

Distributed Data Analysis

CS5811 | Coursework

<Name> | <Roll No> | <Date>

1. Data Description and Research Work:

In this report we analyze the results of using Machine Learning with Life Expectancy data and try to solve the problem of prediction the life expectancy using the different data points in the dataset. The aim of this assignment is to generate value and insight from the processing of heterogeneous data. This will be achieved by implementing several analytic methods/techniques, evaluating them and comparing the effectiveness of the adopted approaches.

The Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries. The datasets are made available to public for the purpose of health data analysis. The dataset related to life expectancy, health factors for 193 countries has been collected from the same WHO data repository website and its corresponding economic data was collected from United Nation website. Among all categories of health-related factors only those critical factors were chosen which are more representative. It has been observed that in the past 15 years, there has been a huge development in health sector resulting in improvement of human mortality rates especially in the developing nations in comparison to the past 30 years. Therefore, in this project we have considered data from year 2000-2015 for 193 countries for further analysis. The individual data files have been merged together into a single dataset. On initial visual inspection of the data showed some missing values. As the datasets were from WHO, we found no evident errors. Missing data was handled in R software by using Missmap command. The result indicated that most of the missing data was for population, Hepatitis B and GDP. The missing data were from less known countries like Vanuatu, Tonga, Togo, Cabo Verde etc. Finding all data for these countries was difficult and hence, it was decided that we exclude these countries from the final model dataset. The final merged file (final dataset) consists of 22 Columns and 2938 rows which meant 20 predicting variables. All predicting variables was then divided into several broad categories:​Immunization related factors, Mortality factors, Economical factors and Social factors.

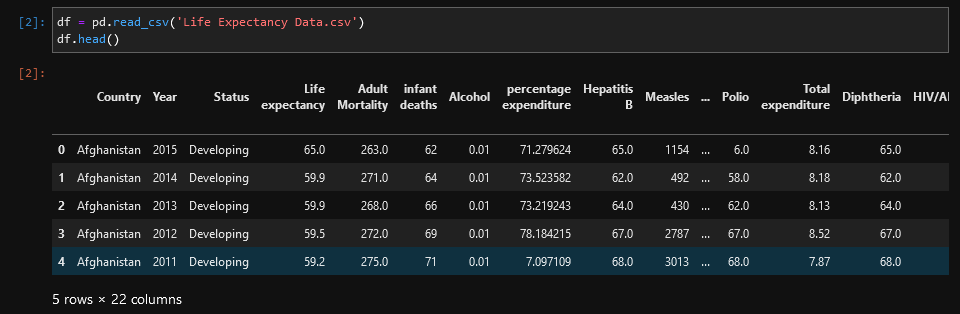
The description of the data-set is given below:

* 1. Dataset Description:

|  |  |
| --- | --- |
| Country | Region |
| Year | Year |
| Status | Developed or developing status |
| Life expectancy | Life expectancy in age |
| Adult Mortality | Adult mortality rates of both sexes (Probability of dying between 15-60 years per 100 population) |
| Infant deaths | Number of infant deaths per 1000 population |
| Alcohol | Alcohol consumption per capita(15+) consumption (in liters of pure alcohol) |
| Percentage expenditure | Expenditure on health as a percentage of Gross domestic Product per capita (%) |
| Hepatitis B | Hepatitis B (HepB) Immunization coverage among 1 year olds (%) |
| Measles | Measles - number of reported cases per 1000 population |
| BMI | Average body mass index of entire population |
| Under five deaths | Number of under-five death per 1000 population |
| Polio | Polio (Pol3) immunization coverage among 1-year olds (%) |
| Total expenditure | General government expenditure on health as a percentage of total government expenditure (%) |
| Diphtheria | Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year olds (%) |
| HIV/AIDS | Death per 1000 live birth HIV/AIDS (0-4 years) |
| GDP | Gross domestic product per capita in USD |
| Population | Population of the country |
| Thinness 1-19 years | Prevalence of thinness among children and adolescents for age 10 to 19 (%) |
| Thinness 5-9 years | Prevalence of thinness among children for age 5-9 years (%) |
| Income composition | Human development index in terms of income composition of resources (index ranging from 0-1) |
| Schooling | Number of years of schooling(years) |

1. Data Preparation and Cleaning:

Firstly, we import pandas and with the help of the library we import the data-set and then we display the data-frame to explore the data.

