Nanodegree Engenheiro de Machine Learning

Proposta de projeto final

Guilherme Thiago Gomes dos Santos

09 de abril de 2018

Proposta

Histórico do assunto

Desde o meu conhecimento de estudos de análises sobre big data, uma coisa que sempre me veio a mente foi o estudo de casos para a manutenção e melhoria dos nossos sistemas de saúde atuais.

Podemos utilizar esses estudos como uma poderosa ferramenta para entender melhor como algumas doenças afetam a nossa sociedade, como forma de propagação, alvos frágeis, entre outras informações. Com isso, nós podemos adquirir mais informações para entender e conseguir prever a evolução de uma possível epidemia e travar uma solução antes que o problema dissemine ainda mais.

Descrição do problema

Minha proposta de projeto baseia-se em obter mais informações a respeito de infecção e disseminação de um problema que assolou fortemente o Brasil e diversos outros países nos últimos anos: O vírus Zika.

Conjuntos de dados e entradas

Através de dados obtidos pelo CDC da América do norte (Centers for Disease Control and Prevention, ou Centro de Controle de doenças e prevenção, em

português), baixado no site kaggle.

- http://www.kaggle.com.

O dataset apresenta metadados sobre o país de infecção, data de reporte da informação, tipo de infecção, entre outros.

Descrição da solução

Com o dataset citado acima, realizarei uma análise sobre o problema, para identificar padrões nos dados que possam nos ajudar em hipotéticas tomadas de decisão, como disseminação da doença e como tipos de infecções e efeitos colaterais ocorreram. As análises serão realizadas em uma visão global, separada por país, e local, especificamente o Brasil.

Também procuro analisar sobre quais tipos de enfermidades a doença causou por região, como Microcefalia, e as mortes por localidade e período.

Modelo de referência (benchmark)

Utilizarei alguns modelos para testar a eficiência deles através do score e do tempo gasto para realizar o treino, o teste e a predição.

Dentre os modelos, utilizarei KNNeighbors, Arvores de decisão e SVM.

Métricas de avaliação

Tentarei fazer que meus modelos de predição analisem qual é a chance de uma nova infecção ocorrer em qual lugar e determinar a possível gravidade dela.

Design do projeto

Primeiramente, irei fazer o download do dataset através do site do kaggle:

- https://www.kaggle.com/cdc/zika-virus-epidemic

Após baixar o dataset, irei adequar os dados para que fiquem melhores de serem tratados, ajustando colunas de enumeradores para que fiquem melhor de

serem analisados.

Em seguida, realizarei algumas análises gráficas sobre os dados gerais, que englobam os países e os estados de cada país. Vou preparar um teste de performance entre os 3 modelos citados acima, e selecionando o que melhor se adequa a nossa realidade.

Na etapa seguinte, irei filtrar os dados par somente trabalhar com informações sobre o Brasil, gerando análises gráficas sobre os dados e utilizando o modelo que teve melhor performance nos dados gerais.