Instituto Atlântico

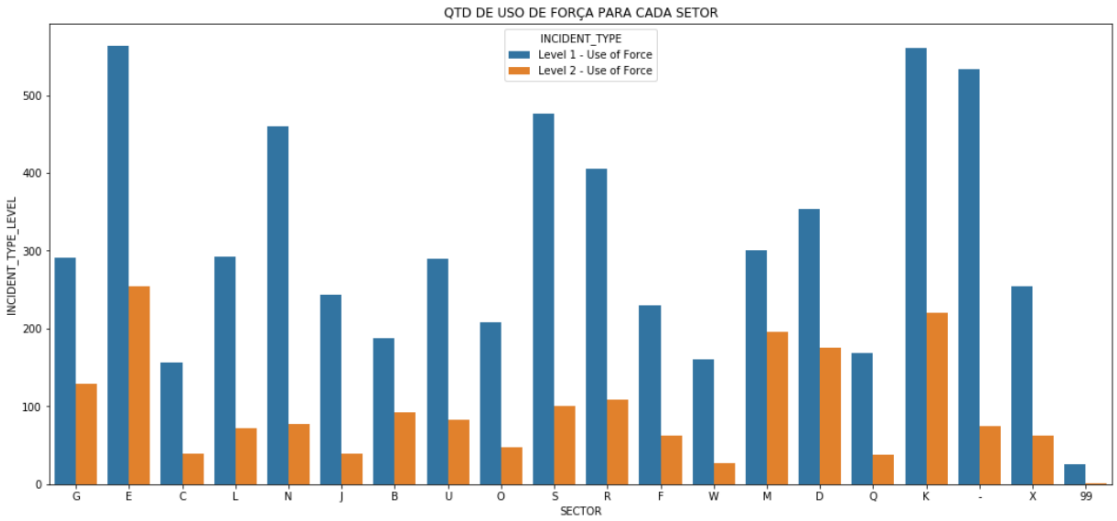
Felipe Rodrigues Dieb

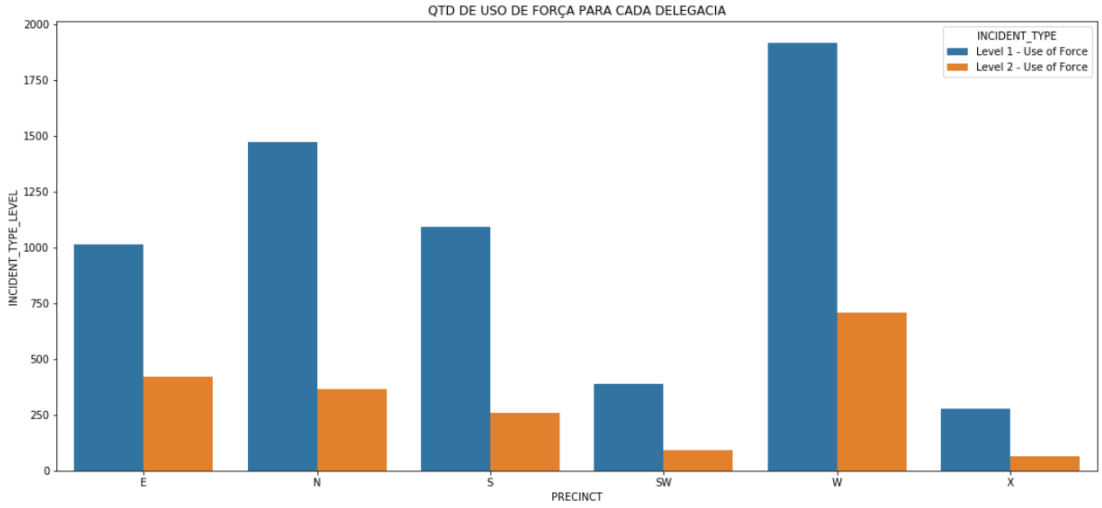
Questão 01

Como é a distribuição do uso de força dentre as delegacias e os setores? Em cada setor, qual o BEAT com maior número de incidentes? Apresente também o ranking dos setores segundo o percentual de incidentes “Level 2” em relação ao total de incidentes do respectivo setor.

