SUICIDE RATE PREDICTION WITH MACHINE LEARNING

SHREYA GOPAL SUNDARI
MACHINE LEARNING & DATA ANALYSIS
PROF. TRAVIS MILLBURN

UNIVERSITY OF NEW HAVEN
1920
CONNECTICUT

INTRODUCTION:

- Suicide is a serious public health problem.
- The World Health Organization (WHO) estimates that every year close to 800 000 people take their own life, which is one person every 40 seconds and there are many more people who attempt suicide.
- Suicide occurs throughout the lifespan and was the second leading cause of death among 15-29-year-olds globally in 2016.
- The objective of this project is to predict the suicide rates using Machine Learning algorithms and to analyzing significant patterns features that result in increase of suicide rates globally.
- The project is done on Google Colaboratory.

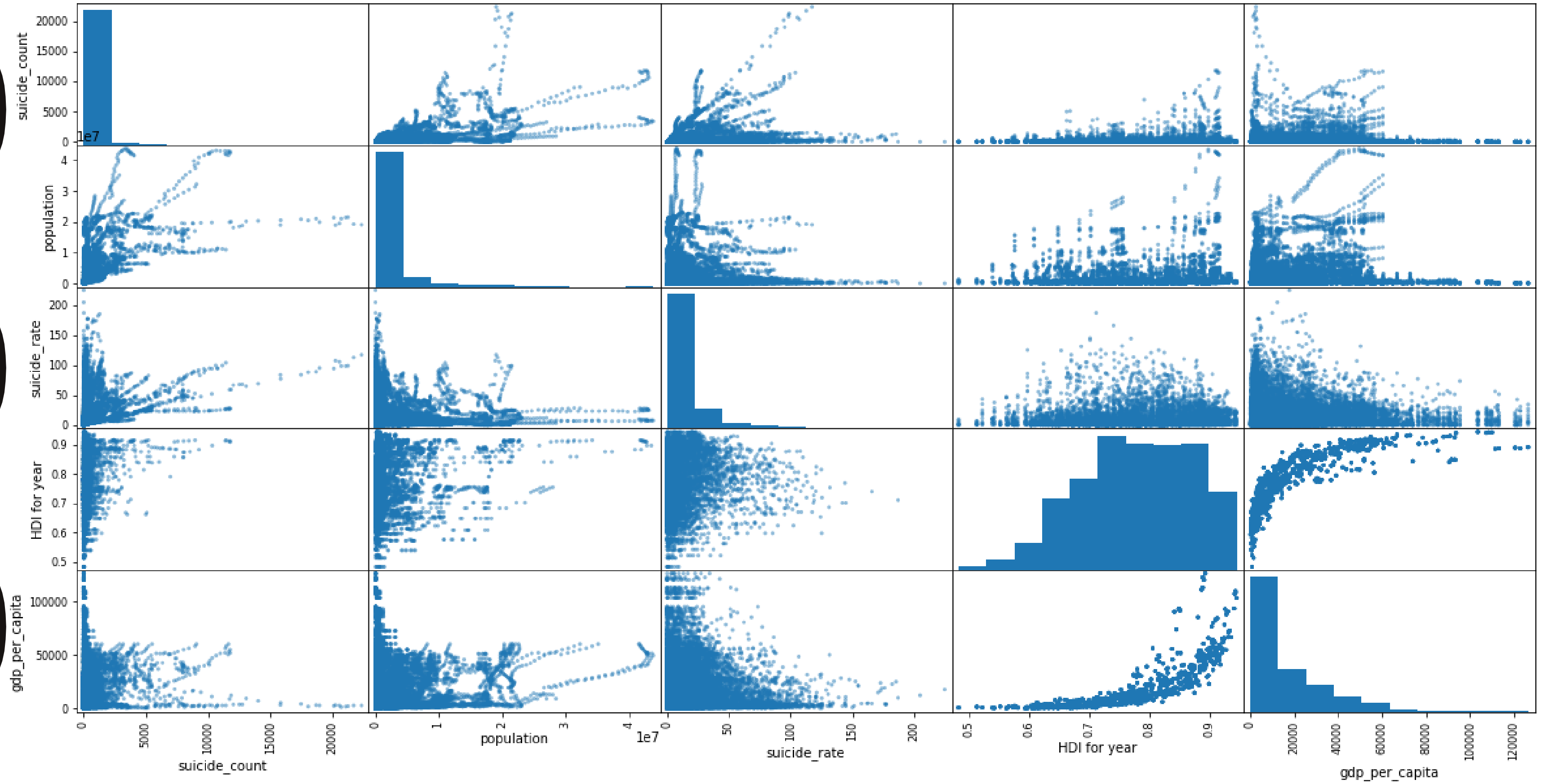
DATASET DETAILS:

