ANALYSIS ON WORLD HAPPINESS REPORT USING MACHINE LEARNING TECHNIQUES

Main Project work submitted to Avinashilingam Institute for Home Science and Higher Education for Women

MASTER OF SCIENCE IN INFORMATION TECHNOLOGY

Submitted By

K.P.KEERTHANA(19PIT004)

Under the guidance of

Mrs.N.Krishnaveni M.sc.,M.phil.,SET.,

Assistant Professor, Department of Information Technology

AVINASHILINGAM INSTITUTE FOR HOME SCIENCE AND HIGHER EDUCATIONFOR WOMEN

SCHOOL OF PHYSICAL SCIENCES AND COMPUTATIONAL SCIENCES

DEPARTMENT OF INFORMATION TECHNOLOGY

COIMBATORE-641043

MAY 2021

DECLARATION

DECLARATION

I hereby declare that the project entitled **"ANALYSIS ON WORLD HAPPINESS REPORT USING MACHINE TECHNIQUES"** is a record of the original work done by K.P.KEERTHANA (19PIT004) under the guidance of Mrs. N. Krishnaveni M.sc., M.phil., SET., Assistant professor Department of Information Technology, school of physical sciences and computational sciences, Avinashilingam Institute for Home Science and Higher Education for Women, in the partial fulfillment for the degree of Master of Science in Information Technology and this project has not formed the basis for any Degree/Diploma/Associates.

Place : Coimbatore

Date:

Signature of the candidate

_____

Countersigned by,

Mrs. N. Krishnaveni M.sc., M.phil., SET.,

Assistant professor Department of Information Technology,

School of physical sciences and computational sciences.

CERTIFICATE

CERTIFICATE

This is to certify that this project work entitled "ANALYSIS ON WORLD HAPPINESS REPORT USING MACHINE LEARNING ALGORITHMS" done by K.P.Keerthana(19PIT004) has been submitted to Avinashilingam Institute for Home science and Higher education for women, Coimbatore-43 in partial fulfillment of the requirement for the award of the degree of MASTER OF SCIENCE IN INFORMATION TECHNOLOGY. This Project has not found the basis for the award of any Degree/Associate/fellowship or similar title to any Candidate of any University. Certified as a bonafied record of the work submitted for the Viva voce held on _____.

Signature of the HOD

Signature of the Guide

Signature of External Examiner

GATEWAY
SOFTWARE SOLUTIONS

ISO 9001:2015

Date: 30/04/2021

TO WHOMSOEVER IT MAY CONCERN

This is to certify the student **Ms. KEERTHANA K P(19PIT004)** pursuing her final year in **MSC INFORMATION TECHNOLOGY** in **AVINASHILINGAM INSTITUTE FOR HOME SCIENCE & HIGHER EDUCATION FOR WOMEN, COIMBATORE** has completed her project entitled **"ANALYSIS ON WORLD HAPPINESS REPORT USING MACHINE LEARNING TECHNIQUES "** in our concern starts from February 2021 to April 2021.

Wish her the best

GATEWAY SOFTWARE SOLUTIONS

GATEWAY SOFTWARE SOLUTIONS GBE-12

Manager

Mobile: 7397078885
E-mail : info@gatewaysoftwaresolutions.com / Webiste : gatewaysoftwaresolutions.com

ACKNOWLEDGEMENT

ACKNOWLEDGEMENT

I would like to express my sincere thanks to God Almighty, for his constant love and grace that he has showered upon me, which kept me in good health, and sound mind without which my project would not have reached a successful end.

I would like to express my deep sense of reverential gratitude and sincere thanks to **Dr. S. P. Thyagarajan, Chancellor**, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for providing all facilities during my course of study.

I owe my great deal of gratitude to **Dr. Premavathy Vijayan M.Sc., M.Ed., Dip. Spl.Edn., M.Phil., Ph.D., Vice Chancellor,** Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for extending all resources that facilitated the smooth conduct of the project study.

I express my gratitude to **Dr. S. Kowsalya, Registrar**, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for providing all facilities and support necessary for the study.

I wish to extend my sincere thanks **Dr. K. Udaya Chandrika M.Sc., M.Phil., Ph.D.,Dean School of Physical Sciences and Computational Sciences**, for her support and valuable guidance.

I take this opportunity to express my profound gratitude to **Dr. D. ShanmugaPriya, M.Sc., M.Phil., Ph.D., Head, Department of Information Technology**, School of Physical Sciences and Computational Sciences, for her valuable guidance and encouragement.

I heartily thank my esteemed project guide **Dr. F. Paulin, M.C.A, M.Phil., Ph.D.,Assistant Professor, Department of Information Technology,** for imparting tremendous assistance and well-timed support for triumph of our project.

I express my honorable thanks to our project coordinator Dr. T. Jayamalar M.C.A,M.Phil., Ph.D, Assistant Professor, Department of Information Technology, for her kind advice and knowledgeable suggestions which helped us to complete our project successfully.

I would like to express my sincere gratitude to all the staff members of the Department of Information Technology, for their constant encouragement and for the opportunity to do our project in this esteemed university. Last yet importantly, I would like to thank my parents, family members, friends and all well-wishers for their kind inspiration, blessings and encouragement during the course of project.

ABSTRACT

ABSTRACT

The World Happiness Report is a Landmark survey of the state of global happiness. The happiness factor varies due to different human perspectives. The factors used in this work include both physical needs and the mental needs of humanity. For example, the educational factor. This work identify more than 90 features that can be used to predict the country happiness. Due to numerous features, it is unwise to rely on the prediction of national happiness by manual analysis.

Therefore, this work used a machine learning technique Decision tree and Random forest used to compare and find the better accuracy. Using data of 187 countries from the UN Development Project, this work is able to identify which factor needed to be improved by a certain country to increase the happiness of their citizens. In Data preprocessing, cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model. Feature extraction involves choosing a set of features from a large collection. These features were supplied to traditional and ensemble machine learning classifiers. Here using decision tree, random forest, decision tree with k-fold, and random forest tree with k-fold. Comparing these four methods and getting a best accuracy.

A random forest is an ensemble model that consists of many decision trees. Predictions are made by averaging the predictions of each decision tree. Here Random forest with k-fold gives better accuracy then decision tree with k-fold technique.

CONTENT

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 World Happiness Report

Happiness has been discussed and pursued since the birth of mankind. Ancient Greek philosophers Aristotle once said: "Happiness is the ultimate goal of all our actions, we do all things are in fact the means." Today when science and technology are highly developed, materialism no longer bothers and occupies people's minds. More and more people have begun the pursuit and discussion of happiness. Happiness to me seems like an individual metric, something that is hard to generalize.

The World Happiness Report is a publication of the United Nations Sustainable Development Solutions Network. It contains articles and rankings of national happiness, based on respondent ratings of their own lives, which the report also correlates with various (quality of) life factors. As of March 2021, Finland was ranked the happiest country in the world four times in a row. The report primarily uses data from the Gallup World Poll. The World happiness Report may be a point of interest survey of the state of worlwide bliss. The joy scores and rankings utilize information from the Gallup World Survey. The scores are based on answers to the most life evaluation address inquired within the survey. This address, known as the Cantril step, asks respondents to think of a step with the most excellent conceivable life for them being a 10 and the most exceedingly bad conceivable life being a and to rate their claim current lives on that scale.

The scores are from broadly agent tests for the a long time 2015-2020 and utilize the Gallup weights to create the gauges agent. The columns taking after the bliss score assess the degree to which each of six variables – financial generation, social back, life anticipation, flexibility, nonattendance of debasement, and liberality – contribute to making life assessments higher in each nation than they are in Dystopia, a theoretical nation that has values rise to to the world's least national midpoints for each of the six variables.

ANNUAL REPORT TOPIC

World Happiness Reports were issued in 2012, 2013, 2015, 2016 , 2017, 2018, 2019, 2020, and 2021. In addition to ranking countries happiness and well-being levels, each report has contributing authors and most focus on a particular theme. The data used to rank countries in each report is drawn from the Gallup World Poll, as well as other sources such as the World Values Survey, in some of the reports. The Gallup World Poll questionnaire measures 14 areas within its core questions: business & economic, citizen engagement, communications & technology, diversity (social issues), education & families, emotions (well-being), environment & energy, food & shelter government and politics, law & order (safety), health, religion & ethics, transportation, and n work.

1.2.1 MACHINE LEARNING METHODS

In machine learning, tasks are generally classified into broad categories. These categories are based on how learning is received or how feedback on the learning is given to the system developed.

Two of the most widely adopted machine learning methods are supervised learning which trains algorithms based on example input and output data that is labeled by humans ,and unsupervised learning which provides the algorithm with no labeled data in order to allow it to find structure within its input data.

1.2.1.1 SUPERVISED LEARNING

In supervised learning, the computer is provided with example inputs that are labeled with their desired outputs. The purpose of this method is for the algorithm to be able to "learn" by comparing its actual output with the "taught" outputs to find errors, and modify the model accordingly.

Supervised learning therefore uses patterns to predict label values on additional unlabeled data. A common use case of supervised learning is to use historical data to predict statistically likely future events. It may use historical stock market information to anticipate upcoming fluctuations, or be employed to filter out spam emails. In supervised learning, tagged photos of dogs can be used as input data to classify untagged photos of dogs.

There are two main types of supervised learning problems: they are classification that involves predicting a class label and regression that involves predicting a numerical value.

- Classification: Supervised learning problem that involves predicting a class label.
- Regression: Supervised learning problem that involves predicting a numerical label.

Both classification and regression problems may have one and more input variables and input variables may be any data type, such as numerical or categorical.

1.2.1.2 UNSUPERVISED LEARNING

In unsupervised learning, data is unlabeled, so the learning algorithm is left to find commonalities among its input data. As unlabeled data are more abundant than labeled data, machine learning methods that facilitate unsupervised learning are particularly valuable. The goal of unsupervised learning may be as straightforward as discovering hidden patterns within a dataset, but it may also have a goal of feature learning, which allows the computational machine to automatically discover the representations that are needed to classify raw data.

Unsupervised learning is commonly used for transactional data. You may have a large dataset of customers and their purchases, but as a human you will likely not be able to make sense of what similar attributes can be drawn from customer profiles and their types of purchases. With this data fed into an unsupervised learning algorithm, it may be determined that women of a certain age range who buy unscented soaps are likely to be pregnant, and therefore a marketing campaign related to pregnancy and baby products can be targeted to this audience in order to increase their number of purchases.

Without being told a "correct" answer, unsupervised learning methods can look at complex data that is more expansive and seemingly unrelated in order to organize it in potentially meaningful ways. Unsupervised learning is often used for anomaly detection including for fraudulent credit card purchases, and recommender systems that recommend what products to buy next. In unsupervised learning, untagged photos of dogs can be used as input data for the algorithm to find likenesses and classify dog photos together.

There are many types of unsupervised learning, although there are two main problems that are often encountered by a practitioner: they are clustering that involves finding groups in the data and density estimation that involves summarizing the distribution of data.

- Clustering: Unsupervised learning problem that involves finding groups in data.
- Density Estimation: Unsupervised learning problem that involves summarizing the distribution of data.

1.2.2 MACHINE LEARNING APPROACHES

As a field, machine learning is closely related to computational statistics, so having background knowledge in statistics is useful for understanding and leveraging machine learning algorithms. For those who may not have studied statistics, it can be helpful to first define correlation and regression, as they are commonly used techniques for investigating the relationship among quantitative variables.

Correlation is a measure of association between two variables that are not designated as either dependent or independent. Regression at a basic level is used to examine the relationship between one dependent and one independent variable. Because regression statistics can be used to anticipate the dependent variable when the independent variable is known, regression enables prediction capabilities. The following approaches are used in this project.

1.2.2.1 RANDOM FOREST REGRESSION:

A random forest is an ensemble model that consists of many decision trees. Predictions are made by averaging the predictions of each decision tree. Or, to extend the analogy—much like a forest is a collection of trees, the random forest model is also a collection of decision tree models. This makes random forests a strong modeling technique that's much more powerful than a single decision tree.

Each tree in a random forest is trained on the subset of data provided. The subset is obtained both with respect to rows and columns. This means each random forest tree is trained on a random data point sample, while at each decision node, a random set of features is considered for splitting. In the realm of machine learning, the random forest regression algorithm can be more suitable for regression problems than other common and popular algorithms.

1.2.2.2 DECISION TREE REGRESSION:

Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.

Discrete output example**:** A weather prediction model that predicts whether or not there'll be rain in a particular day. Continuous output example**:** A profit prediction model that states the probable profit that can be generated from the sale of a product.

1.3 ABOUT THE PLATFORM

1.3.1 ANACONDA NAVIGATOR

Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda distribution that allows you to launch applications and easily manage conda packages, environments, and channels without using command-line commands. Navigator can search for packages on Anaconda.org or in a local Anaconda Repository. It is available for Windows, mac OS, and Linux. In order to run, many scientific packages depend on specific versions of other packages. Data scientists often use multiple versions of many packages and use multiple environments to separate these different versions. The command-line program conda is both a package manager and an environment manager. This helps data scientists ensure that each version of each package has all the dependencies it requires and works correctly. Navigator is an easy, point-and-click way to work with packages and environments without needing to type conda commands in a terminal window. You can use it to find the packages you want, install them in an environment, run the packages, and update them – all inside Navigator. The following applications are available by default in Navigator:

- Jupyter Lab
- Jupyter Notebook
- Spyder
- PyCharm
- VSCode
- Glueviz
- Orange 3 App
- R Studio
- Anaconda Prompt (Windows only)
- Anaconda PowerShell (Windows only)

1.3.2 JUPYTER NOTEBOOK

Jupyter Notebook, an open-source, web-based IDE with deep cross-language integration that allows you to create and share documents containing live code, equations, visualizations, and narrative text. Data scientists and engineers love using Jupyter for data cleaning and transformation, statistical modeling, visualization, machine learning, deep learning, and much more. Jupyter Notebook's format (ipynb) has become an industry standard and can be rendered in multiple IDEs, Git Hub, and other places. Jupyter has support for over 40 programming languages, including Python, R, Julia, and Scale.

Notebooks can be shared easily with others, and code can produce rich, interactive output, including HTML, images, videos, and custom MIME types. It allows you to leverage big data tools such as Spark and explore that same data with pandas, scikit-learn, Tensor Flow, and ggplot2. Jupyter has become an important part of the workflow for data scientists to process, analyze, and manipulate their data and draw insights from it in a pleasant and effective way. The open and standardized Jupyter notebook file format is designed to capture, display, and share natural language, code, and results in a self-contained and powerful computational narrative. In 2014, Fernando Pérez announced a spin-off project from IPython called Project Jupyter. IPython continues to exist as a Python shell and a kernel for Jupyter, while the notebook and other language-agnostic parts of IPython moved under the Jupyter name. In 2015, GitHub and the Jupyter Project announced native rendering of Jupyter notebooks file format (.ipynb files) on the GitHub platform.

CHAPTER 2

LITERATURE REVIEW

Network Learning Approaches To Study World Happiness author MeghnaChaudhariSiddharth & July 21, 2020Predictive Modelling and Bayesian Networks to model the processed historical happiness and index with the accuracy of 70% (GDP per capita). Predictive models for predicting happiness index of a country and BN for exploring causal relationships among variables. Emotion recognition using multi-modal data and ml techniques. Authors are Fiona Marshal, Raymond Bond Ulster University Emotion recognition methods based on multi-channel EEG signals And the accuracy of emotion recognition is 90%. It uses SVM and RF perform better than KNN and NB for EEG-based emotion recognition task are compared. Analysis happiness In EU Countries Using The Multi-Model Classification authors MarcinPelka & 2019. It uses Multidimensional scaling & decision tree to find out factors determine cluster memberships.0.5800769(3 clusters)Model based clustering ensemble to determine selected European countries have similar patterns of happiness**.**

Happiness recognition from mobile phone Data Andrey Bogomolv Bruno Lepri & published in the year 2013 Final machine learning model, based on Random forest classifier with the accuracy of 80.81% Strong predictive power in the source and feature spaces, discuss different approaches, ml models and provide an insight for future research. Measuring Happiness Around World Through Artificial intelligence authors are Rustem Ozakar, Rafet Efe Gazanfe, Y. Sinan Hanay Using an unbiased emotion detector, artificial intelligence (AI) and the accuracy is 95%(confidence interval) Different facial structures which give unintended classifications. further work specialized dataset for purposed idea and different algorithms needs to be explored in terms of performance and accuracy

Detecting Student Emotions in Computer-Enabled Classrooms authors are Nigel Bosch, Sidney K.D'Mello, Ryan S. Baker, Jaclyn Ocumpaugh & in the year 2010 and algorithms are Experimented with supervised classifiers including C4.5 trees and Bayesian classifiers, with the accuray of Boredom-64%, Confusion-74%, Delight-83%, Engagement-64%, Frustration-62%.Important step toward making this vision a reality, by demonstrating the feasibility of automated detection of student affect in a noisy real-world environment a school. Sentiment Analysis of Big Data: Methods, Applications and Open Challenges published on June 28, 2018 with the algorithms of Technical aspect of OMSA and non-tech aspect in the form of application

areas Twitter has been used as a tool for disseminating and propagating information rather than simply a social networking site.

Happiness score identification: a regression approach with the authors of Yichen Ma1 , Andrew Liu2 , Xukai Yuchen Shao in the year of 2016. It mainly used regression approach to explore the major influencing factor with the result of three things Economy-0.638, Health-0.584, Family-0.568 and the R-square value of the economy stayed around 0.62 in the past five years. Investigation of Happiness of Countries with K-Means Clustering and Discriminant Analysis authors are Sadullah celik1 & Necmiye cömertler2 &published in2020.The algorithms are K-Means clustering and discriminant analysis. New visualization approach based discriminant analysis can be proposed in which the K-Means cluster analysis results can be interpreted.Vocal-based emotion recognition using random forests and decision tree the authors are FatemehNooroozi , Dorota kaminshka, Tomaz sapinsiki & 2015. Using the random forests and decision tree models. The average recognition rate obtained by the rate proposed method was also compared with another work using DNN approach by 6.58% where the same dataset was utilized while taking features such as pitch, energy, MFCC and teager energy operator into account. The accuracy of average accuracy rate is 66.28%. The transnational happiness study with Big data Technology the authors are Lingxi peng,Yangang Nei, Ying xie,Ping luo, and Haohaui published in November 2020.Here it use random forest method. Big Data and data mining technology in the social sciences through the research on happiness and demonstrates its promise and , and the results are of practical significance.

Severity Prediction with Machine Learning Methods the authors is Buket giyek Computer Engineering Department, Istanbul Kultur University, Istanbul, Turkey. Multi-layer Perceptron (MLP), Decision Tree classifier, and Random Forest classifier and Naive Bayes classifier methods are used in the process. The accuracy of each methods are Decision tree –80.74%,Random forest – 85.19%, Naïve Bayes classifier-8.40%, and for Multi-layer perceptron is 86.67%. Accidents can be important for estimating accident costs, increasing safety, and determining a strategy. Although it is not possible to stop accidents, it aims to reduce injury levels. Prediction of Mental Health in Human Being Using Machine Learning authors are S.Aparna, S.M Nandhinni Mahalakshm, Kripa.M.Chouhan and published in May 2020. It uses K-nearest neighbors algorithm, Random forest, and Boosting methods for the process. The three methods accuracy are K-Nearest neighbors with 81.65% accuracy and the precision value is

91.62. Then the Random forest gives test accuracy with 83.09% and the precision value is 90.55%. And the last one is Boosting the test accuracy of boosting is 82.46% and the precision ale 87.06%. Performance of different machine learning algorithms such as AdaBoost, KNN and Random Forest were compared. ROC Curve, Confusion matrix and classification report are used to check the quality. Metrics such as accuracy, precision, F1- score, error rate, TP, TN, FP, FN are being used for comparison to find which classifier is efficient in terms of accuracy.

