COVENTRY
UNIVERSITY

Faculty of Engineering, Environment and Computing

School of Computing, Electronics and Mathematics

MSC Data Science

7150CEM-Data Science Project

To Predict UK's GDP using K nearest neighbour (KNN) Machine Learning Technique

Author: Lara Xavier

SID: 11864103

Supervisor: Dr. Tariq Aslam

Submitted in partial fulfilment of the requirements for the Degree of Master of Science in Data Science

Academic Year: 2022/23

Declaration of Originality

I declare that this project is entirely original work of mine and that no portions or all of it have been taken without proper attribution from any other sources. As a result, all references to previously published material (found in books, journals, magazines, the internet, etc.) have been cited within the body of the report to a piece in the References or Bibliography sections. I also consent to this project's electronic copy being saved and utilised for plagiarism detection and prevention.

Statement of copyright

Statement of ethical engagement

I declare that a proposal for this project has been submitted to the Coventry University ethics monitoring website (https://ethics.coventry.ac.uk/) and that the application number is listed below (Note: Projects without an ethical application number will be rejected for marking)

Signed: Lara Xavier Date:09-12-2022

Please complete all fields.

First Name:	Lara
Last Name:	Xavier
Student ID number	11864103
Ethics Application Number	P142756
1st Supervisor Name	Tariq Aslam
2nd Supervisor Name	Zina Jerjees

This form must be completed, scanned and included with your project submission to Turnitin. Failure to append these declarations may result in your project being rejected for marking.

Abstract

Among the indicators of economic factors, the GDP subject has grown in prominence. In the examination of a nation's economy and growth, predicting GDP is a vital task. This paper analyzes the GDP growth of the United Kingdom using Machine learning algorithms using the GDP data from the world data bank. The data shows the annual GDP growth. Using machine learning algorithms such as KNN (k nearest neighbour), Support vector Machine and Random Forest. It was found that all 3 algorithms performed badly due to enough data. The mean squared value for all the 3 algorithms was nearly 14 which is quite high, also this dataset does not take socioeconomic factors into account.

Acknowledgements

I want to sincerely thank Prof. Tariq Aslam for his unwavering assistance and availability during the length of this specific project. At this time, I'd also want to express my gratitude to Rochelle Sassman, the coordinator of our module, for allowing me to work on a project of this scope in the field of data science.

I also want to express my gratitude to my classmates from the Data Science course for their support and encouragement during the project.

Not least, I would want to say my gratitude to my wonderful parents, K P Xavier and Betty Xavier, for their continuous support also belief in me throughout my academic career. They enabled me to pursue my passion of travelling the world.

1. Introduction

The complete value of the products and services generated in the economy is taken into account when calculating the real Gross Domestic Product (GDP), which is a single, all-inclusive measure of economic activity. Even while GDP estimates are often made on an annual basis, they can also be made quarterly. The gross domestic product (GDP) measures the market value of all manufactured products and services produced within a nation's boundaries over a certain time period (GDP). Given that it is a wide measure of total domestic output, it offers a thorough assessment of the state of the economy in a particular country. It determines the monetary value of the finished products and services produced in a country over the course of a certain time period and bought by the final customer.

During a particular period, the commercial value of all outcomes and services manufactured in the United States is known as the GDP. GDP is calculated in three ways, utilise expenditures, income, as well as production.

Three perspectives exist on the GDP.

- Production Method
- The expenditure strategy
- Income Strategy

The GDP is the most important economic indicator because it provides information about the economy

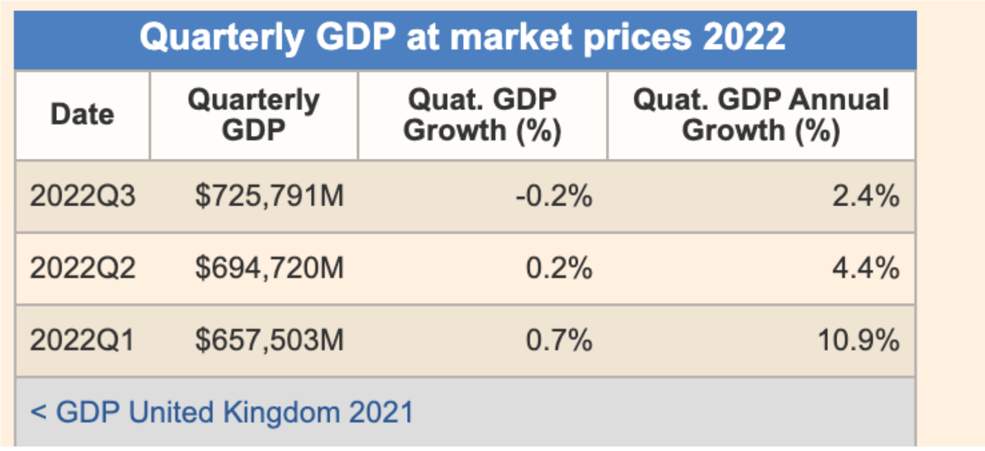
1.1 United Kingdom GDP

The United Kingdom, which encompasses England, Scotland, Wales, and Northern Ireland, has one of the most globally integrated economies in the world. Globally, the UK was the fifth-largest importer and exporter in 2020.

The third quarter of 2022 had seen a 0.2 decrease in the UK's gross domestic product compared to the previous quarter. In comparison to the 4.4% reported in the second quarter of 2022, the GDP growth year over year was 2.4% or 20 tenths with one per cent less.

With a GDP of $725,791 million, the United Kingdom ranks fifth among 53 nations in our quarterly GDP rankings for the third quarter of 2022.

It was $827 higher than the same quarter last year, with a quarterly GDP per capita of $10,895.



Quarterly GDP at market prices 2022

Date	Quarterly GDP	Quat. GDP Growth (%)	Quat. GDP Annual Growth (%)
2022Q3	$725,791M	-0.2%	2.4%
2022Q2	$694,720M	0.2%	4.4%
2022Q1	$657,503M	0.7%	10.9%

< GDP United Kingdom 2021

1.2 Dataset

For this paper, the dataset is taken from the data world bank.
The link given below
https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG?locations=GB

Data Bank is a tool for research and data visualisation that comprises collections of time series data on a range of topics. It has databases of population statistics, gender statistics, education statistics, and statistics on the world's progress, as well as statistical capacity indicators. The data consist of GDP values from the year 1971 to 2021. The data also consist of GDP from other countries.

1.3 Algorithm

Algorithms learn from labelled data in supervised learning. After analyzing the data, the algorithm connects the unlabeled new data to patterns to determine which label to apply.
It is possible to divide supervised learning into two types,which is

- Classification
- Regression

Spam detection, emotion analysis etc. are some examples of classification. Regression can be used to predict house prices, stock prices, heights, and weights.

Type of Classifications

Logistic Regression

It's like linear regression. However, when it is a value that is not the dependent variable like a yes/no response. It performs classification based on the regression. It is used to predict the output of binary numbers.

K-NN

The K-NN approach is used to identify the datasets that are split into different classes in order to predict the classification of the new sample point. K-NN is a lazy learning method that is non-parametric. It assigns new instances to categories based on a similarity measure. It is mostly used for classification algorithms based on the assumption that finding a similar point can be nearby. Why the K-NN method is lazy learning, The computation also takes place when a classification or prediction is formed. The method is also known as instance-based or memory-based learning because It stores all of its training data entirely in memory.

Support vector machine

It is used for classification and regression. The concept of decision planes, which outline choice constraints, serves as its basis. There is a boundary between objects of different classes called a decision plane.

Naive Bayes

The Bayes theorem and predictor independence assumptions form the foundation of the naive Bayes classifier Despite the fact that these traits depend on each other or on the existence of other traits, each of these characteristics also exists on its own.

Decision Tree Classification

A decision tree builds classification or regression models as a tree structure. It simultaneously creates a related decision tree and incrementally splits a dataset into smaller and smaller chunks. The final outcome is a tree containing decision nodes and leaf nodes. The split is chosen by employing the Iterative Dichotomiser 3(ID3) method framework.

1.4 Recession

The peak of the Covid-19 epidemic in August 2020 was the last time the UK had a recession. Due to covid period, people are in an isolated period at that time they lost their job and many companies and enterprises closed, Business goes down. In this case, the country's economy fell by 20.4%. The most popular and longest-standing recession to hit the UK was the Global Recession of 2008. This was brought on by both the collapse of the housing market and rising energy prices. The recession of 2008 was worldwide, hurting all of the G7 nations.

Popular companies went bankrupt overnight, including Lehman Brothers and the Royal Bank of Scotland, which required bailout funding from the UK government.

The unemployment rate in the UK significantly increased to 10%.

To recover the UK economy it took five years with GDP of more than 6% between 2008 and 2009.

Since the recession earnings still not recovered. The customer will pay more for everyday items than the previous year. When inflation rates are high, recessions are more likely because people have less money to spend on things besides basic necessities for their homes.

1.5 History of GDP in United Kingdom

In 2021, the Gross Domestic Product (GDP) of the United Kingdom was worth 3186.86 billion US dollars, according to the World Bank. A total of 2.38 percent of the global economy is accounted for by the United Kingdom's GDP. Between 1960 and 2021, the UK's GDP averaged 1290.78 USD billion.

The output and national wealth of an economy are measured by something called the gross domestic product (GDP). Gross domestic product, or GDP, is the total cost of all completed products and services produced domestically during a defined period of time.

1.6 Structure of British GDP

The British economy relies heavily on domestic demand. Despite exporting significant services, particularly in the finance and IT sectors, the country's consumption capacity keeps the external sector in a largely stable deficit. Due to London's status as one of the world's top financial centers, local businesses usually have an easier time acquiring external capital than those in other nations; thus, credit is more readily available and investment opportunities are greater. This is frequently seen as an essential component of the nation's economic success and as one of the reasons the UK appears to have exited the recent financial crisis on sturdier ground than some of its European counterparts.

1.7 Overview of the project

This project is the analysis of the The UK's GDP (Gross Domestic Product) as calculated using machine learning techniques. An important economic component for a nation's development is its GDP. Gross domestic product (GDP) is the market worth of all manufactured goods and services generated inside a country's borders during a specific time period (GDP). Since it provides a broad estimate of total domestic output, it provides a comprehensive evaluation of the status of the economy in a specific country. It evaluates the economic value of completed goods and services manufactured in a country over the course of a certain time period and bought by the final customer. In this paper the data is taken from the world data bank, the data provide information about the GDP of different countries from the year 1961 to 2021. We want to extract data from the United Kingdom. The UK data is added in excel and then transposed and finally converted into a CSV file. The data is preprocessed where the missing values and duplication of values are dropped. Once the data is cleaned, its ready for Exploratory Data Analysis where the outliers are found, the distribution of data is found and the correlation of the columns is found. After the Exploratory Data analysis, the data is ready for model analysis. The 3 model in this study is KNN (k nearest neighbour) , SVM (support vector machine), and random forest. The mean squared error is found for each model and compared to each other.

2. Literature Review

2.1 Predicting Economic Recessions Using Machine Learning Algorithms

In the first paper Predicting Economic Recessions Using Machine Learning Algorithms which was written by Rickard Nyman and Paul Ormerod, it was found that at the beginning of 2008 economic recession is not being predicted. Two different estimation techniques are used in this paper. They are Ordinary least squares regression and Random forest machine learning. In random forest machine learning is capable of dealing with non-linear, high-dimensional prediction situations. By replacing samples from the data, they build a lot of decision trees during the training process. According to the authors, the random forest family of algorithms produces the best outcomes. The authors point out that the other methods, such as Bayesian and logistic regression, "are not at all competitive." Alessi and Detken (2011) and Alessi et al. (2015) demonstrate positive outcomes using random forest algorithms in the context of earlier warning of financial crises. Alessi and Detken (2011) and Alessi et al. (2015) demonstrate positive outcomes using random forest algorithms in the context of earlier warnings of financial crises. These algorithms are used for UK and USA. Regression model for USA gave R bar sqaure value as 0.009. The random forest approach gave R bar square value as 0.149. For UK the R bar square value was taken as 0.004 and random forest gave 0.246 and for random forest gave 0.149. The country which was targeted was the USA, and UK in the year 2016. Several basic conclusions may be drawn from the literature on predicting accuracy, particularly in light of forecasts for the GDP growth rate one year from now. The very poor track record of recession forecasting is one such issue. The motive of this paper is investigate if machine learning methods can increase predicting precision. This paper focus specifically on short-term projections of real GDP growth in the United States, and more specifically, on the question of whether the recession of the late 2000s might have been anticipated. In this paper, study takes the UK into the analysis.