- The dataset is from Kaggle. This is a compiled dataset pulled from four other datasets linked by time and place from year 1985 to 2016.
- The source of those datasets is WHO, World Bank, UNDP and a dataset published in Kaggle.
- It has 27820 samples and 12 features.
- The features in the dataset are:
  - country, year, sex, age group, country-year, generation (based on age grouping average).
  - count of suicides, population, suicide rate, HDI for year, gdp_for_year, gdp_per_capita.
- The number of countries in the data set are 101.

DATASET DETAILS (CONT):

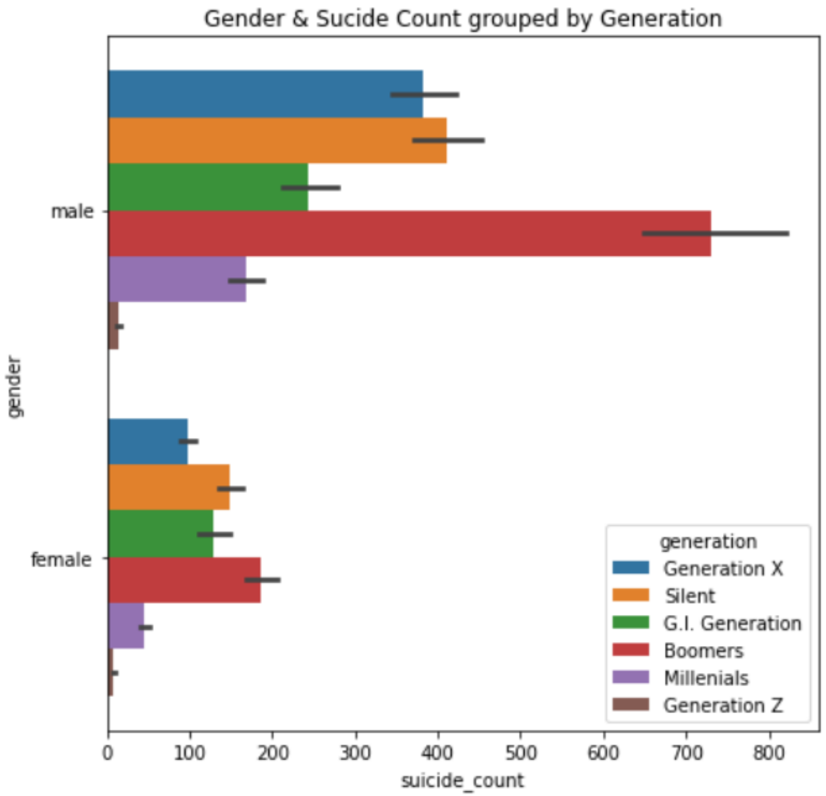
- The age of a person is categorized into 6 age groups.
  - 75+ years
  - 55-74 years
  - 25-34 years
  - 35-54 years
  - 15-24 years
  - 5-14 years
- Similarly generation is also categorized into 6 groups.
  - Generation X
  - Silent
  - Millenials
  - Boomer
  - G.I. Generation
  - Generation Z

