, “Não” para “Rural”, de acordo com as instruções da própria PRF, além da substituição de “(null)” para “ignorado”.

condicao_metereologica

Todos os registros foram normalizados para letra minúscula, substituição do registro vazio para “ignorada”, normalização dos registros “ignorado” e “(null)” para “ignorada”, assim como a substituição de “céu claro” para “ceu claro”.

ano_fabricacao_veiculo

Substituição dos 51777 registros vazios, dos registros em branco e “(null)” para -1. Após isso foi alterado o tipo de object para uint64, uniformizando registros similares. Também foram encontrados anos acima de 2020, nitidamente registros inconsistentes, sendo substituídos pelo registro -1.

tipo_envolvido

Como só existia um registro do tipo “Vítima”, “Proprietário de Carga” e “Proprietário de CNH”, e todos eram condutores, esses registros foram alterados para “Condutor”. Além desses, existiam 05 vítimas com o tipo_envolvido “Ciclista”. Analisei esses casos e verifiquei que as informações importantes para esse projeto (estado_fisico, sexo, idade, etc...) tinham sido registradas da mesma forma – “Ignoradas”. Dessa forma, exclui essas linhas do dataframe.

Com a finalização desses tratamentos, não sobraram mais registros nulos e o novo formato do dataframe ficou com 2.704.727 linhas e 22 colunas:

```
br                        0
causa_acidente            0
condicao_metereologica    0
dia_semana                0
fase_dia                  0
pessoas                   0
sentido_via               0
tipo_acidente             0
tipo_pista                0
tracado_via               0
uf                        0
uso_solo                  0
veiculos                  0
tipo_veiculo              0
ano_fabricacao_veiculo    0
tipo_envolvido            0
estado_fisico             0
sexo                      0
ano                       0
mes                       0
hora                      0
faixa_etaria              0
dtype: int64
Novo formato do DataFrame:  (2704727, 22)
```

Os registros ficaram com os seguintes valores:

Nome da coluna / campo	Valores
br	0,1,2,4,10,20,28,30,37,40,50,60,70,80,84,101,104,110,116,120,122,135,140,146,152,153,154,155,156,158,163,174,178,183,184,186,210,211,221,222,226,230,232,235,241,242,250,251,259,262,265,267,268,270,272,277,280,282,285,287,290,293,304,308,316,317,319,323,324,330,337,342,343,349,352,354,356,359,361,364,365,367,369,373,376,377,380,381,383,386,388,392,393,400,401,402,403,404,405,406,407,408,410,412,414,415,416,417,418,419,420,421,422,423,424,425,426,427,428,429,430,432,433,434,435,436,441,447,448,450,451,452,453,457,459,460,462,463,465,467,468,469,470,471,472,473,474,475,476,477,480,482,484,485,486,487,488,489,493,495,498,499,501,505,552,560,580,591,617,634,648,654,660,661,681,687,719,767,851,884,931
causa_acidente	agressão externa animais na pista avarias e/ou desgaste excessivo no pneu carga excessiva e/ou mal acondicionada defeito mecânico em veículo defeito na via deficiência ou não acionamento do sistema de iluminação/sinalização do veículo

	desobediência às normas de trânsito pelo condutor desobediência às normas de trânsito pelo pedestre dormindo falta de atenção falta de atenção do pedestre fenômenos da natureza ingestão de álcool ingestão de álcool e/ou substâncias psicoativas pelo pedestre ingestão de substâncias psicoativas mal súbito não guardar distância de segurança objeto estático sobre o leito carroçável outras pista escorregadia restrição de visibilidade sinalização da via insuficiente ou inadequada ultrapassagem indevida velocidade incompatível
condição_metereologica	ceu claro chuva garoa/chuvisco granizo ignorada neve nevoeiro/neblina nublado sol vento
dia_semana	segunda terça quarta quinta sexta sábado domingo
fase_dia	amanhecer anoitecer ignorado plena noite pleno dia
sentido_via	Crescente Decrescente Não Informado
tipo_acidente	atropelamento de animal atropelamento de pedestre capotamento colisão com objeto fixo colisão com objeto móvel colisão frontal colisão lateral colisão transversal colisão traseira danos eventuais derramamento de carga engavetamento incêndio queda de motocicleta / bicicleta / veículo saída de pista tombamento

tipo_pista	Dupla ignorado Múltipla Simples
tracado_via	Cruzamento Curva Desvio Temporário Interseção de vias Não Informado Ponte Reta Retorno Regulamentado Rotatória Túnel Viaduto
UF	AC , AL , AM , AP , BA , CE , DF , ES , GO , MA , MG , MS , MT , PA , PB , PE , PI , PR , RJ , RN , RO , RR , RS , SC , SE , SP , TO , (null)
uso_solo	Ignorado, Rural, Urbano
tipo_veiculo	automóvel bicicleta bonde / trem caminhão caminhão-tanque caminhão-trator caminhonete camioneta carro de mão carroça ciclomotor microônibus motocicleta motoneta motor-casa não identificado não informado ônibus outros quadriciclo reboque semi-reboque trator triciclo utilitário
ano_fabricacao_veiculo	-1 1851 1852 1879 1885 1888 1889 1896 1897 1899 1900 1901 1917 1919 1928 1929 1942 1945 1946 1947 1950 1951 1952 1953 1954 1955 1956 1957 1958 1959 1960 1961 1962 1963 1964 1965 1966 1967 1968 1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020
tipo_envolvido	Cavaleiro Condutor Passageiro Pedestre Testemunha

estado_fisico	ileso ferido grave ferido leve óbito
sexo	feminino inválido masculino
faixa_etaria	Ignorada [0-4] [5-14] [15-24] [25-34] [35-44] [45-54] [55-64] [65-74] [+75]

O dataset resultante foi exportado para uma tabela .csv, como uso do seguinte comando em Python: *df.to_csv('C:\TCC\DATASETS\dadostratados.csv')*. Todo o processamento descrito nos itens 2. Coleta de Dados e 3. Processamento/Tratamento de Dados foram realizados no *Jupyter Notebook* **1 – Coleta e Processamento-Tratamento dos Dados**.

4. Análise e Exploração dos Dados

Inicialmente para a exploração dos dados, sugiro responder a perguntas provenientes da natureza do dataset. Número de vítimas por ano, qual o mês e dia da semana com mais vítimas acidentadas, o sexo das vítimas, a hora com mais vítimas, entre outros questionamentos importantes.