In this paper they selected dataset which is on a theoretical basis, without having first looked into how any of the factors linked with GDP growth As is typical in many macroeconomic time series analyses, after the data source was chosen, they made no changes to try to enhance statistical fits. The reason for this was that there is a well-established tradition in economics that economic variables, particularly in the case of severe recessions, are to blame. The financial markets, where there should theoretically be information on the future status of the economy, are where primarily choose our explanatory variables. These variables are both present at the moment and their values are not changed afterwards, which is perhaps more crucial when attempting to recreate a real-world forecasting scenario.

Two estimation model techniques are used in this paper which are Ordinary least squares regression and Random forest machine learning. In a publication with a growing number of citations, Fernandez-Delgado et al. (2014) analyse 179 classification algorithms from 17 "groups," including Bayesian, logistic regression, neural networks and multinomial regression. They evaluate their effectiveness using 121 datasets. Finally, he discovers the optimal algorithm, which belongs to the family of random forest algorithms and produces the best outcomes. Both Alessi and Detken (2011) and Alessi et al. (2015) describe the successful use of random forest algorithms for such early detection of financial crises.

2.2 Nowcasting New Zealand GDP Using Machine Learning Algorithms

In the second paper Nowcasting, New Zealand GDP Using Machine Learning Algorithms which was written by Adam Richardson, Thomas van Florenstein Mulder, Turul Vehb,. They finalize whether forecasts of New Zealand's real GDP growth can be improved using machine learning techniques. In this study, we examine the accuracy of nowcasts of New Zealand's real gross domestic product (GDP) growth for the current quarter obtained using several machine learning (ML) techniques. In order to assess how well these algorithms performed in real time between 2009 and 2018, we used numerous vintages of historical GDP data as well as various vintages of a large feature set, which included over 550 domestic and foreign variables. In the next section, we compare the forecasting accuracy of these algorithms to that of a naive autoregressive benchmark as well as to that of other data-rich strategies such as a factor model, a moderate Bayesian VAR (BVAR), and a set of statistical models used by the RBNZ. We believe that our study is the first to compare the nowcast performance of several ML techniques employed.

The decision-makers take decisions using incomplete data on current economic state. Then they realize that are more possible regressors than there are accessible observations, ML models are especially well adapted for handling huge datasets. They investigate with ML algorithm to predict the GDP growth of New Zealand. These algorithms are applied over 2009-2018 period. The forecasts produced by these algorithms are then compared to the forecasting precision of a benchmark using basic auto - regressive analysis as well as other data-rich techniques including a factor structure, a modest Bayesian VAR (BVAR), and a collection of statistical models employed by the RBNZ.

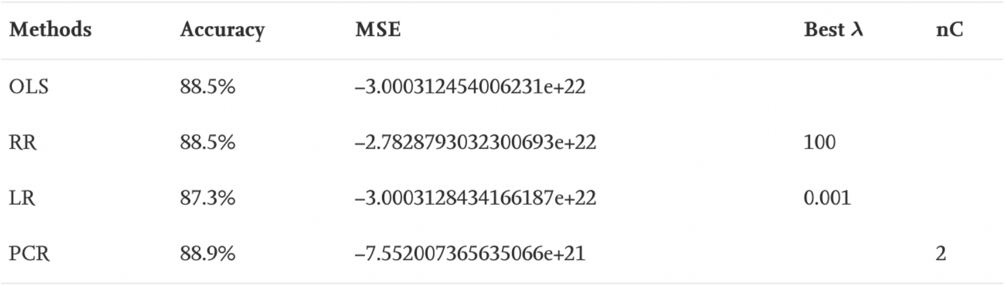
2.3 Forecasting GDP to Macroeconomic Variables

The third paper is Forecasting GDP to Macroeconomic Variables, written by S.C.Agu, F.U. Onu, U.K. Ezemagu, and D.Oden, is the third paper. In this research, macroeconomic indicators allow nations to focus on producing things, providing services, and engaging in other activities that will increase their Gross Domestic Product (GDP). Principal Component Regression (PCR), Ridge Regression (RR), Lasso Regression

(LR), and Ordinary Least Squares (OLS) are four machine learning algorithms to forecast GDP (OLS). In the second stage, identify the significant macroeconomic factors that are most likely to have an effect on GDP growth. In comparison to other approaches, the findings showed that the PCR method had a higher accuracy of 89% and a mean square error of -7.552007365635066e+21 in predicting GDP to macroeconomic data.Techniques for machine learning regression, such as PCR, RR, LR, and OLS.The outcome showed that PCR had a prediction accuracy of 88.9%, which was the highest.The outcome indicate that the performance of OLS, LR, RR, and PCR was 88.5%, 87.3%, 88.5%, and 88.9%.In comparison to RR, LR, and OLS, the PCR reliably forecasts macroeconomic indices with little MSE. Machine learning-Based Prediction and Industrial Structure Analysis of the Local GDP Economy.

The GDP findings may have been predicted using the PCR technique if we were using the "white-box" model for inference, where the goal would be drawing and confirming the conclusion of the GDP result (Gareth et al., 2017Giovanni et al. (2021) asserted in a similar spirit that PCA does not provide an economic interpretation of the data in terms of converting and reducing the number of macroeconomic variables. As a result, while creating the model, we propose using the RR that had the second-best prediction accuracy and smallest mean square error. The rationale for selecting the RR model is that its output would yield accurate predictions of future GDP levels.

Model accuracy



Methods	Accuracy	MSE	Best λ	nC
OLS	88.5%	–3.000312454006231e+22		
RR	88.5%	–2.7828793032300693e+22	100	
LR	87.3%	–3.0003128434166187e+22	0.001	
PCR	88.9%	–7.552007365635066e+21		2

Data

The World Bank has access to the data used in this study through its Macroeconomic Dataset Repository websites. The data used in this study are freely available on the World Bank Macroeconomic Dataset Repository websites, World Bank (2021a, 2021b, and 2021c). The relevant data set consists of missing value-free macroeconomic variables for Nigeria from the World Bank's collection of macroeconomic indicators between 1981 and 2019. The experiment did not utilise the macroeconomic statistics for Nigeria from 1970 to 1980 since the World Bank did not collect them, and it did not use the data for 2020 and 2021 either because they had missing numbers.

We propose the following equation to forecast GDP using the RR model and the macroeconomic factors that positively affect GDP;

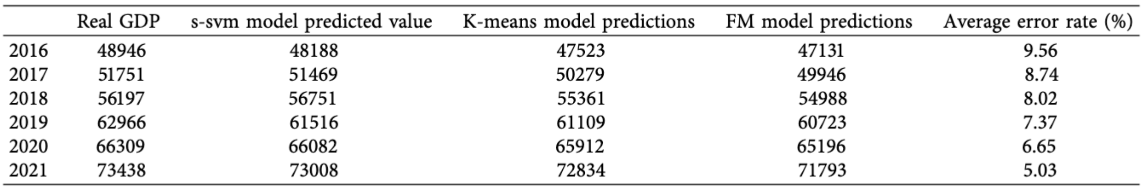
$$GDP = 3,140,795,154.2847595 + pop \text{ x } 6.82614338e + 02 \ + \ fge \text{ x } 2.54062299e + 09 \\ + imp \text{ x } 4.74395349e + 00 + exr \text{ x } 7.90344390e - 01$$

2.4 Prediction and Industrial Structure Analysis of Local GDP Economy Based on Machine Learning

The fourth paper are Prediction and Industrial Structure Analysis of Local GDP Economy Based on Machine Learning which was written by Zhiqiang Jiang. They mainly focused on GDP, SVM, and Random Forest. They predicted SVM models predicted the most accurate results. SVM models predicted values near the actual values.

Introduction

Since 2000, the employment composition of the key industries has been dropping year after year; this tendency is rather apparent, and the reduction is fairly significant. Industrial growth has changed in several ways, but overall the trend remains upward. The service sector employs the most people and has an employment structure that has grown year after year. Its share of total employment typically surpasses that of big industries.

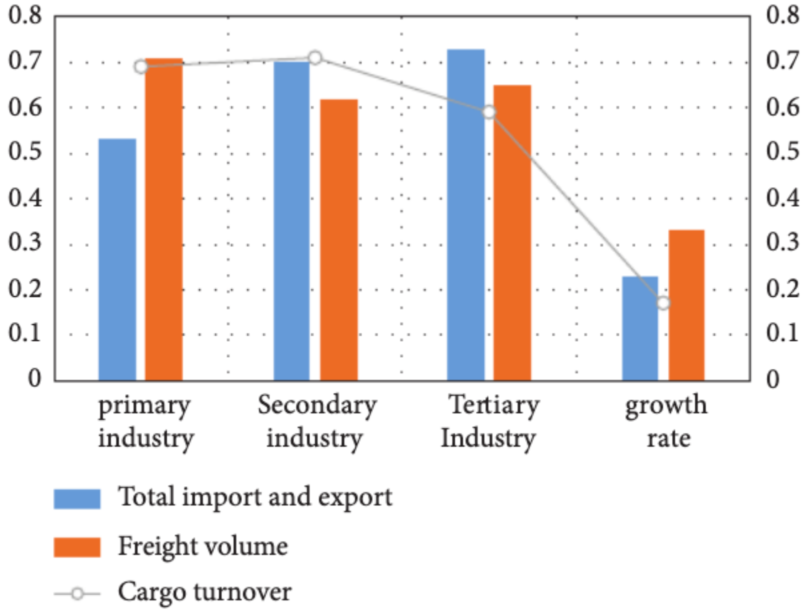


	Real GDP	s-svm model predicted value	K-means model predictions	FM model predictions	Average error rate (%)
2016	48946	48188	47523	47131	9.56
2017	51751	51469	50279	49946	8.74
2018	56197	56751	55361	54988	8.02
2019	62966	61516	61109	60723	7.37
2020	66309	66082	65912	65196	6.65
2021	73438	73008	72834	71793	5.03

Table:2016-2021 GDP forecast and actual GDP under machine learning

The machine learning model estimates that accuracy is 79.46%, reliability is 89.27%, and feasibility is 86.18%. The accuracy, feasibility, and dependability values are 60.14%, 68.24%, and 75.12%, respectively, in accordance with the conventional statistical model. However, the data mining model's accuracy, dependability, and feasibility scores are 68.45%, 75.43%, and 86.18%, respectively. They focused two keys such as relevant features and relevant examples. We outline the advancements achieved in machine learning research—both theoretical and empirical—on these subjects and suggest a broad framework for contrasting various methods. In order to predict brief GDP growth, this paper presented structural model that combine several real-time monthly and quarterly time data.

The true numbers for the first industry are 23%, the second is 33%, and the third is 17%, as can be shown. The percentage of primary industry that the s-svm model predicts is 26.53%, the percentage that the K-means model predicts is 27.38%, and the percentage that the FM model projects is 28.91%. As can be observed, the s-SVM model outperforms the other two in terms of accuracy of prediction.



The evolutionarily long-term rule of regional economic development. Some places could keep expanding over this protracted process of evolution, while others would start to shrink. It requires a lot of time.

