Relatório do Uso da Força por Parte da Polícia – Seattle

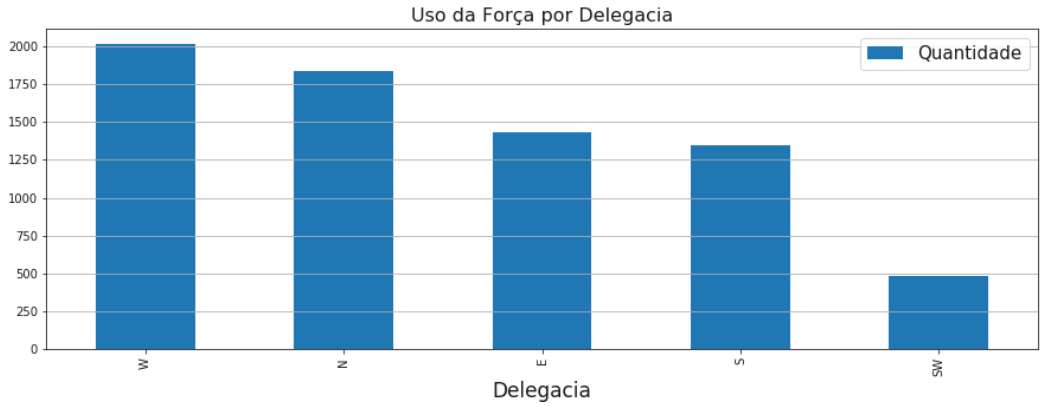
Clairton Menezes

Análise Exploratória de Dados

1. 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.

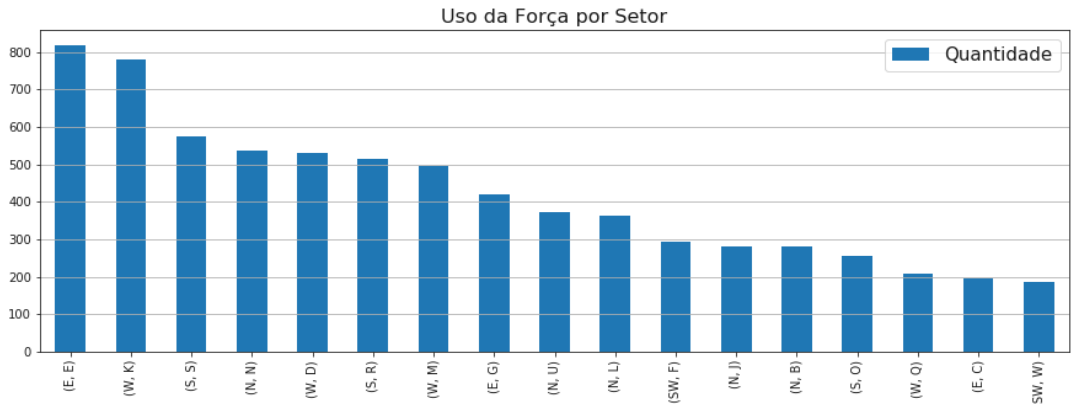
- Uso da força por delegacia

O gráfico a seguir revela que as delegacias de West e North são as que tem maiores números de incidentes. Mas o destaque vai para a delegacia de Southwest, que é a que tem, por bastante vantagem, o menor número de incidentes.