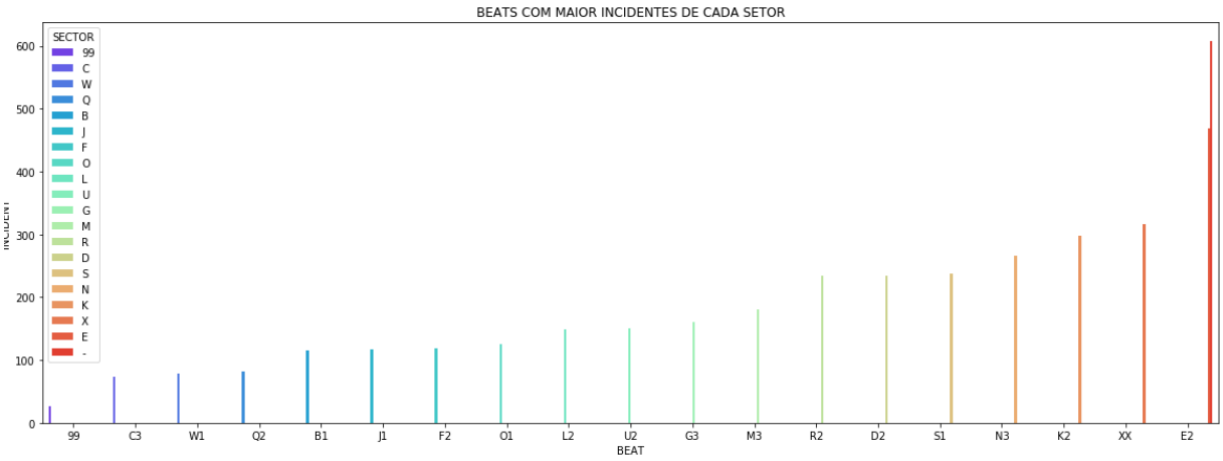
Comentário:

Os gráficos exibidos abaixo mostram a distribuição do uso de força para cada setor e para cada delegacia. Através dessa visualização, nota-se que o uso de força Nível 1 é o mais frequente em todos os cenários.

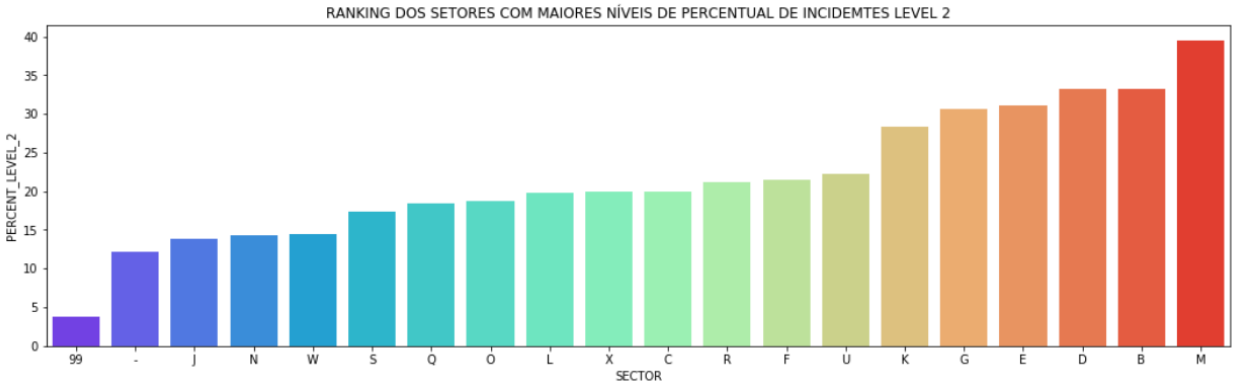




No gráfico abaixo são apresentados os BEATS que possuem maior incidência em cada setor, podendo verificar a quantidade de incidente por cada BEAT selecionado.



O gráfico a seguir descreve o nível de percentual de incidentes ‘Level 2’ de cada setor. A partir desta análise, verifica-se que o setor M possui o maior percentual.



Questão 02

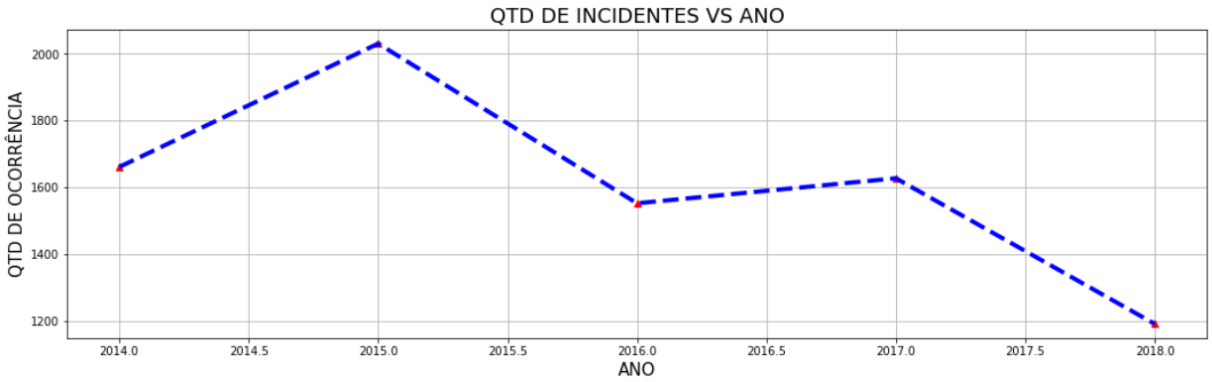
Com relação à distribuição dos incidentes no tempo, é possível encontrar picos ou linhas de tendência dentro dos dias, dos meses, das semanas ou dos anos?