2.5 GDP Growth Prediction of Bangladesh using Machine Learning Algorithm

The fifth paper is GDP Growth Prediction of Bangladesh using Machine Learning Algorithm which was written by Amman Hossain, Md Hossen, Md Mahmudul Hasan. They mainly focused on GDP Growth, ML Algorithms, Relation, Prediction, and Complex, Parameters. The algorithm used random forest and Gradient Boosting regressor. They compared their data to the Canadian GDP. With random forest, the MSE value was obtained at 0.004635. They concentrated country Bangladesh the paper was published in 2021.

Introduction

In this paper the GDP calculated in three ways like Income approach, the Value-added approach final one is expenditure approach. Numerous variables, ranging from economic inequality to declining social standing and a country's workforce, are both influenced by and impacted by the population's uneven growth. The GDP level and growth rates in a nation can both be stimulated by gender equality. In a nation, promoting gender equality in the fields of education, health, and employment produces a productive workforce that may increase GDP and, maybe, speed up the innovation process. Given that it significantly affects a country's development and advancement, the population is crucial. Women make up around half of all groups in Bangladesh. Approximately 75% of people

Content

By constructing confidence intervals, the statistical learning method and quantile loss functions may also be utilised to accurately forecast GDP for a set of time series.

Here using this equation to predict the GDP, Given below

$$GDP = C + I + G + (X - M)$$

There is an increasing need for GDP rate prediction in terms of the male and female population to determine which gender does have a significant impact on the increase of Bangladesh's GDP average price. To this final moment, an analysis that uses machine learning was conducted to demonstrate which gender has much more impact in raising the Gdp growth of current value. They studied the combined population, male and female populations through the years from 1980 to 2019. In this paper they use machine learning techniques including basic linear regression, multi-degree polynomial regression, and SVR.

This graph shows final result of the population growth of Bangladesh.

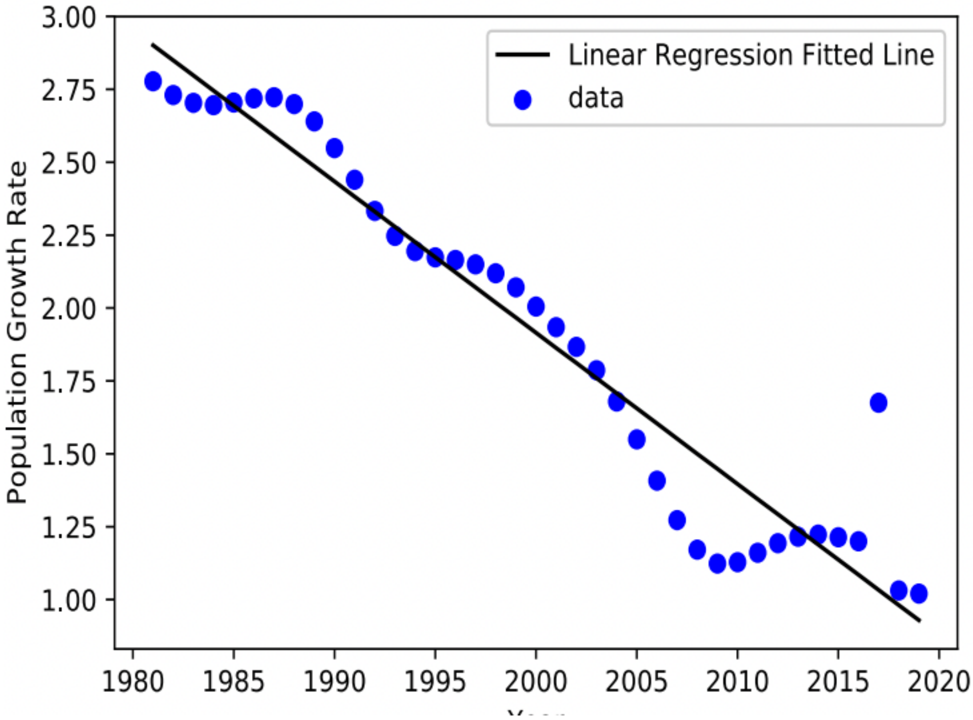
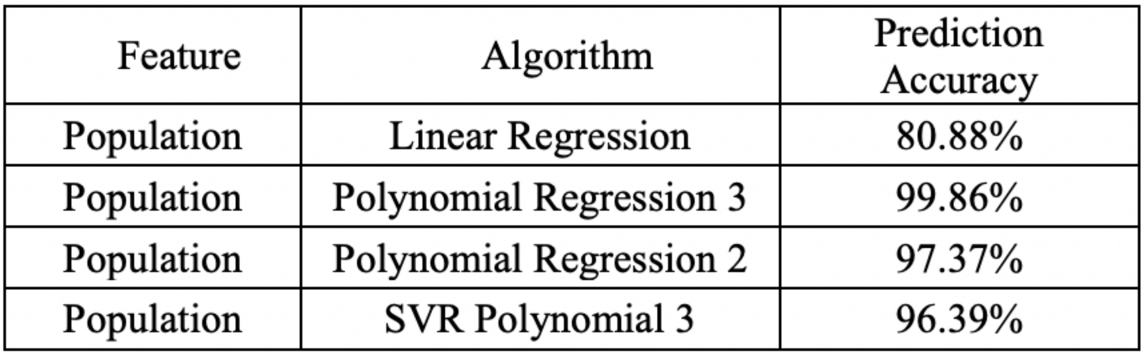


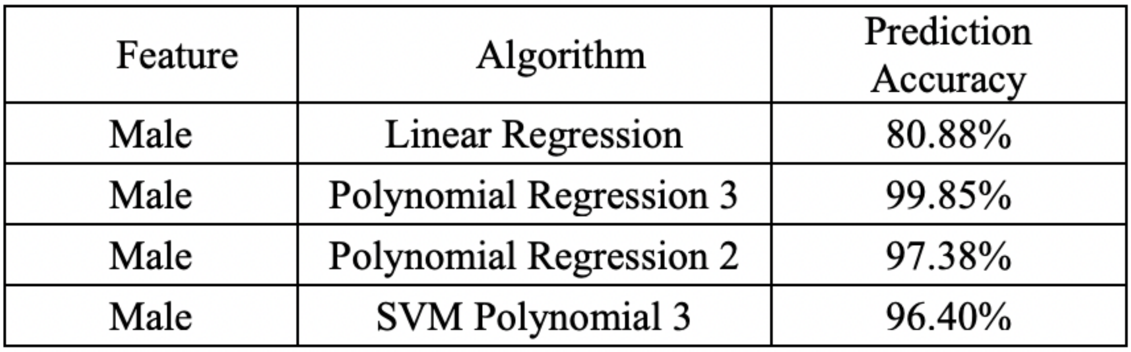
Fig: Population growth years from 1980 to 2020.

Table 1: GDP prediction result of population



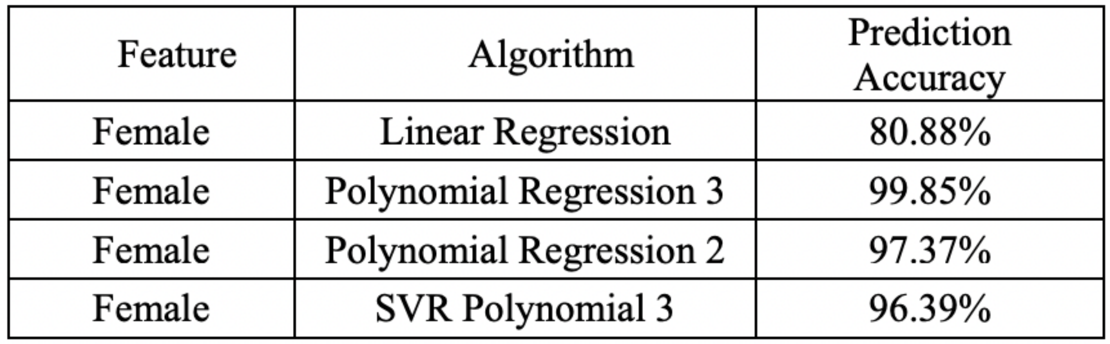
Feature	Algorithm	Prediction Accuracy
Population	Linear Regression	80.88%
Population	Polynomial Regression 3	99.86%
Population	Polynomial Regression 2	97.37%
Population	SVR Polynomial 3	96.39%

Table 2: GDP prediction result of Male



Feature	Algorithm	Prediction Accuracy
Male	Linear Regression	80.88%
Male	Polynomial Regression 3	99.85%
Male	Polynomial Regression 2	97.38%
Male	SVM Polynomial 3	96.40%

Table 3: GDP prediction result of Female



Feature	Algorithm	Prediction Accuracy
Female	Linear Regression	80.88%
Female	Polynomial Regression 3	99.85%
Female	Polynomial Regression 2	97.37%
Female	SVR Polynomial 3	96.39%

Conclusion

Our findings lead us to the conclusion that, in order to support Bangladesh's gdp, the country's female population must continue to grow. The importance of female education, professional advancements, and amenities can no longer be ignored, according to our GDP estimate. Both the male and female populations must collaborate to boost Bangladesh's economic development in order to achieve a significant GDP. Although Bangladesh's growth rate has slowed over the past 40 years, the growing population is extremely concerning.

2.6 machine learning based generic GDP analysis and prediction system

The sixth paper is a machine learning-based generic GDP analysis and prediction system which was written by Nikhil Vyas and jay Patel. Using the customised dataset for Gujarat State, a generic approach to forecast GDP values is presented in this work. Models based on several machine learning techniques, including ARIMA and the Random Forest Regressor, are proposed in this work. Regression and time-series analysis methods are created for the analysis and visualisation of GDP. For exploratory data analysis, the approach makes use of pandas-visual-analysis. They focused algorithm which is Random Forest Regression. It helps to reduce the manpower for predicting GDP. By implementing this model, GDP can be calculated more efficiently

The process can be simplified and paperwork can be reduced. The likelihood of mistakes is greatly reduced since the combination of several algorithms produces extremely high precision.

2.7 Nowcasting Indonesia’s GDP Growth Using Machine Learning Algorithms

The seventh paper are Nowcasting Indonesia’s GDP Growth Using Machine Learning Algorithms which was written as N D Muchisha, N Tamara, Andriansyah and A M Soleh. Since it is useful for determining policy, GDP should be closely watched in real-time. To predict the real-time GDP growth of Indonesia, we developed and compared ML models. Several quarterly macroeconomic and financial market information was utilised to create the 18 variables we used.

The chosen ML models, which include Random Forest, Ridge, SVM, and Neural Network, were finally discovered. GDP may be used to determine a country's economic structure, measure the pace of national economic growth, and assess how other nations are faring economically. The goal of this study is to develop the most effective ML model for real-time Indonesia. RMSE, MAD, and Pearson Correlation Coefficient data are compared to determine GDP growth. The dataset utilised in this study included the years 2009 through 2019.

In this study, using data from Indonesia's GDP growth, we evaluated how well multiple machine learning (ML) algorithms performed in real-time economic analysis. To estimate real-time GDP growth from 2013:Q3 to 2019:Q4, we trained 18 predictor variables in lags 1, 2, 3, and 4 for each algorithm approach. Numerous Indonesian quarterly macroeconomic and financial market data are included among the 18 response variable. Outcomes showed that when compared to the other approaches employed in this study, the real-time GDP growth estimate produced by the forecast combination methodology employing Lasso regression produces superior results.