Em primeiro lugar, importei o dataset gerado no *Jupyter Notebook* **1 – Coleta e Processamento-Tratamento dos Dados**:

```
#Diretório e leitura do dataset
path = r'C:\TCC\DATASETS\dadostratados.csv'
#Importação na codificação "utf-8"
df = pd.read_csv(path, sep = ',', decimal=",", encoding = "utf-8",index_col=0)
```

Para a análise dos dados, utilizei um método personalizado que utiliza o método value_counts do Pandas com a biblioteca gráfica matplotlib.pyplot. Todo o

processamento aqui descrito foi realizado no *Jupyter Notebook* **2 - Análise e exploração dos dados.**

Quantidade de vítimas acidentadas por ano

Em primeiro lugar, analisei a quantidade de vítimas por ano:

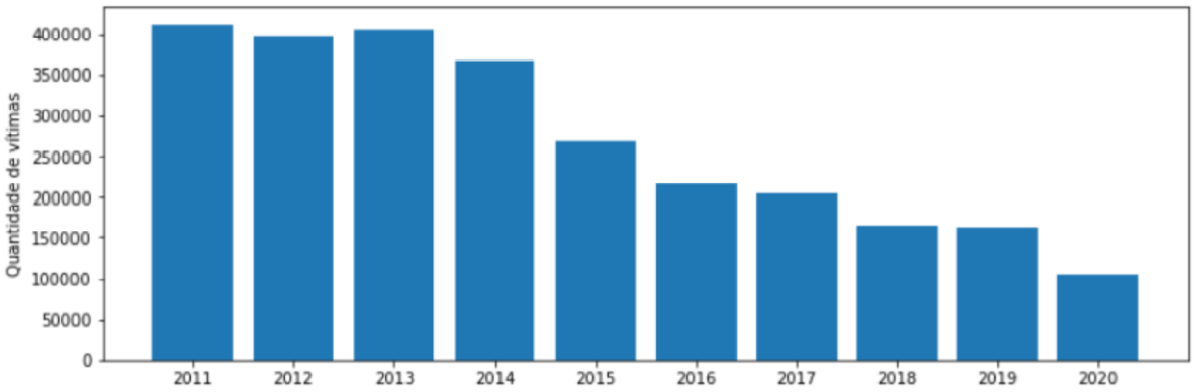
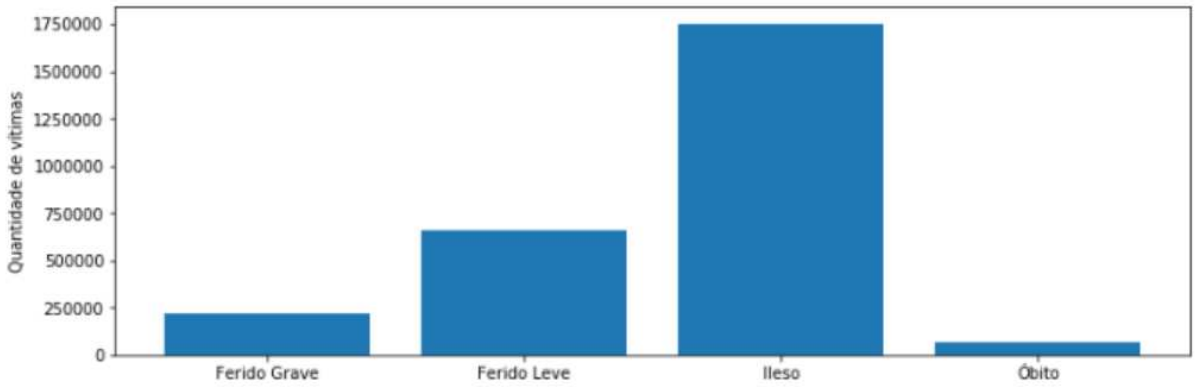


Figura 1: Quantidade de vítimas por ano

Nessa figura conseguimos verificar que a quantidade de vítimas entre os anos de 2011 até 2014 ficaram praticamente estáveis, ocorrendo uma queda considerável a partir de 2015 até o ano de 2018. Essa queda pode ser em decorrência de inúmeras variáveis, aumento na fiscalização, implementação de programas educacionais, melhora nas rodovias, diminuição do tráfego, entre outras possibilidades. Em 2019 o número é praticamente o mesmo de 2018. Já em 2020, não temos como cravar uma redução na quantidade de vítimas, tendo em vista o período abrangido na análise, até setembro de 2020.

Quantidade de vítimas acidentadas nesses 10 anos, de acordo com o seu estado físico

No gráfico abaixo estão o número de vítimas acidentadas nesses dez anos, de acordo com o seu estado físico (óbito, ferido grave, ferido leve e ileso):




```
ileso           0.648323
ferido leve     0.244788
ferido grave    0.081824
óbito           0.025065
Name: estado_fisico, dtype: float64
```

Figura 2: Quantidade de vítimas por ano de acordo com o seu estado físico

Além desses dados, que nos mostram que a quantidade de óbitos e feridos graves são menores que o de feridos leves e ilesos, também fiz uma análise da evolução ano a ano das vítimas, conforme abaixo:

ano estado_fisico	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	estado_fisico
ferido grave	29051	28291	26962	26239	22508	21423	18718	17706	18573	11841	ferido grave
ferido leve	77776	76177	76848	74604	67743	65250	65607	58941	60500	38638	ferido leve
ileso	296780	283780	293584	259451	171934	123191	113843	82883	77866	50226	ileso
óbito	8675	8663	8426	8235	6867	6398	6248	5271	5333	3677	óbito

Figura 3: Evolução ano a ano das vítimas de acordo com seu estado físico

O número de óbitos e de feridos graves vinham reduzindo ano a ano, mas em 2019 esses números voltaram a aumentar e isso pode ter acontecido por inúmeros fatores: redução das fiscalizações, acréscimo do tráfego, condição das rodoviais, etc. Essa análise é de vital importância para que se entenda o motivo desse acréscimo, evitando que isso volte a ocorrer nos próximos anos. Não temos como concluir algo para o ano de 2020, já que ele ainda está incompleto (os dados só vão até o mês de setembro).

Quantidade de vítimas acidentadas por mês

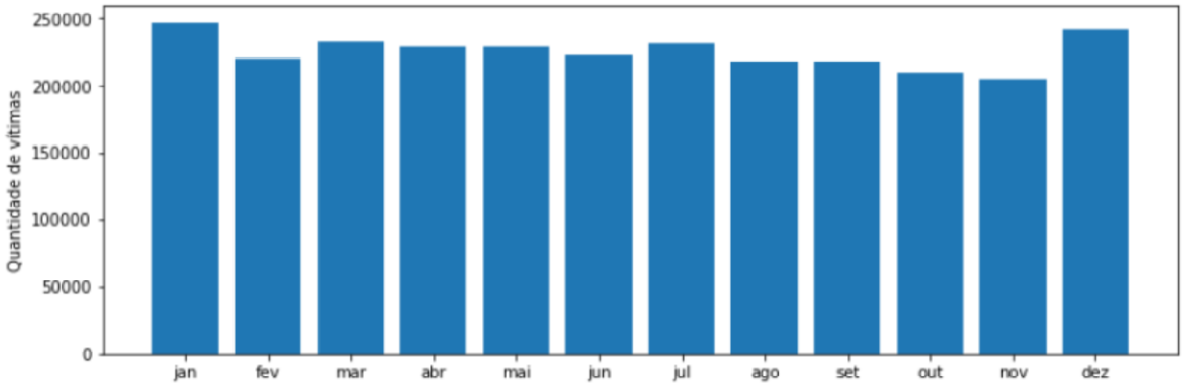


Figura 4: Quantidade de vítimas acidentadas por mês

Aqui conseguimos verificar que os meses das férias escolares (Janeiro, Julho e Dezembro) estão entre os meses com mais vítimas e isso pode indicar um aumento de tráfego nessa época. Apesar disso, a diferença entre o número de vítimas entre os meses não é tão relevante, já que em todos a quantidade de vítimas é bem similar. Mais uma vez é importante observar que os meses de outubro, novembro e dezembro de 2020 não entraram na análise.

estado_fisico	ferido grave	ferido leve	ileso	óbito
mes				
12	19793	60862	155156	6211
7	19123	56001	149561	6086
5	18723	53968	150163	5839
9	18826	54975	138130	5828
8	18815	55283	138530	5797
3	18640	55798	152682	5680
1	18996	60306	162228	5677
6	18650	53455	144635	5672
4	17917	55077	151003	5384
11	16662	50572	132499	5255
10	17687	52110	134952	5224
2	17480	53677	143999	5140

Figura 5: Quantidade de vítimas por mês de acordo com o seu estado físico

Ainda, a título de exemplo, a diferença entre o número de óbitos do mês Fevereiro (com menos óbitos) é menos de 20% em comparação com o mês de Julho (como os valores referentes ao mês de Dezembro/2020 ainda não foram computados, o desconsiderei), ou seja, isso demonstra uma equivalência no número de vítimas entre os meses.