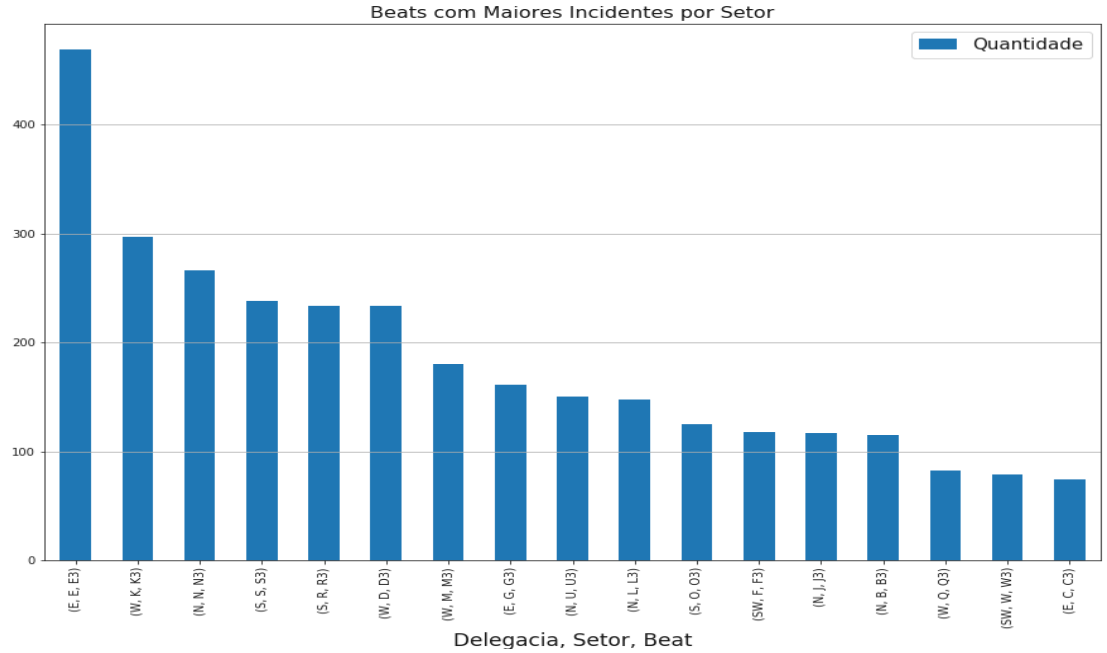
- Uso da força por setor

O próximo gráfico revela que o maior número de incidentes, fica no setor de E da delegacia de East. A delegacia West, a que concentra o maior número de incidentes aparece em segundo lugar com o setor K. Seguindo o gráfico anterior a delegacia de Southwest, setor W, é a que aparece com menor número de incidentes.



- Uso da força por beat

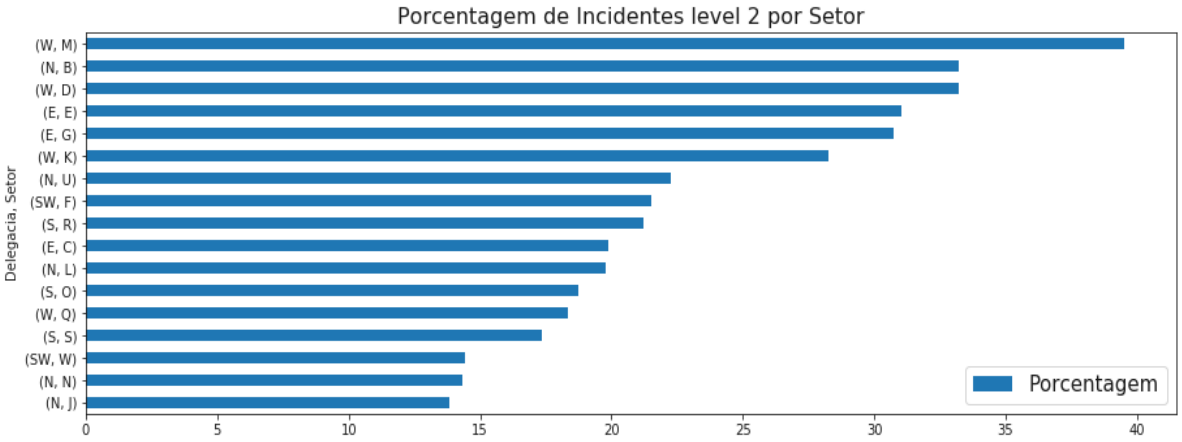
Neste gráfico podemos perceber claramente que o beat E3, é o que puxa o setor E para o topo, como vimos no gráfico anterior de setores. A delegacia Southwest, setor W e beat W3 aparece na penúltima posição desta vez. Contudo é interessante notar que a delegacia East, setor C beat C3 aparece em último lugar neste gráfico.