2.8 A machine learning approach on the relationship among solar and wind energy production, coal consumption, GDP, and CO2 emissions

This study examines the links between the generation of solar and wind energy, the use of coal, economic expansion, and CO2 emissions for these three nations. To achieve this, we apply a cutting-edge machine learning approach to validate the predicted causal links between variables. The target variable for the Causal Direction from the Dependency (D2C) method was CO2 emissions. The results were broken down and calculated in a model of guided prediction. The results, which were supported by three separate Machine Learning techniques, produced an intriguing result. China and the US anticipate lower total carbon emissions due to their extensive usage of renewable energy sources, whereas India has alarming projections of higher CO2 emissions. This demonstrates that reducing CO2 emissions cannot be accomplished by carrying out a thorough transition from fossil to renewable resources, albeit China and the U.S. provide a more hopeful route to sustainability than India. India should increase the use of low-carbon sources in its electricity supply and reduce its reliance on coal as a rising leader in renewable energy.

Conclusion

They used a D2C causality model that can estimate the relationship between the production of energy from coal, solar, and wind, economic growth, and CO2 emissions. This is a sophisticated estimation approach in machine learning. The goal variable has been determined to be CO2 emissions. The findings demonstrate that, in contrast to CO2 emissions, the predicted causal relationship between the usage of coal and economic development is strong. In addition, we performed a disaggregated study in comparison to the findings of the D2C model to determine when the generation of energy from CO2, solar, and wind influenced the CO2 emissions. The empirical data highlighted the significant predicted causal relationship between CO2 emissions and CO2 emissions. We have only seen that outdated technology is present in the case of India. CO2 emissions are the process of making plant panels. The findings of the D2C models' predictions of CO2 emissions through 2023 were then subjected to a machine learning (ML) predictive analysis. The findings, which were supported by three separate studies, revealed that only India had a prediction of an increase in CO2 emissions.

As a result, policy suggestions can be created. First, governments may need to supplement their support for renewable energy with additional measures aimed at lowering the energy-based coal since coal consumption is a significant contributor to CO2 emissions in each of the three nations. Particularly effective strategies for conserving fossil fuels or initiatives to improve energy efficiency can help with this. Second, despite increased investment in solar and wind energy, these sources of power (particularly solar) may become more prevalent. In actuality, non-renewable resources still predominate in their respective proportions, and it is advised to continue implementing low-carbon energy policies. Third, we believe it is possible to achieve secure, low-carbon, and sustainable development by using renewable energy consumption as a direct driver of growth. It has been shown that the wealthier countries have more.

2.9 Nowcasting GDP using machine-learning algorithms: A real-time assessment

We analyse the nowcast performance of common algorithms for predicting macroeconomic variables using machine learning utilising real-time vintages of the New Zealand GDP growth (our target variable) and real-time vintages of over 600 predictors. Our findings demonstrate that machine-learning methods can outperform both a dynamic factor model and a basic autoregressive benchmark. We also show how artificial intelligence (AI) systems may improve upon and, in one case, beat official Reserve Bank of New Zealand projections.

Conclusion

In this article, they assessed how well well-known ML algorithms performed in getting precise nowcasts of real GDP growth for New Zealand in real time. Using numerous vintages of historical GDP data and multiple vintages of a sizable feature set made up of over 600 domestic and foreign variables, they generated a number of ML models for the 2009–2019 timeframe. The accuracy of a dynamic factor model, a naïve AR benchmark, and the official predictions generated by the Reserve Bank of New Zealand were all compared against the projections derived from these models. They discovered that the forecasts from all of the ML models were more precise than those from the AR and dynamic factor models. The findings also imply that the Reserve Bank of New Zealand may have used machine learning (ML) models to increase prediction accuracy. The findings therefore suggest the employment of ML algorithms as supplemental tools to aid in the understanding of the current status of the economy by policy makers.

2.10 Predicting Economic Recessions Using Machine Learning Algorithms

They contrast the actual facts with the forecasts made between 1990 and the present. The programme is unable to outperform the SPF estimates one quarter in advance.

The correlations between actual and expected results three and six quarters in advance are small, but they are very substantially different from zero. A significant slump in the first half of 2009 might have been forecast six quarters in advance in late 2007, despite the timing being somewhat off. The system has never correctly predicted a recession that didn't happen. The random forest machine learning technology has the ability to predict recessions early. They make use of a limited number of explanatory variables from financial markets that a forecaster would have had access to at the time of the forecast. They use the period 1970Q1–1990Q1 to train the algorithm and produce forecasts one, three, and six quarters in advance. They then make new projections using the period between 1970Q2 and 1999Q2, and so forth. We used the algorithm we obtained from package R's default input settings without making any attempt to optimise the predictions.

Conclusion

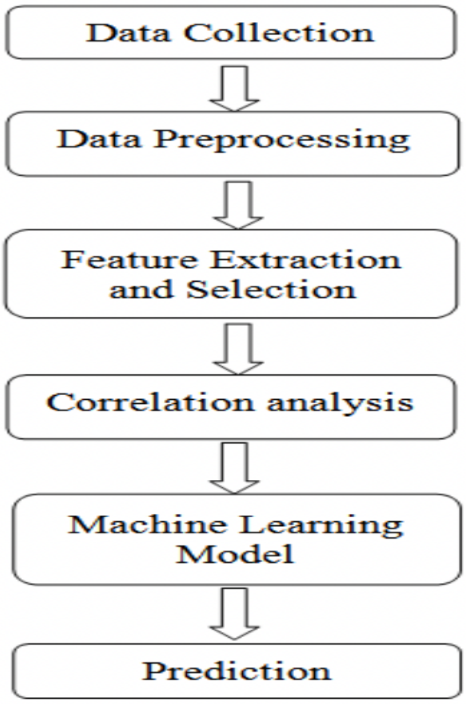
In addition, the regression of actual GDP growth on the mean forecast made three quarters earlier has little explanatory value. The SPF predictions never projected a single quarter of negative growth. This is greatly enhanced by the random forest method.

More shockingly, the random forest method would have predicted in the winter of 2007–2008 that the United States would have a severe recession in 2009 that would end in 2009Q4. To reiterate, they have not made any ex-post optimization attempts to improve these findings.They only employed a few explanatory variables, and they only used the machine learning algorithm's default values for the input parameters. For the UK, we find qualitatively identical results, although the random forest method has significantly greater predictive power than it does for the United States. According to Ormerod and Mounfield (2000), who used modern signal processing techniques, the time series GDP growth data is more heavily influenced by noise than by signal. As a result, there is probably a rather low upper limit on the level of forecast accuracy that may be attained. However, it appears that machine learning approaches have a lot of potential for expanding usable forecasting horizons and giving policymakers better information over such horizons.

3. Methodology

3.1 Flowchart of methodology

Figure 1: Flow Chart representation of the methodology of the study.



3.2 Data Collection

In order to properly estimate various macroeconomic indicators in the nation's economy using artificial intelligence methods, based on machine learning algorithms, the data for this research study were obtained from secondary data sources and acquired using the quantitative data approach.The data used in this paper is contained year and GDP of the UK. The dataset covered year-wise GDP rates in the UK economy. The sample period was from 1971 to 2021. Dataset split in to two sets: training data and testing data.The dataset taken from world data bank refer to fig 2.

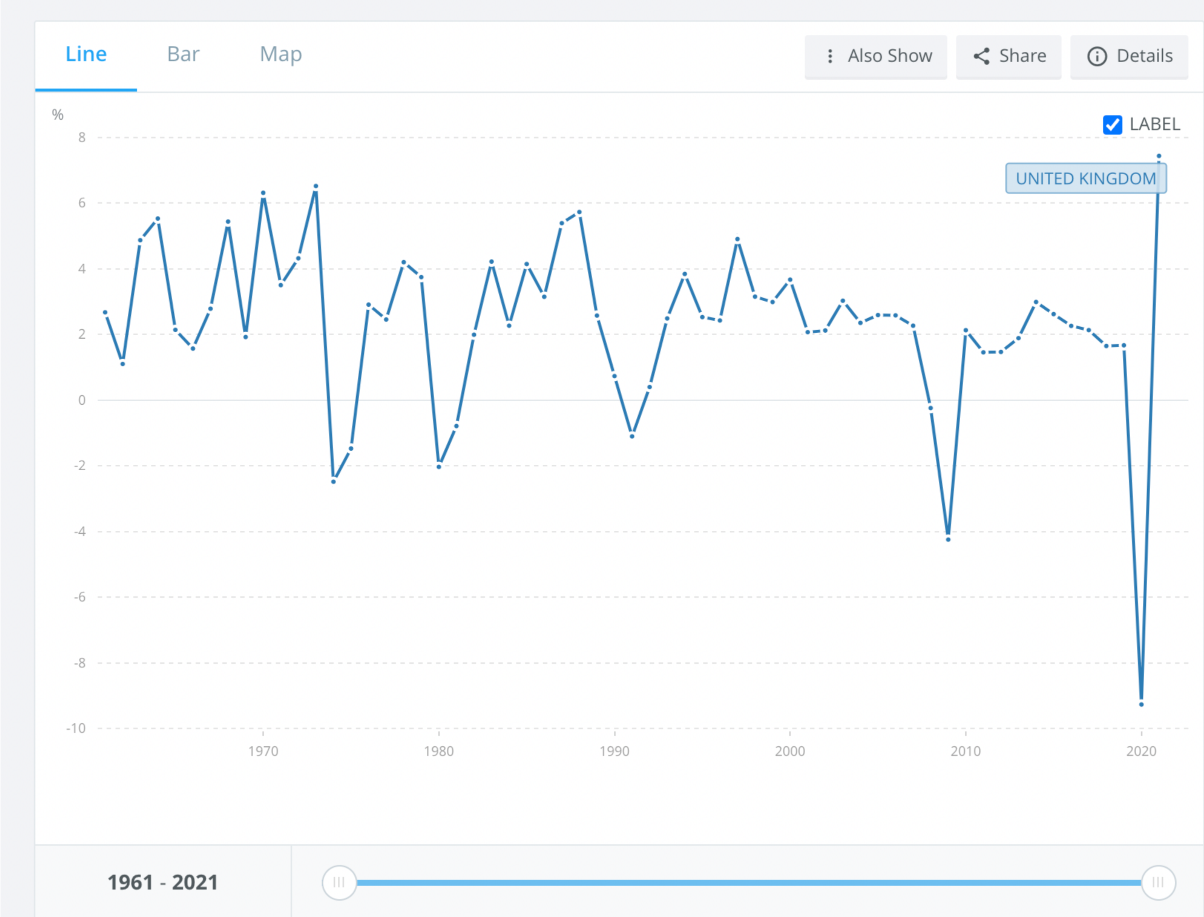


Fig 2; World Bank national accounts data and OECD National Accounts data files

3.3 Machine Learning Algorithm

In this paper, we used the KNN algorithm, Random Forest algorithm and SVM algorithm this three are classifications of machine learning algorithms. This section mainly focused on their different classification thoughts. They are commonly used in different finding scenarios. These algorithms are briefly explained in given below.

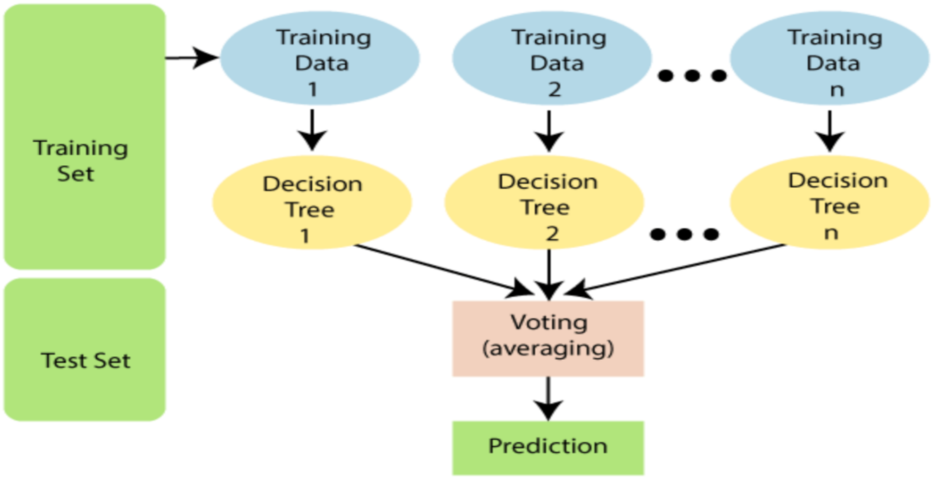
K-Nearest Neighbor (KNN)

A traditional supervised machine learning technique is the KNN algorithm. Both classification and regression tasks may be accomplished with the KNN. The KNN method was employed as a classifier in this study. The closest neighbour of the test sample was located in the train set for prediction and forecasting purposes using the test sample and the train set. KNNs are parameterized by the number of nearest neighbours K. The data point belongs to the same class as the majority of

K's closest neighbours if both belong to the same class and share the same features. The formula for measuring the separation between two data points.