S.no	Title of the year	Author name & year	Algorithm method	Accuracy	Limitation
1	Network Learning Approaches To Study World Happiness	Meghna Chaudhari Siddharth & July 21, 2020	Predictive Modelling and Bayesian Networks to model the processed historical happiness index	70% (GDP per capita)	Predictive models for predicting happiness index of a country and BN for exploring causal relationships among variables
2	Emotion recognition using multi-modal data and ml techniques	Fiona Marshal, Raymond Bond Ulster University	Emotion recognition methods based on multi-channel EEG signals	90%	SVM and RF perform better than KNN and NB for EEG-based emotion recognition task. Compared
3	The geography of world happiness	Jhon F Helliwe, Haifang	To explain the national levels and changes of	-	Life evaluation average 7.4,while bottom the average

		Huang, Shun Wang & 2015	evalutions positive and negative affect.		is less than half that at, 3.4

4	Analysis happiness In EU Countries Using The Multi-Model Classification	Marcin Pełka & 2019	Multidimensional scaling&decision tree to find out factors determine cluster memberships.	0.5800769(3 clusters)	Modelbased clustering ensemble to determine selected European countries have similar patterns of happiness.
5	Happiness recognition from mobilephone Data	Andrey Bogomolv Bruno Lepri & 2013	Final machine learning model, based on Random forest classifier	80.81%	Strong predictive power in the source and feature spaces, discuss different approaches, ml models and provide an insight for future research.
6	Measuring Happiness Around World Through Artificial intelligence	Rustem Ozakar, Rafet Efe Gazanfe, Y. Sinan Hanay	Using an unbiased emotion detector, artificial intelligence (AI)	95%(confidence interval)	Different facial structures which give unintended classifications. further work specialized dataset for purposed idea and different algorithms needs to be explored in terms of performance and accuracy
7	Happiness score	Yichen Ma1 ,	Used regression	Eco-0.638	The R-square

	identification: a regression approach	Andrew Liu2 , Xukai Yuchen Shao	approach to explore the major influencing factor	Health-0.584 Family-0.568	value of the economy stayed around 0.62 in the past five years.

8	Detecting Student Emotions in Computer-Enabled Classrooms	Nigel Bosch, Sidney K.D'Mello, Ryan S. Baker, Jaclyn Ocumpaugh & 2010	Experimented with supervised classifiers including C4.5 trees and Bayesian classifiers,	Boredom-64%, Confusion-74%, Delight-83%, Engagement-64%, Frustration-62%	Important step toward making this vision a reality, by demonstrating the feasibility of automated detection of student affect in a noisy real-world environment a school.
9	Sentiment Analysis of Big Data: Methods, Applications and Open Challenges	June 28, 2018	Technical aspect of OMSA and non-tech aspect in the form of application areas	-	Twitter has been used as a tool for disseminating and propagating information rather than simply a social networking site
10	Clustering countries according to the world happiness report 2019	M.Mujiya Ulhaq and Agra Adyatama	K-metoids selected to illustrate three distinguished clusters	87%	Study expected to give an insight into how to implement clustering algorithm into the real-world entity dataset.

11	Investigation of Happiness of Countries with K-Means Clustering and Discriminant Analysis	Sadullah CELİK1 & Necmiye CÖMERTLER2 &2020	K-Means clustering and discriminant analysis.	80.23%	New visualization approach based discriminant analysis can be proposed in which the K-Means cluster analysis results can be interpreted

12	Vocal-based emotion recognition using random forests and decision tree	FatemehNoroozi, Dorota kaminshka, Tomaz sapinski & 2015	Using random forests and decision tree approach	Average recognition rate 66.28%	The average recognition rate obtained by the rate proposed method was also compared with another work using DNN approach by 6.58% where the same dataset was utilized while taking features such as pitch, energy, MFCC and teager energy operator into account.
13	The transnational happiness study with Big data Technology	Lingxi peng,Yangang Nei, Ying xie,Ping luo, Haohaui lie & Nov 2020	Using random forest method	-	Big Data and data mining technology in the social sciences through the research on happiness and demonstrates its promise and , and the results are of practical significance

14	Severity Prediction with Machine Learning Methods	Buket giyek Computer Engineering Department, Istanbul Kultur University, Istanbul, Turkey	Multi-layer Perceptron (MLP), Decision Tree classifier, and Random Forest classifier and Naive Bayes classifier	Decision tree –80.74% Random forest – 85.19% Naïve Bayes classifier- 8.40% MLP- 86.67%	Accidents can be important for estimating accident costs, increasing safety, and determining a strategy. Although it is not possible to stop accidents, it aims to reduce injury levels.
15	Prediction of Mental Health in Human Being Using Machine Learning	S.Aparna, S.M Nandhinni Mahalakshm, Kripa.M.Chouhan & May 2020	KNN, Random forest, Boosting	KNN- 81.65% Random forest – 83.09% Boosting – 82.46%	Metrics such as accuracy, precision, F1-score, error rate, TP, TN, FP, FN are being used for comparison to find which classifier is efficient in terms of accuracy.

Table 2.1 : Summary of Literature review

From the above referenced papers it is clear that the world happiness report are processed for analyzing different features using predictive modeling. The processed results could be useful in various forms like Prediction of Mental Health using Machine Learning, The geography of world happiness, Network Learning Approaches. To Study World Happiness etc. This paper is based on applying various machine learning techniques for predicting the happiness report using various methods.

CHAPTER 3

METHODOLOGY

Decision Tree and Random forest Regression is used to predict the happiness of the data based on the important features. The validation process using k-fold cross validation technique is used to measure the performance of the data based on the accuracy, sensitivity and specificity values. It is particularly useful for assessing model performance, as it provides a range of accuracy scores across (somewhat) different data sets.

- K-Folds cross validation is one method that attempts to maximize the use of the available data for training and then testing a model.
- It is particularly useful for assessing model performance, as it provides a range of accuracy scores across (somewhat) different data sets.

Here the goal of K-fold technique is to compare the regression model and find the better accuracy. Figure 3.1 shows the frame work of this methodology.

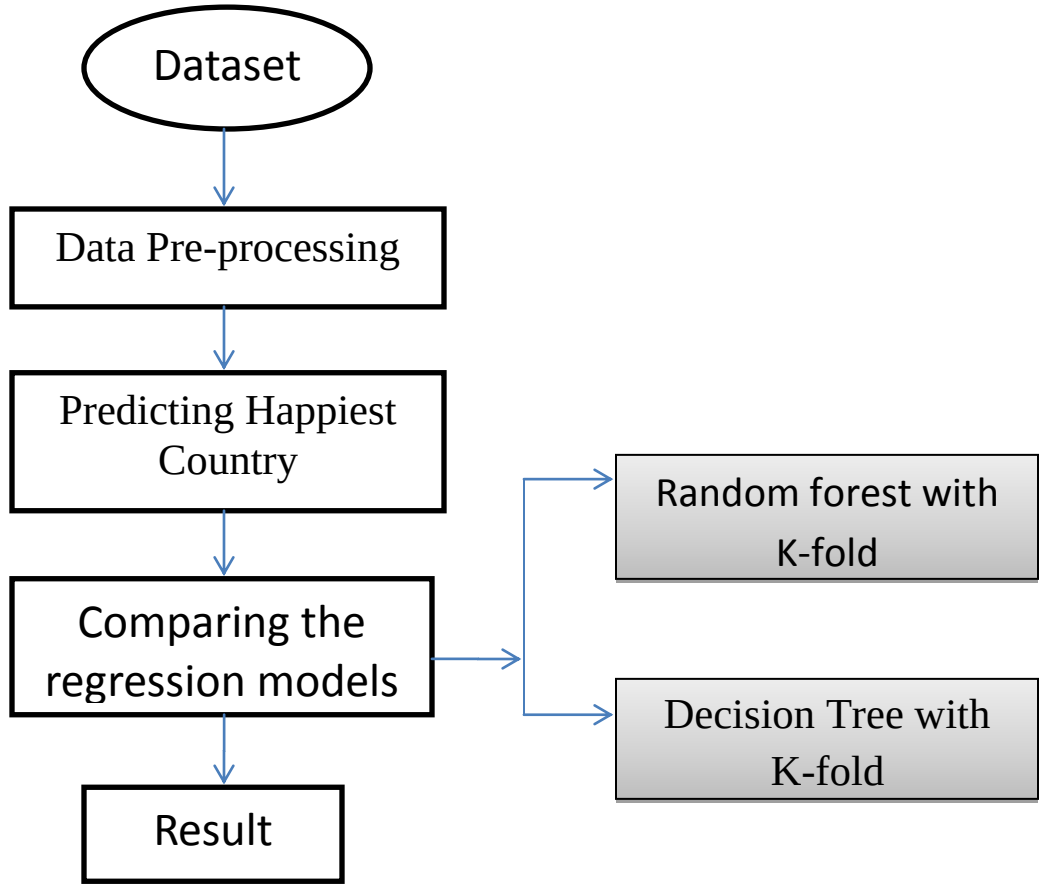


Figure 3.1: Methodology Diagram

3.1 DATA COLLECTION

- Data can be collected using three main types of surveys: censuses, sample surveys, and administrative data. Each has advantages and disadvantages. Data collection is defined as the procedure of collecting, measuring and analyzing accurate insights for research using standard validated techniques.

- A researcher can evaluate their hypothesis on the basis of collected data. Each column represents a particular variable. Each row corresponds to a given member of the dataset in question. It lists values for each of the variables, such as height and weight of an object. Each value is known as a datum.

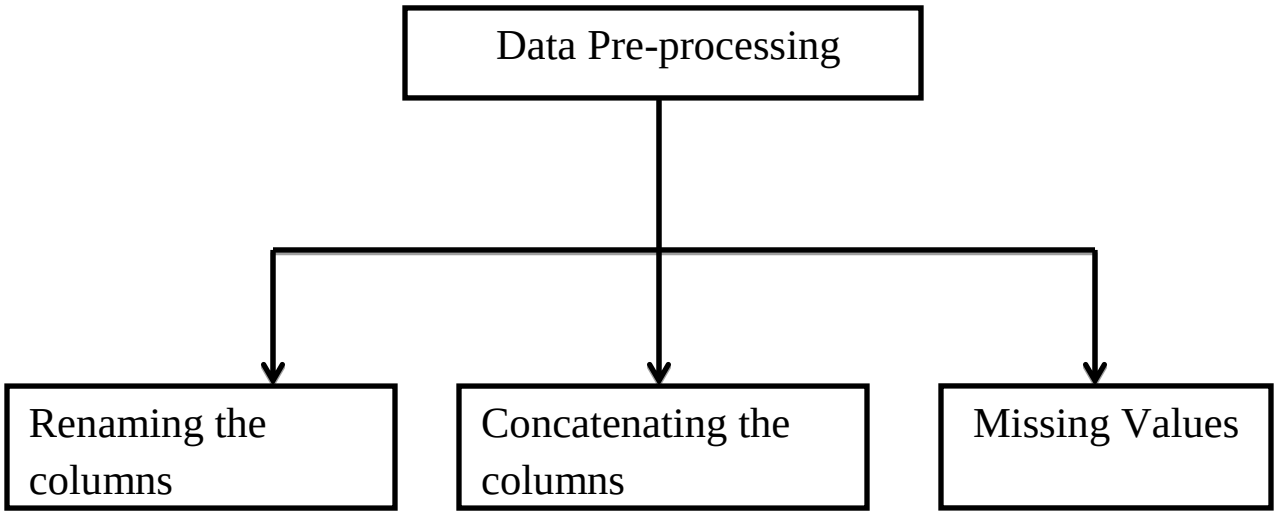
- Data collections from Google-generated data, such as Google Analytics or Google Sheets. A data source based on a CSV file. Metrics and dimensions typed directly into Data Studio.

3.2. DATA PRE-PROCESSING

- To make the process easier, data preprocessing is divided into four stages: data cleaning, data integration, data reduction, and data transformation.
- It is a data mining technique that transforms raw data into an understandable format. Raw data (real world data) is always incomplete and that data cannot be sent through a model. That would cause certain errors. That is why we need to preprocess data before sending through a model.
    - DataCleaning:

      The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data.
    - Regression:

      Here data can be made smooth by fitting it to a regression function. The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).



3.3 FEATURE ENGINEERING

Feature engineering is useful to improve the performance of machine learning algorithms and is often considered as applied machine learning. Feature is also reffered to as ‘variables’ or ‘attributes’ as they affect the output of a process. Features extraction involves choosing set of features from a large collection. From the preprocessed happiness reports, various features are extracted as per the semantics and are converted into probabilistic values. Here, the extracted features are GDP per capita, Healthy Life Expectancy, Social support, Freedom to make life choices, Generosity, Corruption Perception, Residual error.

3.4 COMPARING REGRESSION MODEL

DECISION TREE AND RANDOM FOREST REGRESSION

DECISION TREE:

- Decision Trees are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.
- For instance, in the example below, decision trees learn from data to approximate a sine curve with a set of if-then-else decision rules. The deeper the tree, the more complex the decision rules and the fitter the model.

Some advantages of decision trees are:

- Simple to understand and to interpret. Trees can be visualized. Requires little data preparation. Other techniques often require data normalization, dummy variables need to be created and blank values to be removed. Note however that this module does not support missing values.
- The cost of using the tree (i.e., predicting data) is logarithmic in the number of data points used to train the tree. Able to handle both numerical and categorical data. However scikit-learn implementation does not support categorical variables for now. Other techniques are usually specialized in analyzing datasets that have only one type of variable.
- Able to handle multi-output problems. Uses a white box model. If a given situation is observable in a model, the explanation for the condition is easily explained by Boolean logic. By contrast, in a black box model (e.g., in an artificial neural network), results may be more difficult to interpret.

The disadvantages of decision trees include:

- Decision-tree learners can create over-complex trees that do not generalize the data well. This is called over fitting. Mechanisms such as pruning, setting the minimum number of samples required at a leaf node or setting the maximum depth of the tree are necessary to avoid this problem.
- Decision trees can be unstable because small variations in the data might result in a completely different tree being generated. This problem is mitigated by using decision trees within an ensemble.
- Decision tree learners create biased trees if some classes dominate. It is therefore recommended to balance the dataset prior to fitting with the decision tree.

3.4.1 DECISION TREE REGRESSION:

- Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.
- Discrete output example**:** A weather prediction model that predicts whether or not there'll be rain in a particular day.
- Continuous output example**:** A profit prediction model that states the probable profit that can be generated from the sale of a product.

Algorithm:

Step 1: Select features from a given dataset.
Step 2: Construct a decision tree for each sample and get a prediction result from each decision tree.
Step 3: Perform a vote for each predicted result.
Step 4: Select the prediction result with the most votes as the final prediction.

RANDOM FOREST:

- Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning**,** which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.
- Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

ADVANTAGES OF RANDOM FOREST:

- Random Forest is capable of performing both Classification and Regression tasks.
- It is capable of handling large datasets with high dimensionality.
- It enhances the accuracy of the model and prevents the over fitting issue.

DISADVANTAGES OF RANDOM FOREST:

- It requires much computational power as well as resources as it builds numerous trees to combine their outputs.
- It also requires much time for training as it combines a lot of decision trees to determine the class.
- Due to the ensemble of decision trees, it also suffers interpretability and fails to determine the significance of each variable.

3.4.2 RANDOM FOREST REGRESSION:

- A random forest is an ensemble model that consists of many decision trees. Predictions are made by averaging the predictions of each decision tree. Or, to extend the analogy—much like a forest is a collection of trees, the random forest model is also a collection of decision tree models. This makes random forests a strong modeling technique that's much more powerful than a single decision tree.

- Each tree in a random forest is trained on the subset of data provided. The subset is obtained both with respect to rows and columns.

- This means each random forest tree is trained on a random data point sample, while at each decision node, a random set of features is considered for splitting. In the realm of machine learning, the random forest regression algorithm can be more suitable for regression problems than other common and popular algorithms**.**

Algorithm:

Step 1: Create the training and test data

Step 2: Fit the model on training data and predict list on test data

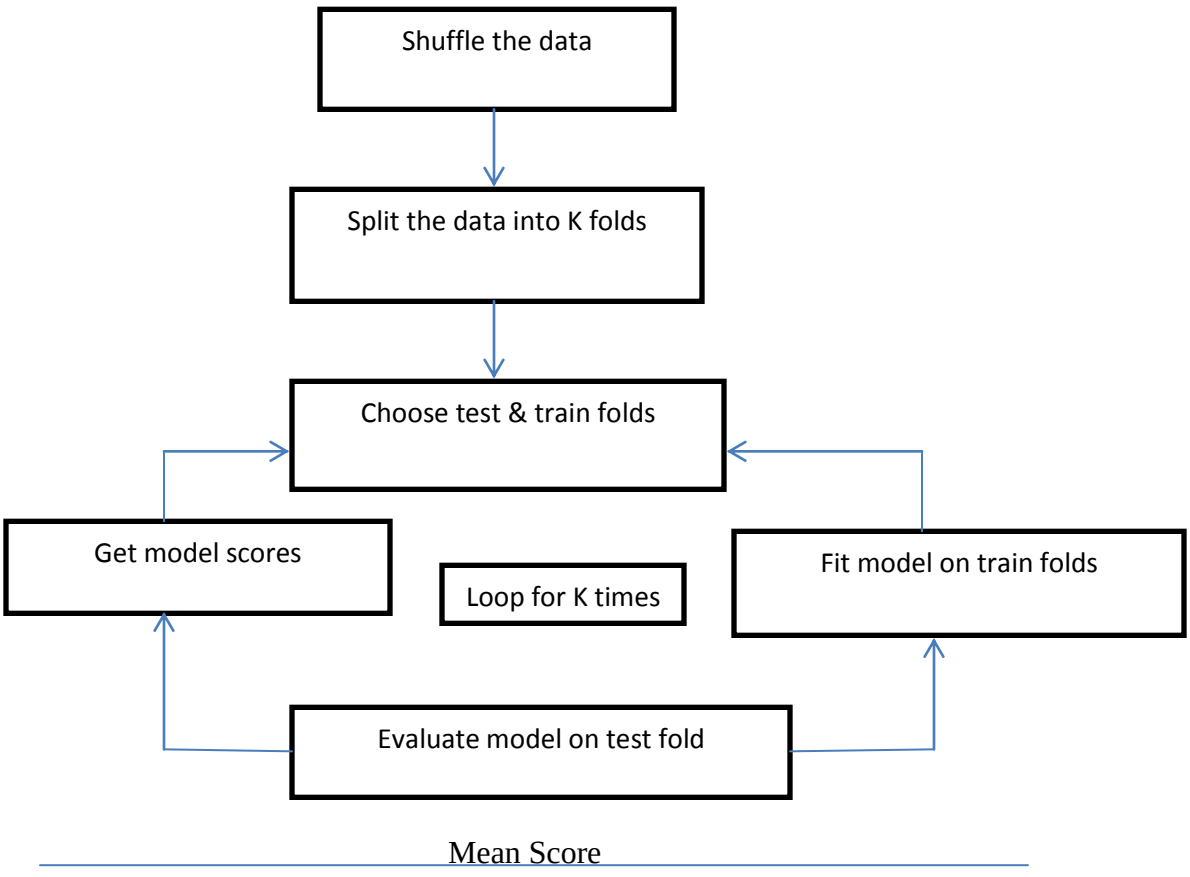
Step 3: Calculate prediction accuracy using happiness score.

3.4.3 DIFFERENCE BETWEEN DECISION TREE AND RANDOM FOREST

DECISION TREE	RANDOM FOREST
Decision Tree is a supervised learning algorithm used in machine learning	It is also used for supervised learning but is very powerful
It operated in both classification and regression algorithms	It is very widely used. The basic difference being it does not rely on a singular decision
It is like a tree with nodes. The branches depend on the number of criteria. It splits data into branches like these till it achieves a threshold unit. A decision tree has root nodes, children nodes, and leaf nodes.	It does not search for the best prediction. Instead, it makes multiple random predictions. Thus, more diversity is attached, and prediction becomes much smoother.