- Rank por Setor

  Observando esse gráfico podemos perceber que os incidentes level 2 são minoria, em alguns casos chegam a menos de 15% do total do setor. Entretanto, os que tem as maiores porcentagens apresentam mais que duas vezes as menores. Vale notar que a

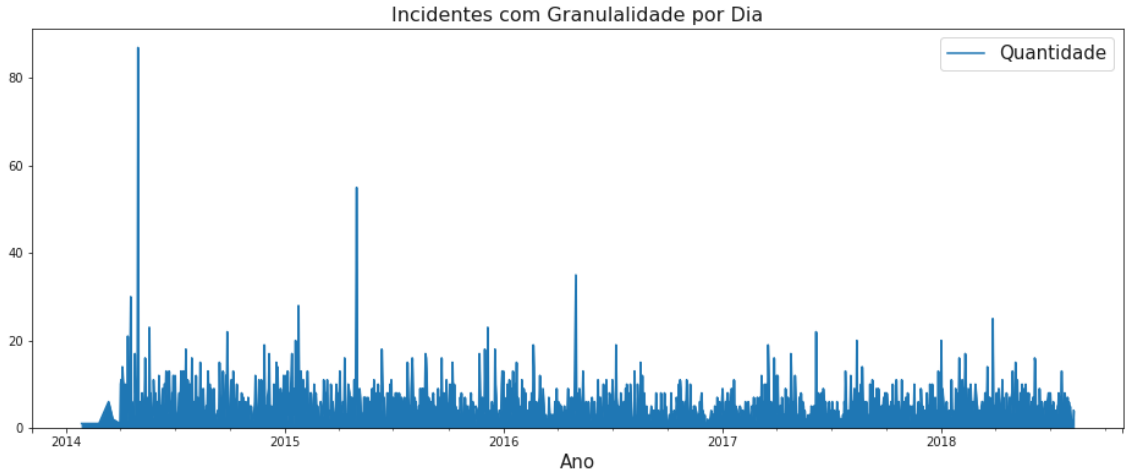
granularidade West, a que tem os maiores números de incidentes, está em primeiro lugar.



2. 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?

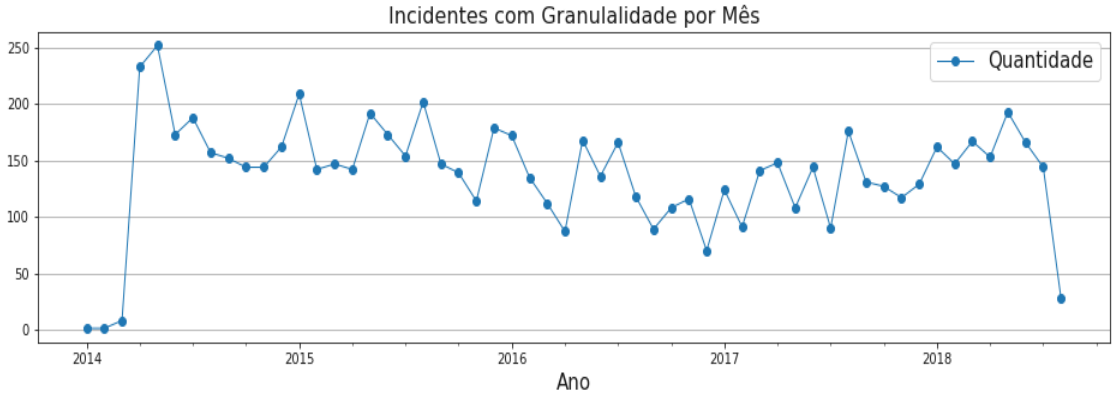
- Incidentes por dia

Este gráfico tem uma granularidade muito alta, contudo podemos perceber que os picos de quantidade de incidentes estão diminuindo ao longo do tempo, ficando cada vez menos frequente.



