**UNIVERSIDADE CATÓLICA DO SALVADOR – UCSAL**
**MBA EM TECNOLOGIAS E APLICAÇÕES DE BUSINESS INTELIGENCE**

# TRABALHO FINAL – MINERAÇÃO DE DADOS

# INFRAÇÕES DE TRÂNSITO

Disciplina: Análise de Dados e Data Mining.
Prof.: Grimaldo Lopes
Aluno: Camila da Silva Oliveira

**Salvador, Abril de 2019**

# Sumário

# Trabalho – Mineração de Dados

## 1. Introdução

Os dados produzidos constantemente pelas instituições públicas ou privadas são a base para obtenção de informações e de conhecimento de extra importância para a tomada de decisões estratégicas e analíticas, assegurando o aumento da assertividade das ações e garantindo vantagem competitiva. No caso das instituições públicas, os dados disponibilizados tem outra função importante: de acordo com a lei nº 12.527 /2011, o governo deverá divulgar os dados para garantir a transparência, criando melhores possibilidades de controle social das ações governamentais.

Em cumprimento da lei de transparência, a Polícia Rodoviária Federal, disponibilizou os dados referente as multas em rodovias federais desde 2007 até 2018. Este trabalho tem como objetivo, efetuar a análise destes dados com objetivo de produzir conhecimento, através do uso da mineração de dados – processo de padrões e tendências em uma base de dados. Existem dois grandes focos usuais para uso da mineração, segundo Pinheiro (2008, p. 97), "O primeiro é a identificação de segmentos com características semelhantes, agrupando os casos, ou as ocorrências, de acordo com as informações descritivas dos mesmos. O segundo foco é a predição de eventos com prévia ciência dos resultados."

Para avaliação das infrações de trânsito, será adotada técnicas para identificação de segmentos com atributos semelhantes. A proposta do estudo é identificar quais as características das infrações de maior gravidade, possibilitando adoção de ações específicas e direcionadas com base nas descobertas realizadas no estudo.

## 2. Objetivo

O principal objetivo deste trabalho é identificar, através da análise e adoção de técnicas de mineração de dados, quais são as principais características de cada gravidade das infrações registradas nas rodovias federais, de acordo, com os dados disponibilizados pela Polícia Rodoviária Federal.

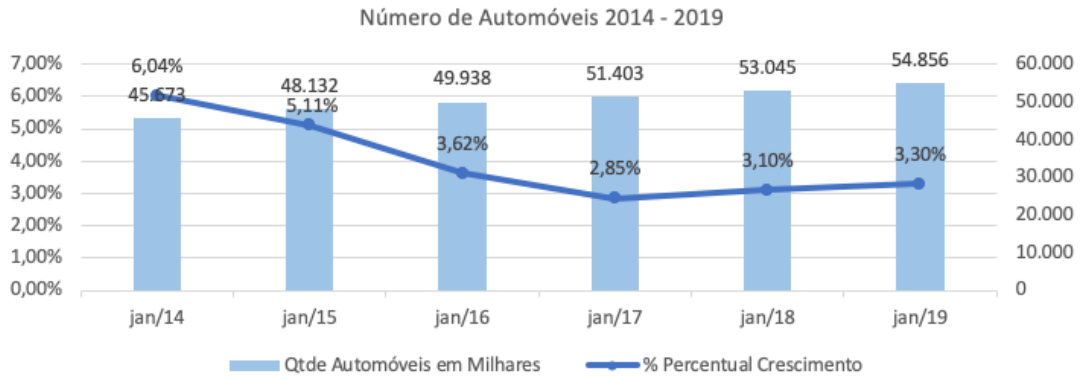
Mediante a análise dos dados, espera-se responder também aos questionamentos adicionais:

| N | PERGUNTAS |
|---|---|
| 1 | Quais as principais características das infrações por gravidade da infração? |
| 2 | Qual melhor método de mineração para análise das infrações por gravidade? |
| 3 | Qual o nível de assertividade do método adotado? |
| 4 | Qual região com maior concentração de multas com maior gravidade? |
| 5 | Qual a predominância da competência das infrações – municipal, estadual e federal? |
| 6 | Existe algum modelo específico de veículo que tem maior características de infração? |
| 7 | Há uma espécie de veículo com maior concentração de infrações graves / gravíssima? |
| 8 | Quem é o maior responsável pelo maior índice de infração? |