Quantidade de vítimas acidentadas por estado

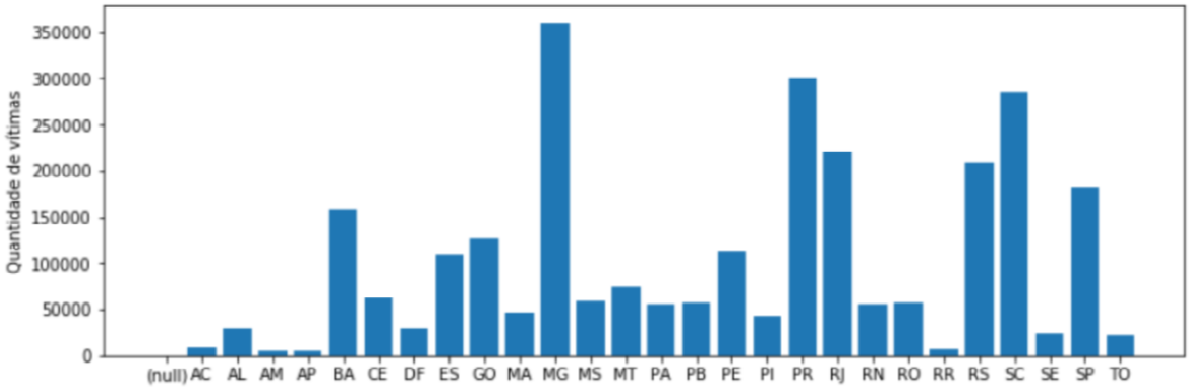


Figura 6: Quantidade de vítimas acidentadas por estado

```
Os dez Estados com mais vítimas de trânsito:
MG    360459
PR    300127
SC    284992
RJ    221061
RS    208679
SP    182162
BA    158397
GO    127010
PE    112549
ES    110133
Name: uf, dtype: int64
```

Figura 7: Os dez estados com mais vítimas de trânsito

Com essa análise conseguimos identificar que os estados de MG, PR, SC, RJ e RS são os estados com mais vítimas, enquanto os estados TO, AC, RR, AP e AM são os estados com menos vítimas. Para entendermos os motivos dessas divergências, poderíamos analisar o número de rodovias que cada estado possuí, quais são os principais veículos que trafegam nesses locais, condições das rodovias, nível de fiscalização nos locais, etc...

Também fizemos uma análise entre os estados e os estados físicos das vítimas:

estado_fisico uf	ferido grave	ferido leve	ileso	óbito
MG	32967.0	103684.0	214340.0	9468.0
PR	24243.0	75024.0	194498.0	6362.0
BA	12993.0	39565.0	99502.0	6337.0
SC	21025.0	76271.0	183127.0	4569.0
RJ	11669.0	47577.0	157634.0	4181.0
GO	12495.0	32558.0	78069.0	3888.0
RS	11924.0	45237.0	147688.0	3829.0
PE	10245.0	23499.0	75077.0	3727.0
SP	7826.0	40970.0	130562.0	2803.0
MA	5647.0	9879.0	28492.0	2590.0

Figura 8: Os dez estados com mais óbitos nos últimos 10 anos

Ou seja, apesar do estado baiano ser o sétimo em número de vítimas acidentadas, ele é o terceiro com mais vítimas fatais, o que chama bastante atenção.

Quantidade de vítimas acidentadas por hora

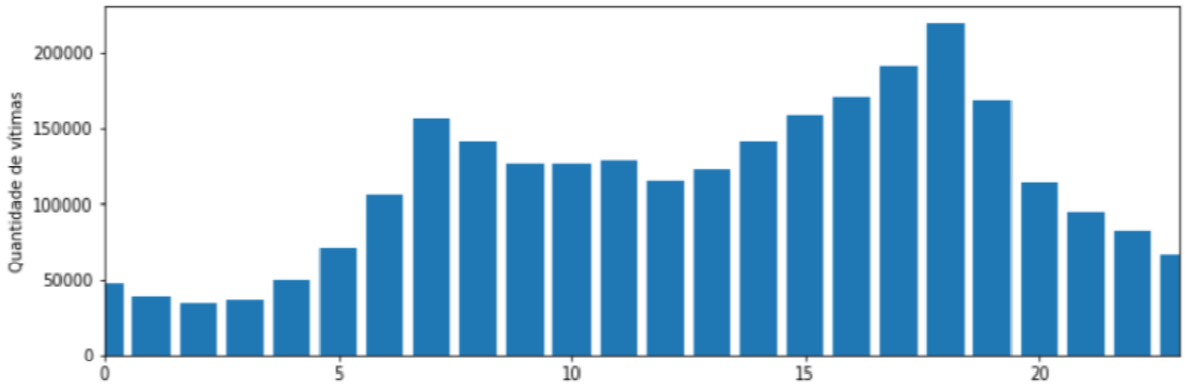


Figura 9: Quantidade de vítimas acidentadas por hora

Interessante que existe um horário de pico em que ocorrem a maior quantidade de acidentes com vítimas, das 15:00 às 19:00.

Os 05 horários com mais óbitos no trânsito:

estado_fisico	ferido grave	ferido leve	ileso	óbito
hora				
18	19004	51298	143130	5542
19	15792	40536	106054	5240
20	11553	28314	69815	4265
21	10021	24234	56744	3628
17	14524	44760	128215	3468

Figura 10: As cinco horas com mais óbitos nos últimos 10 anos

Já com relação aos óbitos nas estradas, é importante observar que eles se concentram no período do fim da tarde, a partir das 17:00 – início do pôr do sol, que pode prejudicar bastante a visibilidade, até às 21:00, quando já anoiteceu.