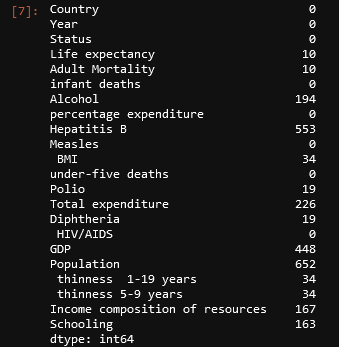


Therefore, then we use **df.info () and df.describe ()** in order to check the data types of the columns and have a little knowledge about the data by seeing the mean, min and max values of every column in the data-set.

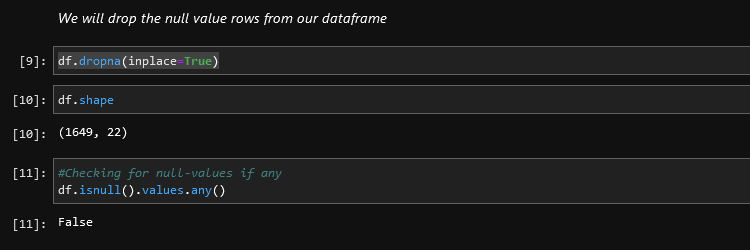
After this we check for any null values present in the dataset using:

**df.isnull().values.any()** – this returns True, which means there are null values.

We use **df.isnull().sum()** in order to display the null values present in each column:



Here we have two options, either to fill these with mean values or drop the null rows, we’ll drop the null values with **df.dropna(inplace=True)** and now if we check the null-values it will return False.

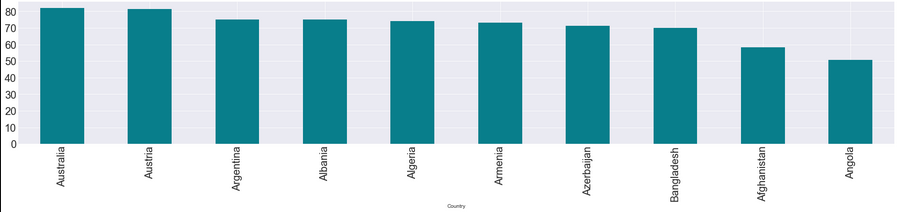


1. Exploratory Data Analysis:

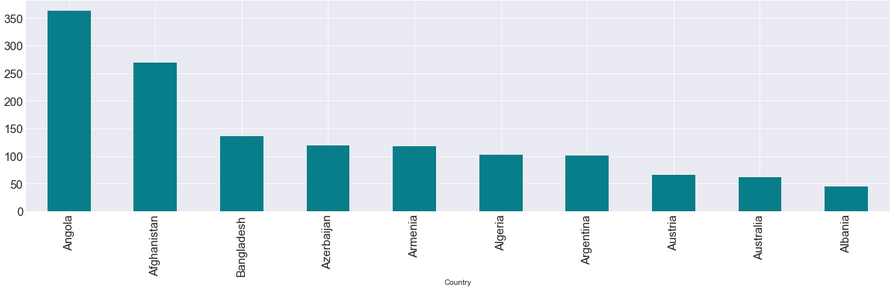
Exploratory data analysis is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods.