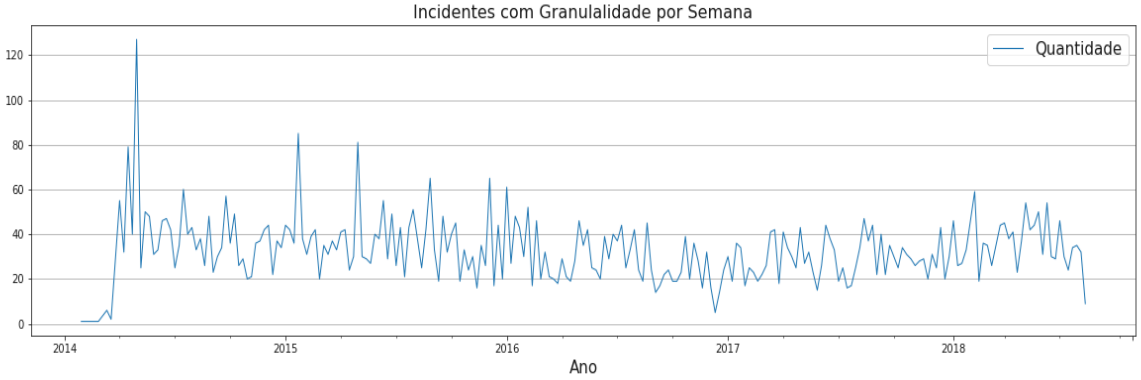
- Incidentes por ano-mês

Podemos perceber neste gráfico que existe um aumento de incidentes no final de todo ano e logo depois há uma diminuída.



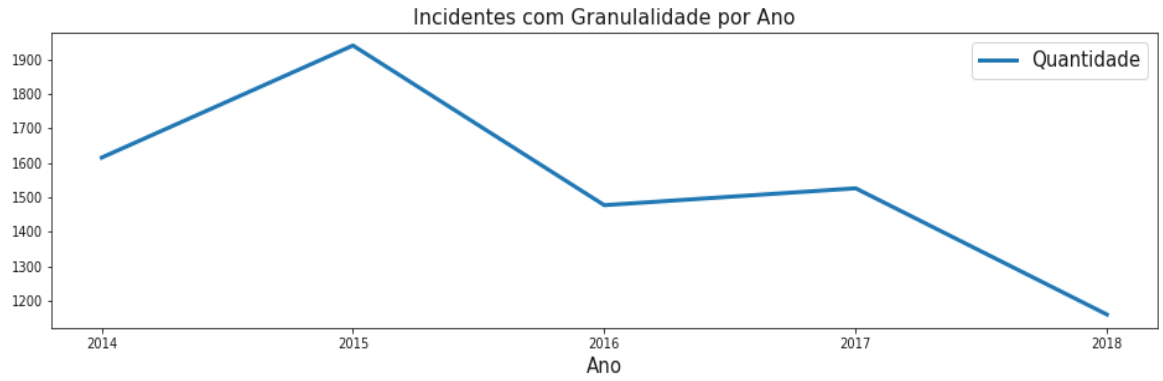
- Incidentes por semana

No seguinte gráfico o ano de 2017 revela ser o mais estável, quase sempre mantendo a quantidade de incidentes por semana entre 20 e 40. Também podemos perceber que o ano de 2018 não está completo.