Random Forest Algorithm

The Random Forest classifier aggregates the outcomes from several decision trees used on diverse subsets of the input dataset in order to improve the estimated accuracy of the input dataset. The below diagram explains how to work Random Forest Algorithm



Fig;1 how to work Random Forest Algorithm

Some decision trees may predict the correct output while others could not, given that the random forest employs a range of trees to estimate the class of the dataset. But when all the trees are included, they successfully forecast the result.

Support Vector Machine (SVM)

Both classification and regression use in support vector machine. Decision planes, which outline decision constraints, serve as its conceptual underpinning. A decision plane is a line separating a group of objects with different class memberships. It is a Parametric method. In this paper SVM is used as regressor. A non-parametric approach called the support vector machine regression model was initially put out by Vapnik (1995). Finding a linear function of the following form is the aim of SVM.

$$F (x) = (w ,x) + b$$

When ww is the weight vector, bb is the bias, and xx is the input or feature vector, making sure that the function is as flat as feasible by aiming for a small ww. One strategy is to reduce the norm, This is a convex optimization problem that can be expressed as:

$$0.5||w||^2 + C\sum_{i=1}^{l}|(y_l - f(x_l))|_\epsilon$$

In which the regularisation parameter C > 0 is. The penalty term that appears initially in the error function grows as the model's complexity rises. The second term is the -insensitive algorithm, which penalises mistakes bigger than and gives the model flexibility.

3.4 Anaconda

Python is an object-oriented and mid-level programming language. Python and Anaconda are using in this project. We can access many environments that let you code in either Python or R after installing Anaconda. It may help you save time, give you more control over your studies, let your creativity be the limit of your research ideas, and maximise the output of your study. Anaconda is the simplest way to use Python/R data science and machine learning on a single PC. You may work with hundreds of open-source packages and libraries thanks to this toolkit. It was developed with lone practitioners in mind. It is the best approach to learn Python and an excellent platform for both inexperienced and seasoned programmers to tackle a variety of jobs.

3.4.1 Downloading and Installing Anaconda

Anaconda provides the website, we download software from the given link for free to Linux, Mac, or Windows. They were given videos and documents for how to install anaconda. This lesson includes a step-by-step installation instruction to make the procedure easier. You must choose your operating system, the Python edition you need, and whether you prefer the 32-Bit or 64-Bit Graphical Installer depending on the kind of CPU in your computer before you can download and install Anaconda. To update the python version every time to get a good result.When the download has finished in the mac operating system given Choosing location for installation and advanced installation option after last step is to give finish installation and wizard. You may find and launch the Anaconda Navigator on your mac when the installation is finished.

The following IDEs are pre-installed with Anaconda and are among the several ones available for Python development.

3.5 Jupyter Notebook

The other methodology we using Jupyter Notebook. A web-based IDE that makes use of your default web browser is Jupyter Notebook. Because each block of code may be executed independently, it is very versatile and simple to experiment with. In all one place, we can use equations, code output, and visualizations. Also used normal text in one place. JupyterLab is the extension of Jupyter Notebook.

3.6 Spyder

Spyder is a powerful IDE created especially for data analysis. It may be thought of as an all-in-one IDE because it includes a Python terminal, a variable explorer for quick data inspection, and an editor for writing code scripts. Powerful tools for code inspection and debugging are also included, enabling users to scrutinise complete or individual lines of code to find and correct mistakes.

4. Implementation and Analysis

4.1 Scikit – learn

A Python package called Scikit-learn contains a number of supervised and unsupervised learning methods. It is built using software that you might be familiar with, like NumPy, pandas,

Scikit-learn supports the following features:

1.Regression, including Linear and Logistic Regression

2.K-Nearest Neighbors Regression Classification

3.Clustering techniques such as K-Means and K-Means++

4.Model preference

5.Min-Max Normalization and Preprocessing

4.2 NumPy

The library known as NumPy, which stands for Numerical Python, offers multidimensional array objects and tools for working with them. Arrays may be used by NumPy for scientific and analytical operations.

4.3 Pandas

A well-known open source Python library for data science, data analysis, and machine learning activities is called Pandas. It is constructed using Numpy, a separate package that supports multidimensional arrays. One of the most well-known tools for manipulating data, Pandas, is frequently included in all Python versions and integrates nicely with a variety of other data science modules.Many of the tedious, repetitive tasks associated with data processing are simple to complete with Pandas.such as:

1. Data cleanup
2. Data entry
3. Normalization of data

merging and joining

5. Displaying data
6. Statistic evaluation
7. Data analysis
8. Data loading and saving

There are 2 parts for Implementation ,the first part is gonna be the data preprocessing and the second part is gonna be the Data Exploratory Analysis.

4.4 Data Preprocessing

Preparing raw data for future processing by any type of processing is known as data preprocessing, which is a subset of data preparation. Historically, it has been an important initial step in the data mining process. Recently, methods for preparing data have been changed to train machine learning and AI models as well as make judgments about them.

Data mining, machine learning, and other data science processes may manage data in a more straightforward and effective way with the help of data preparation. The methodologies are frequently used in the first stages of the machine learning and AI development pipeline to produce dependable results.

Data in the real world is chaotic because it is frequently produced, processed, and stored by a variety of individuals, business processes, and applications. As a result, a data collection may be incomplete, contain mistakes from human entry, or have redundant data or names for the same thing. However, data used to train machine learning or deep learning algorithms must be automatically preprocessed. While these flaws may commonly be discovered and corrected by people in business data, they must be automated. The essential data preparation processes

1.Data Cleaning – Data cleaning is eliminating bad data ,filling in missing data.If the data is a big data ,the rows can be deleted .If the size of the dataset is small ,then we can take the mean of the dataset and fill the missing values.

2.Identify and removing duplicates -There will be data in the dataset ,which may contain duplicated rows.We need to make sure there are no duplication ,if found we need to delete it.

Data Exploratory Analysis

Data scientists typically employ data visualisation techniques when doing exploratory data analysis (EDA), which they use to examine, investigate, and characterise various data sets' key characteristics. It makes it easier for data scientists to find patterns, spot anomalies, test hypotheses, and verify assumptions by providing guidance on how to manipulate data sources efficiently to get the answers needed.

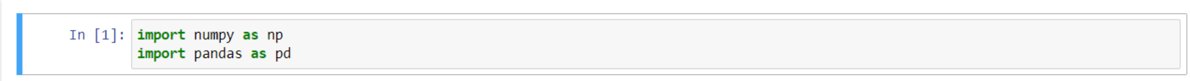
EDA offers a greater knowledge of the variables in the data set and how they interact, and it is usually used to discover what the data may disclose beyond formal model and hypothesis testing operations. It can also assist you in deciding whether the statistical methods you're contemplating using for data analysis are acceptable.

The EDA procedure includes identifying the distribution of the data and any outliers. An outlier is a value in a random sampling of data from a population that differs significantly from the other values. In some ways, this idea leaves it up to the analyst (or a consensus process) to define what behaviour is abnormal. Before anomalous observations can be found, normal observations must first be characterised. Finding the correlation between the columns is also important.

The correlation coefficient is used to calculate the link between two variables.The correlation coefficient can never be -1 or more than 1.1 Means the variables have a perfect linear connection .0 indicates that there is no linear connection between the variables.-1 indicates that the variables have a perfect negative linear connection.

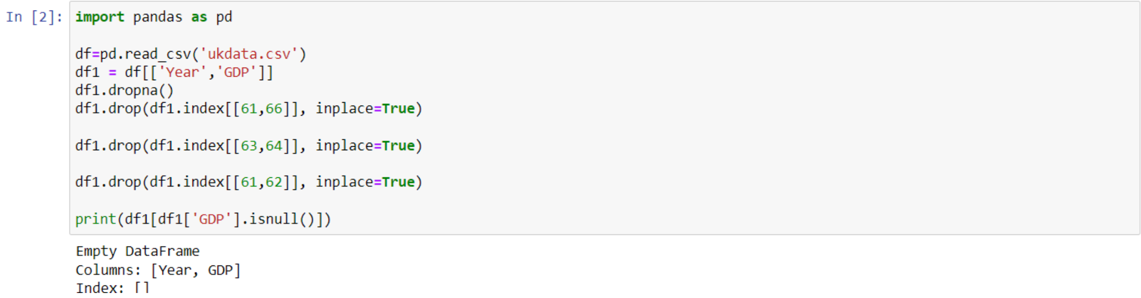
Implementation of code in Jupter Notebook

The first step is to import the necessary library



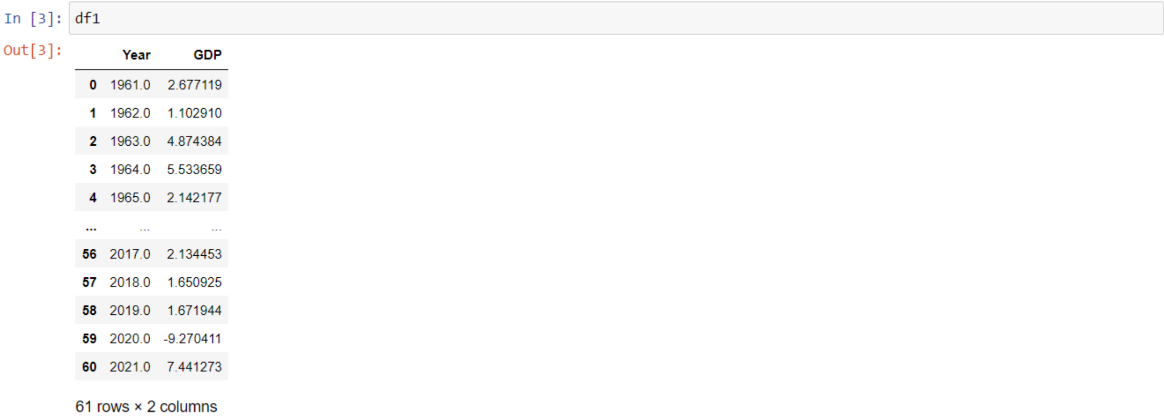
```
In [1]: import numpy as np
        import pandas as pd
```

We are gonna use the read_csv method to call the dataset that is stored in the local computer.Once the data is read we need to remove any missing values ,while transposing data in excel missing values were created ,so we need to use the dropna() function .The Dropna function will find any missing values in the data set ,and it will drop it.



```
In [2]: import pandas as pd

        df=pd.read_csv('ukdata.csv')
        df1 = df[['Year','GDP']]
        df1.dropna()
        df1.drop(df1.index[[61,66]], inplace=True)

        df1.drop(df1.index[[63,64]], inplace=True)

        df1.drop(df1.index[[61,62]], inplace=True)

        print(df1[df1['GDP'].isnull()])

        Empty DataFrame
        Columns: [Year, GDP]
        Index: []
```

As you can see all the missing values have been dropped from the dataset .