Therefore we visualize the data into two types specifically:

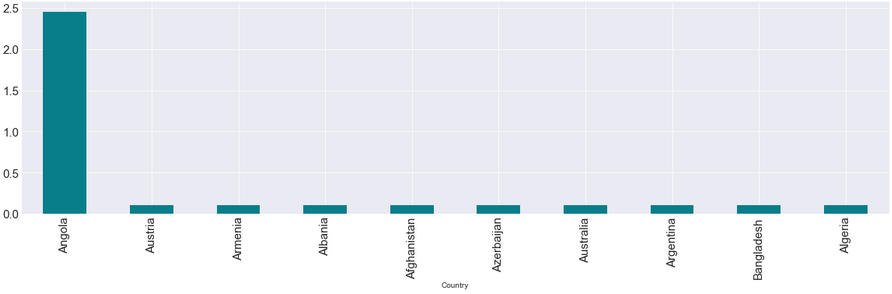
* 1. Exploratory Data Analysis – Country



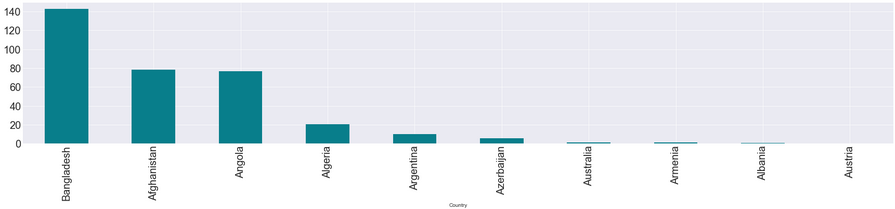
The above graph shows the top ten countries with the highest life expectancy.



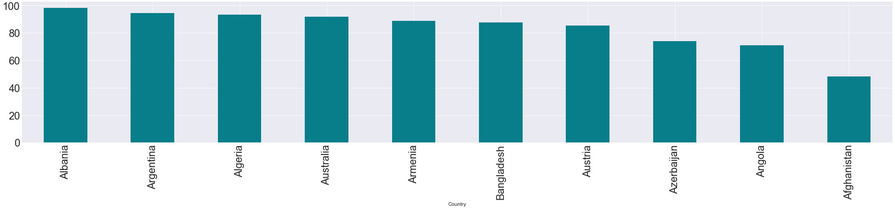
The above graph shows the top ten countries with adult morality.



The above graph shows the top ten countries with AIDS/HIVS.

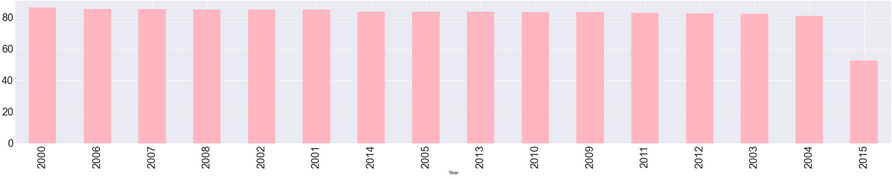


The above graph shows the top ten countries with the most infant deaths.

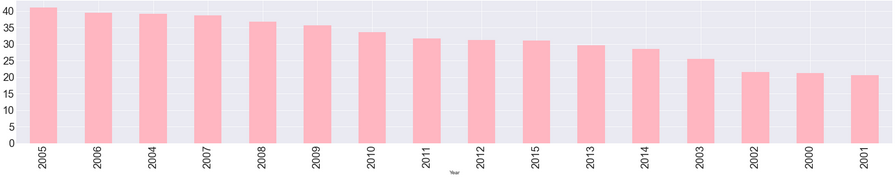


The above graph shows top ten countries with Polio.

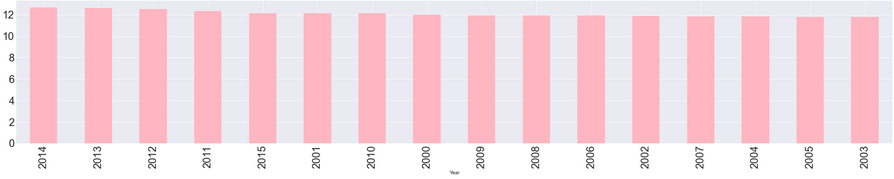
* 1. Exploratory Data Analysis (Year)



The above graph shows the top ten years with polio.