3.4.4 K FOLD CROSS VALIDATION

- Cross validation is an approach that you can use to estimate the performance of a machine learning algorithm with less variance than a single train-test set split. It works by splitting the dataset into k-parts (e.g. k=5 or k=10). Each split of the data is called a fold. The algorithm is trained on k-1 folds with one held back and tested on the held back fold. This is repeated so that each fold of the dataset is given a chance to be the held back test set. After running cross validation you end up with k different performance scores that you can summarize using a mean and a standard deviation. The result is a more reliable estimate of the performance of the algorithm on new data given your test data. It is more accurate because the algorithm is trained and evaluated multiple times on different data.
- The choice of k must allow the size of each test partition to be large enough to be a reasonable sample of the problem, whilst allowing enough repetitions of the train-test evaluation of the algorithm to provide a fair estimate of the algorithms performance on unseen data. For modest sized datasets in the thousands or tens of thousands of records, k values of 3, 5 and 10 are common.



3.5 DATA ANALYSIS

In statistics, data analysis is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily exploratory data analysis is for seeing what the data can tell us beyond the formal modelling or hypothesis testing task. Exploratory data analysis is different from initial data analysis (IDA), which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed.

3.6 DATA VISUALIZATION

Data visualization is the presentation of data in a pictorial or graphical format. A primary goal of data visualization is to communicate information clearly and efficiently via statistical graphics, plots and information graphics. Numerical data may be encoded using dots, lines, or bars, to visually communicate a quantitative message. Effective visualization helps users analyze and reason about data and evidence. It makes complex data more accessible, understandable and usable. Processing, analyzing and communicating this data present ethical and analytical challenges for data visualization.

3.6.1 THE REQUIRED PACKAGES

Matplotlib: This Python package used for data plotting and visualization. It is a useful complement to Pandas, and like it is a very feature-rich library which can produce a large variety of plots, charts, maps, and other visualizations.

sklearn. model selection: It is a Python library that offers various features for data processing that can be used for classification, clustering, and model selection. Model selection is a method for setting a blueprint to analyze data and then using it to measure new data.

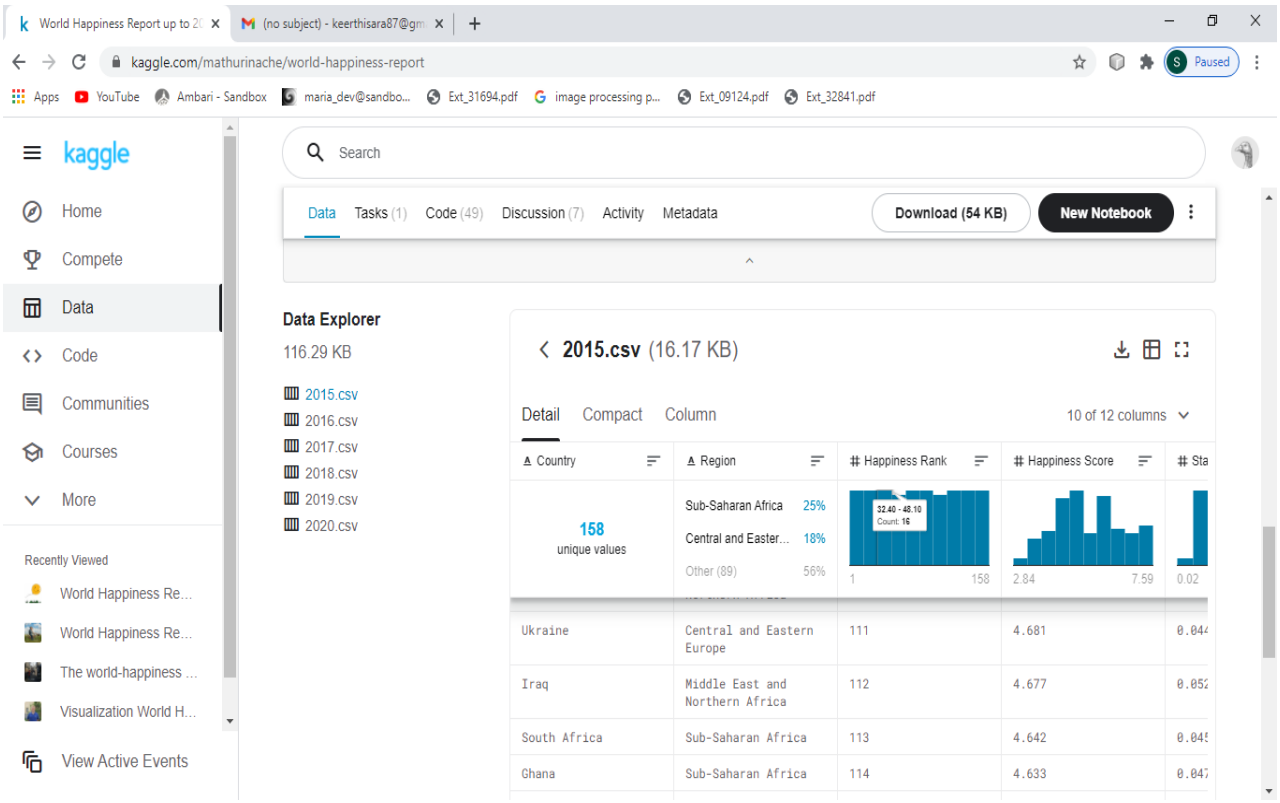
sklearn. metrics: The module implements several loss, score, and utility functions to measure classification performance. Some metrics might require probability estimates of the positive class, confidence values, or binary decisions values.

CHAPTER-4

IMPLEMENTATION AND RESULT

4.1. DATASET COLLECTION:

As mentioned in the proposed section, the data used for this work are gathered from the website www.kaggle.com. Here we use 2015-2020 dataset to predict the happiest country.



DATA SET DESCRIPTION:

- Each contains the Happiness Score for 153 countries along with the factors used to explain the score. The Happiness Score is a national average of the responses to the main life evaluation question asked in the Gallup World Poll (GWP), which uses the Cantril Ladder.
- The Happiness Score is explained by the following factors:
  - GDP per capita
  - Healthy Life Expectancy
  - Social support
  - Freedom to make life choices
  - Generosity
  - Corruption Perception
  - Residual error

Attributes	Type
Overall rank	Number
Country or region	Var char
Score	Float
GDP per capita	Float
Social support	Float
Healthy life expectancy	Float
Freedom to make life choices	Float
Generosity	Float
Perceptions of corruption	Float

EXPLANATION OF COLUMN

COLUMN NAME	EXPLANATION
Country	Name of the country
Region	Region the country belongs to
Happiness Rank	Rank of the country based on the Happiness Score
Standard Error	The Standard Error of the Happiness Score
Economy(GDP per Capita)	The extent to which GDP contributes to the calculation of the Happiness Score.
Family	The extent to which Family contributes to the calculation of the Happiness score
Health(Life Expectancy)	Life expectancy contributed to the calculation of the Happiness Score
Trust(Government Corruption)	Perception of corruption contributed to Happiness Score
Generosity	Generosity contributed to the calculation of the Happiness Score
Freedom	Freedom contributed to the calculation of the Happiness Score
Dystopia Residual	Dystopia Residual contributed to the calculation of the Happiness Score

4.1.1 Data pre-processing

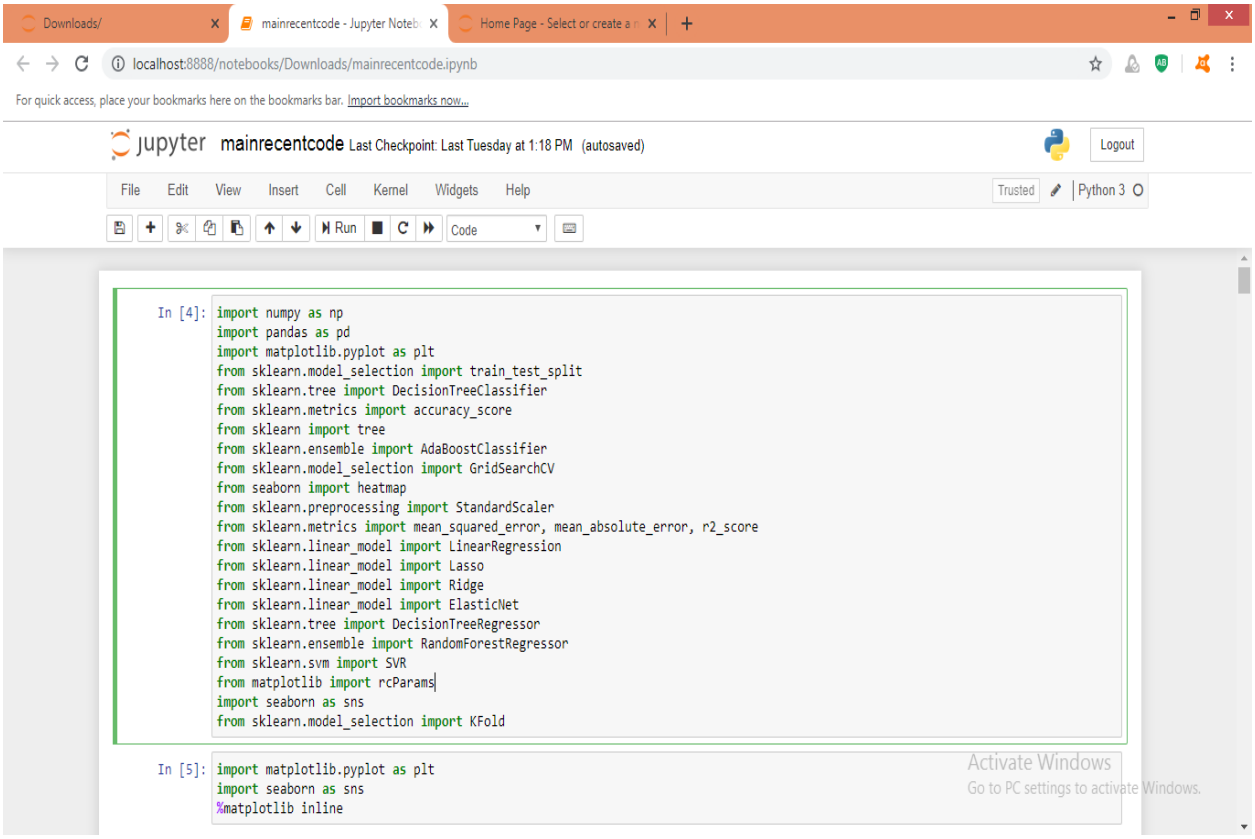
- In order to increase the data quality, the data preprocessing is needed.
- Most important step before we start working on the data is Assess the Raw Data and Clean the Data in this report we have 6 with each excels contains the Report from 2015 to 2020.

- The interesting part is: Not all the Excels have same column Names and 2020 report has few extra columns which the other Datasets doesn't have. So data pre-processing is one of the mandatory step in this project.

IMPORTING PACKAGES:

primary packages are going to be Pandas to work with data, NumPy to work with arrays, scikit-learn for data split, building and evaluating the classification models.

#import numpy as np
#import pandas as pd
#import matplotlib.pyplot as plt
#import seaborn as sns



IMPORTING DATASET:

df_2015 = pd.read_csv('2015.csv')

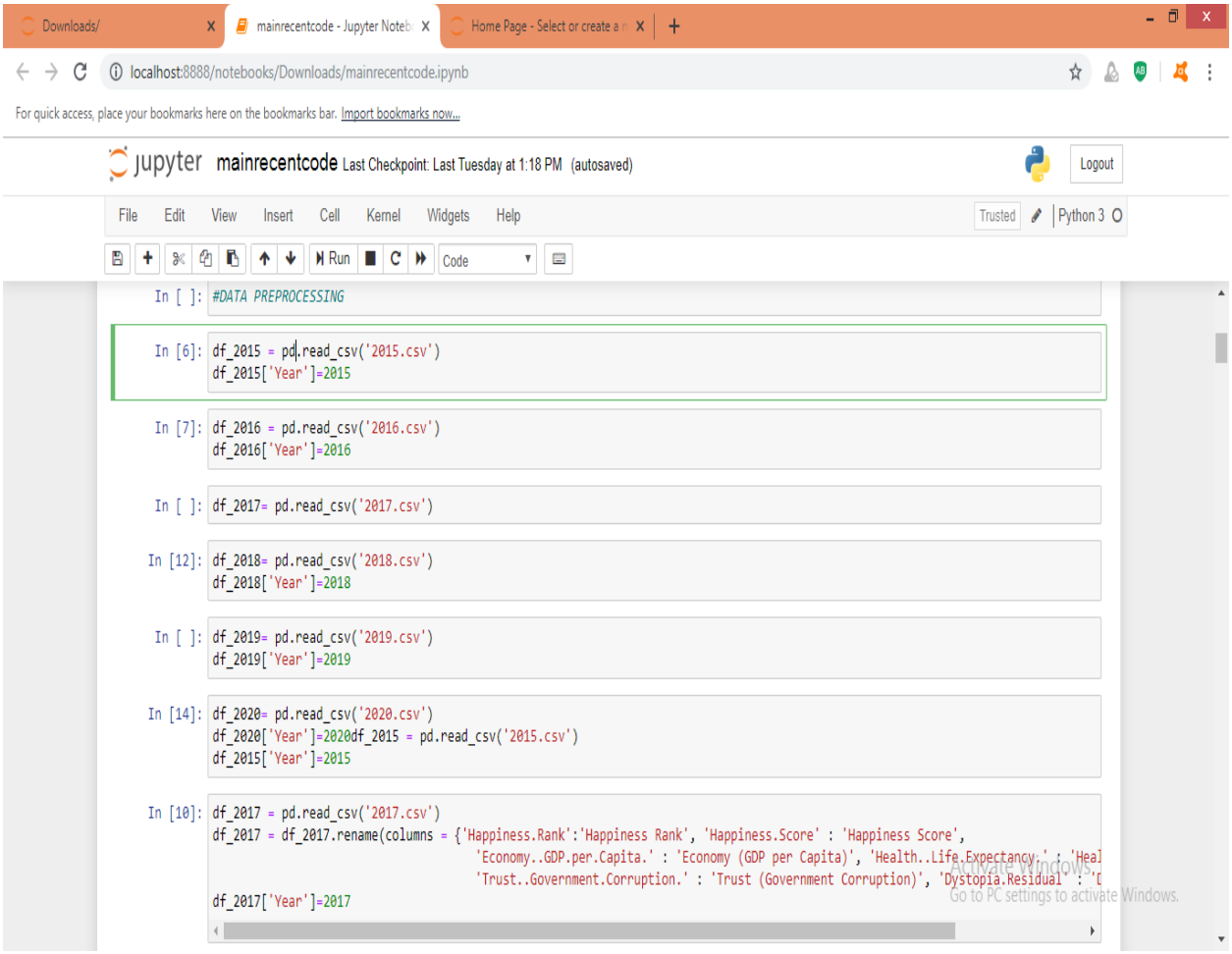
df_2016 = pd.read_csv('2016.csv')

df_2017 = pd.read_csv('2017.csv')

df_2018 = pd.read_csv('2018.csv')

df_2019 = pd.read_csv('2019.csv')

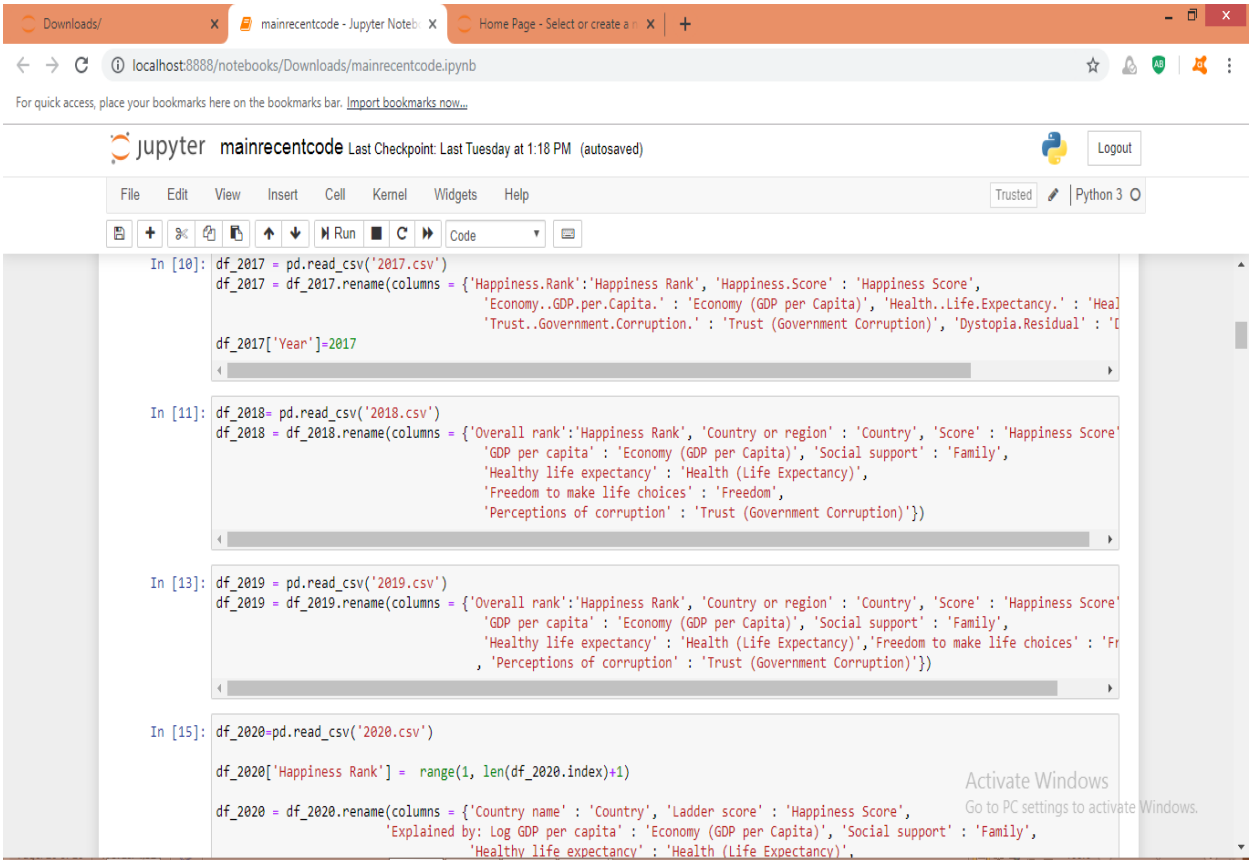
df_2020 = pd.read_csv('2020.csv')



Pre-processing:

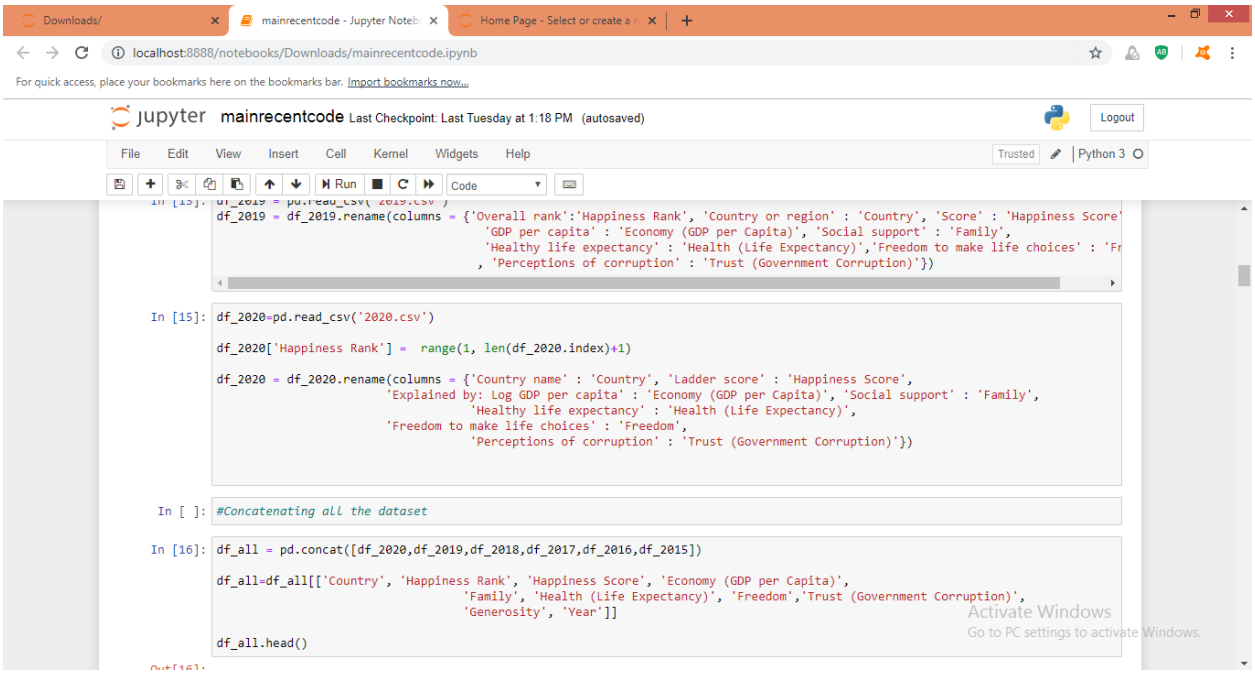
- I combined all happiness report from 2015 to 2020 together to work smoothly with the dataset.
- To combine properly, I created a Year column to sort out in the future and I made sure that each columns of files have the same name which I wanted for my further analysis.
- They are Happiness Rank, Country, Happiness Score, GDP, Social Support, Healthy Life Expectancy, Freedom, Generosity, and Perception of Corruption.