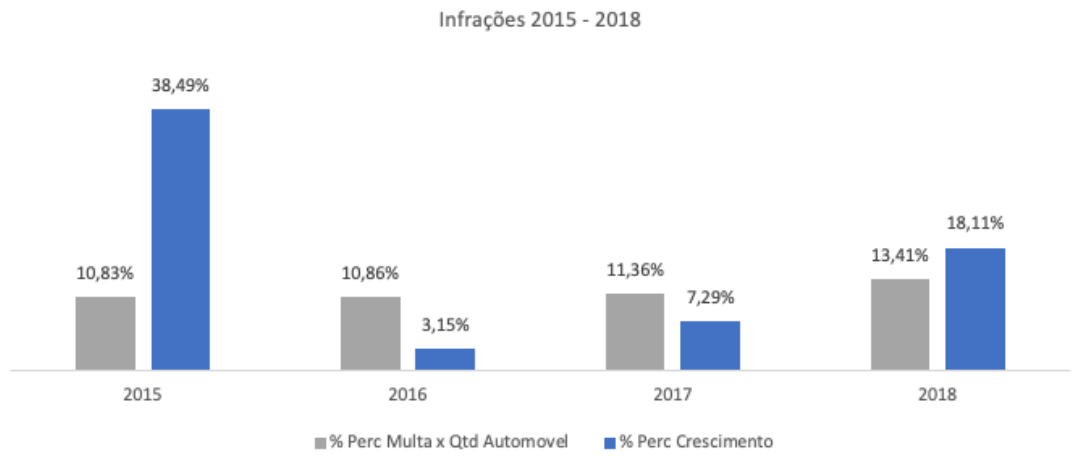
## 3. Justificativa

De acordo com as divulgações oficiais do Departamento Nacional de Trânsito – DENATRAN, o número de automóveis circulando no país cresce em média 3% ao ano, representando um média de 1,8 milhões de veículos, da mesma forma, o número de infrações tem também apresentando um crescimento significado. O crescimento de 2018 para 2017 foi de 18%, mantendo uma relação de 11% em relação ao total de multas pelo montante de automóveis.

Através da análise das dos dados referente as infrações cometidas, buscando descobrir padrões e tendências auxiliará no entendimento das multas registradas. Com o resultado da mineração dos dados, as instituições responsáveis poderão adotar ações mais assertivas e direcionadas, podendo reverter a tendência de crescimento.




Fonte: Denatran




Fonte: Denatran x Polícia Rodoviária Federal

## 4. Conjunto de Dados (Dicionário de Dados)

Para este estudo foi selecionado a base de dados da infração de trânsito, disponibilizados no site da Polícia Rodoviária Federal. As informações contidas no portal de 2007 a 2018 no formato "*csv*", contendo data e hora de registro, modelo e número do veículo, município, estado onde ocorreu a multa e de origem, além de outros dados. Para este trabalho foi utilizado apenas os dados de dezembro de 2018 em decorrência da limitação das ferramentas Weka.

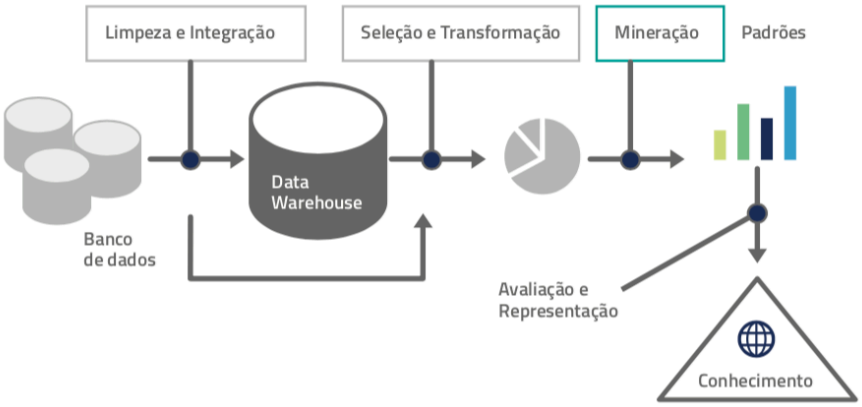
Complementar a base de dados foi incluso informações com detalhamento da infração, os dados foram oriundos de três fontes distintas com os mesmos dados, mas mantendo a informação referente a valor do estado da Bahia. Uma das informações principais é referente a gravidade da infração, competência, enquadramento e responsabilidade.

| **TAB_INFRACAO_BASE** | | | |
|---|---|---|---|
| Tabela que contém todos os dados da infração registradas em dezembro de 2018 | | | |
| Nome | Tipo | Definição | Classificação |
| dat_infracao | DATE | Data da infração no formato dd/mm/aaaa | Atributo |
| tip_abordagem | VARCHAR | Identifica se houve abordagem do veículo: C (houve abordagem), S (não houve abordagem). | Atributo |
| ind_assinou_auto | VARCHAR | Variável que informa se o infrator assinou o auto de infração. S (sim), Vazio (não). | Atributo |
| ind_veiculo_estrangeiro | NUMERIC | Variável que informa se o veículo é estrangeiro. S (sim), N (não). | Atributo |
| ind_sentido_trafego | VARCHAR | Sentido da vida onde ocorreu a infração. C (crescente), D (decrescente). | Atributo |
| uf_placa | VARCHAR | Unidade federativa da placa do veículo. | Atributo |
| uf_infracao | VARCHAR | Unidade federativa do local onde ocorreu a infração. | Atributo |
| num_br_infracao | NUMERIC | Variável com valores numéricos, representando o identificador da BR onde ocorreu a infração. | Métrica |
| num_km_infracao | NUMERIC | Identificação do quilômetro onde ocorreu a infração. | Métrica |
| nom_municipio | VARCHAR | Nome do município onde ocorreu a infração. | Atributo |
| cod_infracao | NUMERIC | Código da infração de trânsito | Atributo |
| descricao_abreviada | VARCHAR | Descrição abreviada da infração. | Atributo |
| enquadramento | VARCHAR | Enquadramento da infração de acordo com o CTB. | Atributo |
| data_inicio_vigencia | DATE | Data no formato dd/mm/aaaa do início da vigência da infração. | Atributo |
| data_fim_vigencia | DATE | Data no formato dd/mm/aaaa do fim da vigência da infração. Quando a infração continuar vigente, a data será igual ao último dia do mês seguinte ao mês com os últimos dados disponibilizados. Ex: os últimos dados disponibilizados são de setembro/2016. A data fim será igual a 31/10/2016. | Atributo |
| med_realizada | INTEGER | Registro da medição realizada em radares, etilômetros, balanças e trenas. | Métrica |
| med_considerada | INTEGER | Medição considerada para o registro da infração. | Métrica |
| exc_verificado | INTEGER | Excesso verificado nas infrações onde são utilizados equipamentos de medição. | Métrica |
| especie | VARCHAR | Espécie do veículo de acordo com o registro. | Atributo |
| nome_veiculo_marca | VARCHAR | Marca do veículo. | Atributo |
| nom_modelo_veiculo | VARCHAR | Modelo do veículo. | Atributo |
| hora | VARCHAR | Hora da infração no formato 00:00:00 | Atributo |