The above graph shows the top ten years with the most infant deaths.

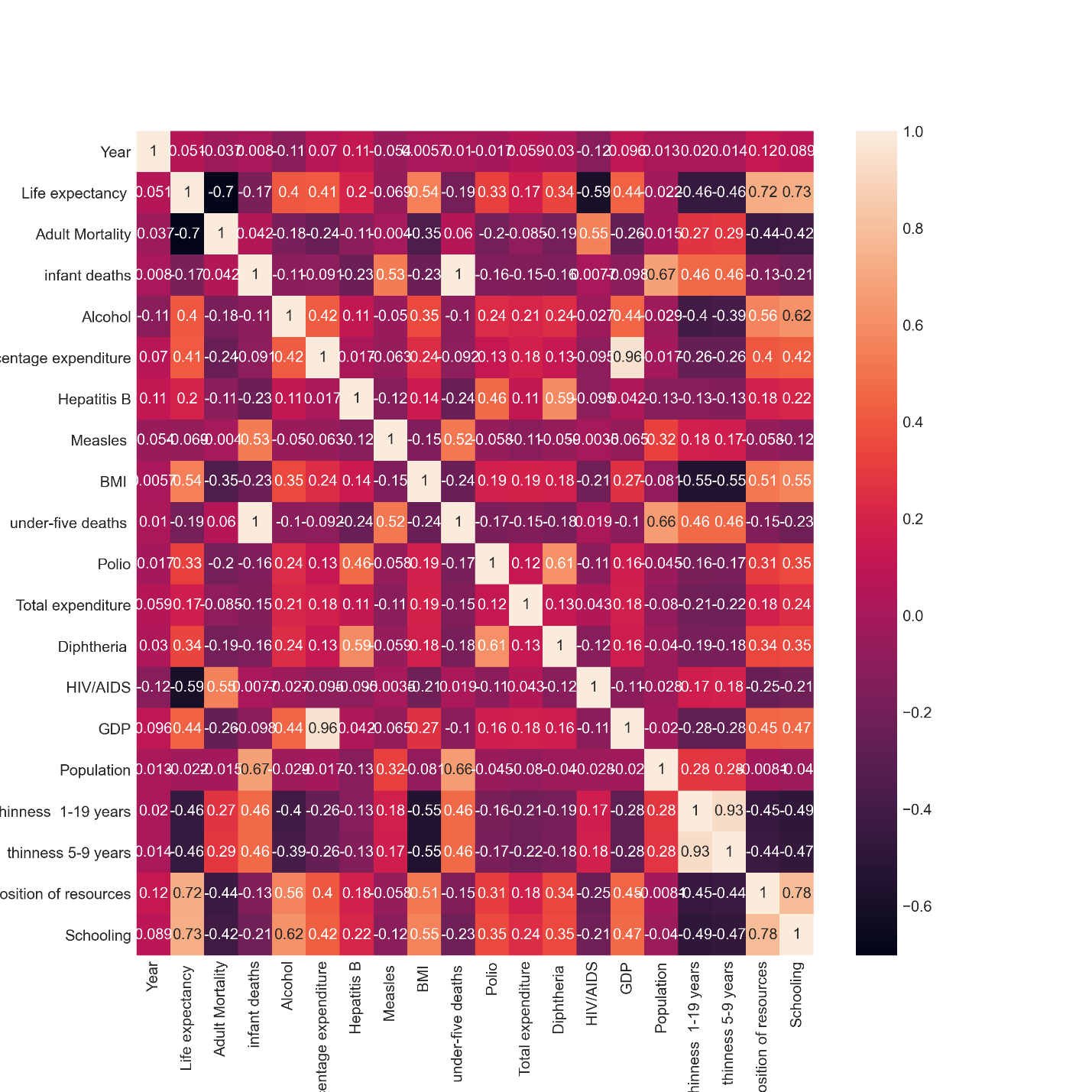


The above graph shows the top ten years of schooling.