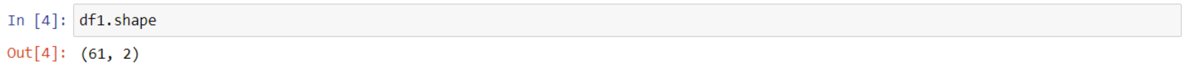


```
In [3]: df1
Out[3]:
```

	Year	GDP
0	1961.0	2.677119
1	1962.0	1.102910
2	1963.0	4.874384
3	1964.0	5.533659
4	1965.0	2.142177
...	...	...
56	2017.0	2.134453
57	2018.0	1.650925
58	2019.0	1.671944
59	2020.0	-9.270411
60	2021.0	7.441273

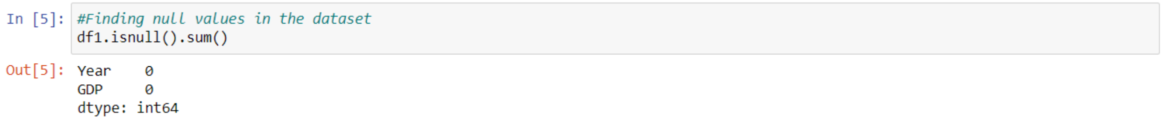
61 rows × 2 columns

The above figure shows the dataset,where there are 2 columns Year and the GDP(Gross domestic product) , the GDP values are the annual growth of United Kingdom from the year 1961 to 2021.So in general there are 61 rows for the analysis.



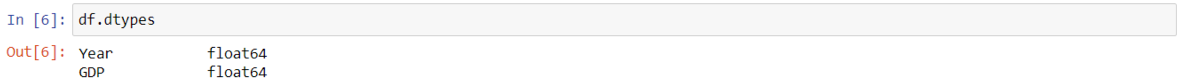
```
In [4]: df1.shape
Out[4]: (61, 2)
```

The shape function tells the shape of the dataset ,so data has 61 rows and 2 columns.



```
In [5]: #Finding null values in the dataset
        df1.isnull().sum()

Out[5]: Year    0
        GDP     0
        dtype: int64
```

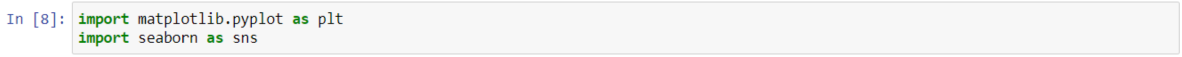
The above figure comfirms that there are no null values in the dataset ,those which were present in the dataset were successfully removed



```
In [6]: df.dtypes
Out[6]: Year          float64
        GDP           float64
```

The column datatype is shown by the dtypes function. At the moment, the Year and GDP are in the float64 datatype. A floating radix point is used to describe a wide dynamic range of numerical values in the double-precision floating-point format, often known as FP64 or float64.

Even though it sacrifices precision, floating point is employed to represent fractional values or when a broader range than fixed point (of the same bit width) is needed. Double precision may be utilised when single precision's range or accuracy is insufficient.



```
In [8]: import matplotlib.pyplot as plt
        import seaborn as sns
```

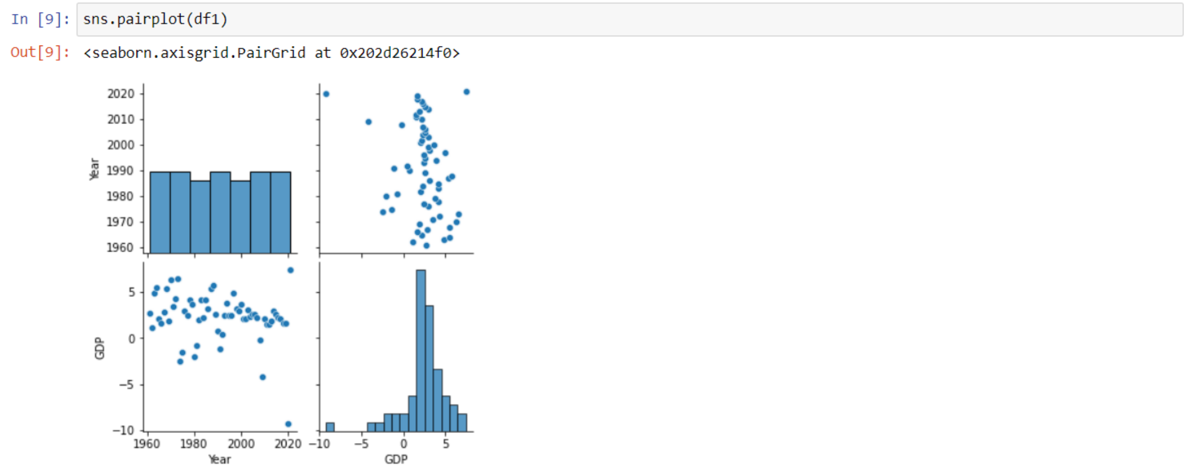
Importing the plotting libraries,here we have imported matplotlib and seaborn library. A cross-platform tool for data visualisation and graphical charting, Matplotlib, is available for Python and its NumPy numerical extension. As a result, it offers an open source replacement for MATLAB. By utilising the APIs of Matplotlib, developers may also include plots in GUI programmes.
(Application Programming Interfaces).

Installing Matplotlib

The python package can be downloaded as pip install matplotlib in the anaconda prompt .For importing matplotlib we use the pyplot method.

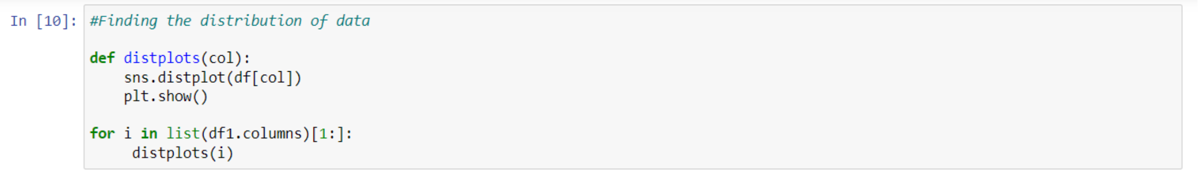
Seaborn - A Python data visualisation software called Seaborn uses the matplotlib library. For making aesthetically beautiful and practical statistical graphics, it provides a high-level interface.

The anaconda prompt contains the code for installing Seaborn using Conda.



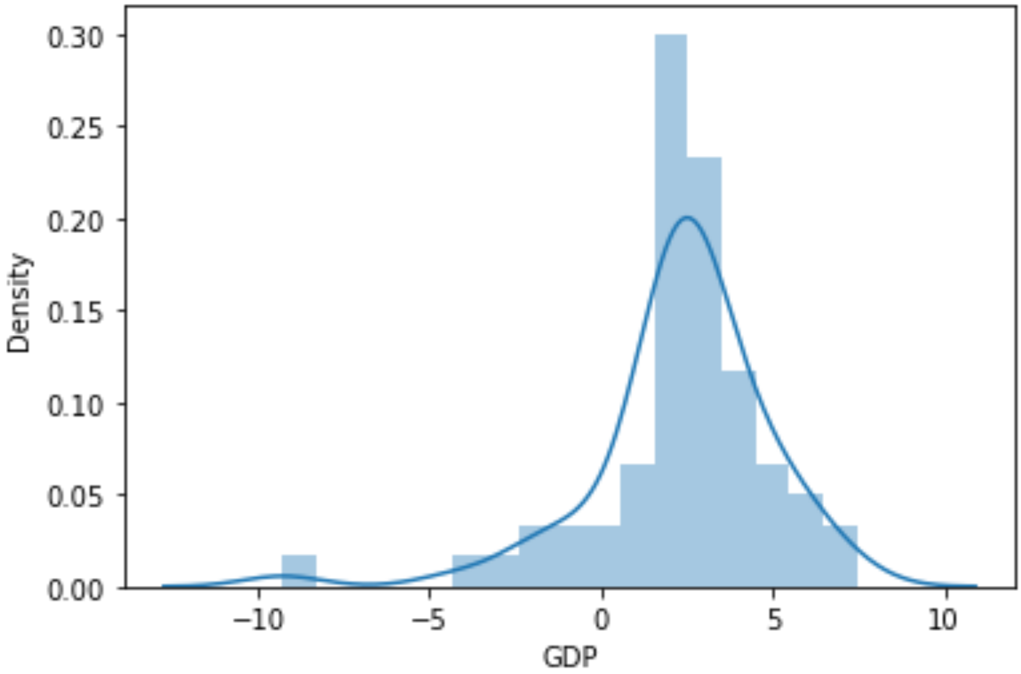
```
In [9]: sns.pairplot(df1)

Out[9]: <seaborn.axisgrid.PairGrid at 0x202d26214f0>
```

Using the seaborn library we used the pair plot method with df1 as the input argument .Pairplot gives the visualization of the given data and find relationship between the variables .In the above figure the diagonal are univariate histogram of each attribute.And the scatterplot shows distribution of data points for each attribute.



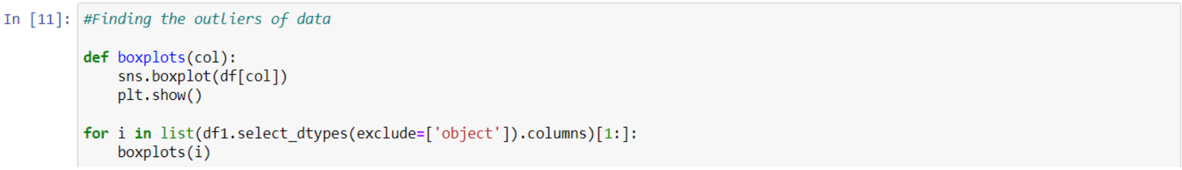
```
In [10]: #Finding the distribution of data

         def distplots(col):
             sns.distplot(df[col])
             plt.show()

         for i in list(df1.columns)[1:]:
              distplots(i)
```

We need to check the distribution of the data , first a function is created for distribution plot which is named as distplot with an input argument of column as col .Then a for loop is written to iterate through the list of columns and calling the distplot function.

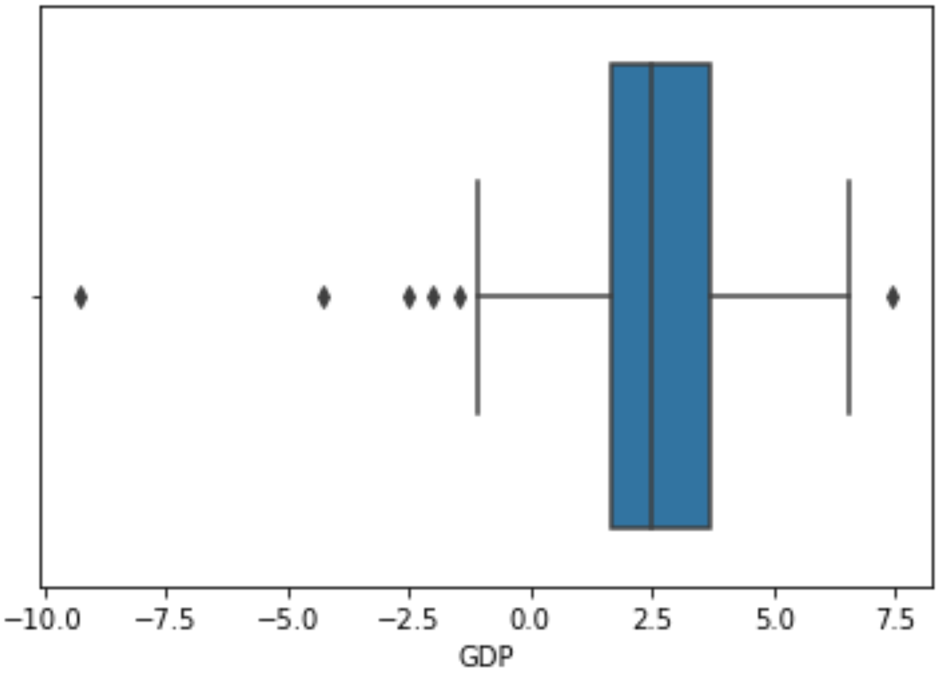


The above graph shows the distribution of the GDP column using an Histogram.Most of the values lies between 0 and 5.