| TAB_DADOS_INFRA | | | |
|---|---|---|---|
| Tabela com os dados complementar referente as infrações de trânsito. | | | |
| Nome | Tipo | Definição | Classificação |
| num_infracao | INTEGER | Código da infração de trânsito | Atributo |
| descricao | VARCHAR | Descrição abreviada da infração. | Atributo |
| enquadramento | VARCHAR | Número artigo / lei que fornece o amparo legal para aplicação da infração de acordo com código de trânsito brasileiro. | Atributo |
| gravidade | VARCHAR | Gravidade da infração, podendo ser: leve, moderada, grave e gravíssima. | Atributo |
| responsabilidade | VARCHAR | Responsável pela infração registrada, valores registrados. | Atributo |
| competencia | VARCHAR | Órgão responsável por autuar a infração | Atributo |
| valor | DECIMAL | Valor cobrado pela infração | Métrica |

## 5. Algoritmos e Etapas da Mineração de Dados

Para efetuar a mineração dos dados da infração foi adotado o processo de extração de conhecimento, conhecido como KDD - Knowledge Discovery in Databases. O KDD é constituído pelas fases: limpeza e integração, seleção e transformação, mineração de dados e por fim avaliação dos resultados identificados.




Fluxo mineração de dados – Fonte: BI como deve ser.

### 5.1. Limpeza e Integração

O processo de limpeza e integração dos dados foi realizada utilizando aplicativos distintos, por etapas. A primeira etapa foi a carga dos dados no banco de dados MySQL, utilizando o Pentaho, criando transformações específicas para cada tipo de dado. Cada transformação efetua a cargo do arquivo csv no banco.

Arquivos csv utilizados:

- Base de dados das infrações de dezembro de 2018.
- Base de dados das relações de infrações de Foz de Iguaçu, primeira base adotada.

Transformação realizada no Pentaho para carga no banco de dados

As bases foram carregadas no banco MySQL e em análise inicial das tabelas, verificou-se no cruzamento das duas tabelas que muitas informações complementares de infrações ficaram incompletas ou com poucos dados, apenas 16% com informação.

Para assegurar que os dados ficassem completos, foram utilizadas mais duas fontes de dados que, após pesquisa, verificou-se que o código de infração é o mesmo, assim como a descrição. As informações foram convertidas em uma fonte única, priorizando os dados de preço oriundo do site da Bahia. As análises foram feitas no aplicativo Calc do Libre Office. O processo de transformação do Pentaho foi efetuado novamente para os dados complementar.

Após a carga uma segunda análise foi realizada, os dados referentes as informações complementares passaram para 90,9%, considerado aceitado para seguir para a segunda etapa: transformação dos dados e uma única base, a tabela *Infra_MD*, resultado da junção dos dados com a inclusão da informação Região para enriquecer mais a base de dados. O processo de transformação foi realizado através de uma consulta SQL, com a carga das informações necessárias. Para conclusão da etapa foi criada a transformação no Pentaho para extração da base, criando o arquivo base.csv para análise na ferramenta de mineração Weka.

Transformação realizada no Pentaho para extração dos dados

Para finalizar o processo de limpeza, ao efetuar o teste de abertura do arquivo no Weka, a ferramenta sinalizou alguns erros na base que foram corrigidas manualmente: algumas linhas possuíam quebra desnecessária e foi necessário excluir as aspas “” que constavam no campo de descrição. Após finalização da limpeza e integração, os dados foram carregados corretamente no Weka.

