Sou Alessandra de assis barbosa e como trabalho de conclusão do curso de pós-graduação latu sensu em ciência de dados e big data da puc minas conduzi o estudo que teve por propósito avaliar a capacidade de previsão da expectativa de vida da população utilizando algoritmos de machine learning a partir de informações socioeconômicas para isso foram implementados modelos supervisionados de aprendizado de máquina.

O objetivo deste trabalho é realizar uma análise exploratória que permitirá de individuar os fatores correlacionados a redução ou aumento da expectativa de vida, identificar um modelo matemático que possa ser usado para a previsão da expectativa de vida, comparar a performance de vários modelos de machine learning e a realização de cenários que possam auxiliar na tomada de decisões de investimentos público e privado.

O objetivo maior deste projeto não é somente prever a expectativa de vida nos países mas identificar fatores diretamente vinculados**.**

**Além disso esse trabalho também visa responder as seguintes hipótese:**

●Os vários fatores de previsão escolhidos inicialmente realmente afetam a expectativa de vida?

● Quais são as variáveis de previsão que realmente afetam a expectativa de vida?

●Um país com expectativa de vida menor (<65) deve aumentar seus gastos com saúde para melhorar sua expectativa de vida média?

●Como as taxas de mortalidade de bebês e adultos afetam a expectativa de vida?

●A expectativa de vida tem correlação positiva ou negativa com hábitos alimentares, estilo de vida, exercícios, fumo, bebida alcoólica etc.

● Qual é o impacto da escolaridade na expectativa de vida dos humanos?

●A expectativa de vida tem uma relação positiva ou negativa com o consumo de álcool?

●Países densamente povoados ou altamente populosos tendem a ter menor expectativa de vida?

● Qual é o impacto da cobertura de imunização na expectativa de vida?

●Países mais poluídos apresentam uma expectativa de vida menor?

As informações utilizadas neste projeto são dados dos anos 2010 à 2015 e foram coletados em diferentes fontes: órgão mundial da saúde e das nações unidas, dados do IBGE, organização Our World in Data e API de geolocalização. São dados de 193 países que foram subdivididos em seis grandes áreas: fatores relacionados à economia, fatores ambientais, fatores demográficos, fatores de mortalidade, fatores relacionados à imunização e fatores relacionados à saúde

Essa fase do projeto foi dividida em várias seções de forma a descrever com cura os dados obtidos.

A criação de features uma das técnicas de feature engineering envolve derivar novas features das existentes

percentual de mulheres e emissão de gás por tamanho da população. A ideia era ajustar a taxa de emissão em cada país e testar a hipótese que países com um percentual maior de mulheres tem uma expectativa de vida maior.

O database final contém 32 colunas e 2938 registros .

A análise descritiva dos dados ajuda a entender melhor a natureza dos dados, ver melhor os erros e entender melhor o processo e auxilia no tratamento dos dados e na escolha do algoritmo.

No resumo da tabela e histogramas dos dados numéricos podemos notar que a maioria dos dados tem como mediana NaN devido a dados faltantes que impossibilitaram o cálculo. Vemos por exemplo o comportamento da variável resposta e a média da expectativa de vida dos anos 2000 a 2015 (69,22 anos) e a sua amplitude (52,70) . Isso indica que existe uma disparidade muito grande entre o mínimo e o máximo da expectativa de vida. Esse comportamento se observa em outras variáveis.

Pode-se notar também uma kurtosis e um Skew perto do zero o que indica uma aproximação à curva normal nos dados. Se è verifica também um possível erro na coleta dos dados de GDP que indica como mínimo um Gdp muito baixo para uma nação e porcentagens superior a 100%.