Boxplots- A five-number summary of a set of data is displayed in a box and whisker plot, sometimes referred to as a box plot. The five-number summary is composed of the minimum, first quartile, median, third quartile, and maximum. Boxplots also reveal the presence of outliers in the data. Below, you can see the Boxplot of our data.

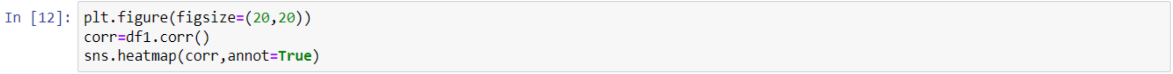


```
In [11]: #Finding the outliers of data

         def boxplots(col):
             sns.boxplot(df[col])
             plt.show()

         for i in list(df1.select_dtypes(exclude=['object']).columns)[1:]:
             boxplots(i)
```



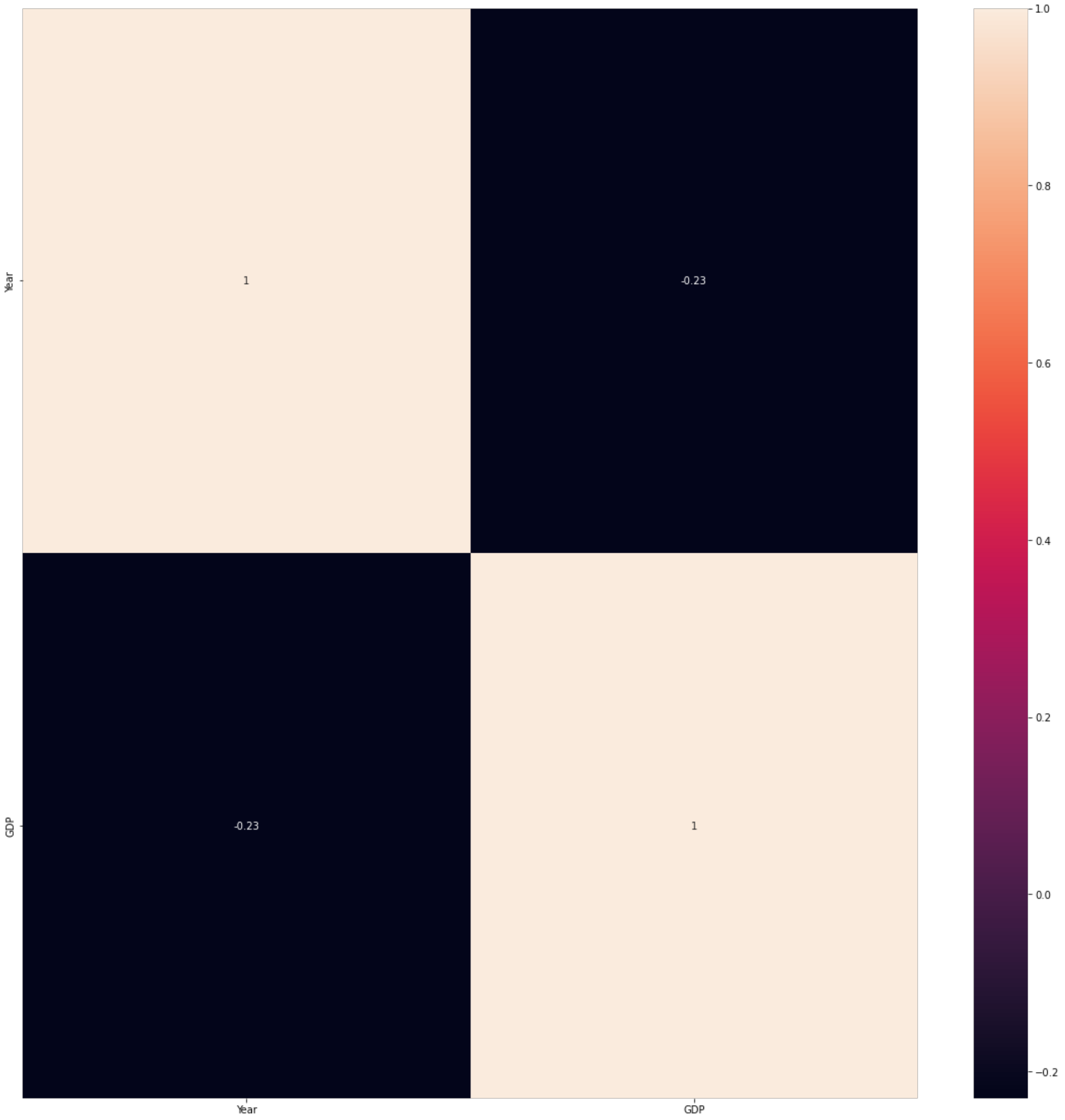
The above boxplot shows there are outliers , but since are GDP can be in negative value ,We cannot take the dotted points as outliers.

Next we find the correlation between the 2 attributes. The heatmap is used find the correalation.



```
In [12]: plt.figure(figsize=(20,20))
         corr=df1.corr()
         sns.heatmap(corr,annot=True)
```

We used the seaborn library to plot the heatmap


1
-0.23
-0.23
1
Year
GDP
Year
GDP
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

5. Testing

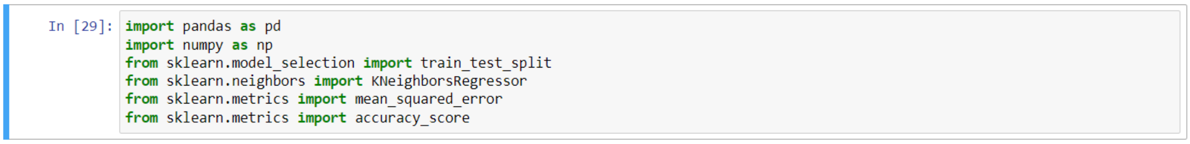
The Data that have been preprocessed and analysed after Exploratory Data Analysis, is ready to be fitted in the model.The3 models that we gonna use are KNN (k nearest neighbours), SVM, and Random Forests. All the algorithms are taken from the scikit-learn library

The first step towards testing of the models is to import all the necessary libraries. Some of the libraries we are gonna use train_test_split, accuracy_score , mean_squared_error.

Train_test_split- The train test split technique splits a single dataset into training and testing halves. Your model is developed using the testing subset. By applying the model on untested data, the testing subset is utilised to assess the model's performance.

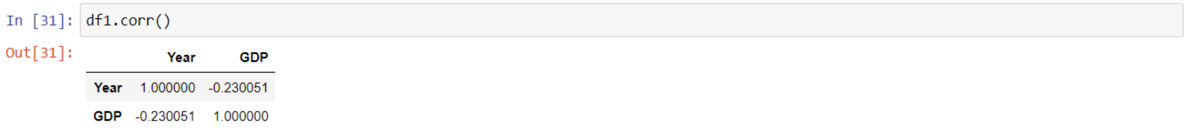
Accuracy_score - A combination of predicted labels and actual labels are combined to get an accuracy score using the accuracy score function of the Python learn.metrics package.

Mean_squared_error - An estimation of a regression line's proximity to a set of points is given by its mean squared error (MSE). The distances between the points and the regression line are squared to achieve this (the "errors"). Positive signals must be eliminated using squaring. Greater weight is also placed on larger differences. The phrase "mean squared error" refers to how you calculate the average of a number of errors. Forecasting accuracy increases with decreasing MSE.



```
In [29]: import pandas as pd
         import numpy as np
         from sklearn.model_selection import train_test_split
         from sklearn.neighbors import KNeighborsRegressor
         from sklearn.metrics import mean_squared_error
         from sklearn.metrics import accuracy_score
```

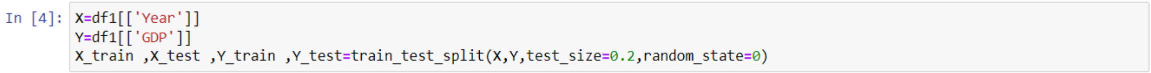
The above figure shows the required libraries to be imported.



```
In [31]: df1.corr()
Out[31]:
```

	Year	GDP
Year	1.000000	-0.230051
GDP	-0.230051	1.000000

The coorealation matrix shows a negative correlation which is -0.23

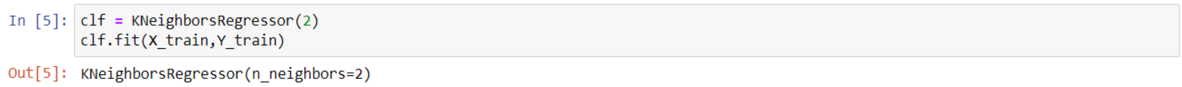


```
In [4]: X=df1[['Year']]
        Y=df1[['GDP']]
        X_train ,X_test ,Y_train ,Y_test=train_test_split(X,Y,test_size=0.2,random_state=0)
```

The data will now be divided into two categories: training data and testing data. Since the test size is set to 0.2, 20% of the data will be used for testing, while the remaining 80%

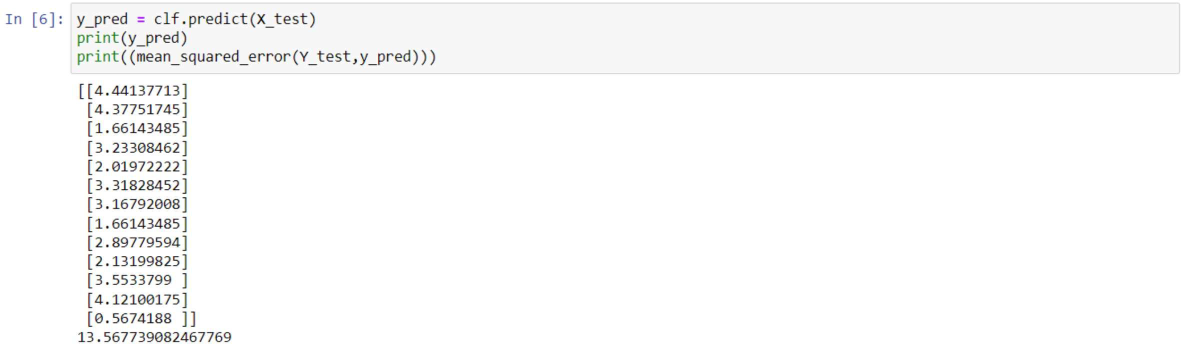
will be used for training data. The variables that have been set are X_train as x variable training data ,Y_train which is Y variable training data ,X_test is X variable testing data and Y_test is Y variable testing data.

Model 1



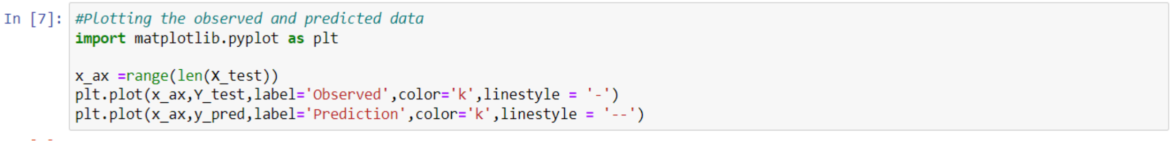
```
In [5]: clf = KNeighborsRegressor(2)
        clf.fit(X_train,Y_train)

Out[5]: KNeighborsRegressor(n_neighbors=2)
```

The first model in this analysis the KNN (k nearest neighbour) ,using the KneighboursRegressor ,a model named clf is created.TheX_train and Y_train data are fitted into the model.



