World happiness prediction

Author - Name: Ponmani Manoharan

Table of Contents

[I) Problem Definition: 3](#_Toc69571699)

[Overview of World Happiness data: 3](#_Toc69571700)

[Importance of the report: 3](#_Toc69571701)

[Brief Analyses of the dataset: 3](#_Toc69571702)

[II) Data Analysis: 3](#_Toc69571703)

[Loading the dataset: 4](#_Toc69571704)

[Summary statistics: 5](#_Toc69571705)

[Exploratory Data Analysis: 6](#_Toc69571706)

[EDA Concluding Remark: 11](#_Toc69571707)

[III) Pre-Processing Pipeline. 12](#_Toc69571708)

[IV) Building Machine Learning Models. 22](#_Toc69571709)

[Finding the best Algorithm: 22](#_Toc69571710)

[V) Concluding Remarks. 23](#_Toc69571711)

Table of Figures :

[Figure 1: importing libraries 4](#_Toc69571651)

[Figure 2: Read first 5 records 5](#_Toc69571652)

[Figure 3: Column names 5](#_Toc69571653)

[Figure 4: check the shape 5](#_Toc69571654)

[Figure 5: summary statistics 6](#_Toc69571655)

[Figure 6: count plot 7](#_Toc69571656)

[Figure 7 : scatter plot 8](#_Toc69571657)

[Figure 8: correlation matrix 11](#_Toc69571658)

[Figure 9: Code for correlation 11](#_Toc69571659)

[Figure 10: pair plot 12](#_Toc69571660)

[Figure 11: check null values 13](#_Toc69571661)

[Figure 12: heat map for null values 14](#_Toc69571662)

[Figure 13: Data types 14](#_Toc69571663)

[Figure 14: Code for label encoding 15](#_Toc69571664)

[Figure 15: Data after label encoding 15](#_Toc69571665)

[Figure 16: Density plot 16](#_Toc69571666)

[Figure 17: power transform 20](#_Toc69571667)

[Figure 18: box plot 20](#_Toc69571668)

[Figure 19: z score 21](#_Toc69571669)

[Figure 20: zscore removal 21](#_Toc69571670)

[Figure 21: standard scaler 22](#_Toc69571671)

[Figure 22: Linear regression 23](#_Toc69571672)

[Figure 23: DecisionTreeRegressor 23](#_Toc69571673)

[Figure 24: RandomForestRegressor 23](#_Toc69571674)

[Figure 25: AdaBoostRegressor 23](#_Toc69571675)

[Figure 26: Ada Boost Regressor Learning Model 24](#_Toc69571676)

[Figure 27: predicted values 24](#_Toc69571677)

[Figure 28: prediction 25](#_Toc69571678)

# Problem Definition:

## Overview of World Happiness data:

In this article, a step-by-step analysis of happiness score prediction is defined. The world happiness model prediction revolves around six factors like the economy, production, social support, life expectancy, freedom, absence of corruption and generosity. The world happiness report consists of surveys from the state of global happiness. The report was first published in 2012. For fresh and accurate information, the 2017 report of world happiness is measured in this article.

The 2017 report consists of 155 countries and they are ranked based on the happiness levels. The report was released on 20th of March at the UN at an event commemorating International Day of Happiness.

## Importance of the report:

Many governments, organizations and civil society are highly interested to use the happiness indicators. These indicators improve the policy-making decision, Experts in the field of economics, health, public policy, and more think that measurement of well being can be used as a valuable tool to measure the progress of a nation. This report also explains the personal and national variations in happiness.

## Brief Analyses of the dataset:

The data was collected from the Gallup World Poll. The questions on the poll were known as Cantril ladder, ask the participants to think of a ladder with the best possible life for them being a 10 and the worst possible life being a 0.

The scores for the years 2013-2016 is used from the national representative samples. Following the happiness score the next six factors – economic production, social support, life expectancy, freedom, absence of corruption, and generosity, these factors contribute to make life evaluations higher in each country than they are in Dystopia. Dystopia is an imaginary a hypothetical country, a locale with world’s least-happy people. They have no impact on the total score reported for each country, but they do explain why some countries rank higher than others.

The residuals, or unexplained components, differ for each country, reflecting the extent to which the six variables either over- or under-explain average 2014-2016 life evaluations. These residuals have an average value of approximately zero over the whole set of countries.