Quantidade de vítimas por tipo de acidente

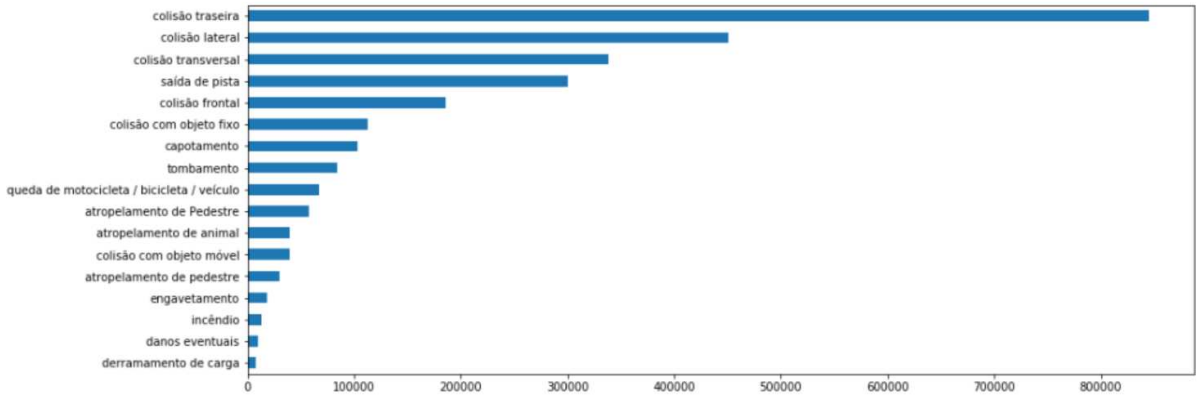


Figura 11: Quantidade de vítimas por tipo de acidente

Aqui verificamos que as colisões traseira, lateral e transversal são os tipos mais comuns de acidentes nas rodovias federais.

Além disso, também analisamos o estado físico das vítimas por tipo de acidente, conforme abaixo:

estado_fisico tipo_acidente	ferido grave	ferido leve	ileso	óbito
colisão frontal	36973	52854	74304	21341
atropelamento de pedestre	17279	19284	39957	10914
saída de pista	27267	110208	155079	7347
colisão transversal	35025	90962	206816	6142
colisão traseira	31469	126998	680025	6118
colisão lateral	22751	78796	344483	4458
capotamento	11314	46861	42769	2706
colisão com objeto fixo	8929	31784	70007	2414
tombamento	8009	38050	36485	1806
queda de motocicleta / bicicleta / veículo	12831	41844	11116	1747
colisão com objeto móvel	4824	8819	24463	1723
atropelamento de animal	3663	10118	25473	847
engavetamento	460	3602	14242	100
danos eventuais	277	970	8398	55
derramamento de carga	138	463	7449	42
incêndio	103	471	12472	33

Figura 12: Quantidade de vítimas por tipo de acidente de acordo com o seu estado físico

Dessa forma, conseguimos identificar que a colisão frontal, apesar de ser o 5º maior tipo de acidente com vítimas, é o com maior número de vitimas fatais (praticamente o dobro do segundo colocado) e lesões graves. Além dele, os outros acidentes que causam mais óbitos nas estradas federais são atropelamento de pedestre, saída de pista, colisão transversal e colisão traseira.

Quantidade de vítimas acidentadas por causa de acidente

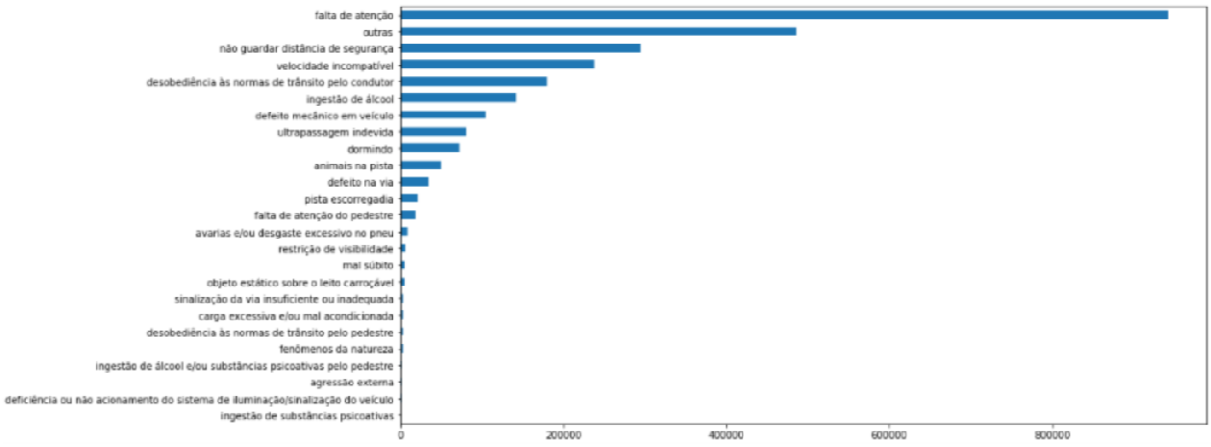


Figura 13: Quantidade de vítimas acidentadas por causa de acidente

A falta de atenção é a principal causa para acidentes com vítimas nas rodovias federais.

estado_fisico causa_acidente	ferido grave	ferido leve	ileso	óbito
outras	43782	112090	312446	17927
falta de atenção	65964	211967	649945	14186
velocidade incompatível	25003	76497	127319	9312
ultrapassagem indevida	11081	21079	43496	5540
desobediência às normas de trânsito pelo condutor	20141	48786	105766	4960

Figura 14: As cinco causas de acidentes com mais óbitos nos últimos 10 anos

Com base na planilha acima, verificamos que a causa com maior incidência de óbitos na estrada é **outras**. Como essas informações são preenchidas pela PRF, não temos como saber o significado dessa ocorrência, apenas podemos supor que são diversas das demais.

Quantidade de vítimas acidentadas por faixa etária

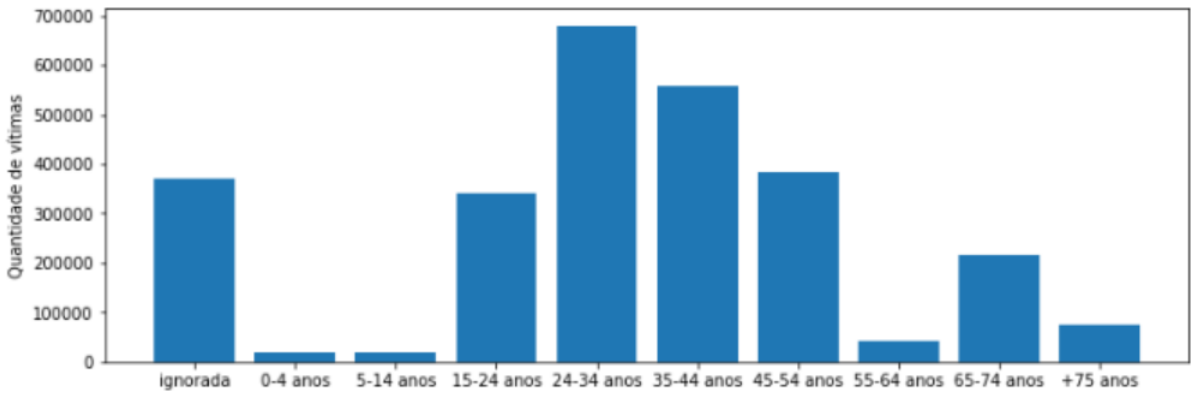


Figura 15: Quantidade de vítimas acidentadas por faixa etária

Conforme gráfico acima, conseguimos verificar que a maioria das vítimas se concentram nas idades entre 25 a 54 anos.

estado_fisico faixa_etaria	ferido grave	ferido leve	ileso	óbito
[25-34]	54087	167117	444402	14584
[35-44]	40034	118600	386578	12931
[45-54]	27375	76955	269692	10196
[15-24]	42060	127021	161815	9900
[55-64]	15309	43762	149905	6743

Figura 16: As 5 faixas etárias com mais óbitos nos últimos 10 anos

Também é nessa faixa de idades (25 a 54 anos) que se concentram os acidentes fatais nas rodovias federais.