#RENAMING THE COLUMNS



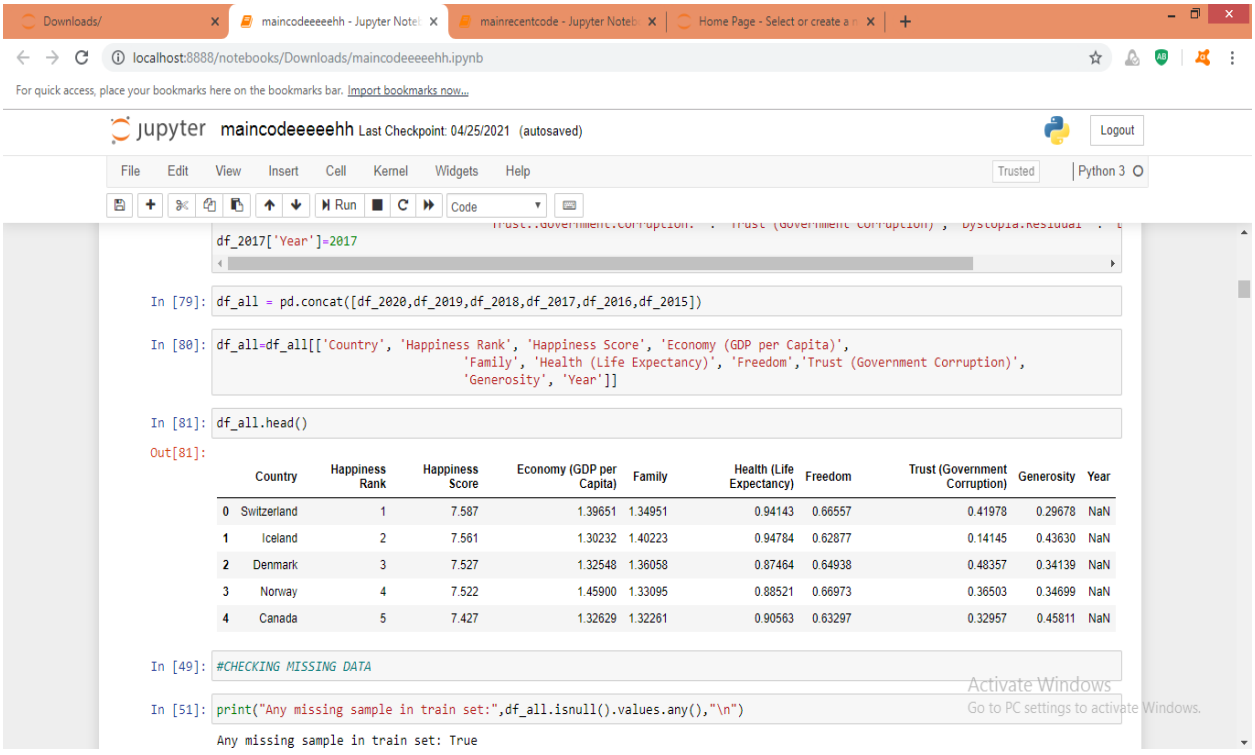
```
In [10]: df_2017 = pd.read_csv('2017.csv')
         df_2017 = df_2017.rename(columns = {'Happiness.Rank':'Happiness Rank', 'Happiness.Score' : 'Happiness Score',
                                             'Economy..GDP.per.Capita.' : 'Economy (GDP per Capita)', 'Health..Life.Expectancy.' : 'Heal
                                             'Trust..Government.Corruption.' : 'Trust (Government Corruption)', 'Dystopia.Residual' : 'D
         df_2017['Year']=2017

In [11]: df_2018= pd.read_csv('2018.csv')
         df_2018 = df_2018.rename(columns = {'Overall rank':'Happiness Rank', 'Country or region' : 'Country', 'Score' : 'Happiness Score'
                                             'GDP per capita' : 'Economy (GDP per Capita)', 'Social support' : 'Family',
                                             'Healthy life expectancy' : 'Health (Life Expectancy)',
                                             'Freedom to make life choices' : 'Freedom',
                                             'Perceptions of corruption' : 'Trust (Government Corruption)'})

In [13]: df_2019 = pd.read_csv('2019.csv')
         df_2019 = df_2019.rename(columns = {'Overall rank':'Happiness Rank', 'Country or region' : 'Country', 'Score' : 'Happiness Score'
                                             'GDP per capita' : 'Economy (GDP per Capita)', 'Social support' : 'Family',
                                             'Healthy life expectancy' : 'Health (Life Expectancy)','Freedom to make life choices' : 'Fr
                                            , 'Perceptions of corruption' : 'Trust (Government Corruption)'})

In [15]: df_2020=pd.read_csv('2020.csv')

         df_2020['Happiness Rank'] =  range(1, len(df_2020.index)+1)

         df_2020 = df_2020.rename(columns = {'Country name' : 'Country', 'Ladder score' : 'Happiness Score',
                                'Explained by: Log GDP per capita' : 'Economy (GDP per Capita)', 'Social support' : 'Family',
                                'Healthy life expectancy' : 'Health (Life Expectancy)',
```


#CONCATENATE ALL THE YEARS DATA INTO A SINGLE DATA FRAME



```
In [15]: df_2020=pd.read_csv('2020.csv')

         df_2020['Happiness Rank'] =  range(1, len(df_2020.index)+1)

         df_2020 = df_2020.rename(columns = {'Country name' : 'Country', 'Ladder score' : 'Happiness Score',
                                'Explained by: Log GDP per capita' : 'Economy (GDP per Capita)', 'Social support' : 'Family',
                                'Healthy life expectancy' : 'Health (Life Expectancy)',
                                'Freedom to make life choices' : 'Freedom',
                                'Perceptions of corruption' : 'Trust (Government Corruption)'})

In [ ]: #Concatenating all the dataset

In [16]: df_all = pd.concat([df_2020,df_2019,df_2018,df_2017,df_2016,df_2015])

         df_all=df_all[['Country', 'Happiness Rank', 'Happiness Score', 'Economy (GDP per Capita)',
                        'Family', 'Health (Life Expectancy)', 'Freedom','Trust (Government Corruption)',
                        'Generosity', 'Year']]

         df_all.head()
```

#PRINT TOP 5 ROWS FROM THE DATASET



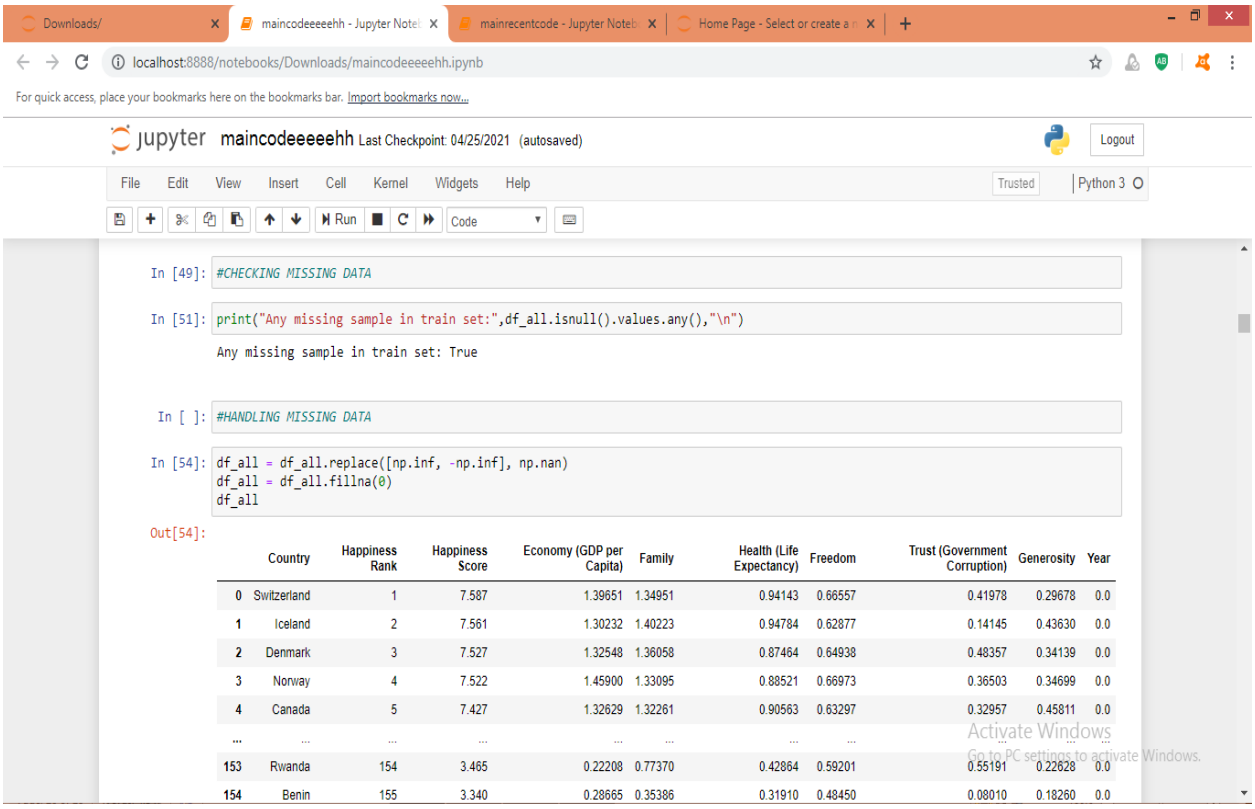
```
df_2017['Year']=2017

In [79]: df_all = pd.concat([df_2020,df_2019,df_2018,df_2017,df_2016,df_2015])

In [80]: df_all=df_all[['Country', 'Happiness Rank', 'Happiness Score', 'Economy (GDP per Capita)',
                        'Family', 'Health (Life Expectancy)', 'Freedom','Trust (Government Corruption)',
                        'Generosity', 'Year']]

In [81]: df_all.head()
```

Out[81]:

	Country	Happiness Rank	Happiness Score	Economy (GDP per Capita)	Family	Health (Life Expectancy)	Freedom	Trust (Government Corruption)	Generosity	Year
0	Switzerland	1	7.587	1.39651	1.34951	0.94143	0.66557	0.41978	0.29678	NaN
1	Iceland	2	7.561	1.30232	1.40223	0.94784	0.62877	0.14145	0.43630	NaN
2	Denmark	3	7.527	1.32548	1.36058	0.87464	0.64938	0.48357	0.34139	NaN
3	Norway	4	7.522	1.45900	1.33095	0.88521	0.66973	0.36503	0.34699	NaN
4	Canada	5	7.427	1.32629	1.32261	0.90563	0.63297	0.32957	0.45811	NaN

```
In [49]: #CHECKING MISSING DATA

In [51]: print("Any missing sample in train set:",df_all.isnull().values.any(),"\n")

Any missing sample in train set: True
```

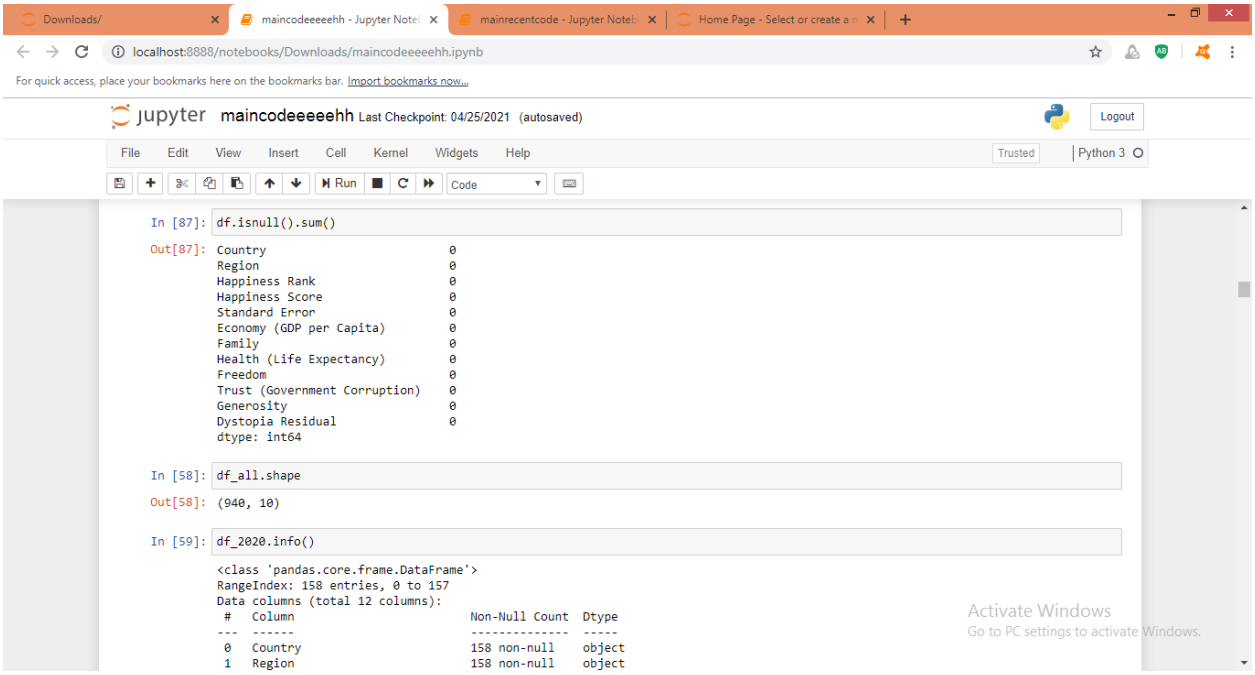

#CHECKING MISSING DATA AND HANDLING MISSING DATA



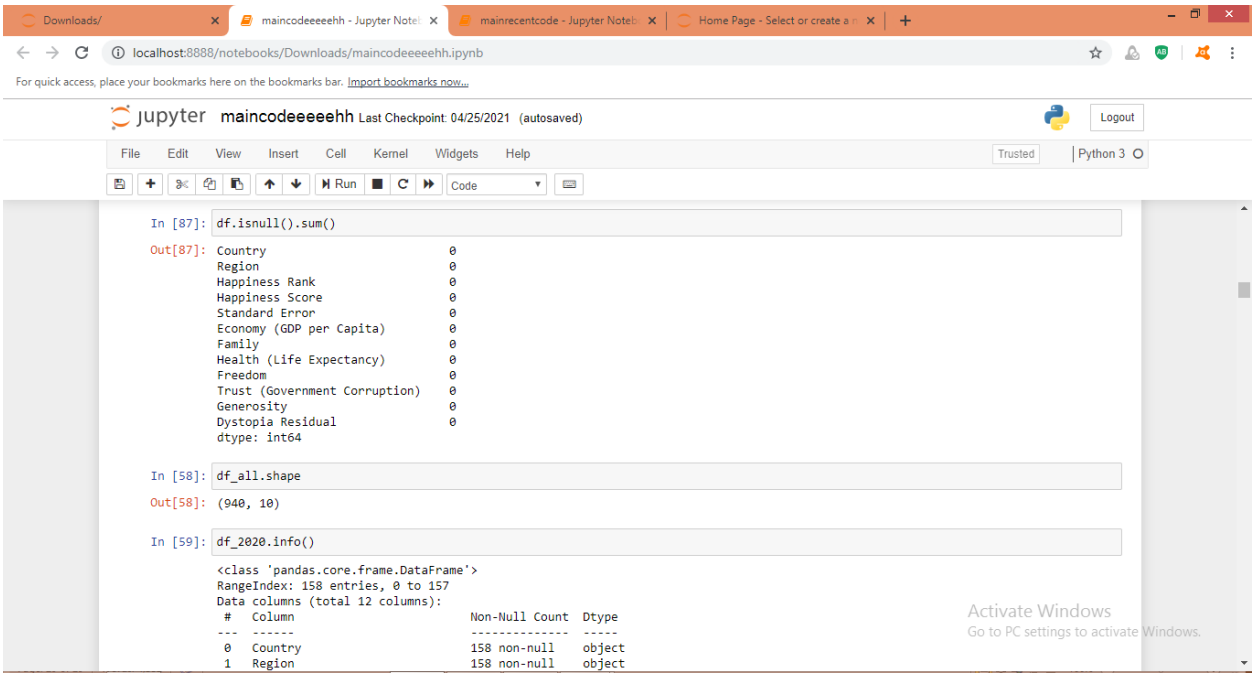
```
In [49]: #CHECKING MISSING DATA

In [51]: print("Any missing sample in train set:",df_all.isnull().values.any(),"\n")

Any missing sample in train set: True

In [ ]: #HANDLING MISSING DATA

In [54]: df_all = df_all.replace([np.inf, -np.inf], np.nan)
         df_all = df_all.fillna(0)
         df_all
```

Out[54]:

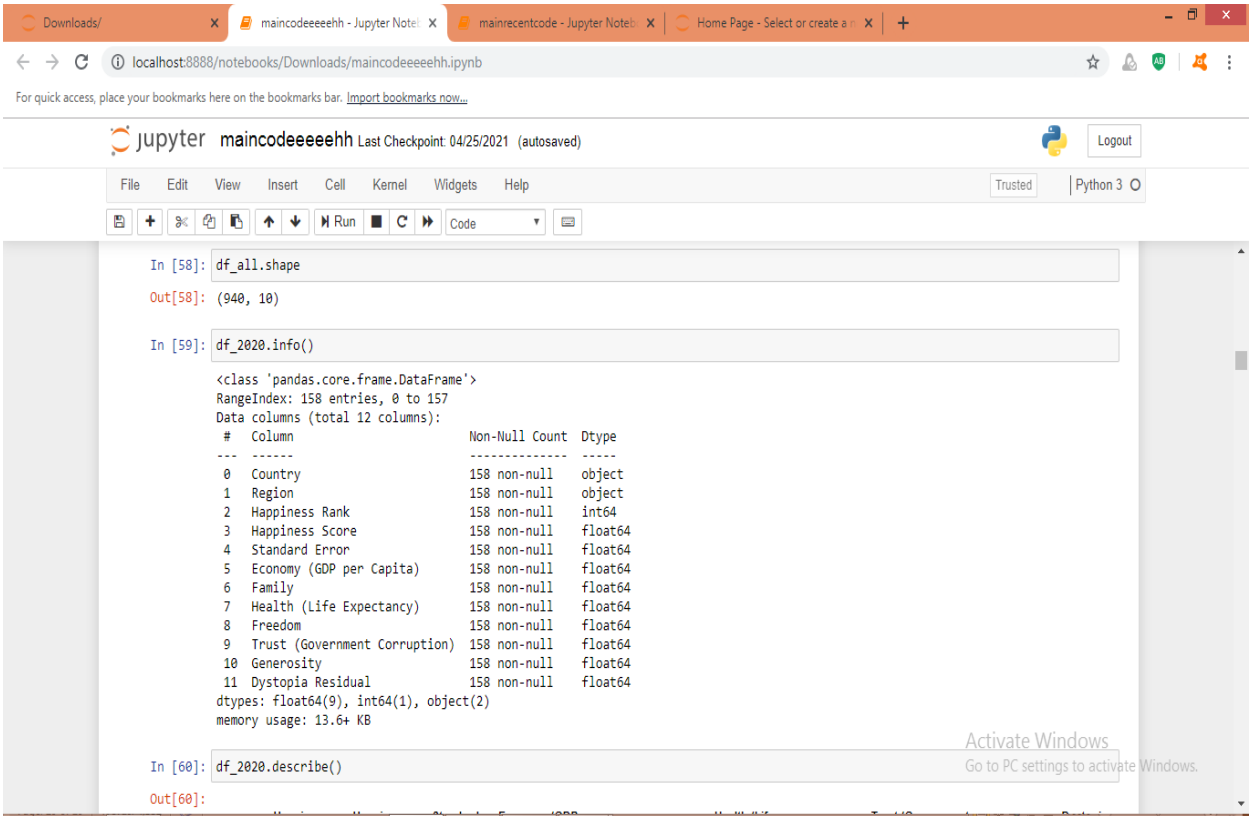
	Country	Happiness Rank	Happiness Score	Economy (GDP per Capita)	Family	Health (Life Expectancy)	Freedom	Trust (Government Corruption)	Generosity	Year
0	Switzerland	1	7.587	1.39651	1.34951	0.94143	0.66557	0.41978	0.29678	0.0
1	Iceland	2	7.561	1.30232	1.40223	0.94784	0.62877	0.14145	0.43630	0.0
2	Denmark	3	7.527	1.32548	1.36058	0.87464	0.64938	0.48357	0.34139	0.0
3	Norway	4	7.522	1.45900	1.33095	0.88521	0.66973	0.36503	0.34699	0.0
4	Canada	5	7.427	1.32629	1.32261	0.90563	0.63297	0.32957	0.45811	0.0
...	...	...	...	...	...	...	...	...	...	...
153	Rwanda	154	3.465	0.22208	0.77370	0.42864	0.59201	0.55191	0.22628	0.0
154	Benin	155	3.340	0.28665	0.35386	0.31910	0.48450	0.08010	0.18260	0.0



#FINDING NULL VALUES AND SHAPE

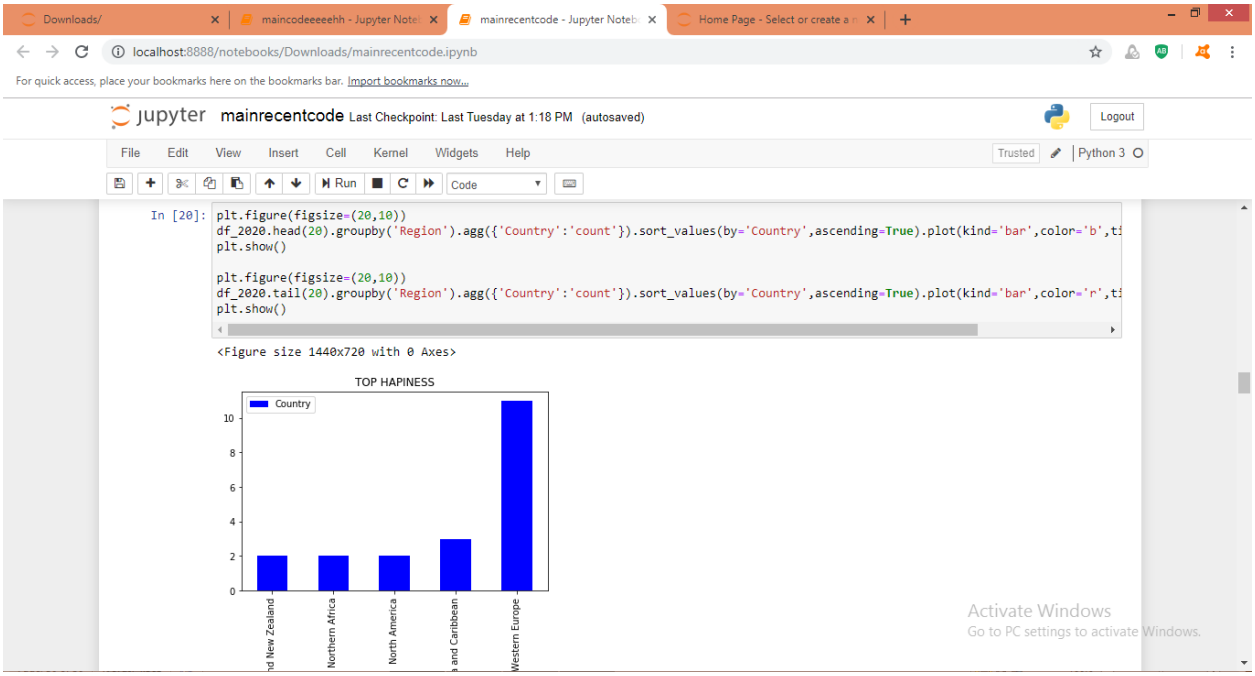


#DESCRIBING THE DATASET:

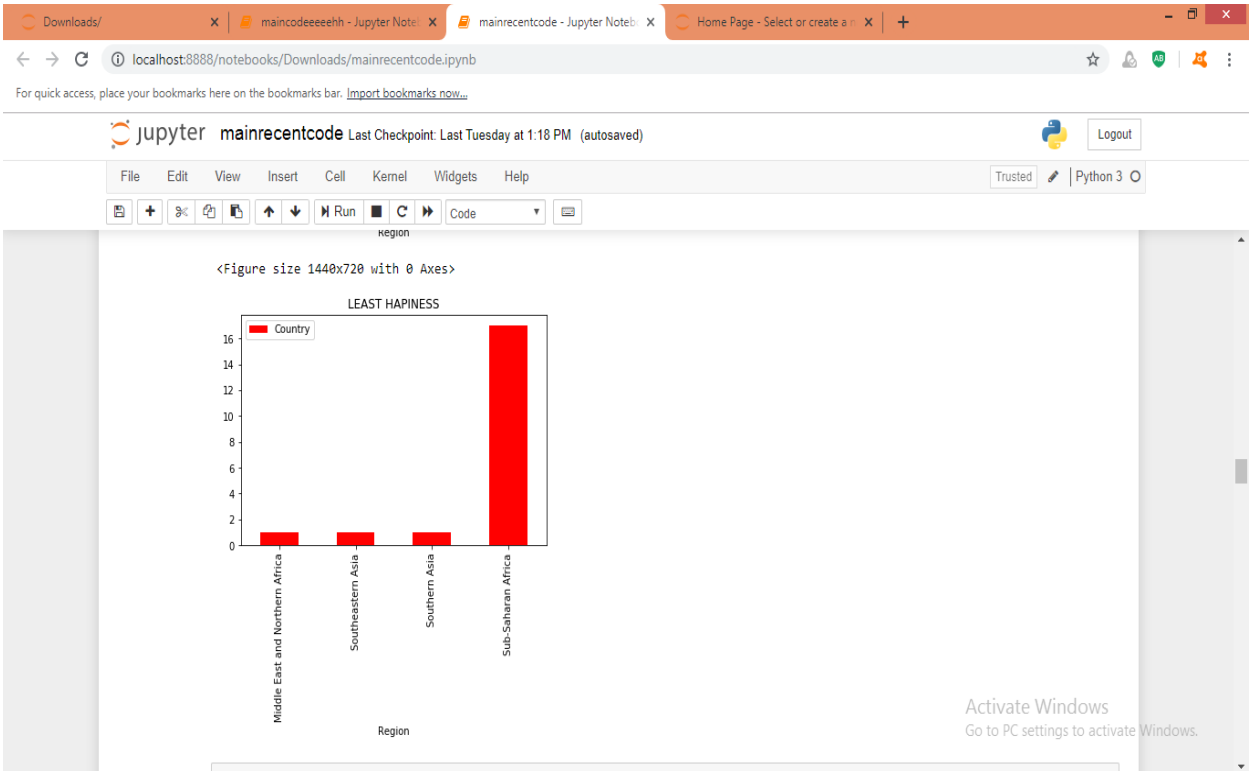


4.3 PREDICTING HAPPIEST COUNTRY

VISUALIZING REGIONS WITH TOP HAPPINESS SCORE



#VISUALIZING REGIONS WITH LEAST HAPPINESS SCORE

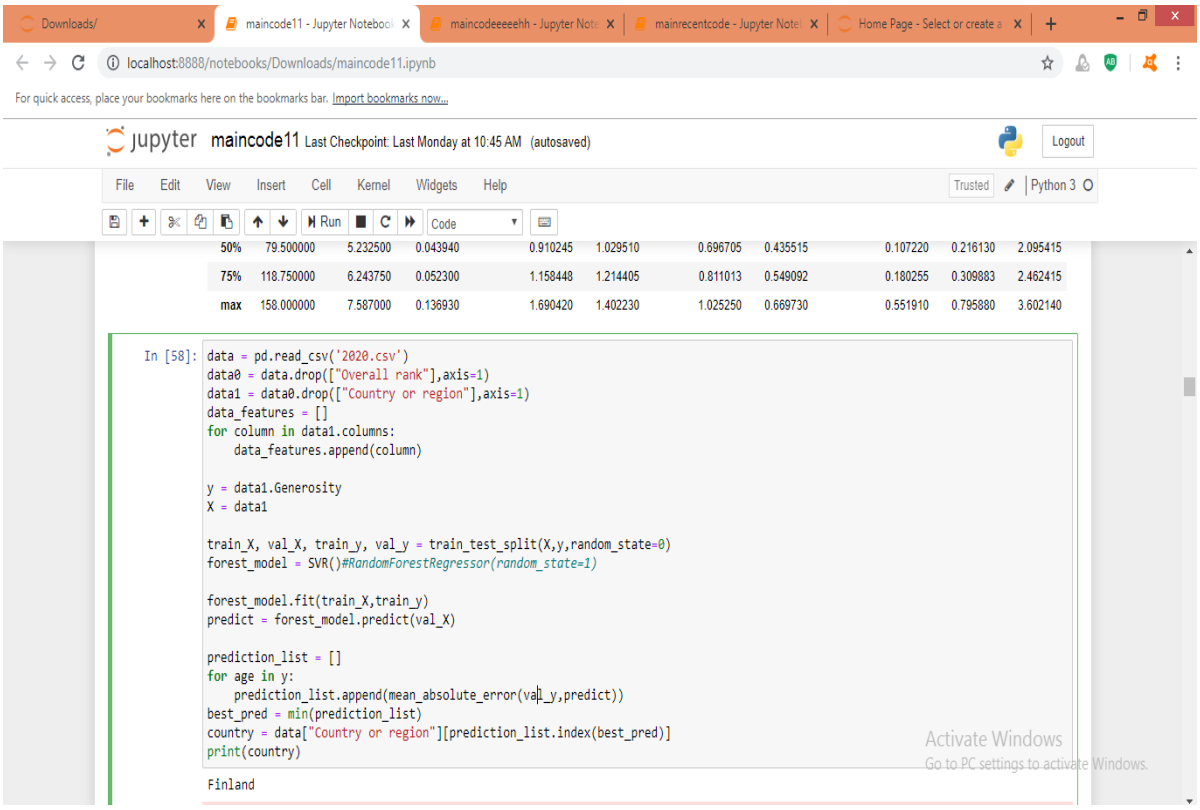


DATA SPLITTING:

- In this process, going to define the independent (X) and the dependent variables (Y).
- Using the defined variables, will split the data into a training set and testing set which is further used for modelling and evaluating.
- Split the data easily using the 'train_test_split' algorithm in python.

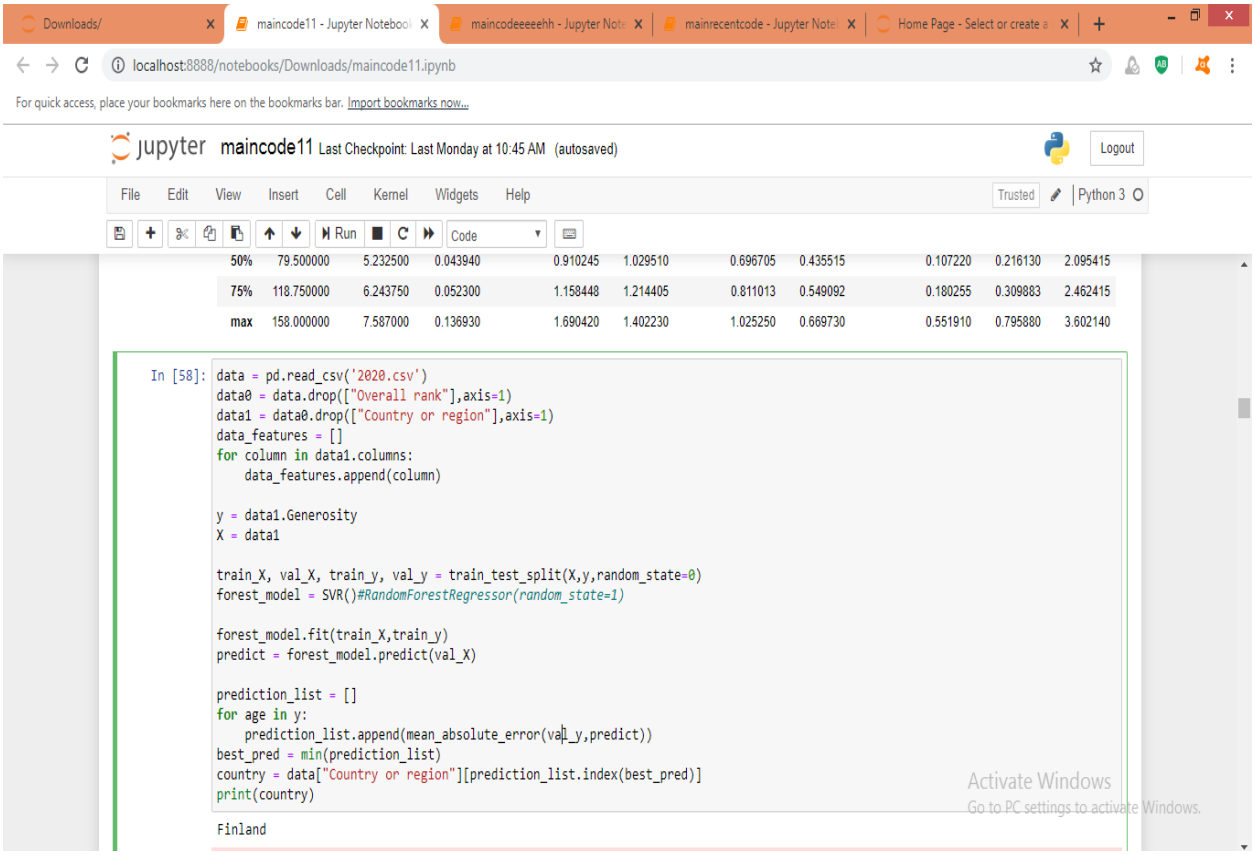
train_X, val_X, train_y, val_y = train_test_split(X,y,random_state=0)
forest_model = SVR()#RandomForestRegressor(random_state=1)

forest_model.fit(train_X,train_y)
predict = forest_model.predict(val_X)

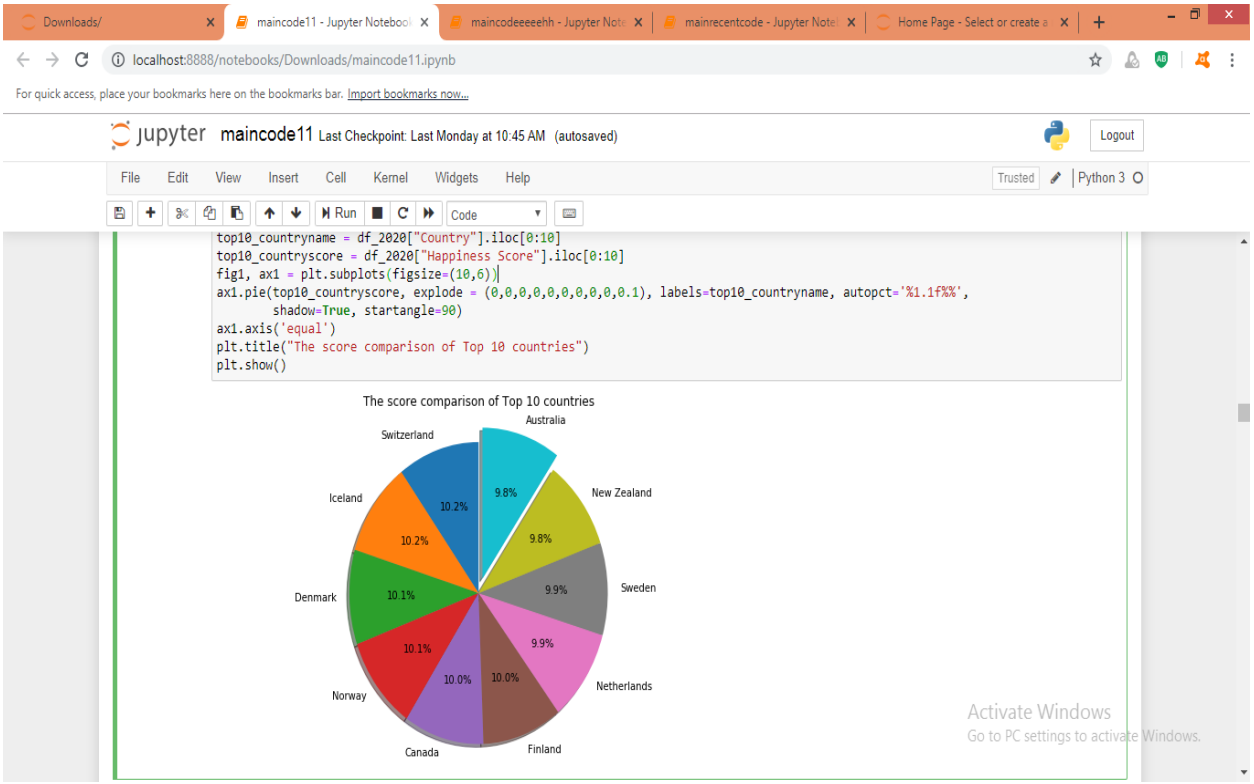


#HAPPIEST COUNTRY [FINLAND]

```
prediction_list = []

for age in y:

    prediction_list.append(mean_absolute_error(val_y,predict))

best_pred = min(prediction_list)

country = data["Country or region"][prediction_list.index(best_pred)]

print(country)
```