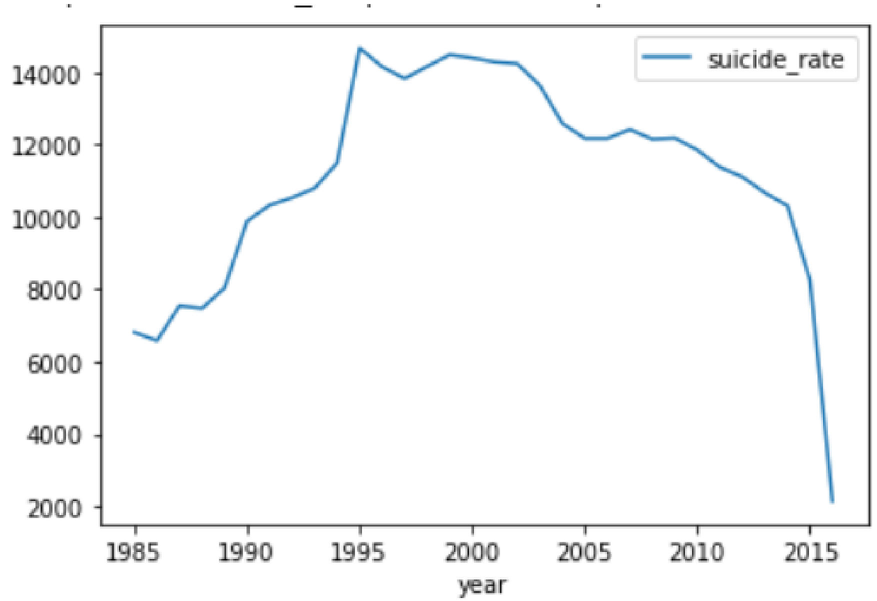
DATASET SCATTER MATRIX

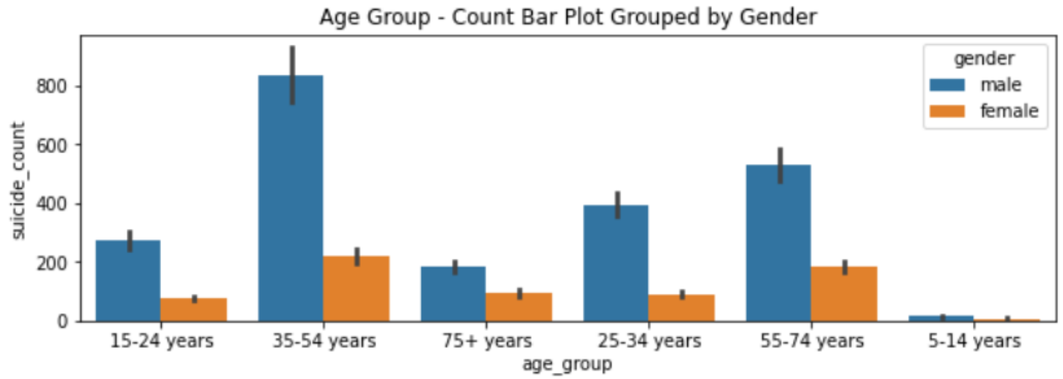


From the scatter matrix, it is obvious that the data has outlier & these are addressed during data preprocessing.

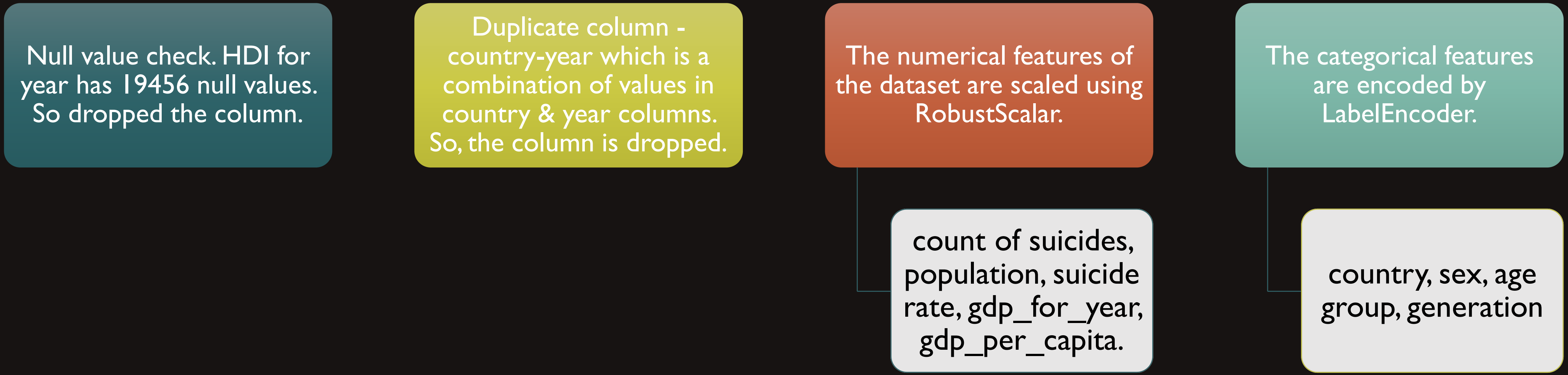
DATA VISUALIZATIONS







DATA PREPROCESSING:



Null value check. HDI for year has 19456 null values. So dropped the column.

Duplicate column - country-year which is a combination of values in country & year columns. So, the column is dropped.

The numerical features of the dataset are scaled using RobustScalar.

- count of suicides, population, suicide rate, gdp_for_year, gdp_per_capita.