Quantidade de vítimas acidentadas por br

As rodovias federais de números: 101, 116, 381, 40, 153, 364, 163 são as com maior número de acidentes.

Figura 17: Quantidade de vítimas acidentadas por br

Quantidade de vítimas acidentadas por Condição Meteorológica

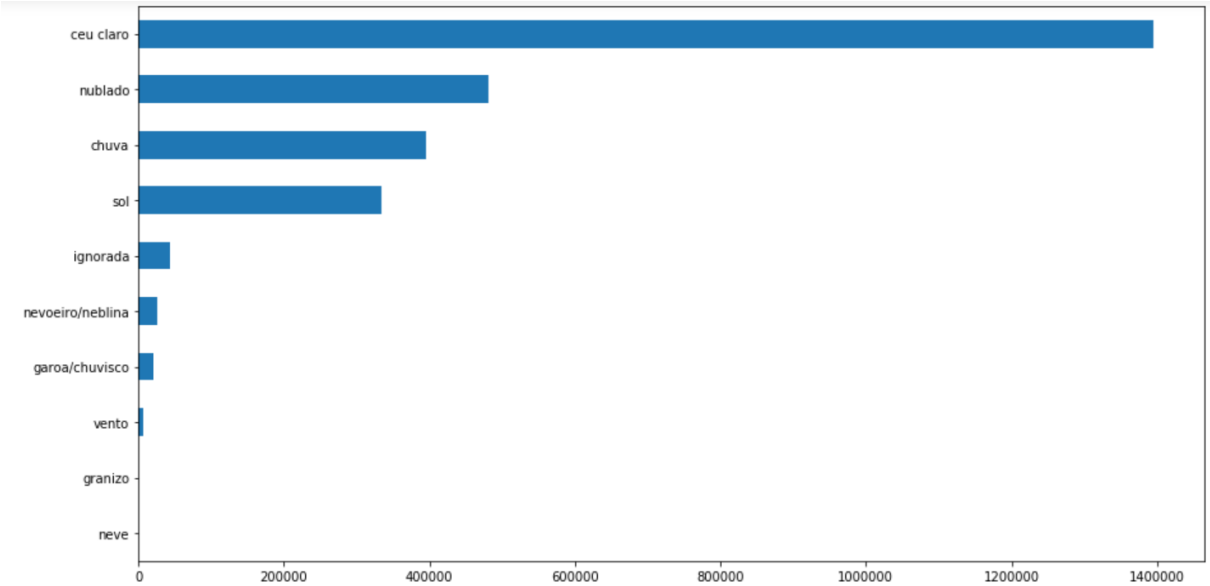


Figura 18: Quantidade de vítimas acidentadas por condição meteorológica

A maioria dos acidentes ocorrem com uma excelente condição meteorológica, céu claro, o que pode trazer alguns questionamentos sobre o motivo desses acidentes: imprudência dos motoristas? Condições ruins das rodovias? É importante que esses questionamentos sejam realizados em conjunto com os principais horários desses acidentes, das 15:00 às 19:00.

Quantidade de vítimas acidentadas por Sexo

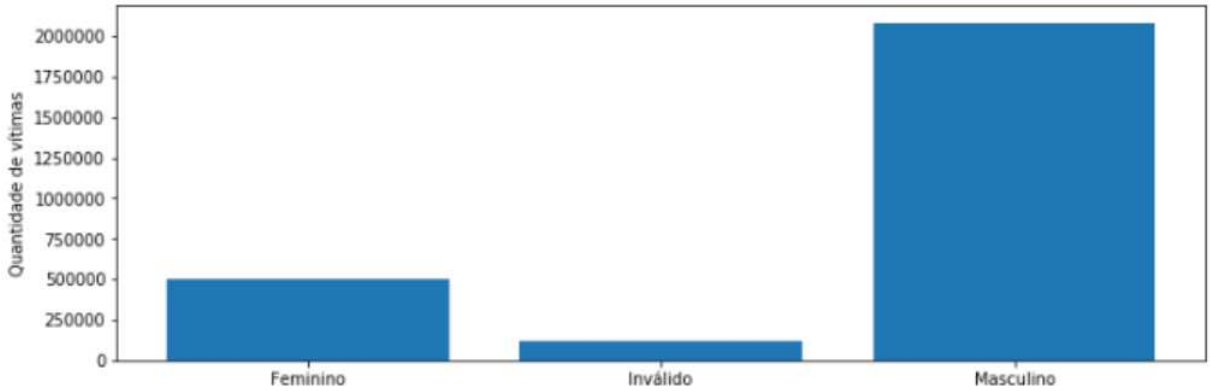


Figura 19: Quantidade de vítimas acidentadas por sexo

A quantidade de vítimas acidentadas do sexo masculino é absurdamente maior e isso pode gerar inúmeros questionamentos: o homem é mais imprudente no trânsito? Qual a quantidade de homens e mulheres nas rodovias federais?

ano	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
sexo										
feminino	65919	64666	68769	63788	51727	45842	44057	37212	37155	21587
inválido	15158	14283	15177	14099	10703	8849	13190	10290	10336	7564
masculino	331205	317962	321874	290642	206622	161571	147169	117299	114781	75231

Figura 20: Quantidade de vítimas acidentadas por sexo ano a ano

Apesar de termos um número bem maior de vítimas masculinas nos últimos 10 anos, verificamos que houve uma redução relevante nesses números desde 2011 (principalmente a partir de 2014). Em 2019, o número de vítimas masculinas foi quase um terço do registrado em 2011. Essa redução acentuada não aconteceu nas vítimas do sexo feminino (reduziu quase 44% de 2014 a 2019). Enquanto em 2011, para cada cinco vítimas, uma era do sexo feminino, em 2019 esse número foi reduzido para três. Mesmo assim, a diferença ainda é bem elevada.