# Data Analysis:

The analysis starts by importing libraries, Initially only the most important libraries such as NumPy, pandas, seaborn, and matplotlib are imported.

Graphical user interface, application

Description automatically generated

Figure 1: importing libraries

### Loading the dataset:

The dataset is read into the notebook, since the report is a csv file, a pandas function read\_csv is utilized. After readind the dataset into a variables it is converted into a dataframe.

Firstly, we can analyse the first 5 records of the dataframe using the method head() to gain a brief insight of that happiness data.

Table

Description automatically generated

Figure 2: Read first 5 records

Secondly, we check the name of the columns using the ***DataFrame.columns*** attribute we can also review the columns names to check if all the column variables are available in the dataset.

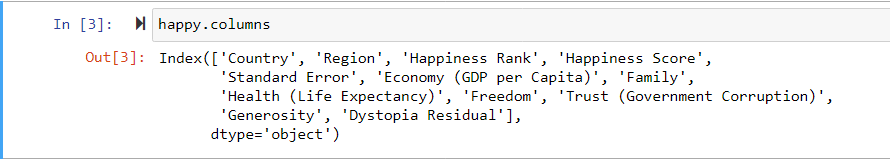


Figure 3: Column names

After this step we see that there 12 columns present in the dataset.

Next, We use the ***DataFrame.shape*** attribute to view the shape of the data frame.

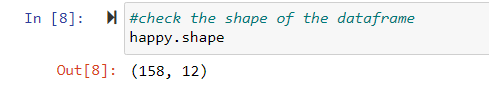


Figure 4: check the shape

The happiness data frame has 158 records and 12 columns.

### Summary statistics:

Now, we have a fundamental knowledge of how the dataset is perceived. The next step is to conduct Summary statistics.

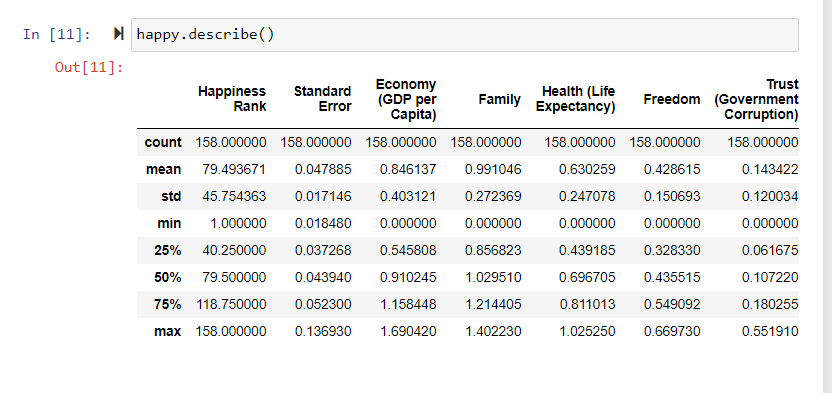


Figure 5: summary statistics

Note: Only the numeric data is measured in the summary statistics. We can include the categorical data into the summary statistics with ***include=['object']***

There seems to be a low standard deviation, a standard deviation measure that is quiet less than mean, this indicates that the data points are very closer to the mean. The min max range is quite large indicating skewness and outlier and this might affect the prediction.

Thus, to manage the difference in the range and standard deviation and to reduce variations in prediction, Exploratory Data Analysis and pre-processing is implemented.

### Exploratory Data Analysis:

There are three ways by which we can proceed with EDA.

Univariate analysis : By Using one column/ attribute we can analyse their significance to the dataset. We can also find a trend or the individual importance of the column. Some common visualizations are column charts, bar charts, histograms and more.

In the happiness dataset we are using categorical variables first: Consider the column “Region”

Chart, bar chart

Description automatically generated

Figure 6: count plot

From the count plot we see that the highest population participated in the survey, belongs to Sub-Saharan Africa followed Central and Eastern Europe. Whereas, Australia, New Zealand, and North America have the least participant count.

Bivariate analysis: By using two columns/ attributes we can analyse the relationship between these attributes and their significance with the target value. We can find the correlation among the variable and decide to evaluate on which column is more important than the other based on the multicollinear characteristics. Some common visualizations are scatter plots, line plots, area plots, and more.