* 1. Heat Map:

A heat map is a data visualization technique that shows magnitude of a phenomenon as color in two dimensions. The variation in color may be by hue or intensity, giving obvious visual cues to the reader about how the phenomenon is clustered or varies over space.

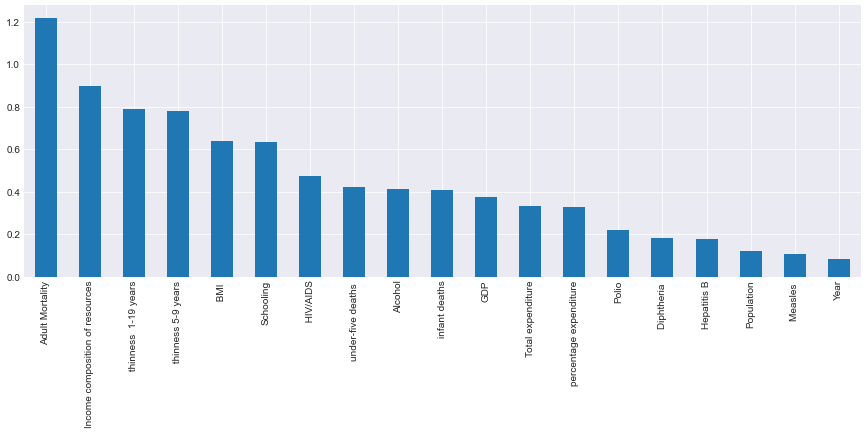
We use seaborn to visualize the heat-map for our dataset.



1. Machine Learning Prediction:
   1. Feature Engineering:

The first step is to apply feature engineering on the dataset in order to get the features which will help us in getting accurate result to predict the life expectancy. We are solving a regression problem, therefore we’ll only select the numerical columns for prediction. We’ll use sklearn and divide our data-set into train set and test set.

We then apply mutual information gain to the train set in-order to get the most valuable columns for our prediction.



We use sklearn library **sklearn.feature\_selection import SelectKBest** to select the top 5 features for our prediction.

* 1. Linear Regression:

Linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables. The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression.

We use sklearn library **sklearn.linear\_model import LinearRegression** to train our linear regression model.

The model gives an accuracy score of 72% on the test data.



4.2 Linear Regression

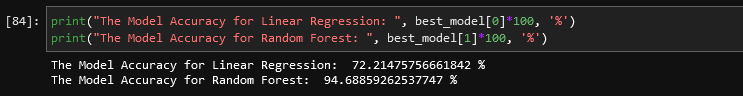
Thus by looking at the graph, we can see that there are outliers in our data, if we want to make the predictions more precise we can eliminate these outliers and then re-train the model to get more accurate predictions.

* 1. Random Forest Regressor:

A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

We use **sklearn.ensemble import RandomForestRegressor** for training our model.

The model gives an accuracy of 94% on the test data.



The above screenshot shows the accuracy of both the models on the test data.

1. High Performance Computation:

PySpark is the Python API for Apache Spark, an open source, distributed computing framework and set of libraries for real-time, large-scale data processing.

We’ll use pyspark for the high performance computing.

We import the following libraries:

* from pyspark.sql import SparkSession
* from pyspark import SparkContext

We first build a spark session: spark = SparkSession.builder.appName('IntroPySpark').getOrCreate()

Therefore, we read in the csv and then display the schema of the csv. We are using the Life Expectancy csv, therefore we’ll perform the same data cleaning and data visualization techniques discussed in above section 2 and 3.

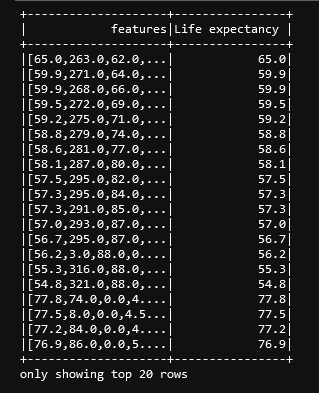
We use pyspark.ml.feature import StringIndexer to transform the status column into integer i.e 0 and 1.