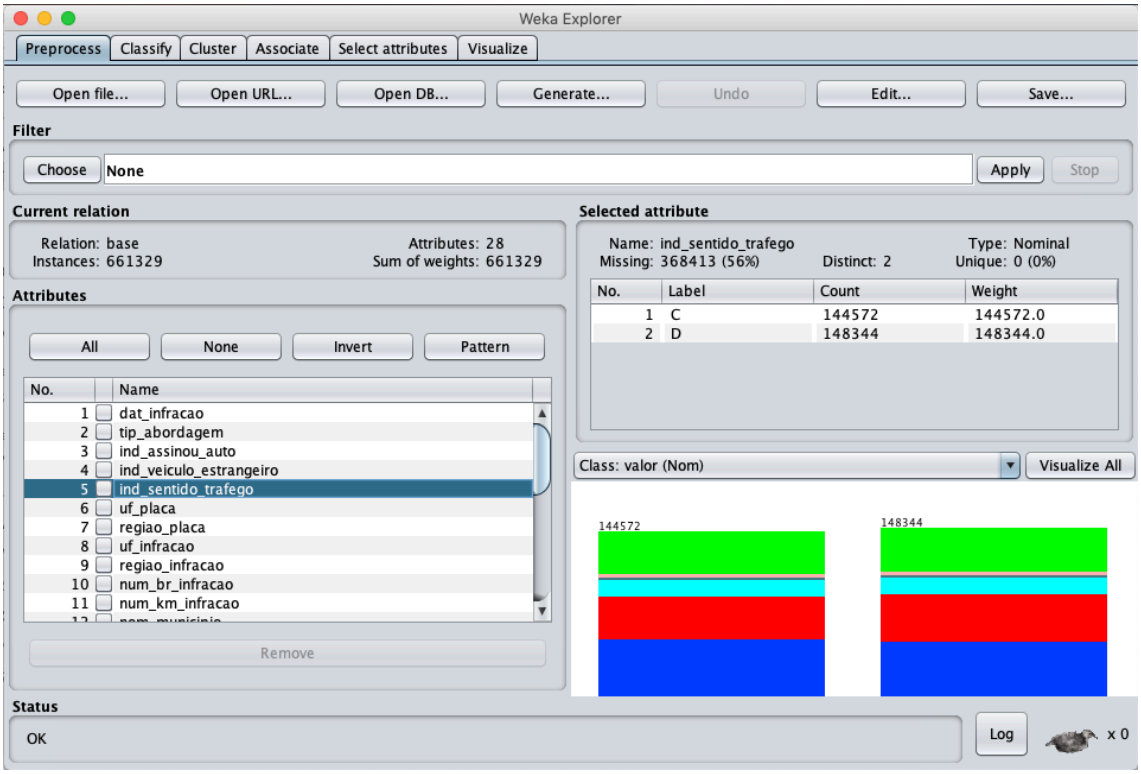
## 5.2. Seleção e Transformação

Os dados foram carregados corretamente no Weka e foram realizados os primeiros testes de mineração de dados, não trazendo resultados positivos – campos sem dados, com informações em granularidade alta entre outros pontos, sendo necessário seguir com a etapa para selecionar e transformar os campos da tabela *Infra_MD*, conforme quadro descritivo abaixo.

| **INFRA_MD** | | | |
|---|---|---|---|
| Tabela com a base de infração para mineração de dados – detalhamento da análise efetuada em cada campo. | | | |
| N | Campo | Ação | Motivo da Ação |
| 1 | dat_infracao | Transformar | Converter as datas em período: 1ª a 4ª semana |
| 2 | tip_abordagem | Manter | |

| INFRA_MD | | | |
|---|---|---|---|
| Tabela com a base de infração para mineração de dados – detalhamento da análise efetuada em cada campo. | | | |
| N | Campo | Ação | Motivo da Ação |
| 3 | ind_assinou_auto | Excluir | Todas as infrações registradas foram assinadas, não havendo relevância para o estudo. |
| 4 | ind_veiculo_estrangeiro | Excluir | Não consta carro estrangeiro na base de dados, não havendo relevância para o estudo. |
| 5 | ind_sentido_trafego | Excluir | Informação do sentido da via da infração, volume de missing elevado – 56%. |
| 6 | uf_placa | Excluir | O dado é diretamente relacionado a região, em decorrência da granularidade, optou-se por efetuar a avaliação considerando a região. |
| 7 | regiao_placa | Manter | |
| 8 | uf_infracao | Excluir | O dado é diretamente relacionado a região, em decorrência da granularidade, optou-se por efetuar a avaliação considerando a região. |
| 9 | regiao_infracao | Manter | |
| 10 | num_br_infracao | Excluir | Dado referente a localização da estrada onde ocorreu a infração, para dado de localização optou-se por utilizar a região. |
| 11 | num_km_infracao | Excluir | Dado referente a localização da estrada onde ocorreu a infração, para dado de localização optou-se por utilizar a região. |
| 12 | nom_municipio | Excluir | O dado é diretamente relacionado a região, em decorrência da granularidade, optou-se por efetuar a avaliação considerando a região. |
| 13 | cod_infracao | Manter | |
| 14 | descricao_abreviada | 2ª Base | O dado será utilizado para fazer uma análise específica direcionada a análise de texto. |
| 15 | enquadramento | Excluir | Variável relacionada diretamente ao tipo de infração. O dado será analisado por texto de acordo com a descrição da infração. |
| 16 | data_inicio_vigencia | Excluir | O campo possui apenas duas informações, sendo 99% uma data específica, não sendo necessário para análise. |
| 17 | data_fim_vigencia | Excluir | Campo nulo. |
| 18 | med_realizada | Excluir | As medições não possuem informação de que tipo de medição foi realizada, não fazendo sentido para análise. |
| 19 | med_considerada | Excluir | As medições não possuem informação de que tipo de medição foi realizada, não fazendo sentido para análise. |
| 20 | exc_verificado | Excluir | As medições não possuem informação de que tipo de medição foi realizada, não fazendo sentido para análise. |
| 21 | especie | Excluir | Grande concentração de missing (66%) |
| 22 | nome_veiculo_marca | Excluir | Dados manual, preenchimento incompleto sem padronização. Realizada tentativa de corrigir, mas não foi possível. |
| 23 | nom_modelo_veiculo | Excluir | Dados manual, preenchimento incompleto sem padronização. Realizada tentativa de corrigir, mas não foi possível. |
| 24 | hora | Transformar | Converter em manhã, tarde e noite. |
| 25 | gravidade | Manter | |
| 26 | responsabilidade | Transformar | Ajustar o campo – corrigir acento. |

| INFRA_MD | | | |
|---|---|---|---|
| Tabela com a base de infração para mineração de dados – detalhamento da análise efetuada em cada campo. | | | |
| N | Campo | Ação | Motivo da Ação |
| 27 | competencia | Transformar | Corrigir – muitos dados com campos incorretos. |
| 28 | valor | Manter | |



Tela Weka com análise do campo sentido_trafego

Para agrupamento do campo modelo de veículo, efetuada a correção na ferramenta Calc do Libre Office, extração de um arquivo csv com o depara, carga no banco de dados e após atualização da base original através de consulta SQL. A atualização dos demais campos foi realizada também através de consulta SQL. Após conclusão da versão final da tabela, efetuada a extração apenas dos campos selecionados para a mineração de dados através da ferramenta Pentaho, através de uma transformação específica.