Chart, line chart

Description automatically generated

Figure 7 : scatter plot

The first plot is based on happiness score(target) versus the happiness rank, The scatter plot indicates that as the happiness rank increases the happiness score decreases. The smaller the rank the higher the happiness score. There is negative linear relationship.

Countries with the lowest happiness rank have the highest happiness score. Mainly the African and south Asian continents have poor happiness score and are ranked to be the last.

##### Using For loop for Bivariate analysis:

For the desirable columns segregated from list of columns, we use a for loop and plot reg-plot with a regression line plotted between the maximum and minimum of happiness score(target), Observations indicate that,

Chart, line chart

Description automatically generated Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

happiness rank has a negative linear relationship with happiness score : As happiness rank increase the happiness score decreases.

* Standard- error has no proper relationship with the target
* Economy increases as the happiness score increase (positive linear relationship)
* Family happiness increases as the overall happiness score increase (positive linear relationship)
* Health score increases as the overall happiness score increase (positive linear relationship)
* Freedom increases as the overall happiness score increase (positive linear relationship)
* The trust, generosity, dystopia residual have a weak relationship with the target.

Multivariate analysis: By using more than two columns we can analyse multiple relationships and trends. These save time and can be more efficient for the reader to understand. They can be used to Identify null values, erroneous data variation, poor or strong correlation, and more. Some common visualizations are heatmap, pair plot and more.

##### Heatmap :

For the happiness score data we can do multivariate analysis with the heatmap:

Note : The heat map is plotted to find correlation among variables and hence the categorical variables will be omitted.

Chart, treemap chart

Description automatically generated

Figure 8: correlation matrix

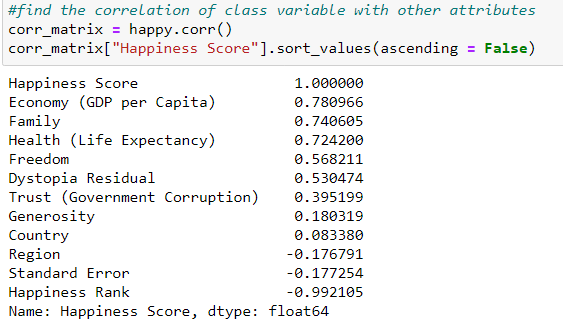


Figure 9: Code for correlation

From the correlation matric we can identify the economy, family, health, freedom, trust, generosity, and dystopia residual are in high correlation with the target variables.

##### Pair plot:

A picture containing chart

Description automatically generated

Figure 10: pair plot

From the pair plot we can understand that there is skewness present in each column most of the columns are unique and there is no multicollinearity within the columns.

## EDA Concluding Remark:

**The exploratory Data analysis findings suggest that:**

Counties in the first few ranks are having a good happiness score > 6:

* Western Europe
* North America
* Australia and New Zealand

Countries with happiness ranks between 100 to 160 seems to having a poor happiness score < 4. 5

* Sub-Saharan Africa
* Southern Asia

Countries with happiness ranks between 20 to 10 seems to having a happiness score between >4.5 and < 6.5

***1. happiness score is high when the happiness rank is small***

Columns like economy, family, health and freedom are positively related to the target variables

***2. Happiness score is high when the economy, family, health and freedom are high***

# Pre-Processing Pipeline.

Pre-processing data is an important step before prediction, pre-processing allows the analyst to develop an accurate and intuitive predictions. Often, the data we obtain is deficient, noisy, and unreliable. By pre-processing the quality of data improves. Incomplete data may be obtained due to wrong data collection methods or defects in the medium through which the data was collected, it can also be due to misunderstanding of the actual purpose of the proposed question. Noisy data occurs when we have wrong information that are irrelevant to the attribute this is mainly caused due to human error, some time it may be induced due to transition or transformation of data from one file format to another. To make the data more relevant for the machine learning algorithm we can pre-process the data.

Data pre-processing consist of few steps, each of which are done only if needed, they are : cleaning the data, amalgamation of the data, transformation of the data and reducing the data.

For the Happiness Dataset I have used few of the pre-processing pipeline steps:

#### 1. Data cleaning:

Here, the data is examined for missing values, inconsistencies like “?” instead of NaN, after which methods like simple imputer is applied to replace the null values and inconsistent data. If there is a chance to drop the records which have null values drop NA can be used.