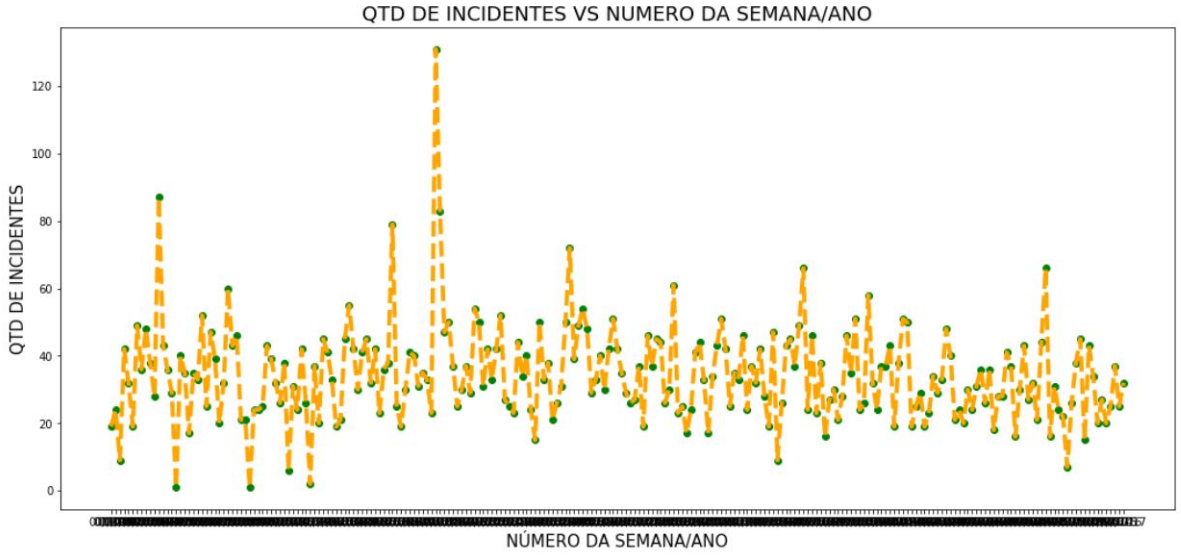
Comentário:

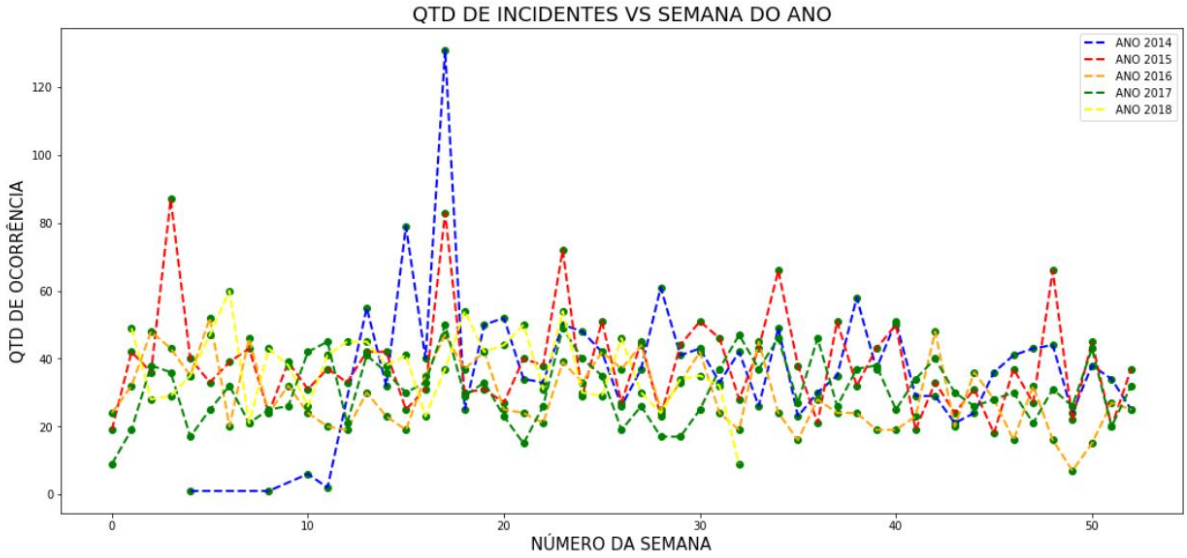
Foi feita uma análise através do gráfico de linha, utilizando a quantidade de incidentes por ano, e observou-se uma tendência decrescente. Isso significa que, a cada ano, ocorre uma redução no número de incidentes. De 2015 para 2018 a redução chega a aproximadamente 41%, passando de 2031 incidentes para 1190.