## 5.3. Mineração de Dados

Para efetuar a análise da base de dados foram testadas as tarefas de mineração: classificação J48, agrupamento por EM, agrupamento por simples K Means e redes neurais.

### 5.3.1. Classificação J48

A primeira análise foi realizada através de árvore de decisão que possibilita criar segmentação a partir de um conjunto de dados, que poderá ser utilizada para a predição de alguma parte da informação. O algoritmo utilizado produz validações mais completas e integradas do que as outras técnicas de data mining.

Na primeira execução do modelo o resultado do fato Kappa, método estatístico para avaliar o nível de concordância ou reprodutibilidade entre dois conjuntos de dados. Inclusa a regra para podar a árvore. O resultado apresentado foi de 0,9993. Na análise do resultado, verificou-se que o valor é um indicador extremamente correlacionado com a gravidade da multa, os valores maiores são concentrados para as multas graves e gravíssima. Além disso, verificou que não há necessidade de manter as variáveis: região da placa, o resultado apresentado ficou confuso, por já haver um atributo com resultado parecido – região_infração, responsabilidade e competência que se encontra concentrado mais de 80% em apenas um dos resultados.

Resultado da tarefa Weka:

```
Number of Leaves  :     90

Size of the tree :      118


Time taken to build model: 3.25 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 1.41 seconds

=== Summary ===

Correctly Classified Instances      661058               99.959  %
Incorrectly Classified Instances       271                0.041  %
Kappa statistic                          0.9993
Mean absolute error                      0.0004
Root mean squared error                  0.0136
Relative absolute error                  0.1279 %
Root relative squared error              3.5735 %
Total Number of Instances           661329

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,998    0,000    1,000      0,998    0,999      0,999    1,000     1,000     GRAVE
                 1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     MEDIA
                 1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     GRAVISSIMA
                 0,999    0,000    0,981      0,999    0,990      0,990    1,000     0,997     LEVE
Weighted Avg.    1,000    0,000    1,000      1,000    1,000      0,999    1,000     1,000

=== Confusion Matrix ===

      a      b      c      d   <-- classified as
 155935    125      0    129 |      a = GRAVE
     13 382877      0      0 |      b = MEDIA
      0      0 115429      0 |      c = GRAVISSIMA
      4      0      0   6817 |      d = LEVE
```

Na segunda tentativa de executar a tarefa sem as variáveis definidas, o resultado do Kappa statistic foi de 0,373, ficando abaixo do resultado esperado de 0,8.

```
=== Summary ===

Correctly Classified Instances      448245               67.7794 %
Incorrectly Classified Instances    213084               32.2206 %
Kappa statistic                          0.373
Mean absolute error                      0.2311
Root mean squared error                  0.3399
Relative absolute error                 79.9049 %
Root relative squared error             89.3896 %
Total Number of Instances           661329

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,026    0,010    0,460      0,026    0,050      0,062    0,615     0,322     GRAVE
                 0,963    0,500    0,726      0,963    0,828      0,542    0,761     0,750     MEDIA
                 0,652    0,126    0,522      0,652    0,580      0,483    0,803     0,462     GRAVISSIMA
                 0,000    0,000    ?          0,000    ?          ?        0,689     0,023     LEVE
Weighted Avg.    0,678    0,314    ?          0,678    ?          ?        0,733     0,591

=== Confusion Matrix ===

      a      b      c      d   <-- classified as
   4131  99631  52427      0 |      a = GRAVE
    863 368823  13204      0 |      b = MEDIA
   3721  36417  75291      0 |      c = GRAVISSIMA
    266   3235   3320      0 |      d = LEVE
```

Analisando o resultado da árvore de decisão, observa-se o seguinte padrão nos casos de infrações sem abordagem, os maiores casos de ocorrência aconteceram na região Sul com multa de média gravidade. Quando se refere a infrações gravíssima, as ocorrências se concentraram na região Norte, na 4ª semana. Vale ressaltar que há uma concentração de ocorrências grave e gravíssima no período da noite nessa região.

Nos casos das infrações registradas com abordagem, a maior tendência concentra-se na região nordeste. Observa-se uma concentração significativa de ocorrência das infrações com abordagem na quarta semana, possivelmente em decorrência das viagens que devem acontecer neste período devido as festas de final de ano.