In the happiness data we have identified zero null values I used a combination of isnull() and sum() methods to identify the count to null values in the dataframe.

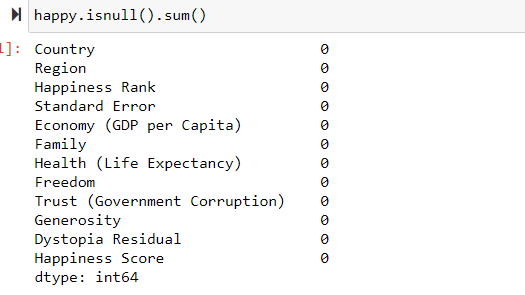


Figure 11: check null values

I also visualized the data frame from null values using the seaborn heatmap function

Chart, bar chart, histogram, waterfall chart

Description automatically generated

Figure 12: heat map for null values

No null values

After analysing the dataframe for null values, the data was checked to identify the datatype of the columns

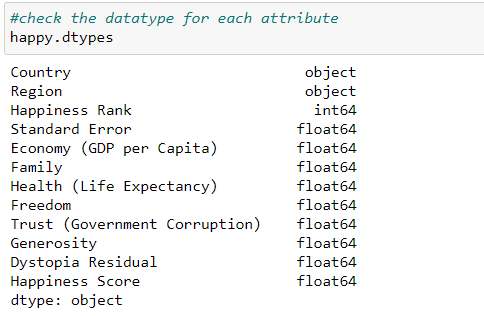


Figure 13: Data types

The ***DataFrame*.*dtypes*** attribute was used. The observations indicate that there are 2 columns of type “object” namely “Country” and “Region”, The other columns are of numeric type with either “float” or “int.

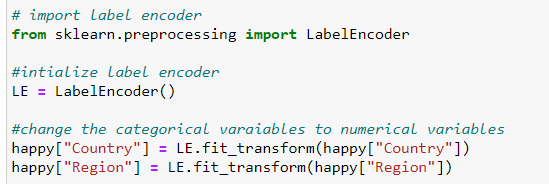


Figure 14: Code for label encoding

Note: Since machine learning models does not take object types as input we use label encoding

For “object” type columns the Label Encoder is used it is very efficient encoding tool. The Label Encoder encode new labels with a value between and n-1 where n is the number of distinct labels.

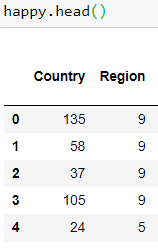


Figure 15: Data after label encoding

Encoded.

#### 2. Data Transformation

Before beginning the transformation the data is spilt into dependent and independent columns namely x and y data. Where x represents the independent columns and y represent the target column.

##### Skewness check and removal:

A normal distribution of the data in the columns improve the accuracy and efficiency of machine leaning models

Diagram, histogram

Description automatically generated

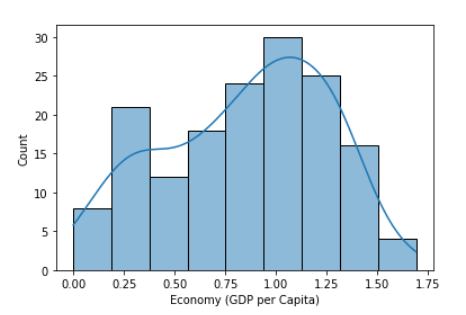
Figure 16: Density plot

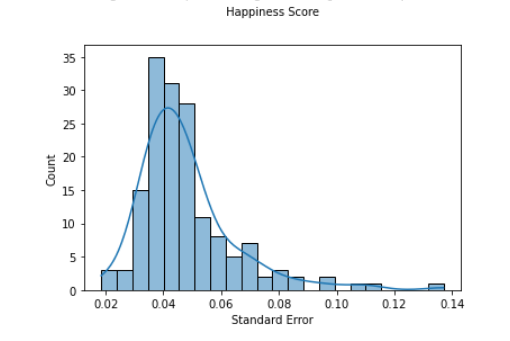
After encoding the data, we check for the density of the columns. The data distribution needs to be normalized because there seems to be skewness in some of the attributes.