Quantidade de vítimas acidentadas por tipo de veículo

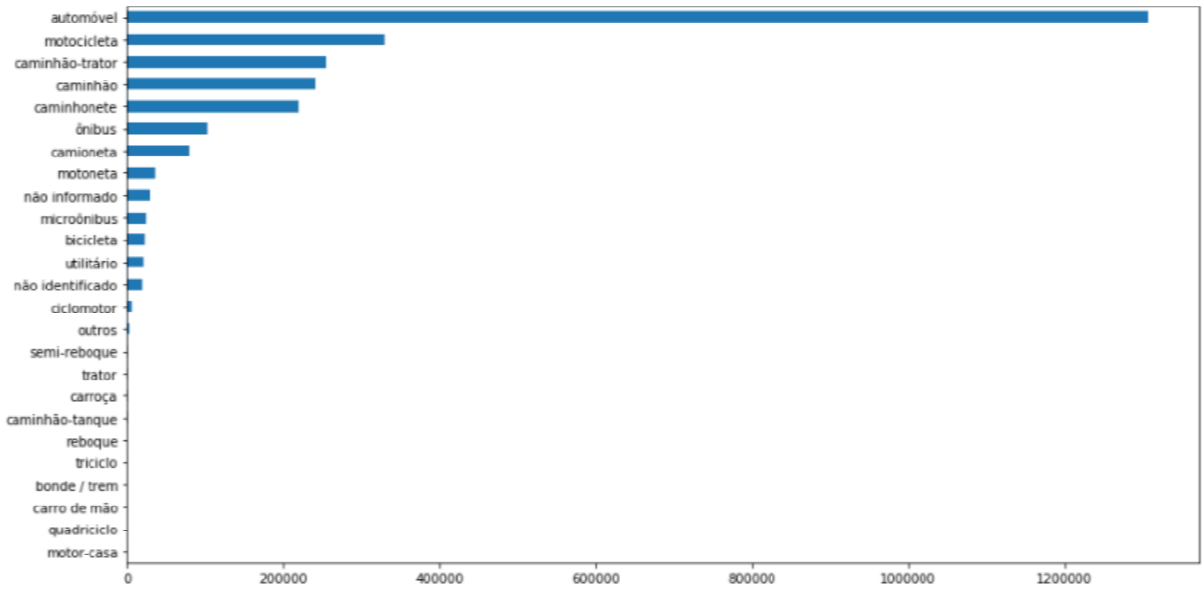


Figura 21: Quantidade de vítimas acidentadas por tipo de veículo

Como já era esperado, os automóveis e as motocicletas são os veículos com mais vítimas acidentadas.

Quantidade de vítimas acidentadas por tipo envolvido

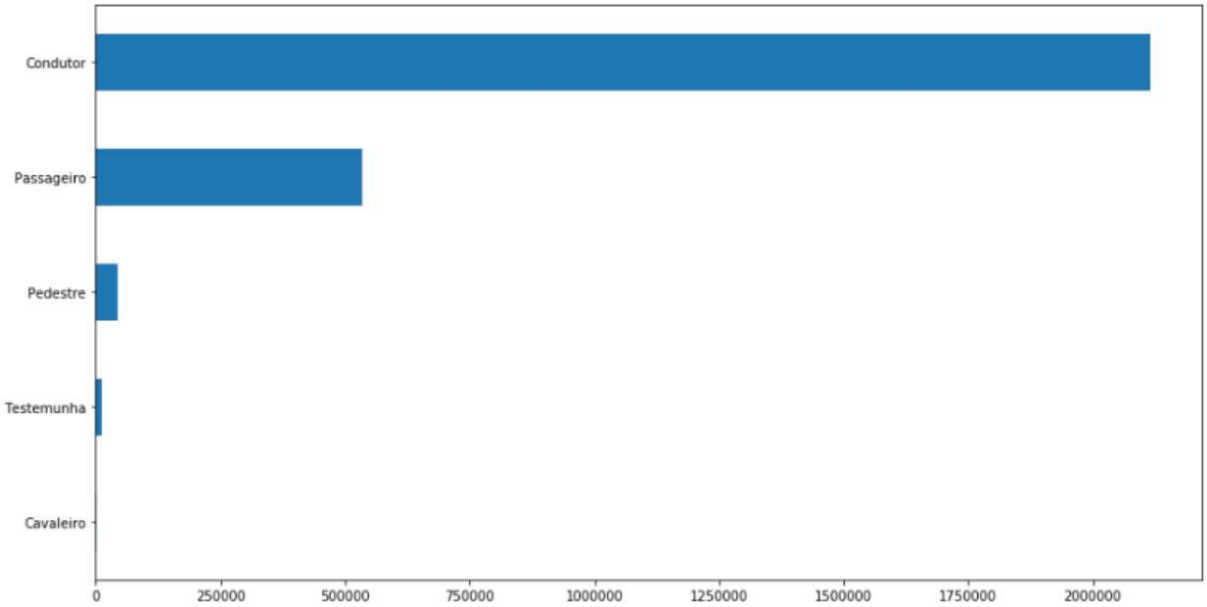


Figura 22: Quantidade de vítimas acidentadas por tipo envolvido

estado_fisico tipo_envolvido	ferido grave	ferido leve	ileso	óbito
Condutor	132469.0	386667.0	1555832.0	39005.0
Passageiro	72770.0	260929.0	183106.0	17895.0
Pedestre	15931.0	14296.0	1683.0	10828.0
Cavaleiro	142.0	192.0	158.0	64.0
Testemunha	NaN	NaN	12759.0	1.0

Figura 23: Estado físico das vítimas acidentadas por tipo envolvido

Com base nas figuras 22 e 23, conseguimos verificar que os principais tipos envolvidos nos acidentes com vítimas nas rodovias federais são os condutores e passageiros. Apesar disso, é importante observar que a incidência de vítimas fatais e ferimentos graves nos pedestres é bem maior. Essa é uma preocupação constante nas rodovias brasileiras, os pedestres, principalmente no entorno de cidades às margens das rodovias.

Além das análises realizadas acima, poderíamos fazer inúmeras outras, levando em consideração as outras variáveis do dataset, como por exemplo: tipo de pista, traçado da via, zona urbana ou rural, número de veículos, ano de fabricação do veículo, etc...

5. Criação de Modelos de Machine Learning

Após a fase de pré-processamento e entendimento dos dados, iniciamos a construção dos modelos de Machine Learning. Todo o processamento aqui descrito foi realizado no *Jupyter Notebook* **3 – Classificação**.

Com base no dataset trabalhado até agora, iremos tentar prever o estado físico da vítima acidentada com um certo grau de confiança. Como estamos diante de um problema de classificação com várias classes – aprendizado supervisionado, iremos aplicar os seguintes algoritmos de classificação: *DummyClassifier, GaussianNB, LinearDiscriminantAnalysis* e *DecisionTreeClassifier.*

Após a importação do dataset, verificamos que ele é composto de 2.704.727 linhas e 22 colunas. Iremos criar mais uma coluna, que será o nosso atributo alvo, conforme abaixo:

```
#Criando a coluna que será utilizando como atributo alvo
df.loc[df['estado_fisico'] == 'óbito', 'target'] = 0
df.loc[df['estado_fisico'] == 'ferido grave', 'target'] = 1
df.loc[df['estado_fisico'] == 'ferido leve', 'target'] = 2
df.loc[df['estado_fisico'] == 'ileso', 'target'] = 3
```

Com a criação da coluna *target*, conseguimos verificar que nos últimos dez anos, das 2.704.727 vítimas acidentadas: 1.753.538 saíram ilesos, 662.084 com ferimentos leves, 221.312 com ferimentos graves e 67.793 faleceram nas rodovias federais. Essa é a representatividade do nosso atributo alvo, ou seja, há um desbalanceamento entre o estado físico das vítimas:

```
3    0.648323
2    0.244788
1    0.081824
0    0.025065
Name: target, dtype: float64
```

Para finalizar os atributos que serão utilizados na classificação, excluímos os que não serão uteis: ano e estado_fisico, deixando dataframe com 2.704.727 linhas e 21 colunas. Ainda, utilizamos o utilitário LabelEncoder da biblioteca scikit-learn para converter os valores categóricos em valores numéricos.

```
def df_labelencoder(df):
    for column in df.columns:
        if df[column].dtype == type(object):
            le = LabelEncoder()
            df[column] = le.fit_transform(df[column].astype(str))
    return df
df_final = df_labelencoder(df)
```