Árvore de decisão:

```
J48 unpruned tree
------------------

tip_abordagem = S
|   regiao_infracao = Sudeste: MEDIA (228424.0/52701.0)
|   regiao_infracao = Sul: MEDIA (98730.0/30529.0)
|   regiao_infracao = Nordeste: MEDIA (96515.0/31372.0)
|   regiao_infracao = Centro-Oeste: MEDIA (79361.0/22184.0)
|   regiao_infracao = Norte
|   |   periodo_infracao = NOITE
|   |   |   dia_infracao = SEM_1: GRAVISSIMA (49.0/24.0)
|   |   |   dia_infracao = SEM_2: GRAVE (74.0/40.0)
|   |   |   dia_infracao = SEM_3: GRAVISSIMA (110.0/62.0)
|   |   |   dia_infracao = SEM_4: GRAVE (195.0/106.0)
|   |   periodo_infracao = MANHA: MEDIA (3155.0/1579.0)
|   |   periodo_infracao = TARDE
|   |   |   dia_infracao = SEM_1: MEDIA (492.0/259.0)
|   |   |   dia_infracao = SEM_2: MEDIA (594.0/261.0)
|   |   |   dia_infracao = SEM_3: MEDIA (835.0/398.0)
|   |   |   dia_infracao = SEM_4: GRAVISSIMA (1464.0/883.0)
tip_abordagem = C
|   regiao_infracao = Sudeste
|   |   dia_infracao = SEM_1
|   |   |   periodo_infracao = NOITE: GRAVISSIMA (1455.0/793.0)
|   |   |   periodo_infracao = MANHA: GRAVE (1390.0/765.0)
|   |   |   periodo_infracao = TARDE: GRAVISSIMA (1744.0/908.0)
|   |   dia_infracao = SEM_2
|   |   |   periodo_infracao = NOITE: GRAVISSIMA (1764.0/996.0)
|   |   |   periodo_infracao = MANHA: GRAVE (1491.0/778.0)
|   |   |   periodo_infracao = TARDE: GRAVISSIMA (1539.0/749.0)
|   |   dia_infracao = SEM_3
|   |   |   periodo_infracao = NOITE: GRAVE (2307.0/1282.0)
|   |   |   periodo_infracao = MANHA: GRAVISSIMA (2314.0/1222.0)
|   |   |   periodo_infracao = TARDE: GRAVISSIMA (2737.0/1369.0)
|   |   dia_infracao = SEM_4: GRAVISSIMA (12065.0/5960.0)
|   regiao_infracao = Sul: GRAVISSIMA (38877.0/19502.0)
|   regiao_infracao = Nordeste: GRAVISSIMA (47500.0/20207.0)
|   regiao_infracao = Centro-Oeste
|   |   dia_infracao = SEM_1
|   |   |   periodo_infracao = NOITE: GRAVE (787.0/429.0)
|   |   |   periodo_infracao = MANHA: GRAVISSIMA (999.0/550.0)
|   |   |   periodo_infracao = TARDE: GRAVE (992.0/492.0)
|   |   dia_infracao = SEM_2
|   |   |   periodo_infracao = NOITE: GRAVISSIMA (868.0/491.0)
|   |   |   periodo_infracao = MANHA: GRAVE (1091.0/588.0)
|   |   |   periodo_infracao = TARDE: GRAVISSIMA (993.0/559.0)
|   |   dia_infracao = SEM_3: GRAVISSIMA (6161.0/2969.0)
```

```
|   |   dia_infracao = SEM_4: GRAVISSIMA (10550.0/5083.0)
|   regiao_infracao = Norte
|   |   dia_infracao = SEM_1
|   |   |   periodo_infracao = NOITE: GRAVISSIMA (395.0/203.0)
|   |   |   periodo_infracao = MANHA: GRAVISSIMA (614.0/330.0)
|   |   |   periodo_infracao = TARDE: GRAVE (654.0/370.0)
|   |   dia_infracao = SEM_2: GRAVISSIMA (1981.0/1066.0)
|   |   dia_infracao = SEM_3: GRAVISSIMA (3361.0/1764.0)
|   |   dia_infracao = SEM_4: GRAVISSIMA (6702.0/3261.0)
```

Seguiu-se com a tentativa de apurar resultado com outras tarefas para obtenção de um resultado mais favorável.

### 5.3.2. Agrupamento EM

O algoritmo EM atribui uma distribuição de probabilidade a cada instância que indica a probabilidade de pertencer a cada um dos clusters. O EM pode decidir quantos clusters criar por validação cruzada ou você pode especificar quantos clusters para gerar, ou seja, quando maior o número apresentado, maior a probabilidade de ocorrer o atributo.

Para execução do algoritmo foi realizada as seguintes adaptações no algoritmo: ajustado o número de cluster para 4, sendo realizada o modo para avaliação de classes para cluster em dados de treinamento.

Resultado da execução do algoritmo de cluster por EM apresentou uma margem de erro de 57,34%, extremamente elevada. Avaliando o resultado por cluster, apresentou o seguinte agrupamento:

- Cluster 2, infração Grave – A maior probabilidade de infração ocorrer é sem abordagem, na região Sudeste, na amostra apurada a tendência é de acontecer na última semana do mês, no período da tarde. Em decorrência da amostra ter ocorrido em dezembro, há uma grande probabilidade de ter sido influenciada pelos feriados de final de ano. Sendo necessário efetuar uma nova análise ampliando o período de avaliação.
- Cluster 0, infração Gravíssima – No caso da análise da infração gravíssima, a probabilidade de ocorrer com ou sem abordagem é basicamente a mesma. A maior tendência da infração Gravíssima acontecer na região Nordeste, sendo o segundo maior a região Sul. A maior probabilidade das infrações gravíssima acontecerem pela tarde e também pela manhã. As ocorrências também se concentração na 4ª semana do mês de dezembro, reforçando a necessidade de uma nova análise com a inclusão de novos meses.