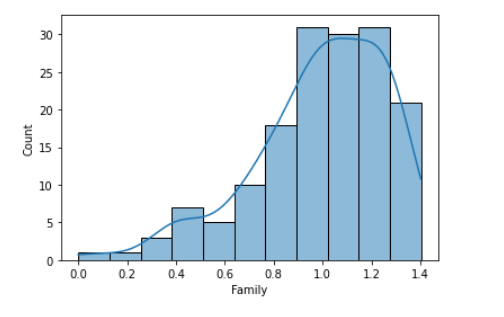
##### Histograms:

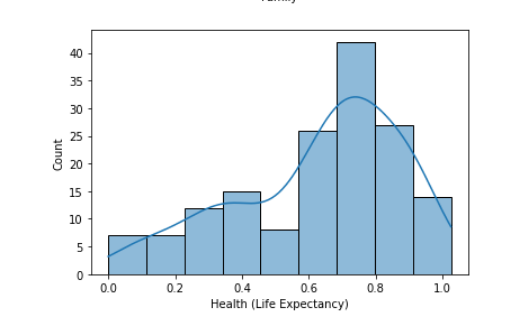
Histograms are used to find the distribution of data and understand whether the data is skewed to one side of the distribution, either positively skewed or negatively skewed. Skewed data may cause inaccurate predictions .

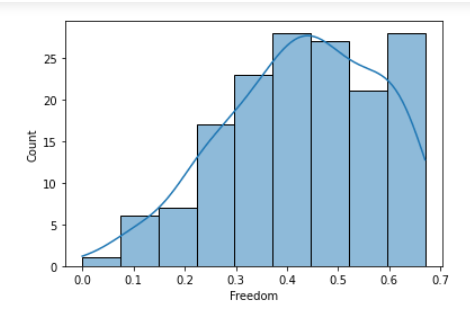
Using loops to plot histograms:

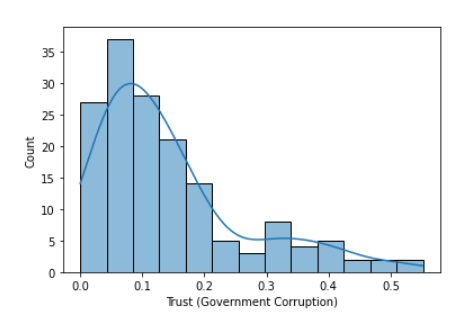


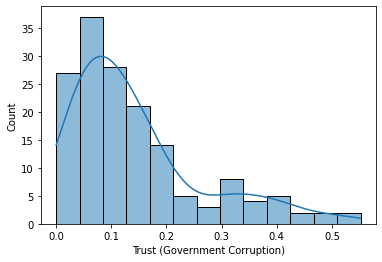


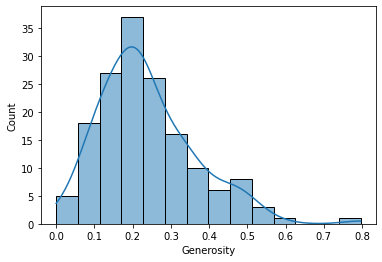


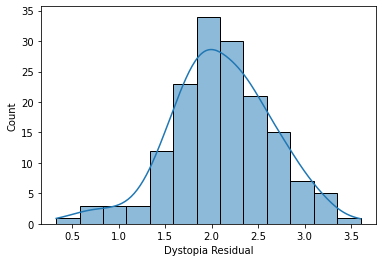
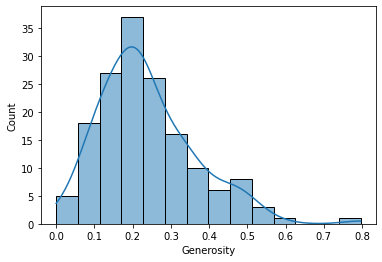












Country and region are categorical variables and they do not have skewness, the other numeric columns with skewness are normalized using the sklearn. preprocessing – power transform

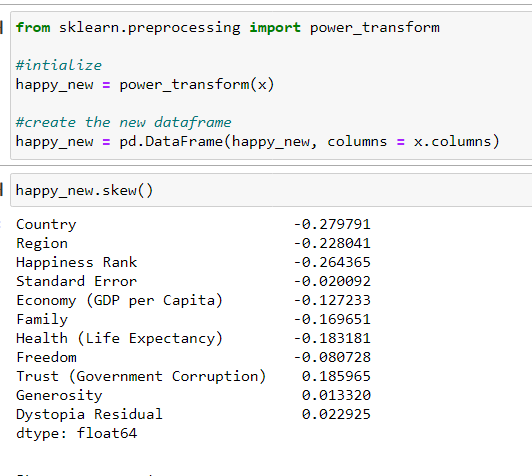


Figure 17: power transform

##### Outlier check and removal:

After removing the skewness the data is analysed for outliers, the box plot is used to find the outlier in the data.