Os algoritmos utilizados nesse projeto foram os seguintes:

- ***DummyClassifier*:** Utilizado como uma linha de base simples para comparar com outros classificadores. A linha base é comum no processo de criação de modelos pois serve como um ponto de referência para poder estimar com modelos mais robustos a eficiência gerada em cima dessa base. Apresentou uma acurácia bem baixa,

assim como valores extremamente baixos para as métricas de classificação das classes 0, 1 e 2.

```
---- DummyClassifier----
Acurácia (base de treinamento): 0.4874582963802714
Acurácia de previsão: 0.487714115641857
              precision    recall  f1-score   support

           0       0.03      0.03      0.03     13546
           1       0.08      0.08      0.08     44526
           2       0.24      0.25      0.24    132087
           3       0.65      0.65      0.65    350787

   micro avg       0.49      0.49      0.49    540946
   macro avg       0.25      0.25      0.25    540946
weighted avg       0.49      0.49      0.49    540946
```

- ***GaussianNB:*** Aumentou consideravelmente a acurácia, mas ainda apresentou valores extremamente baixos para as métricas de classificação das classes 0 e 1. Melhorou bastante os valores referente à classe 2, mas ainda continuam bem baixos.

```
---- GaussianNB----
Acurácia (base de treinamento): 0.6490837104124678
Acurácia de previsão: 0.6493014090130993
              precision    recall  f1-score   support

           0       0.12      0.17      0.14     13546
           1       0.10      0.03      0.05     44526
           2       0.48      0.38      0.42    132087
           3       0.74      0.85      0.79    350787

   micro avg       0.65      0.65      0.65    540946
   macro avg       0.36      0.36      0.35    540946
weighted avg       0.61      0.65      0.62    540946
```

- ***LinearDiscriminantAnalysis:*** Melhorou um pouco a acurácia em comparação com o modelo anterior, mas piorou a revocação das classes 0, 1 e 2.

```
---- LinearDiscriminantAnalysis----
Acurácia (base de treinamento): 0.6709805659630064
Acurácia de previsão: 0.6714607373009506
              precision    recall  f1-score   support

           0       0.26      0.13      0.17     13546
           1       0.26      0.01      0.01     44526
           2       0.49      0.26      0.34    132087
           3       0.70      0.93      0.80    350787

   micro avg       0.67      0.67      0.67    540946
   macro avg       0.43      0.33      0.33    540946
weighted avg       0.61      0.67      0.61    540946
```

- ***DecisionTreeClassifier:*** Para esse algoritmo primeiro rodamos ele com os parâmetros *default,* conforme resultado abaixo:

```
---- DecisionTreeClassifier----
Acurácia (base de treinamento): 0.9948774852907942
Acurácia de previsão: 0.6859390770982686
              precision    recall  f1-score   support

           0       0.19      0.21      0.20     13546
           1       0.25      0.27      0.26     44526
           2       0.49      0.50      0.50    132087
           3       0.84      0.83      0.83    350787

   micro avg       0.69      0.69      0.69    540946
   macro avg       0.45      0.45      0.45    540946
weighted avg       0.69      0.69      0.69    540946
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=0,
            splitter='best')
```

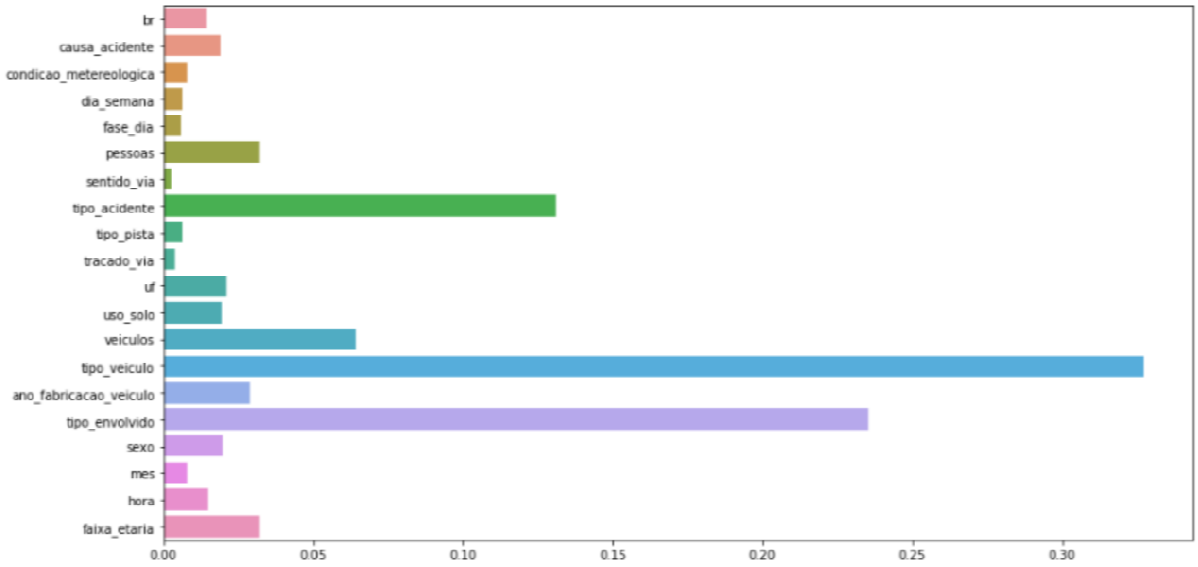
Depois utilizamos o GridSearchCV, para encontrar os melhores parâmetros entre os seguintes: 'criterion':('gini', 'entropy'), 'min_samples_leaf':[50, 100, 200]. O critério escolhido foi o **{'criterion': entropy', 'min_samples_leaf': 100}:** Com esse algoritmo encontramos a melhor acurácia e os melhores valores para as métricas de classificação das classes 0, 1, 2 e 3.

```
---- DecisionTreeClassifier----
Acurácia (base de treinamento): 0.7631100374760662
Acurácia de previsão: 0.7552010736746366
              precision    recall  f1-score   support

           0       0.44      0.13      0.21     13546
           1       0.39      0.11      0.17     44526
           2       0.57      0.62      0.59    132087
           3       0.84      0.91      0.88    350787

   micro avg       0.76      0.76      0.76    540946
   macro avg       0.56      0.44      0.46    540946
weighted avg       0.73      0.76      0.73    540946
{'criterion': 'entropy', 'min_samples_leaf': 100}
```

Após a escolha do melhor algoritmo (*DecisionTreeClassifier)*, foi feito um gráfico detalhando a importância das variáveis utilizadas na definição do atributo alvo:



Com base nessa análise, conseguimos identificar os atributos mais importantes para a classificação do estado físico das vítimas, com o uso do algoritmo *DecisionTreeClassifier.*

Após análise desse algoritmo, verificamos que era importante usarmos o parâmetro min_samples_leaf, que determina o número mínimo de observações que cada folha deve ter, para prevenir o *overfitting.* Como a classe é desbalanceada, não podemos usar valores altos.

O resultado alcançado pelo *DecisionTreeClassifier* teve uma acurácia bem melhor que os demais algoritmos, mas isso ocorreu em grande parte por causa da classe 3 (Ileso), tendo em vista a recorrência desse estado físico, 1.753.538 registros de um total de 2.704.727 (64,8%).