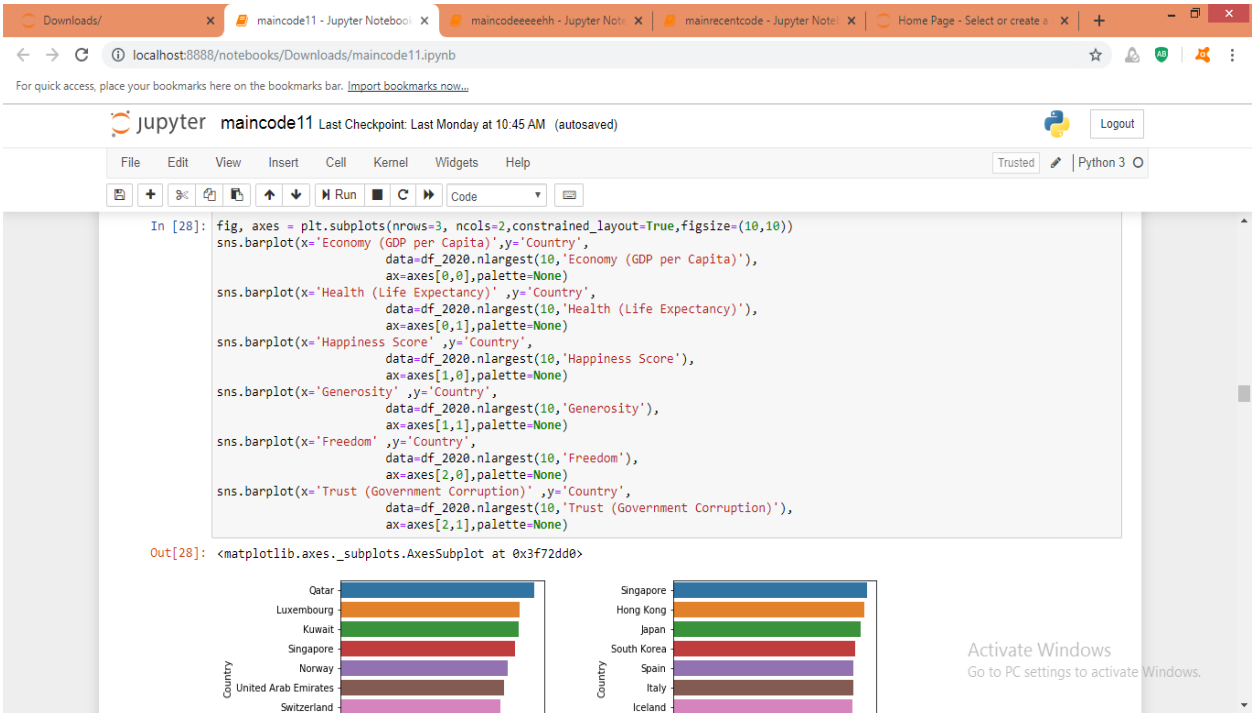


#VISUALIZING THE TOP 10 HAPPIEST COUNTRIES USING PIE CHART

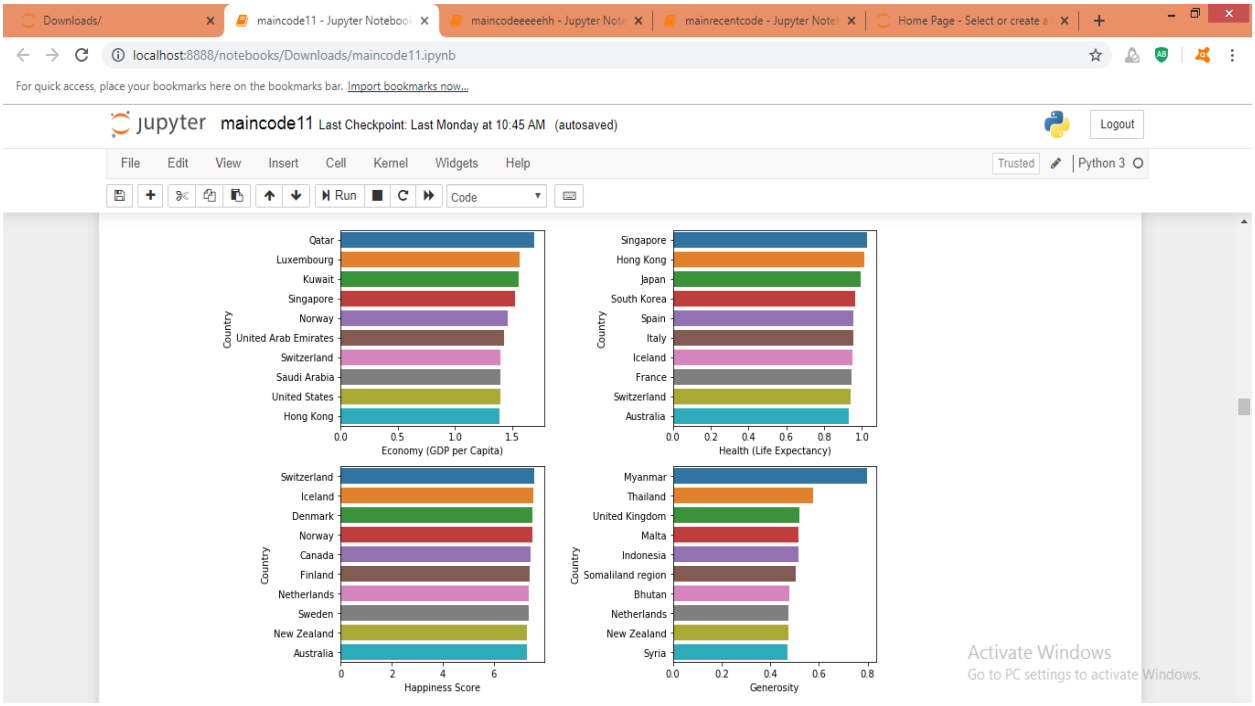


#PLOT AND VISUALIZE TOP 10 COUNTRIES FOR EACH FEATURE

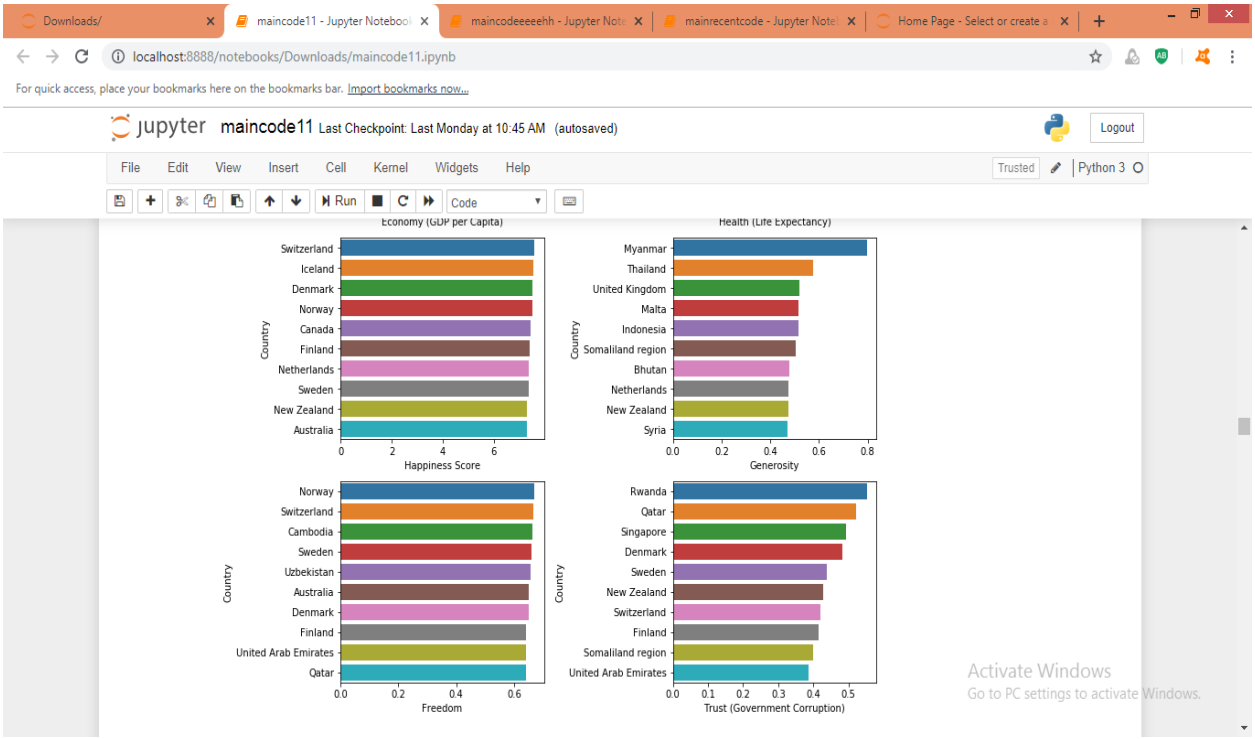
ECONOMY GDP PER CAPITA AND HEALTH [LIFE EXPECTANCY]



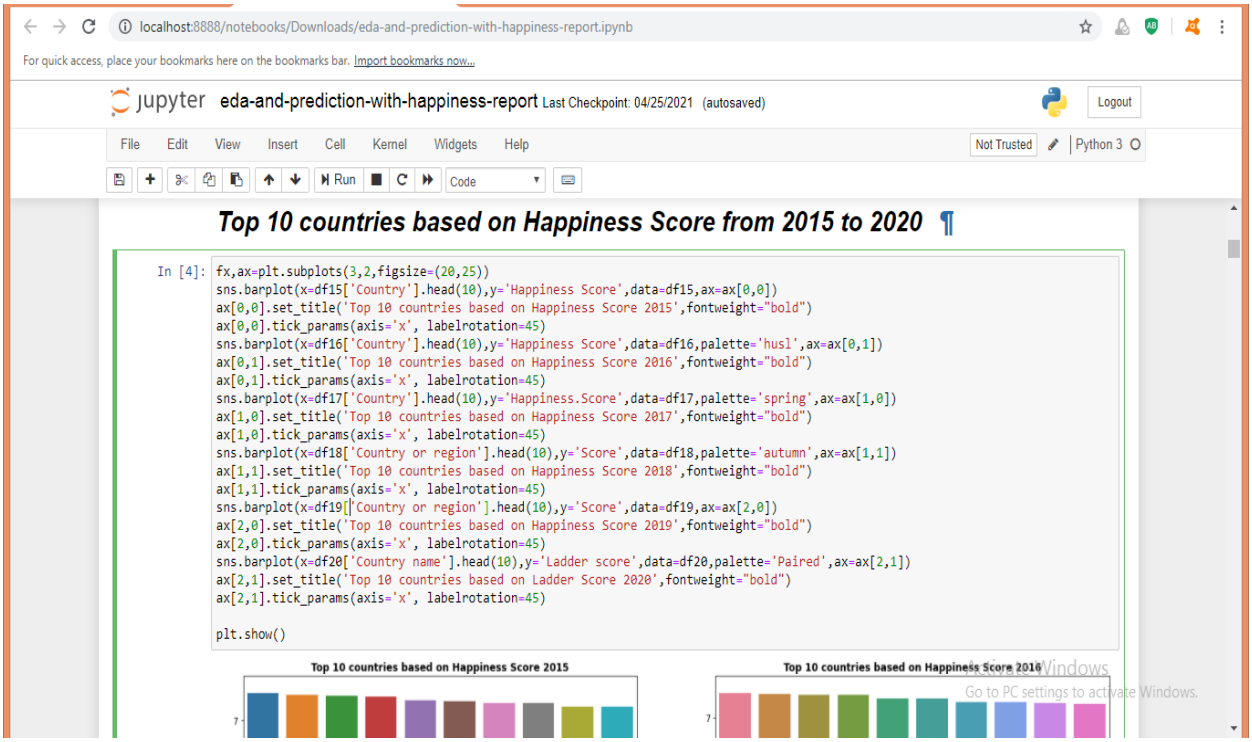
#HAPPINESS SCORE AND GENEROSITY



#FREEDOM AND TRUST [GOVERNMENT CORRUPTION]