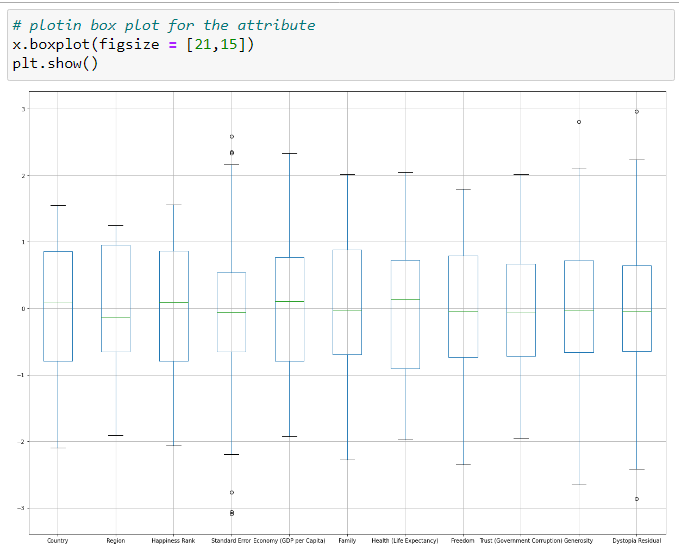


Figure 18: box plot

The results from happiness score prediction data indicates that there are very few outliers in the dataset and they are removed from the records with z-score method. The threshold set for z-score is 3 and if any record with a z-score more than 3 is identified then it will be considered as an outlier.

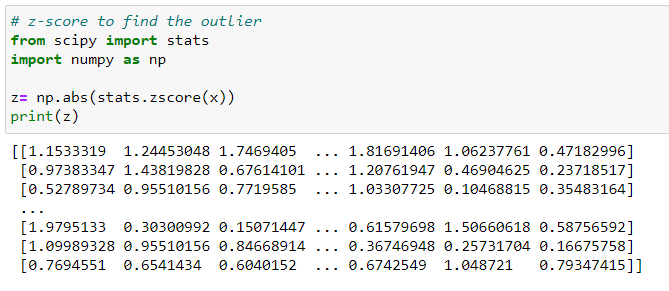


Figure 19: z score

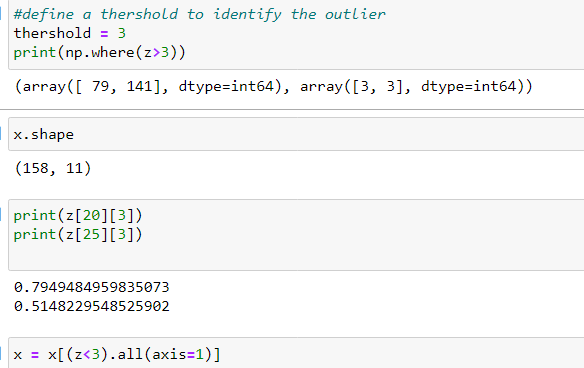


Figure 20: zscore removal



The outliers are removed.

The shape of the x - data is changed to 156 records from 158 records and 11 columns where all the 11 column are independent

The shape of the y – data is changed to 156 records and where all the records belong to happiness score column (target column).

#### 3. Data Normalization:

Normalization scales the data in the columns between a certain interval, in the happiness dataset, the standard scalar method is used. Therefore the interval is set between [0,1]

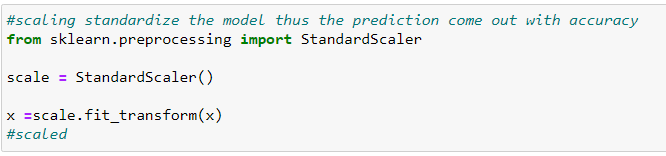


Figure 21: standard scaler

Data is scaled and ready of ML modelling.