Ao analisar a semana de cada ano, confirma-se uma pequena tendência de redução no decorrer dos anos, sendo possível observar a existência de um outlier.

Os gráficos abaixo foram desenvolvidos para a análise dessa questão:







Questão 03

A polícia deseja dar início a uma investigação interna para verificar se existem policiais excessivamente violentos. No entanto, o prazo para o término desta investigação é bastante limitado. Elabore um script capaz de elencar os policiais em ordem decrescente de chance de violência excessiva com base no número de incidentes dos quais eles participaram.

Comentário:

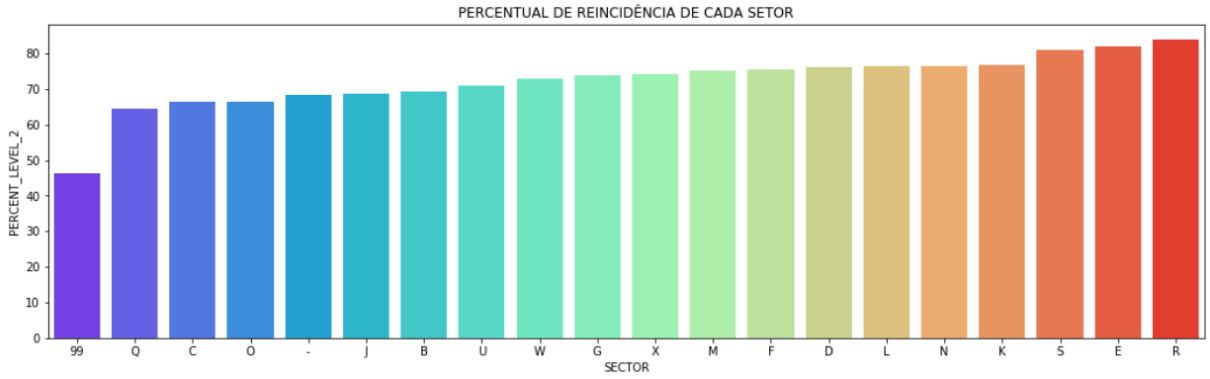
Script está no código fonte.

Questão 04

Uma métrica interessante para a polícia é o grau de reincidência por parte dos civis. Apresente o percentual de casos reincidentes em relação ao total de incidentes em cada setor e verifique se há correlação entre esta métrica e o percentual de incidentes "Level 2" calculado na questão 1. Que interpretação pode ser dada a este resultado?

Comentário:

O gráfico desenvolvido auxilia na análise da quantidade de reincidentes em cada setor. Através dele é possível observar que o setor 99, o qual possui menor número de reincidentes, já apresenta uma taxa muito alta, de aproximadamente 45%.



Questão 05

A liderança do Departamento de Polícia de Seattle manifestou o interesse em uma aplicação que classifica os incidentes em "Level 1" ou "Level 2" com base em outras colunas da tabela e lhe requisitou um parecer sobre esta proposta. Descreva os desafios envolvidos, enumere fatores que fomentem a criação deste classificador e sugira um modelo estatístico para executar esta tarefa, justificando a sua escolha.

Comentário:

Alguns desafios foram encontrados durante a análise, como por exemplo, quais colunas fariam parte do treinamento, no entanto ao ser observado um baixo ganho de informação nas colunas, sendo o valor mais alto de 30%, optou-se por treinar com todas as informações possíveis para evitar a perda da qualidade no modelo.

Após a seleção das colunas verificou-se quais algoritmos seriam escolhidos. Devido à distribuição não linear dos dados, decidiu-se utilizar os algoritmos *RandomForest* e *RegressionLogistic*, uma vez que eles se adaptam bem a dados não lineares.

O modelo *RandomForest*, se baseia no algoritmo de *DecisionTree*. Logo, ele utiliza um conjunto de árvores de decisões com o objetivo de reduzir a possibilidade de *overfitting*. Esse modelo possui duas técnicas importantes, o *BootStrap* e o *Bagging*.