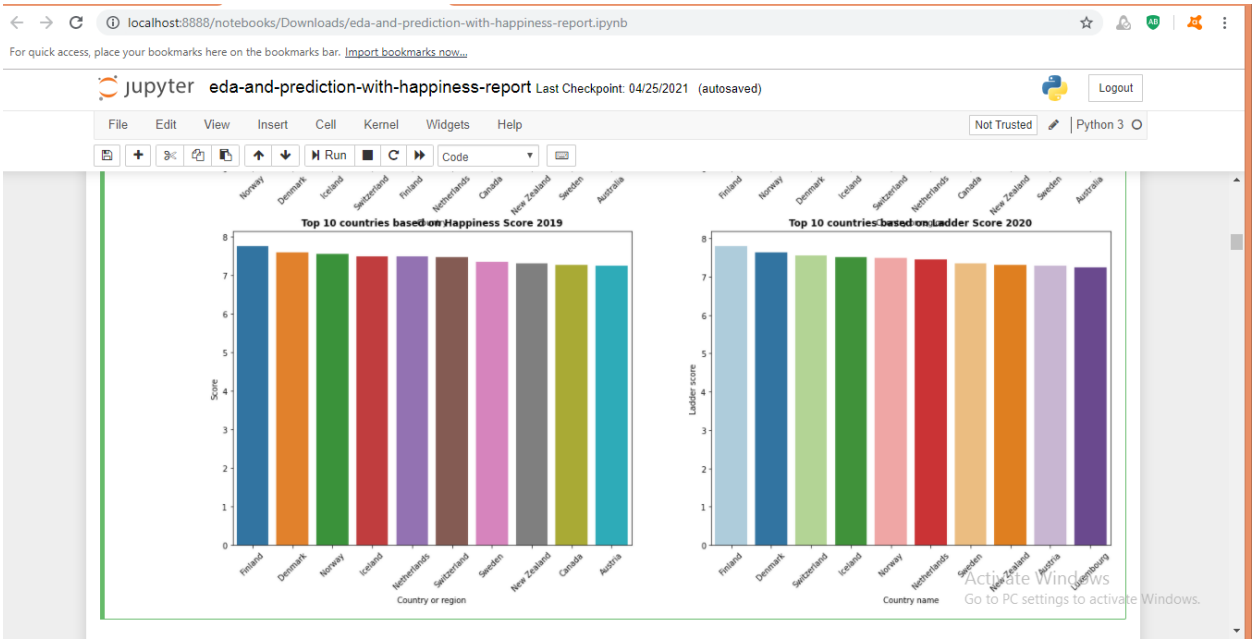


#TOP 10 COUNTRIES BASED ON HAPPINESS SCORE FROM 2015 TO 2020

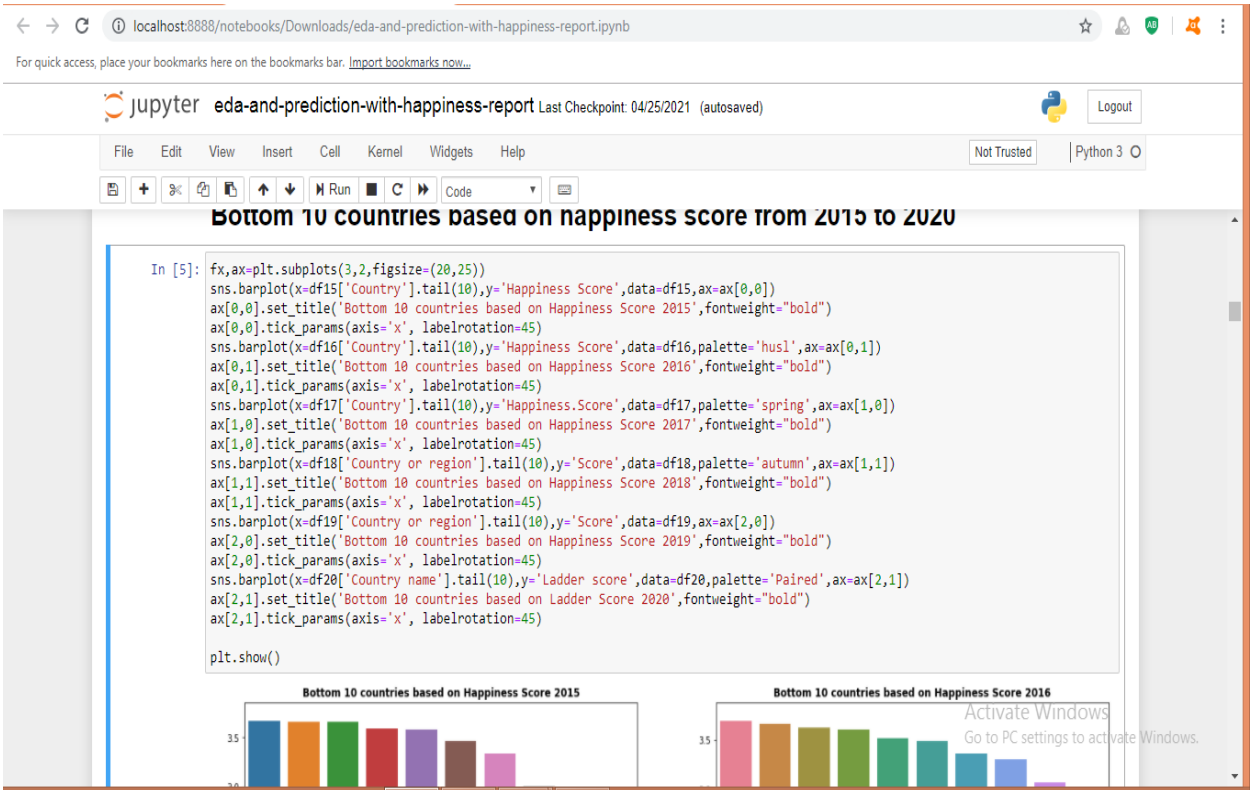


HAPPINESS SCORE 2017 TO 2018

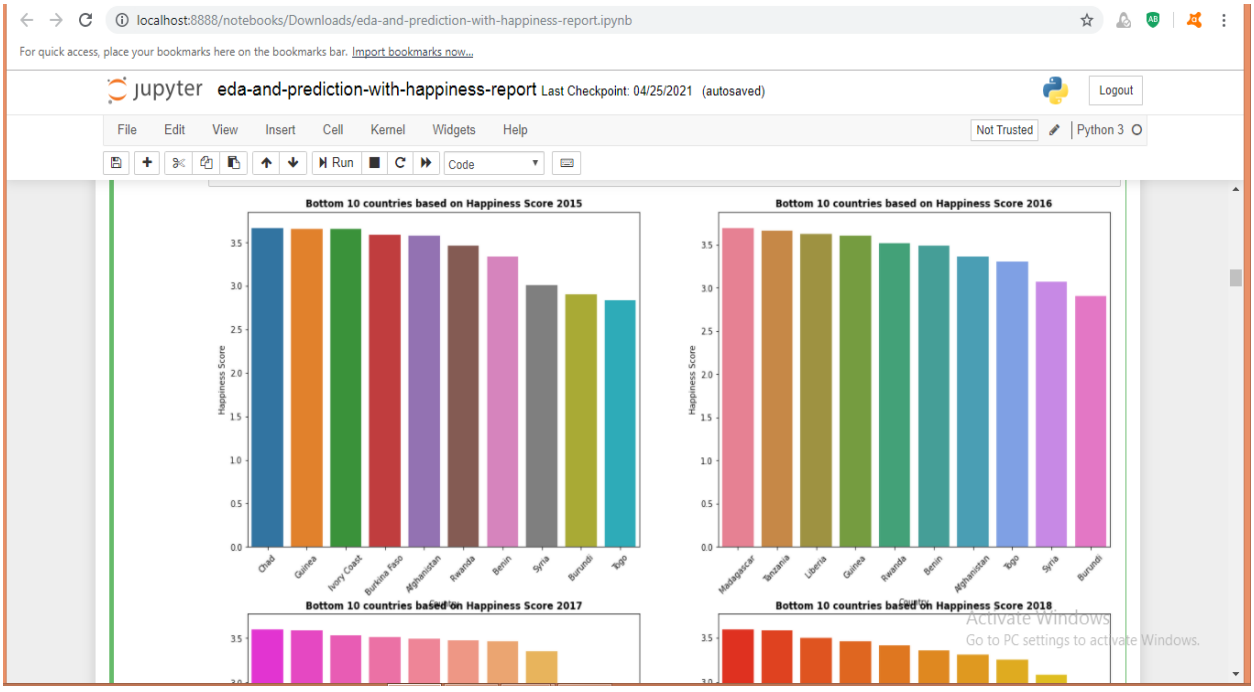
HAPPINESS SCORE FROM 2019 TO 2020



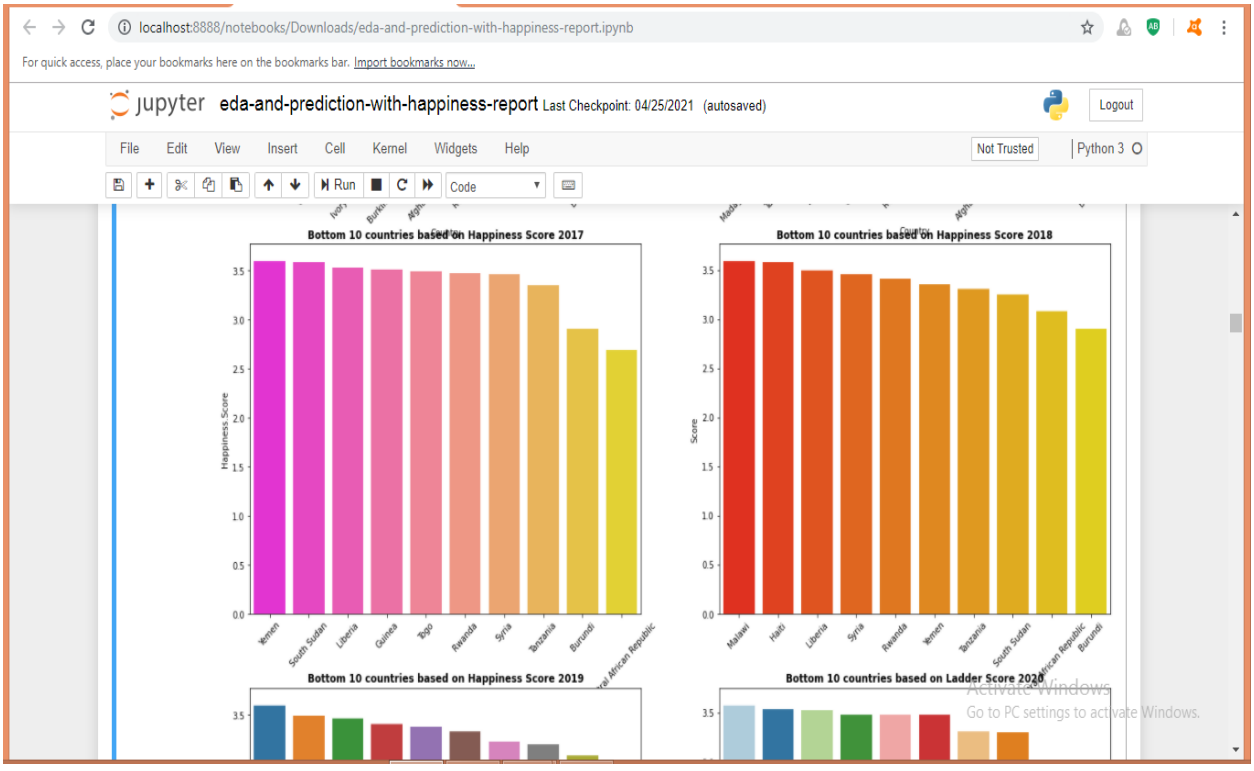
BOTTOM 10 COUNTRIES BASED ON HAPPINESS SCORE FROM 2015 2020



HAPPINESS SCORE FROM 2015 TO 2016

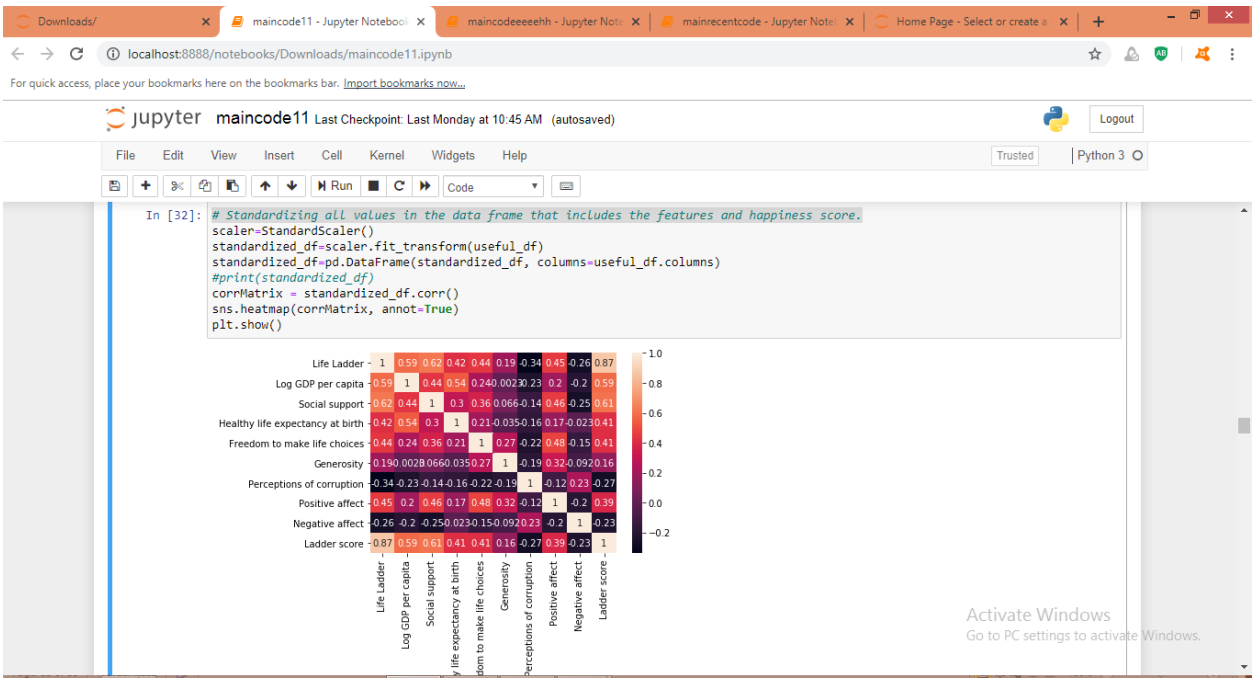


HAPPINESS SCORE FROM 2017 TO 2018



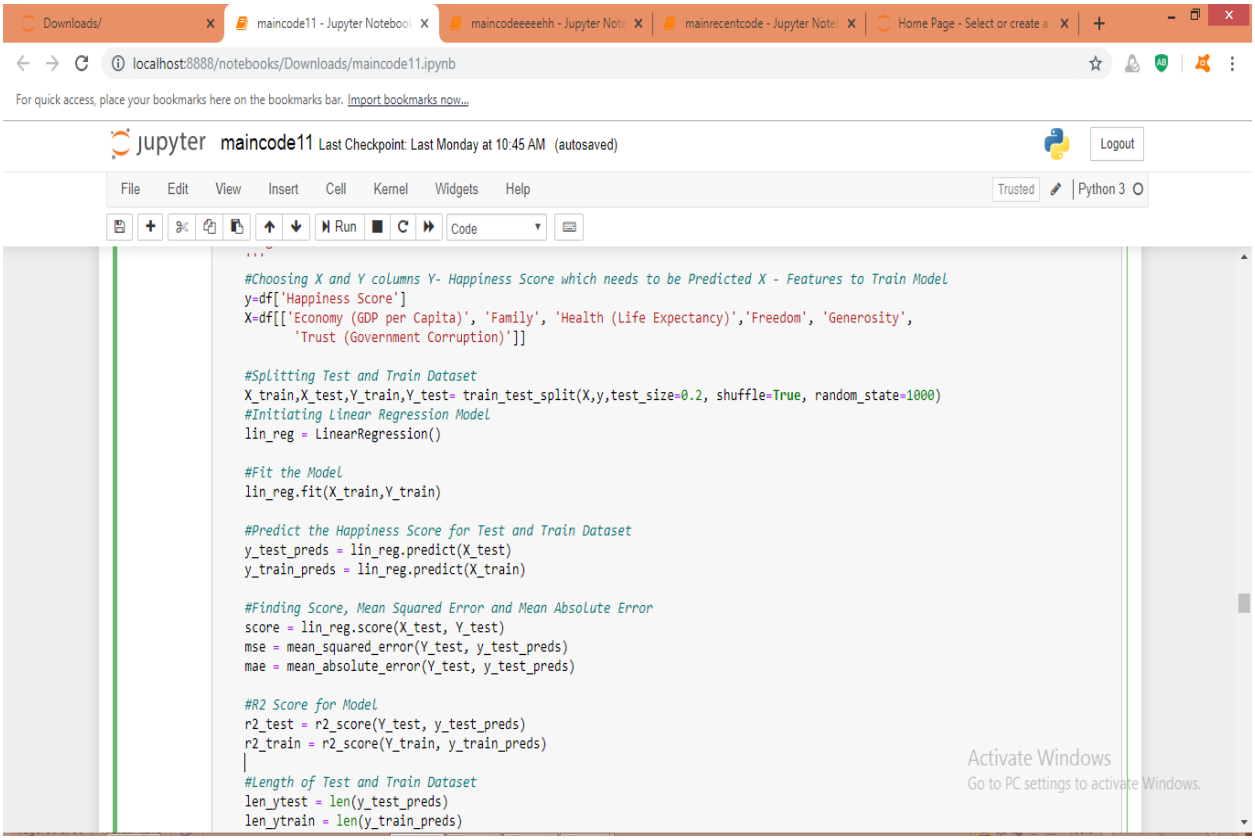
HAPPINESS SCORE FROM 2019 TO 2020

STANDARDIZING ALL VALUES IN THE DATA FRAME THAT INCLUDES THE FEATURES AND HAPPINESS SCORE

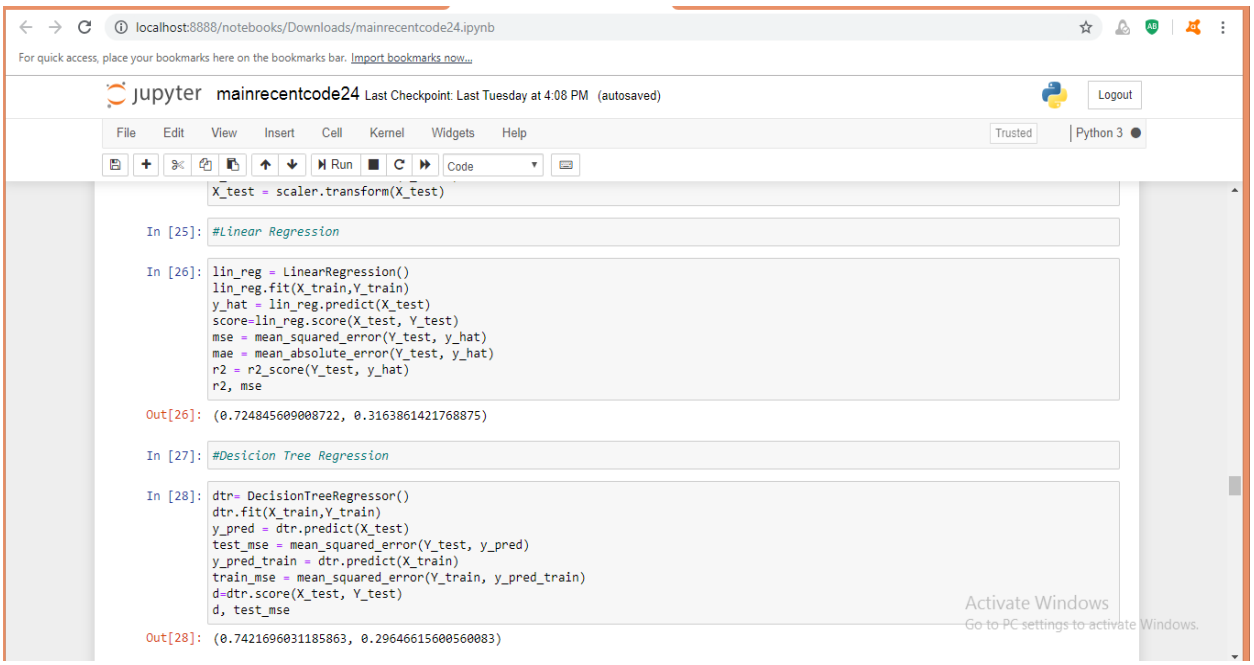


4.4 COMPARING THE REGRESSION MODELS

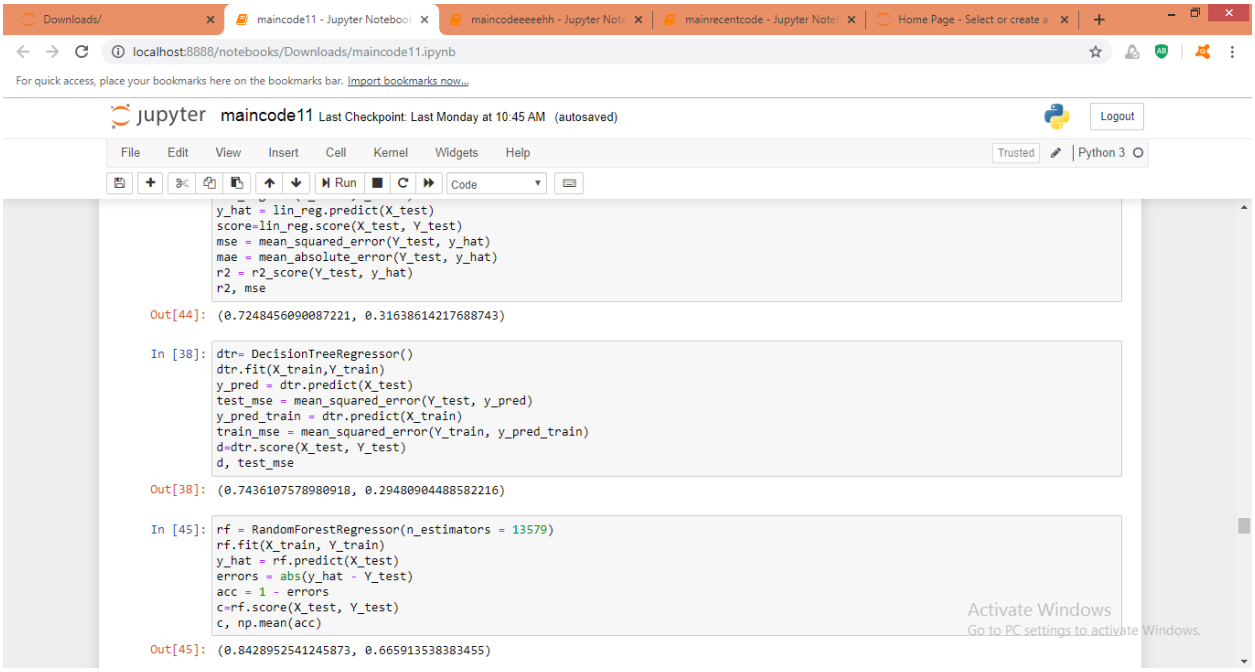
```
#Splitting Test and Train Dataset:

  X_train,X_test,Y_train,Y_test= train_test_split(X,y,test_size=0.2, shuffle=True, random_state=1000)

 #Initiating Linear Regression Model

   lin_reg = LinearRegression()

  #Fit the Model

   lin_reg.fit(X_train,Y_train)

  #Predict the Happiness Score for Test and Train Dataset

    y_test_preds = lin_reg.predict(X_test)

   y_train_preds = lin_reg.predict(X_train)

 #Finding Score, Mean Squared Error and Mean Absolute Error

   score = lin_reg.score(X_test, Y_test)

   mse = mean_squared_error(Y_test, y_test_preds)

   mae = mean_absolute_error(Y_test, y_test_preds)

 #R2 Score for Model

   r2_test = r2_score(Y_test, y_test_preds)

   r2_train = r2_score(Y_train, y_train_preds)

#Length of Test and Train Dataset

   len_ytest = len(y_test_preds)

   len_ytrain = len(y_train_preds)  return r2_test, len_ytest, r2_train, len_ytrain
```