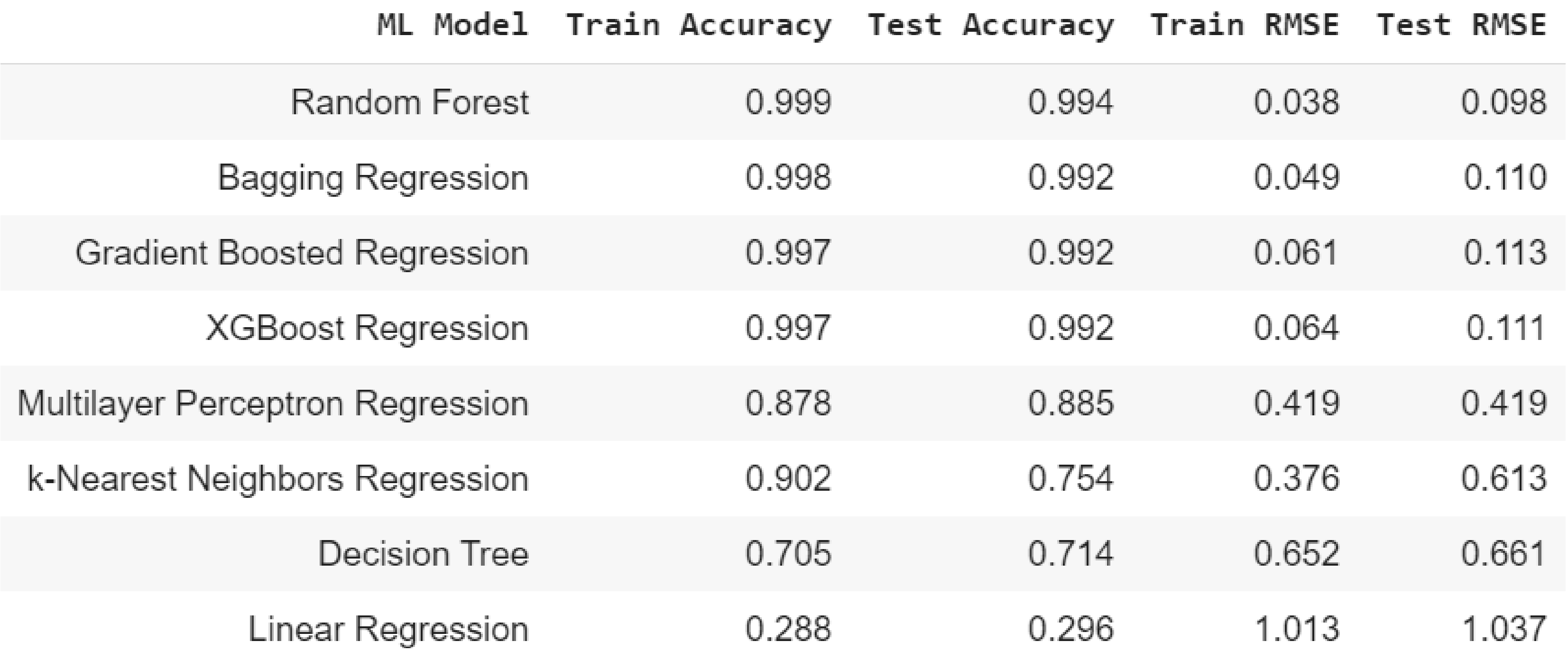
The categorical features are encoded by LabelEncoder.

- country, sex, age group, generation

TRAINED MODELS:

- The Supervised machine learning models trained on this dataset are:
  - k-Nearest Neighbors Regression
  - Linear Regression
  - Decision Tree
  - Random Forest
  - Gradient Boosted Regression
  - Multilayer Perceptron Regression
  - XGBoost Regression
  - Bagging Regression
- Few more interesting models needs to be trained on the dataset which includes custom ensemble model comprising of the low performance models from the models mentioned above.

MODEL RESULTS:



ML Model	Train Accuracy	Test Accuracy	Train RMSE	Test RMSE
Random Forest	0.999	0.994	0.038	0.098
Bagging Regression	0.998	0.992	0.049	0.110
Gradient Boosted Regression	0.997	0.992	0.061	0.113
XGBoost Regression	0.997	0.992	0.064	0.111
Multilayer Perceptron Regression	0.878	0.885	0.419	0.419
k-Nearest Neighbors Regression	0.902	0.754	0.376	0.613
Decision Tree	0.705	0.714	0.652	0.661
Linear Regression	0.288	0.296	1.013	1.037

THANK YOU