| Attribute | Cluster 0 (0.44) | 1 (0.26) | 2 (0.2) | 3 (0.1) |
|---|---|---|---|---|
| tip_abordagem | | | | |
| S | 146731.8159 | 168809.7107 | 132146.6753 | 62313.7981 |
| C | 146030.4624 | 95.1084 | 27.2575 | 5182.1717 |
| [total] | 292762.2783 | 168904.8191 | 132173.9329 | 67495.9698 |
| regiao_infracao | | | | |
| Sudeste | 51561.724 | 92357.2053 | 65358.3717 | 47956.6989 |
| Sul | 79842.6548 | 27098.0461 | 26149.7979 | 4520.5012 |
| Nordeste | 95831.4714 | 31094.478 | 17086.4311 | 6.6195 |
| Centro-Oeste | 44860.1538 | 18352.946 | 23579.2022 | 15013.6981 |
| Norte | 20669.2743 | 5.1436 | 3.13 | 1.4521 |
| [total] | 292765.2783 | 168907.8191 | 132176.9329 | 67498.9698 |
| dia_infracao | | | | |
| SEM_1 | 40394.7069 | 46343.2757 | 4270.773 | 13386.2444 |
| SEM_2 | 42738.4933 | 56373.9803 | 2581.4227 | 13796.1037 |
| SEM_3 | 74919.5178 | 65104.2875 | 516.2178 | 15100.9769 |
| SEM_4 | 134711.5603 | 1085.2756 | 124807.5194 | 25214.6447 |
| [total] | 292764.2783 | 168906.8191 | 132175.9329 | 67497.9698 |
| periodo_infracao | | | | |
| NOITE | 59499.3762 | 8213.717 | 65.5723 | 66312.3345 |
| MANHA | 111990.9228 | 78556.8724 | 63987.444 | 110.7608 |
| TARDE | 121272.9793 | 82135.2297 | 68121.9165 | 1073.8745 |
| [total] | 292763.2783 | 168905.8191 | 132174.9329 | 67496.9698 |

Cluster 0 <-- GRAVISSIMA
Cluster 1 <-- MEDIA
Cluster 2 <-- GRAVE
Cluster 3 <-- LEVE

### 5.3.3. Agrupamento por simples K-Means

A segunda tarefa utilizada para análise dos dados do modelo foi o agrupamento, tentativa de criar conjunto de dados com características similares, desta forma é possível o reconhecimento de padrões e uma análise mais profunda dos dados. Para esta análise foi utilizado o algoritmo k-means.

A análise através do algoritmo evidenciou um percentual elevado de erro: 57,76%. Mesmo com resultado pouco expressivo, é possível verificar a criação de grupos específicos para os diferentes tipos de gravidades. Principais características:

- Gravíssima, maior probabilidade de ocorrência: com abordagem, na região nordeste na terceira semana e no período da tarde.
- Grave: tendência de ocorrer na região sudeste, concentrando-se na primeira semana, período da tarde, mas sem abordagem.

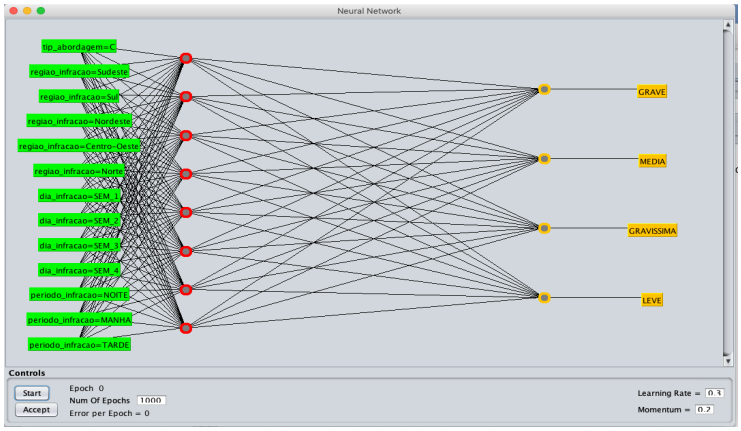
```
Final cluster centroids:
                                Cluster#
Attribute           Full Data          0          1          2          3
                   (661329.0) (335929.0) (186689.0)  (87099.0)  (51612.0)
=============================================================================
tip_abordagem               S          S          S          C          S
regiao_infracao       Sudeste        Sul    Sudeste   Nordeste   Nordeste
dia_infracao            SEM_4      SEM_4      SEM_1      SEM_3      SEM_2
periodo_infracao        TARDE      MANHA      TARDE      TARDE      TARDE


Cluster 0 <-- MEDIA
Cluster 1 <-- GRAVE
Cluster 2 <-- GRAVISSIMA
Cluster 3 <-- LEVE
```

### 5.3.4. Redes Neurais

Após a aplicação de três algoritmos com resultados do Kappa e probabilidade de erro elevada, aplicou-se a técnica de rede neurais segundo Pinheiro, esta técnica implementa padrões de detecção e algoritmos de aprendizado de máquina para construir modelos de predição para bases de dados históricos em larga escala. São utilizadas em um modo de aprendizado não supervisionado para a criação de grupos.

Para adoção da técnica foram ajustados os seguintes critérios:

- GUI = true. Exibição do usuário de interface gráfica.
- SED = 100. Quantidade de neurônios utilizados.
- Traingtime = 1000. Quantidade de treinamento.
- Epoch = 150. Quantidade de interações padronizadas.



O resultado apresentado também foi abaixo do esperado, com resultado da Kappa de 0,3735. Com percentual de erro de 32,34%, também elevado. O resultado da rede neural será os pesos que poderá ser aplicado para predição das próximas infrações. Por exemplo, é possível predizer a gravidade de uma infração, inserindo as informações de tipo de abordagem, região, dia da infração e o período.

```
=== Summary ===

Correctly Classified Instances      447419               67.6545 %
Incorrectly Classified Instances    213910               32.3455 %
Kappa statistic                          0.3735
Mean absolute error                      0.2286
Root mean squared error                  0.3399
Relative absolute error                 79.0523 %
Root relative squared error             89.3851 %
Total Number of Instances           661329

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,000    0,000    ?          0,000    ?          ?        0,620     0,333     GRAVE
                 0,960    0,495    0,727      0,960    0,828      0,541    0,766     0,766     MEDIA
                 0,691    0,139    0,512      0,691    0,588      0,494    0,811     0,485     GRAVISSIMA
                 0,000    0,000    ?          0,000    ?          ?        0,697     0,022     LEVE
Weighted Avg.    0,677    0,311    ?          0,677    ?          ?        0,739     0,607

=== Confusion Matrix ===

      a      b      c      d   <-- classified as
      0  99075  57114      0 |      a = GRAVE
      0 367653  15237      0 |      b = MEDIA
      0  35663  79766      0 |      c = GRAVISSIMA
      0   3215   3606      0 |      d = LEVE
```

### 5.3.5. Word Tags – Nuvem de tags

Para análise dos tipos de multas aplicadas em dezembro foi aplicada a apresentação visual da nuvem tag, onde os resultados com maiores incidências será a de maiores tamanhos. No caso das análises das infrações, ao invés da descrição dos dados foi adotado o código da multa, para melhorar a visualização.