6. Apresentação dos Resultados

Primeiro iremos apresentar um resumo do que foi realizado nesse projeto, com o uso do modelo Canvas desenvolvido por Jasmine Vasandani:

<table>
<tr><td colspan="3">Título: Análise das vítimas dos acidentes de trânsito ocorridos nos últimos 10 anos nas rodovias federais brasileiras</td></tr>
<tr><td>Definição do Problema

Analisar datasets referentes a acidentes de trânsito nas rodovias federais nos últimos 10 anos, com o objetivo de identificar informações relevantes.</td><td>Resultados e Previsões:

O objetivo desse trabalho é identificar nos datasets disponibilizados pela PRF padrões nos acidentes de trânsito com o estado físico das vítimas.</td><td>Aquisição de Dados:

Os dados foram obtidos do site da PRF – acidentes agrupados por ocorrência e acidentes agrupados por pessoa.</td></tr>
<tr><td>Modelagem:

Foram realizadas inúmeras análises nos dados importados, com o uso da linguagem Python, no intuito de se preparar um dataset para aplicação de diferentes algoritmos de classificação.</td><td>Avaliação do Modelo:

Os algoritmos de classificação serão avaliados por meio da matriz de confusão e do relatório de classificação com foco nas métricas por classe.</td><td>Preparação dos Dados:

Todos os registros foram tratados após a junção dos datasets disponibilizados pela PRF. Como cada registro possui sua especificidade, a preparação foi feita de forma pontual para cada registro.</td></tr>
</table>

Os dados disponibilizados pela PRF são ricos em informações, com inúmeros dados interessantes. Nos acidentes agrupados por ocorrência, o foco é o acidente, enquanto nos acidentes agrupados por pessoa, o foco está nas pessoas acidentadas. Com a junção desses *datasets*, conseguimos incluir informações importantes para uma análise mais apurada desses acidentes.

Após essa junção e o tratamento dos dados, conseguimos identificar várias situações interessantes nos acidentes com vítimas nas rodovias federais e que podem ajudar no trabalho de prevenção de acidentes, como por exemplo:

- Meses de janeiro e dezembro concentram a maior parte das vítimas acidentadas;
- MG, PR, SC, RJ e RS são os estados com mais vítimas, enquanto os estados TO, AC, RR, AP e AM são os estados com menos vítimas.
- O estado da Bahia, apesar de ser o sétimo em número de vítimas, é o terceiro com mais vítimas fatais.

- O horário das 15:00 às 19:00 é o que concentra a maior quantidade de vítimas acidentadas. Já com relação aos acidentes com vítimas fatais, eles se concentram no horário das 17:00 às 21:00.
- A colisão traseira é o acidente com mais vítimas acidentas, mas a colisão frontal é o que possui mais vítimas fatais.
- A falta de atenção é o principal motivo para acidentes com vítimas.
- A maioria das vítimas de trânsito se concentram nas idades entre 25 a 54 anos. Também é nessa faixa que se concentra os acidentes fatais.
- As rodovias federais de números 101, 116, 381 são as com maior número de acidentes.
- Na maior parte dos acidentes com vítimas a condição meteorológica é de céu claro.
- Os homens são as maiores vítimas dos acidentes de trânsito.
- Os automóveis e as motocicletas são os veículos mais presentes nos acidentes com vítimas.

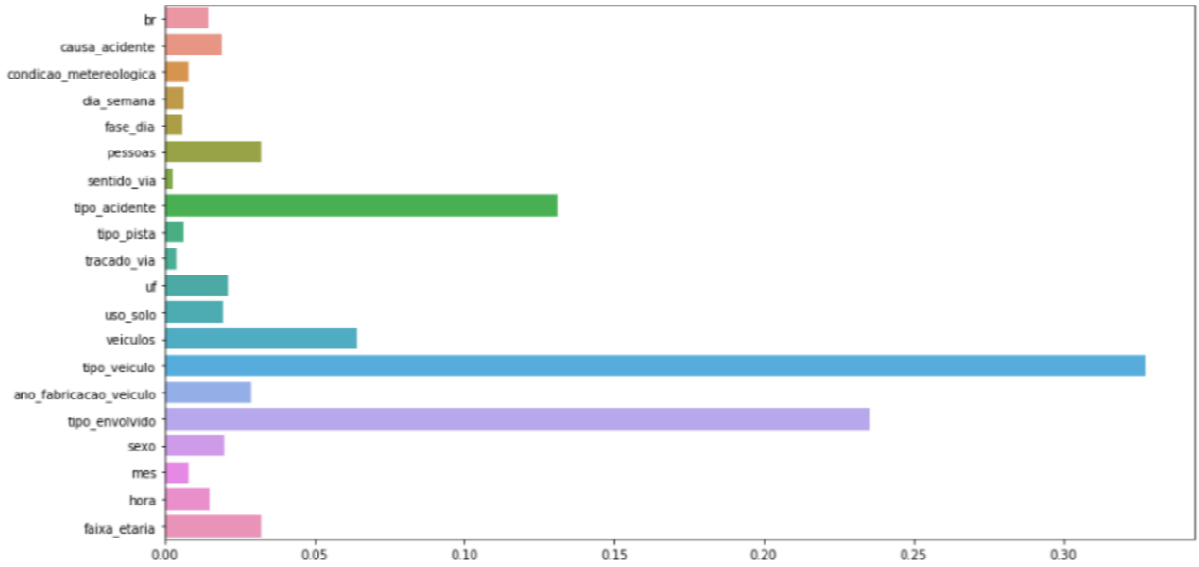
O *dataset* final construído com esses dados foram submetidos a um conjunto de algoritmos classificadores: *GaussianNB*, *LinearDiscriminantAnalysis* e *DecisionTreeClassifier*.

O algoritmo mais preciso foi o *DecisionTreeClassifier*, que, conforme apresentado acima, apresentou como tributos mais importantes para a classificação do estado físico das vítimas acidentadas, o tipo de veículo utilizado (automóvel, motocicleta, caminhão, etc...), o tipo envolvido no acidente (condutor, passageiro, pedestre, etc...) e o tipo de acidente (colisão traseira, lateral, frontal, etc...).

```
---- DecisionTreeClassifier----
Acurácia (base de treinamento): 0.7631100374760662
Acurácia de previsão: 0.7552010736746366
              precision    recall  f1-score   support

           0       0.44      0.13      0.21     13546
           1       0.39      0.11      0.17     44526
           2       0.57      0.62      0.59    132087
           3       0.84      0.91      0.88    350787

   micro avg       0.76      0.76      0.76    540946
   macro avg       0.56      0.44      0.46    540946
weighted avg       0.73      0.76      0.73    540946
{'criterion': 'entropy', 'min_samples_leaf': 100}
```



Apesar de ter tido a melhor acurácia, as métricas por classe de previsão e revocação foram bem baixos. Como o *dataset* é bem desbalanceado, fica difícil classificar quando a vítima faleceu, teve ferimento grave ou ferimento leve:

7. Links

Link para o vídeo: https://youtu.be/B1dflLjeC1g

Link para o repositório: **github.com/danunes84/tcc-puc-minas**.