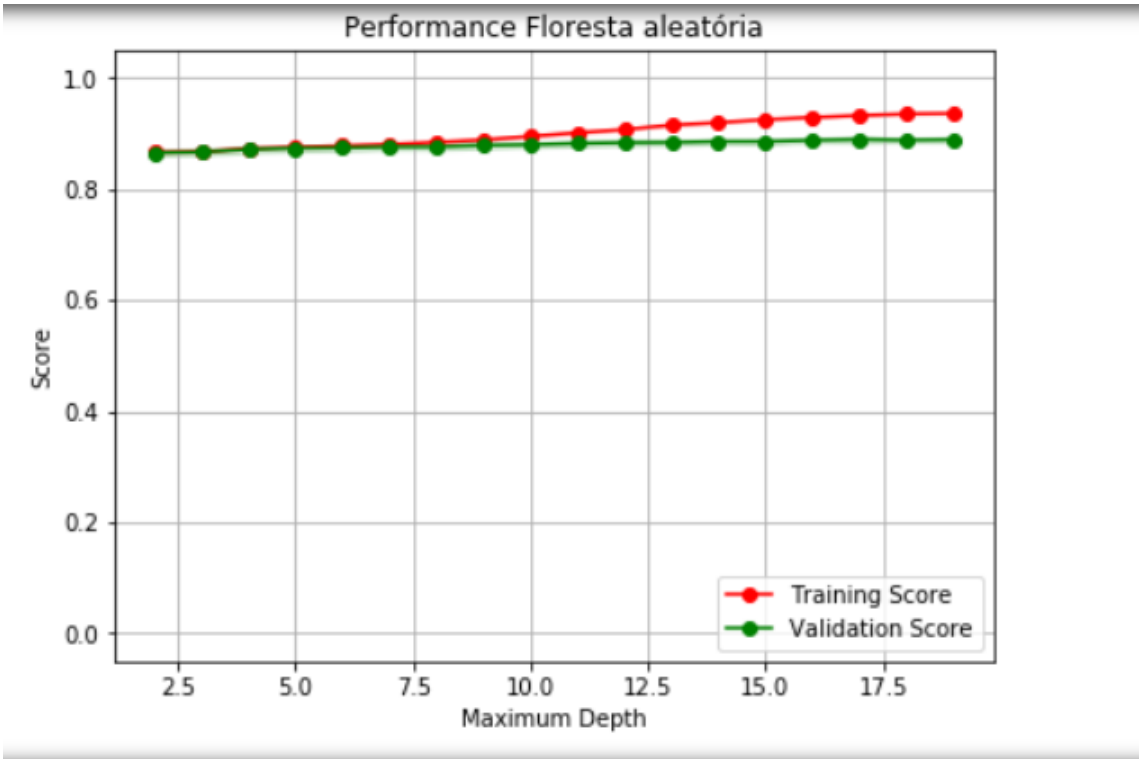
O *Bootstrap* é um método de geração de amostras, as quais serão utilizadas nas possíveis árvores do modelo, sendo que cada amostra possui a mesma quantidade de dados. O *Bagging*, utiliza as amostras separadas pelo *Bootstrap* para o treinamento das árvores, objetivando fazer a média dos resultados de cada árvore para calcular a predição final do modelo.

A Regressão Logística é um modelo no qual seu resultado consiste em probabilidade, ou seja, o valor de saída é entre 0 e 1. Isso ocorre porque a função utilizada é de achatamento, onde o resultado se dá por meio de probabilidade. Esta função é chamada de função logística ou sigmoide.

O modelo que mais se adequou ao problema foi o *RandomForest*, obtendo 77% de *f1-score* e 80% de acurácia, tendo sido melhor avaliado nas duas métricas em relação à Regressão Logística. Além disso, ao analisar a matriz de confusão dos dois modelos, percebeu-se que houve bastante erro para a classificação de Level 2 por meio da Regressão Logística. Diante disso, o modelo teve um resultado um pouco aceitável, devido à base consistir, em sua maior parte, das informações do Level 1.

Ao analisar cada modelo, comparando as métricas com a base de treinamento e teste, é visto que o *RandomForest* tende a ter um *overfitting* à medida que a profundidade aumenta. Isso se comprova, ao observar que o modelo tem uma tendência de melhoria na base de treinamento, enquanto na base de validação possui uma tendência de piora.

O gráfico a seguir refere-se à representação da análise anterior:


Performance Floresta aleatória
Score
Maximum Depth
Training Score
Validation Score
1.0
0.8
0.6
0.4
0.2
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5