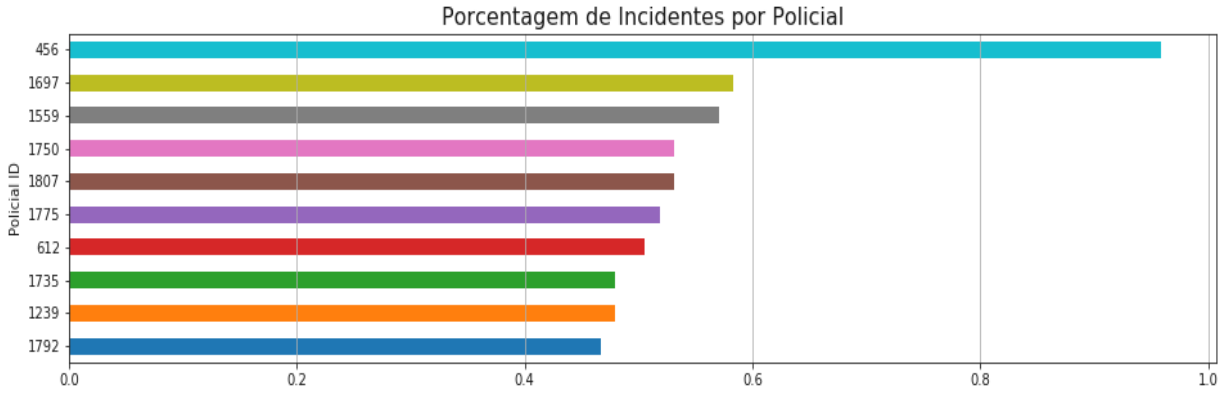
- Incidentes por ano

Neste gráfico é mais evidente a redução dos incidentes ao longo dos anos, porém não temos dados suficientes do ano de 2018 para ter uma conclusão mais precisa. É interessante notar que o ano de 2015 ficou em bastante destaque pelo número de incidentes.



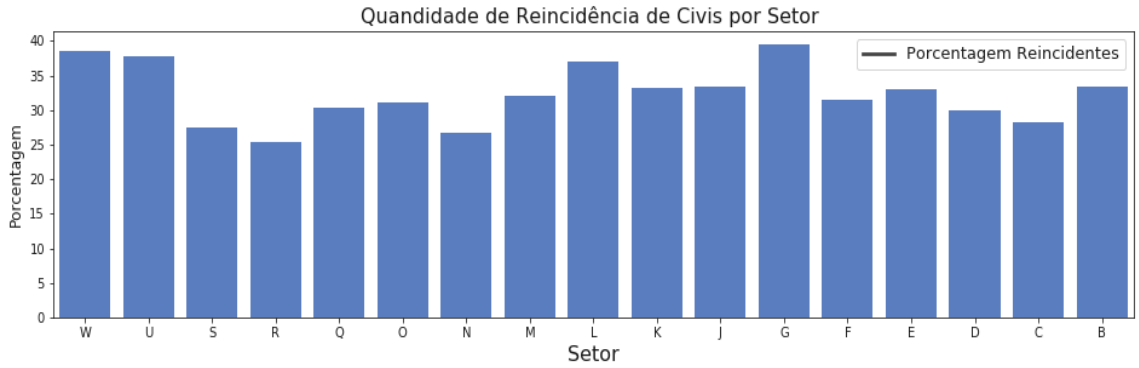
3. A polícia deseja dar início a uma investigação interna para verificar se existem policiais excessivamente violentos. No entanto, o prazo para o término desta investigação é bastante limitado. Elabore um script capaz de elencar os policiais em ordem decrescente de chance de violência excessiva com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

Podemos perceber por este gráfico que o Policial 456 é o que tem maior número de incidentes, chegando a quase 1% do total. Enquanto, o segundo colocado não chega a 0,6%.

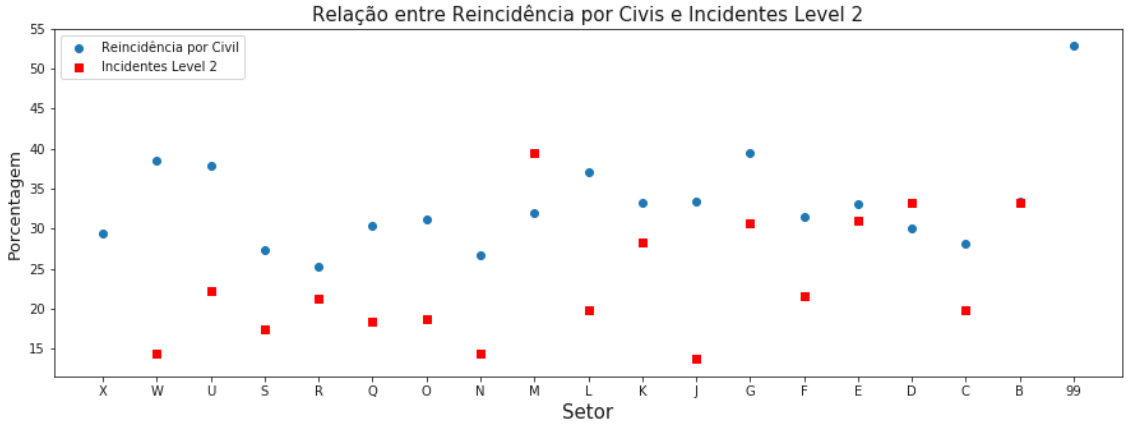


4. Uma métrica interessante para a polícia é o grau de reincidência por parte dos civis. Apresente o percentual de casos reincidentes em relação ao total de incidentes em cada setor e verifique se há correlação entre esta métrica e o percentual de incidentes "Level 2" calculado na questão 1. Que interpretação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

No gráfico a seguir podemos notar que existem diferenças entre os setores na porcentagem de reincidentes porém não são diferenças muito acentuadas.