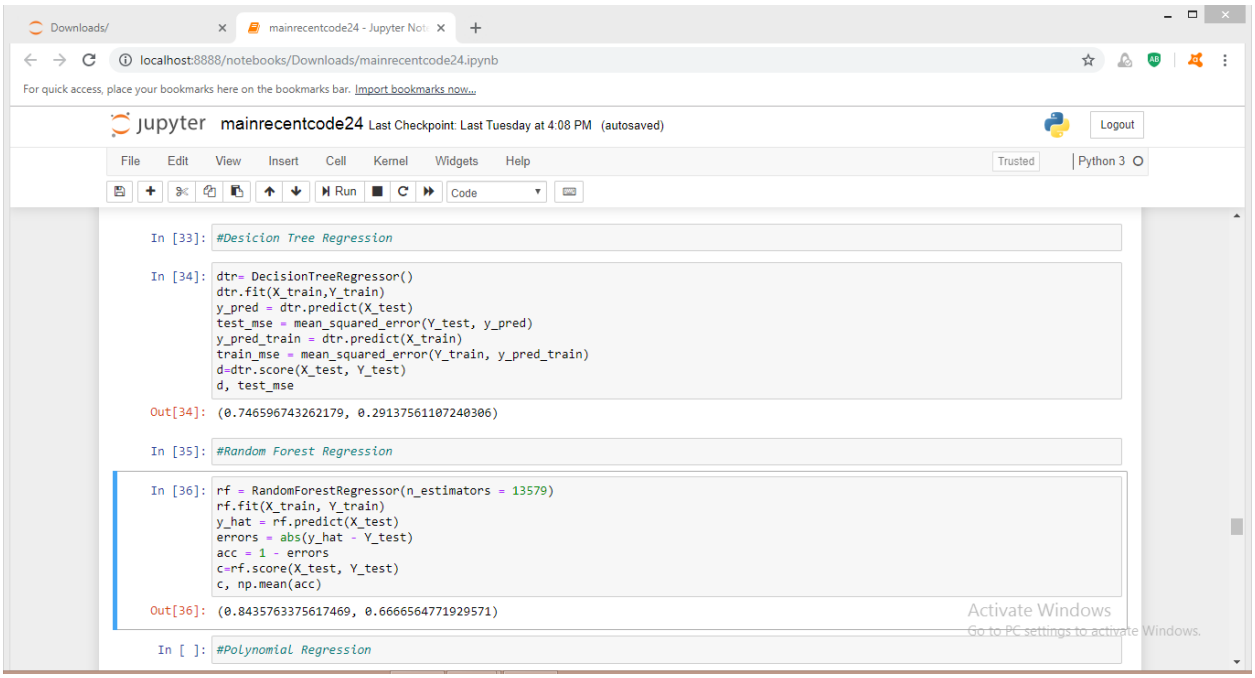
LINEAR REGRESSION



DECISION TREE REGRESSION

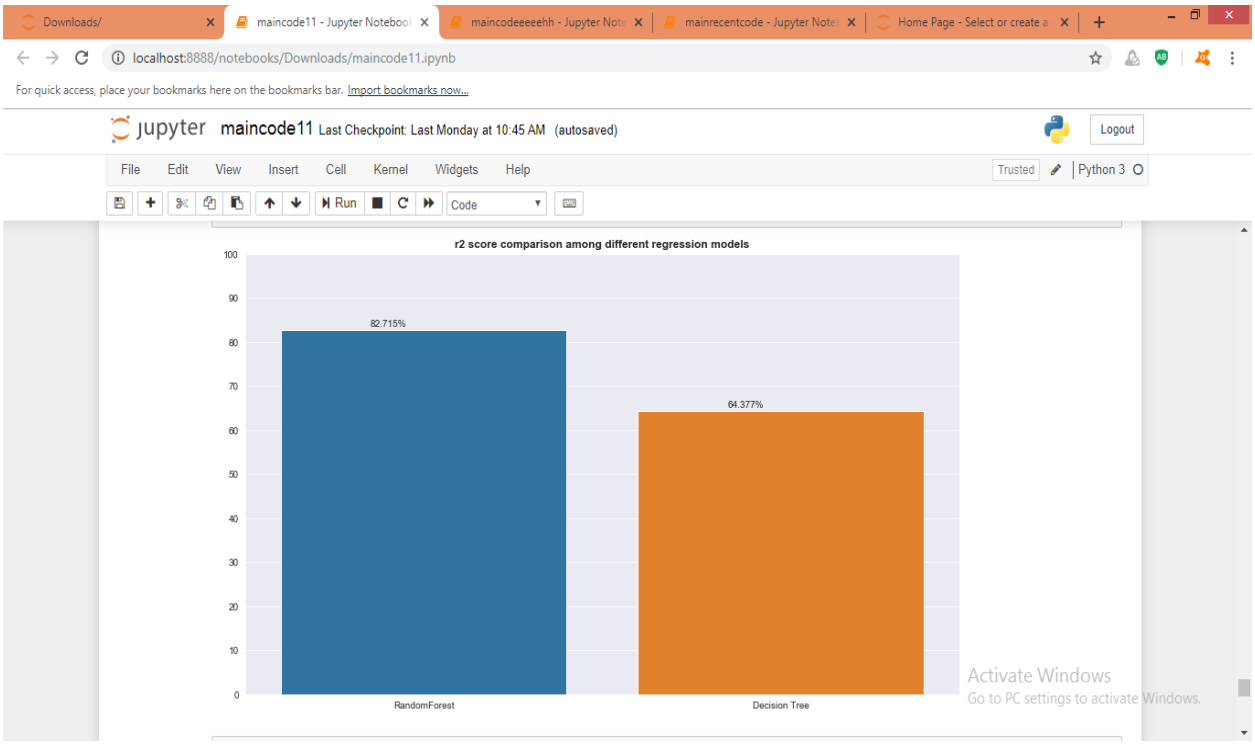


RANDOM FOREST REGRESSION



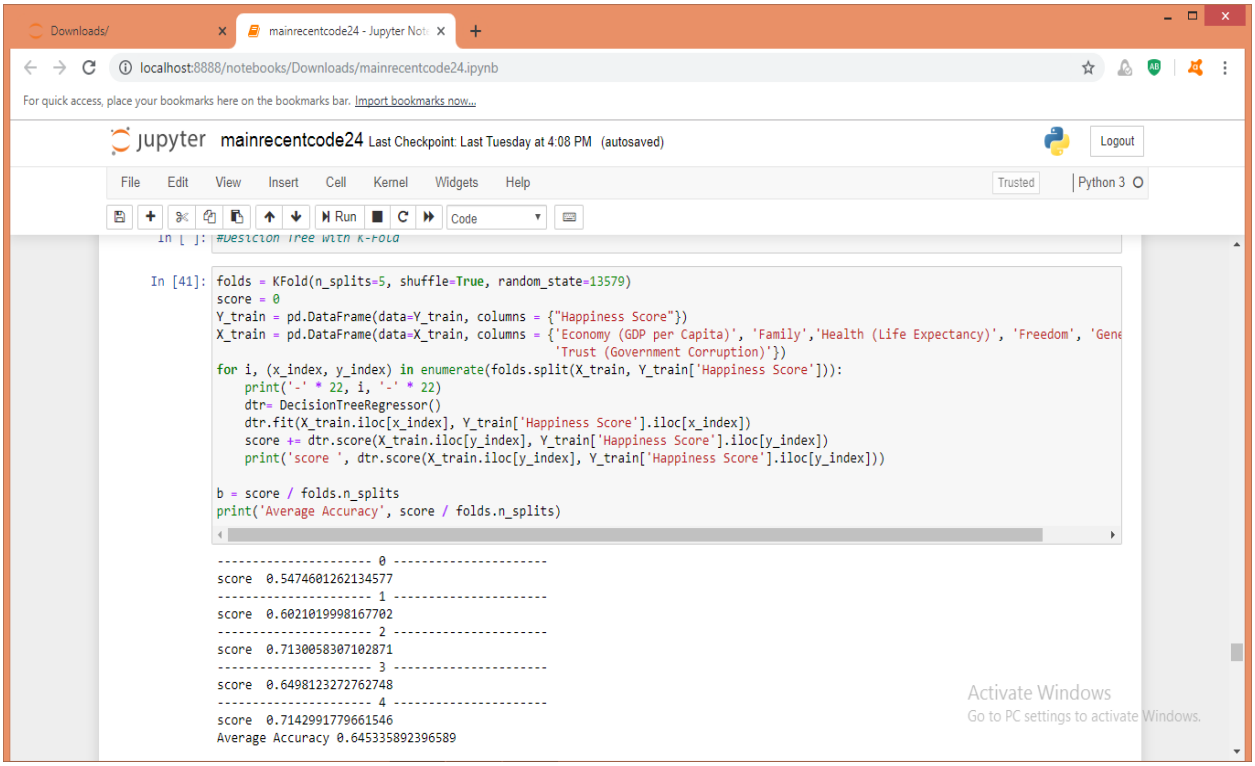
4.5 COMPARING THE DECISION TREE AND RANDOM FOREST REGRESSION:

```
labelList = ['RandomForest', 'Decision Tree']

mylist2 = [a, b]

for i in range(0,len(mylist2)):

    mylist2[i]=np.round(mylist2[i]*100,decimals=3)

print(mylist2)
```

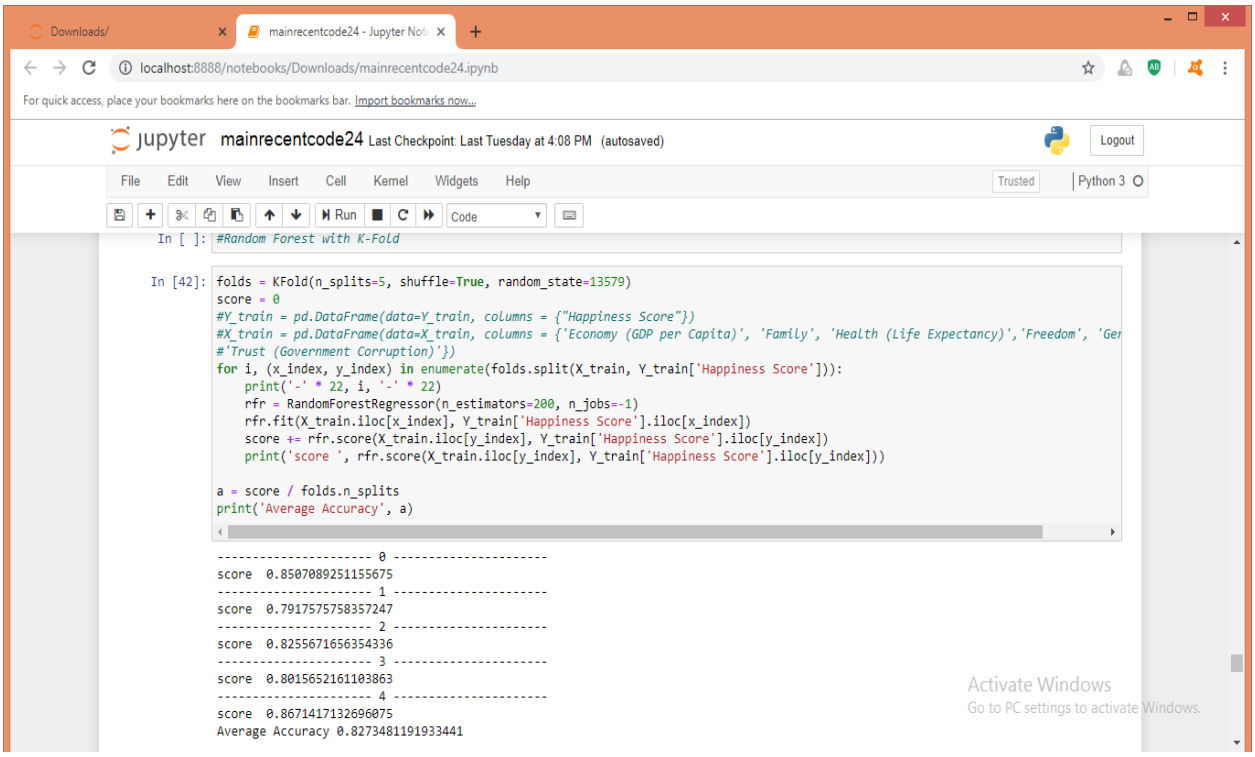


DECISION TREE WITH K-FOLD

```
for i, (x_index, y_index) in enumerate(folds.split(X_train, Y_train['Happiness Score'])):
    print('-' * 22, i, '-' * 22)
    dtr= DecisionTreeRegressor()
    dtr.fit(X_train.iloc[x_index], Y_train['Happiness Score'].iloc[x_index])
    score += dtr.score(X_train.iloc[y_index], Y_train['Happiness Score'].iloc[y_index])
    print('score ', dtr.score(X_train.iloc[y_index], Y_train['Happiness Score'].iloc[y_index]))
b = score / folds.n_splits
print('Average Accuracy', score / folds.n_splits)
```



RANDOM FOREST WITH K-FOLD TECHNIQUE



4.5 COMPARING DECISION AND RANDOM FOREST K-FOLD TECHNIQUES

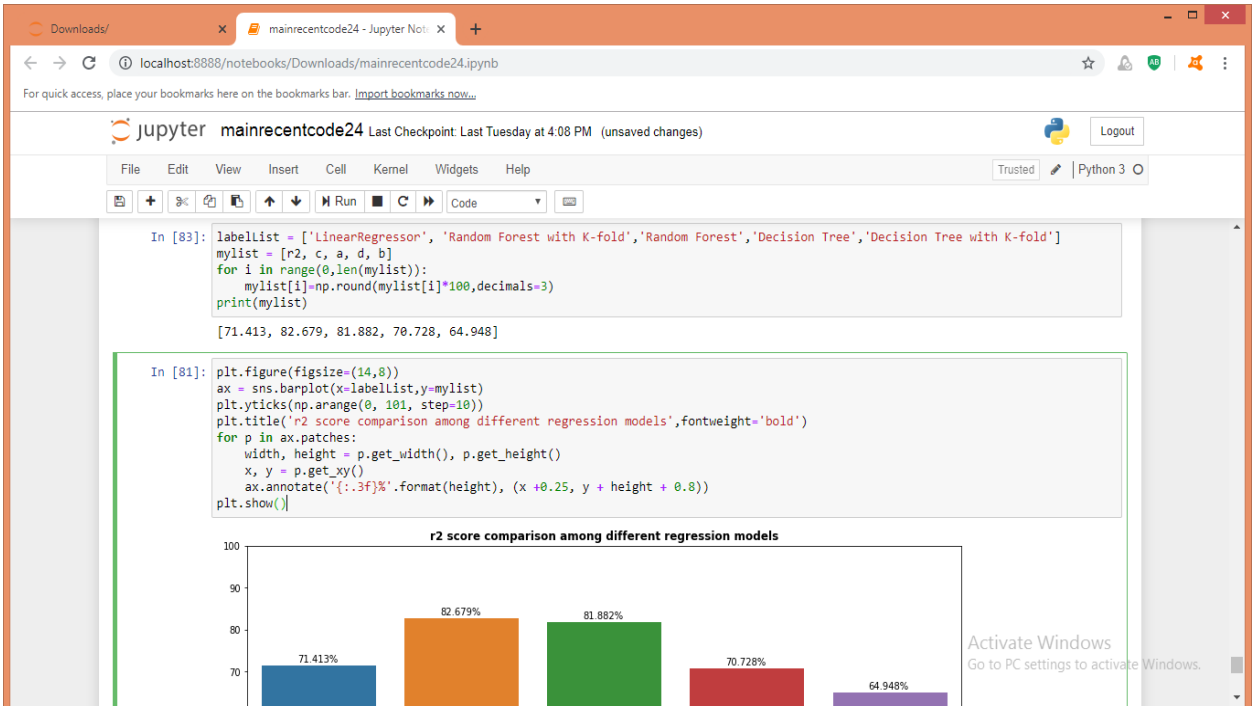
labelList = ['Linear regression', 'Random Forest with K-fold','Random Forest','Decision Tree','Decision Tree with K-fold']

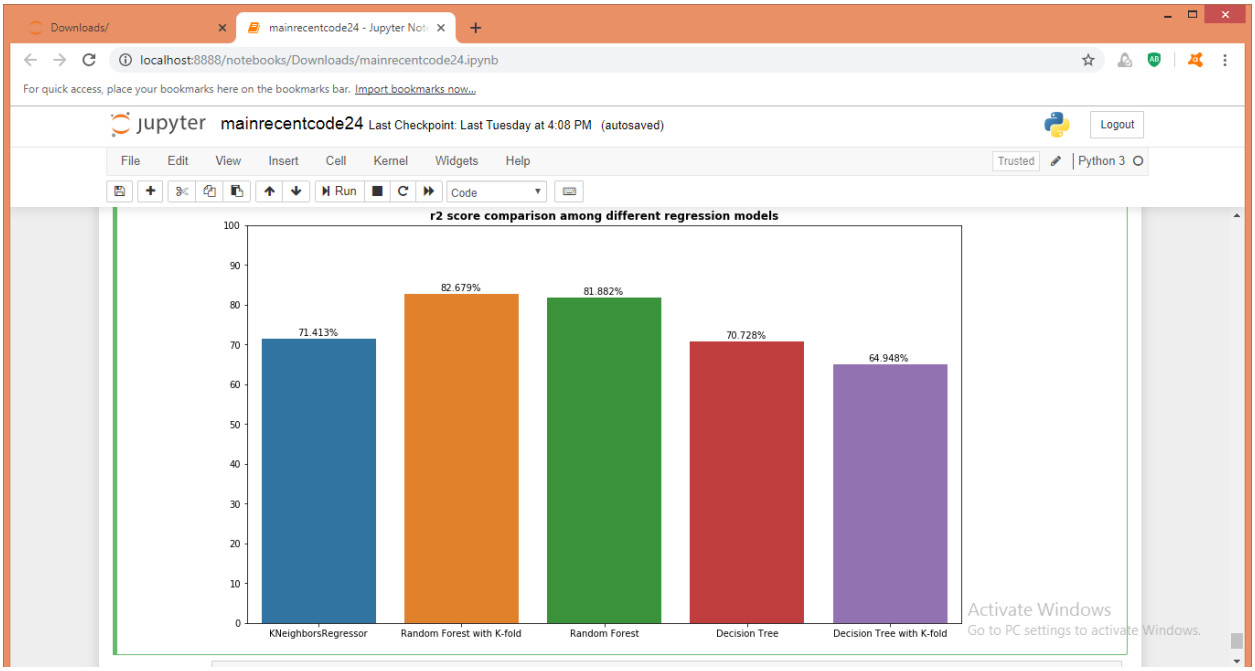
mylist = [r2, c, a, d, b]

for i in range(0,len(mylist)):

    mylist[i]=np.round(mylist[i]*100,decimals=3)

print(mylist)





Random Forest ->81.882%

Decision Tree->70.728%

Random forest with K-fold->82.679%

Decision Tree with K-fold->70.728%

4.6 PARAMETERS FOR ANALYSIS

The following parameters are used to:

- Predictive accuracy: It is defined as the percentage of correct prediction made by a classification algorithm.

4.6.1 PREDICTIVE ACCURACY

- Predictive accuracy is expressed as the correlation between the prediction and actual score. Accuracy is often the starting point for analyzing the quality of predictive model.

ALGORITHMS	ACCURACY
RANDOM FOREST	81.882%
DECISION TREE	70.728%
RANDOM FOREST WITH K-FOLD	82.679%
DECISION TREE WITH K- FOLD	70.728%

The above table describes the accuracy value of four machine Learning regression algorithms on predicting happiest country. The accuracy level is calculated by using the formula:

Accuracy = (TP+TN)/TP *100

TP- True positive

TN- True negative

CHAPTER 5

CONCLUSION

Happiest Countries have Higher GDP Per Capita, Social Support from Family, Trust in Government, More Life Expectancy and people have more Freedom to make Life choices compared to Unhappiest Countries. Finland ranked 1st for the past 3 Years with the Highest Happiness Score of 7.5 .India has been ranked 139 out of 149 countries.

Some of the Countries have seen an increase in 2020 but many countries faced a downfall of the Happiness score never seen in 5 years. Trust in the government and support of family has fallen over the years. There is a vast difference in the economic state of countries, which may be a big factor impacting the happiness of the countries. The 2020 report features the happiness score averaged over the years 2017–2019. Finland is the happiest country in the world, followed by Denmark, Switzerland, Iceland, and Norway. The data comes from the Gallup World Poll, based entirely on survey scores and answers to the main life evaluation question asked in the poll.

A random forest is an ensemble model that consists of many decision trees. Predictions are made by averaging the predictions of each decision tree. Or, to extend the analogy—much like a forest is a collection of trees, the random forest model is also a collection of decision tree models. This makes random forests a strong modeling technique that's much more powerful than a single decision tree. Each tree in a random forest is trained on the subset of data provided. The subset is obtained both with respect to rows and column.

Here Random forests gives (81.822%) consists of multiple single trees each based on a random sample of the training data. They are typically more accurate than decision tree (70.728%)**.**Cross validation is an approach that you can use to estimate the performance of a machine learning algorithm with less variance than a single train-test set split. Random forest with K-fold gives (82.679%) accuracy and decision tree with K-fold gives (70.728%).

BILIOGRAPHY

[1] Ezgi Tascı Ankara, Ankara, Turkey https://www.kaggle.com/ezgitasci/eda-and-model-prediction **https://www.kaggle.com/ezgitasci/eda-and-model-prediction**

[2]MihaelaGRIGORE Paris, Île-de-France, France **https://www.kaggle.com/mishki**

[3]Sunku Sowmya Sree Senior Software Engineer at HCL Bengaluru, Karnataka, India **https://www.kaggle.com/sowmya96**

[4] Siti Khotijah Artificial Intelligence and Data Science Enthusiast at Freelance Bandung, West Java, Indonesia **https://www.kaggle.com/khotijahs1**

[5] Thomas SchlecticAspiring Data Scientist at Seeking Employment

Baltimore, Maryland, United States **https://www.kaggle.com/thomasschlectic**

[6] Accuracy, Recall, Precision, F-Score & Specificity **https://towardsdatascience.com/accuracy-recall-precision-f-score-specificity-which-to-optimize-on-867d3f11124**

[7]https://en.m.wikipedia.org/wiki/World_Happiness_Report#:~:text=The%20World%20Happiness%20Report%20is,(quality%20of)%20life%20factors

[8] Changing world happiness **https://worldhappiness.report/ed/2019/changing-world-happiness/**

[9]WorldHappinessReport/EDA/LR,DT,Bag,RF,Boost,KNN **https://www.kaggle.com/harshmistry97/world-happiness-report-eda-lr-dt-bag-rf-boost-knn**

[10] Happiness Data Analysis 2015-2019 **https://www.kaggle.com/kojisera/happiness-data-analysis-2015-2019**

[11] Happiness Report-Logistic Regression **https://www.kaggle.com/aysuncag/happiness-report-logistic-regression**

[12] Happiness prediction **https://www.kaggle.com/serhatkaraman/happiness-prediction**