# Building Machine Learning Models.

Since the target variables is numeric variable and it is continues in nature, the ML model used for prediction will be ***supervised regression model***. Evaluation metrics measure the model performance and its effectives in delivery a good prediction with by analysing the appropriate relationships

***The evaluation metric that can be used are :***

R2 score – This metric represent how strong the relationship between the ml model and the dependent variable are. Good R^2 delivers a good model

MSE – Mean Squared Error, it is the average of the square difference between the predicted and actual values, MSE captures larger errors

RMSE – Root Mean Square Error, it is the squared root of the average difference between predicted and actual values, RMSE captures larger errors.

MAE – Mean Absolute Error, The absolute difference between the target values and the value predicted by the model. MAE does not capture larger errors.

### Finding the best Algorithm:

Using a for loop and linear regression as the key model we find the best random state to be 108. Applying the random state to be 108 , the independent and dependent data is split into train and test set.

The machine leaning models used in this project are Linear Regression, Decision Tree Regressor, Random Forest Regressor, and Ada Boost Regressor

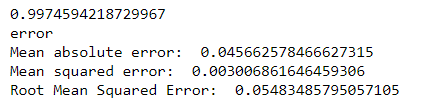


Figure 22: Linear regression

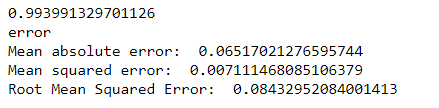


Figure 23: DecisionTreeRegressor

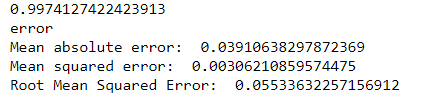


Figure 24: RandomForestRegressor

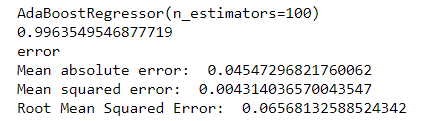


Figure 25: AdaBoostRegressor

The highest accuracy score belongs to the ada boost regressor model and linear regression model. 100% This may be a result of overfitting. To make the model more accurate cross validation is used.

##### Cross validation:

**The cross validation score of Linear regression model : 0.9918715400748456**

**The cross validation score of Decision Tree regression model : 0.9944580595530308**

**The cross validation score of Random Forest regression model : 0.9964867765817613**

**The cross validation score of AdaBoost regression model : 0.9957777821616949**

After calculating the difference between actual r2 score and cross validation score. The Ada Boost regressor, Random forest regressor and Decision tree regressor give best results.

I have chosen Ada boost regressor to be the best model and have applied hyper parameter tuning to tune the features. Which is optional considering the model performance to be well and the accuracy is more than 99% without hyper parameter tuning.

# Concluding Remarks.

The model is saved using joblib package dump method, and is transformed into pickled file for further development.



Figure 26: Ada Boost Regressor Learning Model

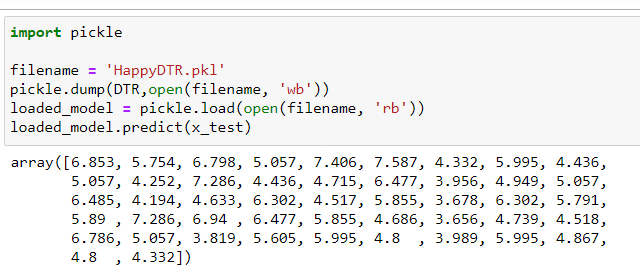


Figure 27: predicted values

The happiness score prediction revolves around the independent variables . The higher the score the lower the rank. Therefore If any new data is given as input to this project it will render a good prediction of happiness score, Considering to have followed all the steps necessary such as Exploratory data analysis, Pre-processing pipelines and Multiple machine learning models with proper cross validation. The model will render a good prediction enabling interested parties to take good decisions.

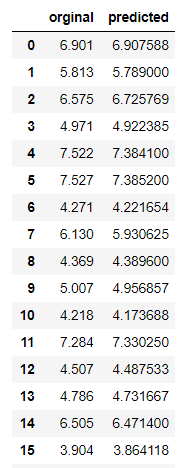


Figure 28: prediction