Como podemos perceber no gráfico a seguir: não conseguimos achar uma correlação entre a reincidência por civis e incidentes level 2. A porcentagem de reincidentes por civis é normalmente maior que a porcentagem de incidentes level 2.



Aprendizado de Máquina

5. 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. P.S.: Sua justificativa deve conter explicação teórica de ao menos dois algoritmos, um benchmark destas soluções candidatas de tempo e performance, os experimentos e análise do bias variance threshold.

A criação de um classificador para determinar se um incidente é level 1 ou 2 tem alguns desafios, porém, também tem vantagens notórias.

Como desafios podemos enumerar:

1. Confiabilidade dos dados, é extremamente importante que os dados usados sejam precisos e livres de segundas intenções. Se os dados de entrada não forem precisos, os resultados também não serão.
2. Testar diversos algoritmos e configurações para chegar no modelo que seja mais eficaz é uma parte importante. Para chegar no melhor modelo, que mais reflete a realidade, devemos submeter os dados de entradas aos mais diversos algoritmos e suas configurações.
3. É importante ressaltar que nenhum modelo é perfeito e pode acontecer de alguma predição esteja errada.

Como vantagens:

1. O processo de definição de tipo de incidente será mais rápido e automático.
2. Eliminação do erro humano ou possíveis definições subjetivas sobre o tipo de incidente.
3. Aprimoramento contínuo, com o passar do tempo vamos ter mais dados e dados mais diversos para aperfeiçoar o modelo.

Análise dos experimentos:

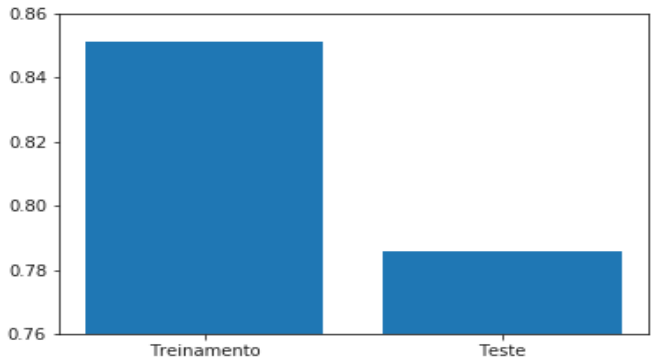
- KNN

KNN, K Nearest Neighbors, usa um conjunto de pontos próximos ao ponto alvo, vizinhos, para determinar a classe a que esse ponto pertence. Podemos definir quantos pontos serão usados para classificar o ponto. Não existe um número padrão de vizinhos, cada problema tem a sua particularidade e, através de testes, podemos encontrar o número que vai fornecer a maior acurácia.

Para o cálculo da distância entre os pontos vizinhos e o ponto a ser classificado, podem ser usadas várias métricas de distâncias, porém a distância euclidiana é a mais comum.

Underfitting e overfitting

Podemos perceber que os dados de treinamento tiveram um melhor desempenho, mas essa diferença foi pequena, revelando um pequeno overfitting.



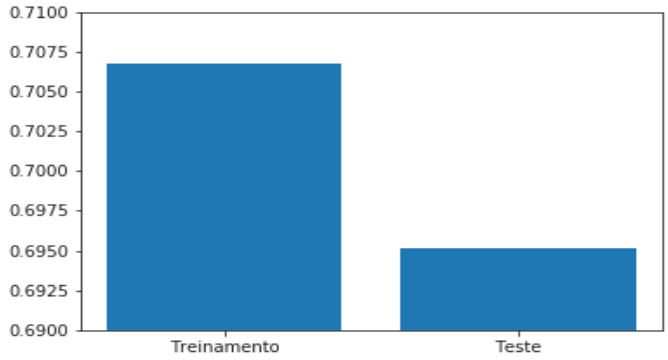
- Naive Bayes

Naive Bayes se baseia no teorema de Bayes e assume que as características do conjunto dos dados são independentes, ou seja, eles não estão relacionados uns com os outros. O teorema de Bayes utiliza informações de eventos anteriores para calcular a probabilidade de um evento acontecer.