* 1. Linear Regression:

We use:

* pyspark.ml.linalg import Vectors
* pyspark.ml.feature import VectorAssembler

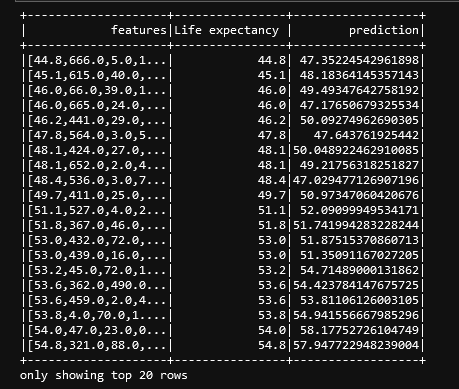
for converting the columns into features.



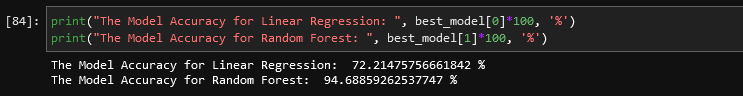
* 1. Random Forest Regressor:

We use random forest for training the model and then fitting that model on our test data.

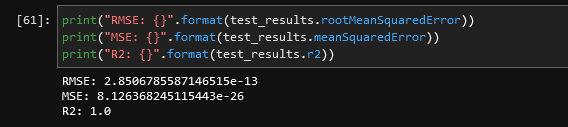
We get the following predictions on the features:



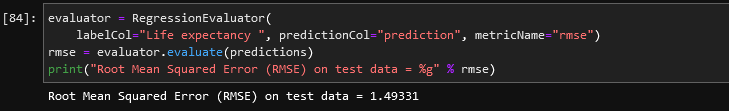
1. Performance evaluation and comparison of methods



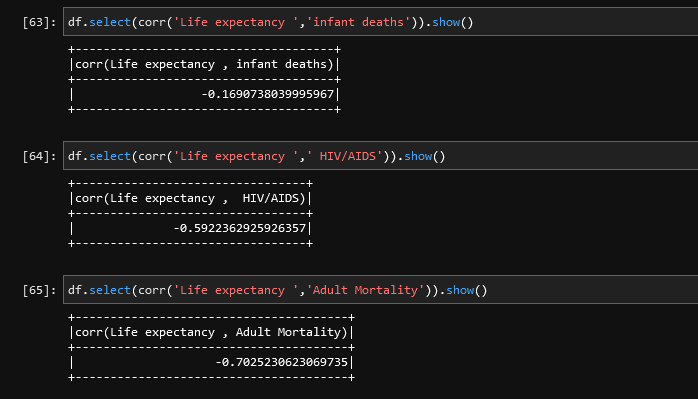
The above screenshot shows the accuracy of the machine learning models using the sklearn library to train and test.



The above graph shows the Root Mean Squared Error, Mean Squared Error and R2 score of the linear regression model that was build using the pyspark MLib library.



The above graph shows the RSME for the Random Forest Regressor.



The above results show negative correlation, negative correlation is a relationship between two variables in which one variable increases as the other decreases, and vice versa. For instance the increase in HIV/AIDS will result in low life expectancy.

1. Discussion of Findings:

Following are the findings we obtain from the data:

* Angola, Afghanistan, Bangladesh and Azerbaijan are the top countries with adult morality.
* Bangladesh has the most infant deaths comparative to other Asian countries.
* Year 2000 had the peak of polio in the entire world.
* Australia and Austria has the highest percentage of adult morality.
* The graph of schooling shows that as the years proceed people started considering to send their children to schools as the schooling growth over the years.
* The heat map shows the negative correlation among life expectancy and adult morality, it also shows negative correlation to infant deaths and HIVS/AIDS which validate our evaluation from the outputs.
* The negative correlation shows that if infant deaths, adult morality and AIDS will increase than the adult morality will decrease.
* BMI has a positive relation with life expectancy which means a good BMI results in high life expectancy.