As ocorrências com maiores incidências foram:

- 74550 – Transitar em velocidade superior à máxima permitida em até 20%, gravidade média.
- 74630 – Transitar em velocidade superior à máxima permitida em mais de 20% até 50%, gravidade Grave.
- 59670 – Ultrapassar pela contramão linha de divisão de fluxos opostos, contínua amarela, gravidade gravíssima.
- 65992 – Conduzir o veículo registrado que não esteja devidamente licenciado, gravidade gravíssima.

As três maiores ocorrências são referentes a velocidade transitando na via, condução perigosa como ultrapassagem irregular, evidenciando a má condução dos motoristas. Reforçando a necessidade de ações educativas, punitivas.

## 6. Conclusão

A adoção da mineração de dados é uma excelente forma de avaliar os dados históricos para descobrimento de padrões e adoção de ações direcionadas. No entanto, para um resultado confiável e assertivo, é necessário que a base de dados seja confiável e possua dados que descrevam que forma mais completa possível as ações do passado.

No caso da base de dados disponibilizado pela Polícia Rodoviária Federal são dados colhidos de forma manual, muitos dos atributos grande volume de missing, dados cadastrados de forma não padronizadas e informações com falta de dados de referência da unidade de valor. Para melhorar a qualidade dos dados coletados e possibilitar a criação de modelos preditivos no futuro, recomenda-se treinamento ou adoção de ferramentas de fiscalização ou coleta de dados com informações uniformizadas.

No estudo realizado, em decorrência da limitação sistêmica, optou-se por analisar a base referente a um mês de infrações. Devido a restrição da base, alinhado com o problema de qualidade detectado, houve uma queda significativa do resultado das técnicas adotadas impactando na qualidade do resultado. Ainda assim, foi possível detectar alguns padrões interessantes que poderão ser confirmados após a execução de uma técnica com uma base mais completa.

A maior parte das análises aponta a região nordeste como a maior probabilidade de área de ocorrência de infrações gravíssima, concentradas no período da tarde. Nos três modelos a região Nordeste representa uma grande concentração de infrações gravíssima. A região Sudeste concentra-se grande número de incidências de ocorrência grave. Na análise de mineração de texto foi possível detectar quais o maior número de multas aplicadas.

O conhecimento obtido no estudo sinaliza quais as regiões de incidência das infrações mais graves e qual período que elas ocorrem, desta forma é possível adotar ações como definir quais áreas precisam de maior policiamento, como o nordeste e sudeste. Pode-se também direcionar as campanhas educativas, para estas áreas, em especial destinada a campanha de redução da velocidade.  No período de feriados, como a 4ª semana de dezembro, há um incremento de ocorrências, portanto, vale reforçar as campanhas educativas já vinculadas nas imprensas locais, além de penalização alternativas para os condutores reincidentes.

## 7. Bibliografia


Oliveira, Diego Elias; Oliveira, Grimaldo Lopes de. BI Como Deve Ser – O Guia Definitivo. 2ª ed. Salvador: 2016.

Pinheiro, Carlos André Reis. Inteligência Analítica – Mineração de Dados e Descoberta de Conhecimento. Rio de Janeiro: Editora Ciência Moderna Ltda., 2008.

Dados Abertos de Infrações. Polícia Rodoviária Federal, ministério da justiça e segurança pública < https://www.prf.gov.br/portal/dados-abertos/infracoes> Acessado em 16/04/19.

Relação das Infrações. Governo do Município de Foz de Iguaçu. < http://www.pmfi.pr.gov.br/Portal/VisualizaObj.aspx?IDObj=11179> Acessado em 16/04/19.

Relação das Infrações e Multas. Detran de Tocantins. <https://central3.to.gov.br/arquivo/320049/> Acessado em 16/04/19.

Tabela de infrações. Superintendência de Trânsito e Transporte de Salvador - TRANSALVADOR <http://www.transalvadorantigo.salvador.ba.gov.br/arquivos/tabela_de_infracoes.pdf> Acessado em 16/04/19.

Montagem e treinamento de redes neurais perceptron para identificação de proteínas efetoras. Gabriel S. OLIVEIRA; Leonardo F. MOREIRA; Claudinei O DOTTO; Gustavo J. SILVA < https://jornada.ifsuldeminas.edu.br/index.php/jcmch4/jcmch4/paper/viewFile/3013/2384> Acessado em 19/04/19.

Rede Neural Artificial Multilayer Perceptron para previsão de tendências de fechamento do IBOVESPA: índices que influenciam o IBOVESPA <https://www.researchgate.net/publication/299543593_Rede_Neural_Artificial_Multilayer_Perceptron_para_previsao_de_tendencias_de_fechamento_do_IBOVESPA_indices_que_influenciam_o_IBOVESPA> Acessado em 19/04/19.