Apesar de ter características simples, o Naive Bayes é conhecido pelo seu desempenho em aplicações do mundo real, como classificação de documentos e filtro de spam. Além de ser rápido no aprendizado e na classificação comparado com outros classificadores.

Underfitting e overfitting

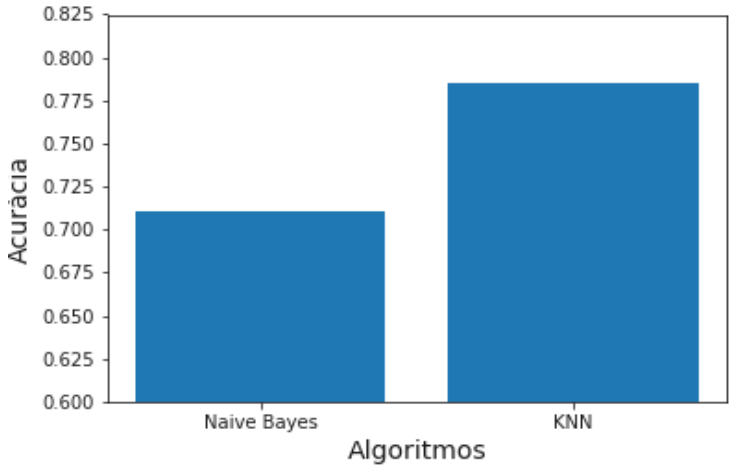
Diferentemente do KNN a acurácia dos dados de testes foi maior que do dado de treinamento, ou seja aconteceu um underfitting. Contudo, a diferença não foi muito grande entre as acurácias.



- Comparando os resultados do Naive Bayes e do KNN

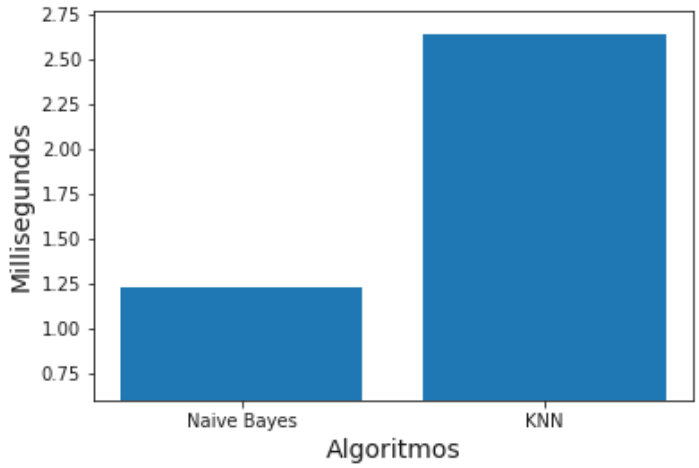
Acurácia

O KNN teve um melhor desempenho na acurácia, em torno de 8% de diferença



Tempo de Processamento

O Naive Bayes foi bem mais rápido que o KNN, mais que o dobro.



Analisando os resultados obtidos com o Naive Bayes e o KNN podemos concluir:

1. Deletando linhas de dados faltosos, em geral há um melhor desempenho nos algoritmos, mesmo com a significativa redução do conjunto de dados.
2. Entre os dois algoritmos, o KNN foi o que obteve teve melhor acurácia, tanto com os dados substituídos quanto com os apagados. O Naive Bayes foi bem mais rápido do que o KNN, contudo o tempo de processamento de ambos não foi elevado, ainda que o KNN tenha demorado mais que o dobro. Levando isso em consideração e também o fato que a acurácia é um fator muito importante, o KNN é o algoritmo escolhido, pois tem o modelo com melhor custo-benefício.