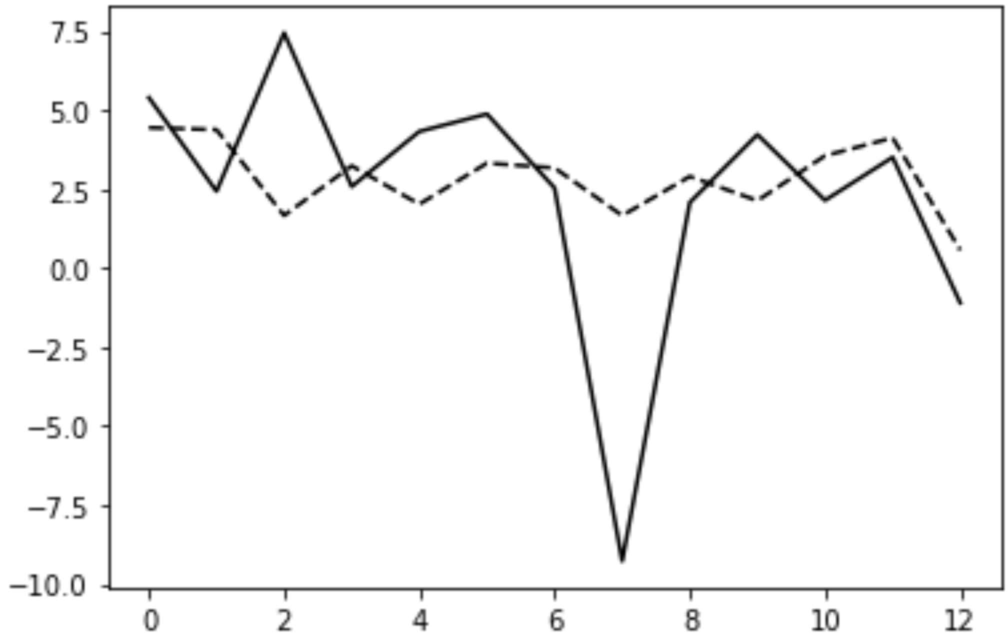
```
In [6]: y_pred = clf.predict(X_test)
        print(y_pred)
        print((mean_squared_error(Y_test,y_pred)))

        [[4.44137713]
         [4.37751745]
         [1.66143485]
         [3.23308462]
         [2.01972222]
         [3.31828452]
         [3.16792008]
         [1.66143485]
         [2.89779594]
         [2.13199825]
         [3.5533799 ]
         [4.12100175]
         [0.5674188 ]]
        13.567739082467769
```

Now using the predict method ,the X_test data is fitted to predict the y_predict,and also we find the mean_squared_error to find the error .The mean_squared_error is foun to be 13.5 which is very high. The mean squared error should be lower to be a better model.



```
In [7]: #Plotting the observed and predicted data
        import matplotlib.pyplot as plt

        x_ax =range(len(X_test))
        plt.plot(x_ax,Y_test,label='Observed',color='k',linestyle = '-')
        plt.plot(x_ax,y_pred,label='Prediction',color='k',linestyle = '--')
```

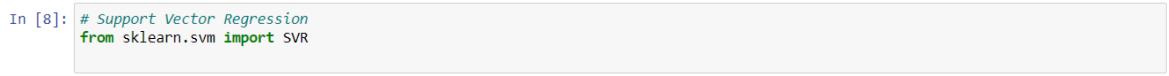
Using the matplotlib.pyplot we are gonna plot the observed and the predicted value.The below Graph shows the plot.



As from the plot,the prediction value couldn't predict very accurately .It couldn't predict the 2008 recession which was huge drop in GDP.

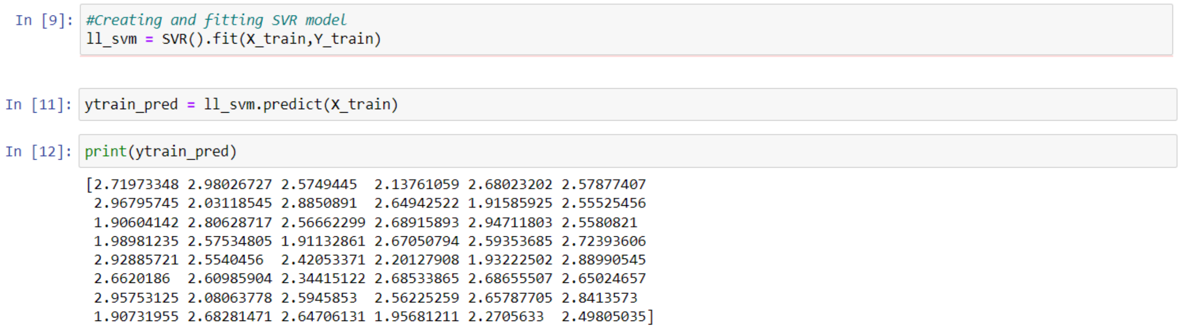
Model 2

For the second model we are gonna use the Support Vector Regressor .The SVR library is imported from sklearn.svm



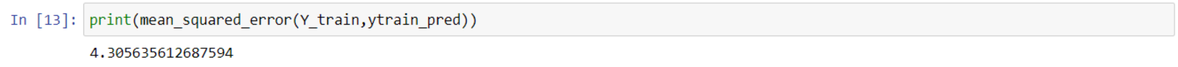
```
In [8]: # Support Vector Regression
        from sklearn.svm import SVR
```

The next thing is to fit the model ,and fitting the X_train and Y_train into it



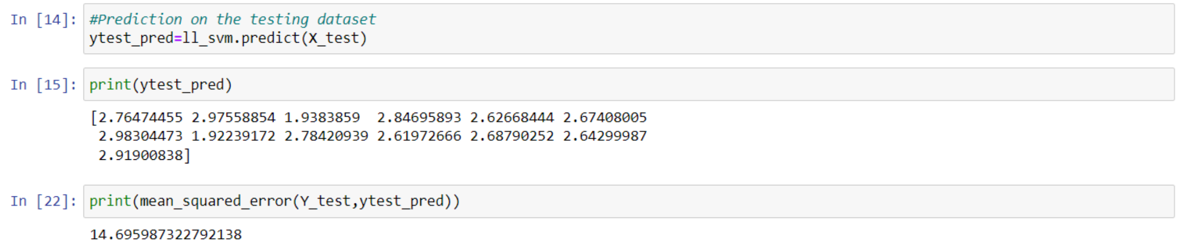
```
In [9]: #Creating and fitting SVR model
        ll_svm = SVR().fit(X_train,Y_train)

In [11]: ytrain_pred = ll_svm.predict(X_train)

In [12]: print(ytrain_pred)

         [2.71973348 2.98026727 2.5749445  2.13761059 2.68023202 2.57877407
          2.96795745 2.03118545 2.8850891  2.64942522 1.91585925 2.55525456
          1.90604142 2.80628717 2.56662299 2.68915893 2.94711803 2.5580821
          1.98981235 2.57534805 1.91132861 2.67050794 2.59353685 2.72393606
          2.92885721 2.5540456  2.42053371 2.20127908 1.93222502 2.88990545
          2.6620186  2.60985904 2.34415122 2.68533865 2.68655507 2.65024657
          2.95753125 2.08063778 2.5945853  2.56225259 2.65787705 2.8413573
          1.90731955 2.68281471 2.64706131 1.95681211 2.2705633  2.49805035]
```

The above figure shows the ytrain_prediction values in an array



```
In [13]: print(mean_squared_error(Y_train,ytrain_pred))

         4.305635612687594
```

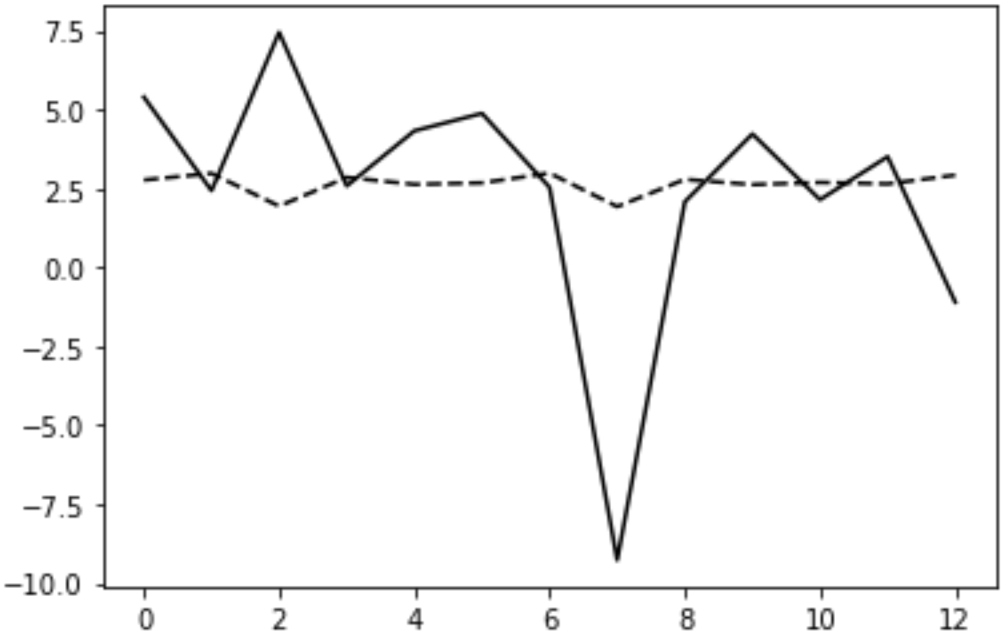
The mean squared error for the training data is found to 4.30 which is a good value.



```
In [14]: #Prediction on the testing dataset
         ytest_pred=l1_svm.predict(X_test)

In [15]: print(ytest_pred)

         [2.76474455 2.97558854 1.9383859  2.84695893 2.62668444 2.67408005
          2.98304473 1.92239172 2.78420939 2.61972666 2.68790252 2.64299987
          2.91900838]

In [22]: print(mean_squared_error(Y_test,ytest_pred))

         14.695987322792138
```

The above code shows prediction for the testing data and the mean squared error was found to be 14.6 ,which is again a very bad model.

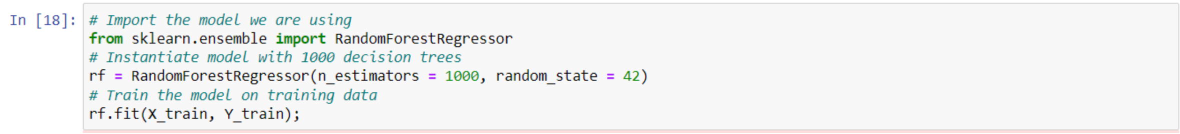
Below is the plot for the observed and predicted values.



As the graph shows the predicted values do not match the observed value at all.

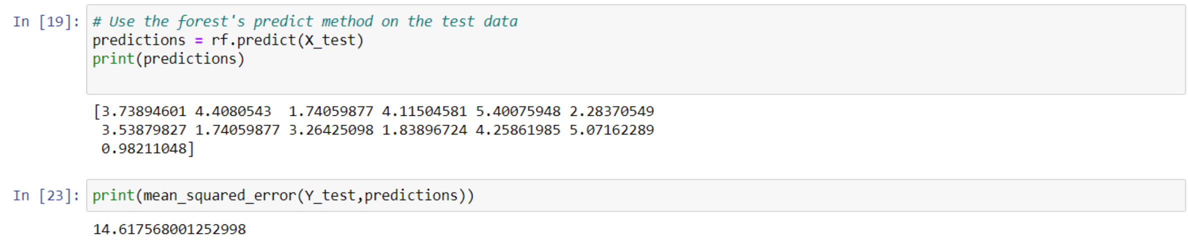
Model 3.

For Model 3 we will be using Random Forest algorithm .The Random forest regressor library is imported from sklearn.ensemble



```
In [18]: # Import the model we are using
         from sklearn.ensemble import RandomForestRegressor
         # Instantiate model with 1000 decision trees
         rf = RandomForestRegressor(n_estimators = 1000, random_state = 42)
         # Train the model on training data
         rf.fit(X_train, Y_train);
```

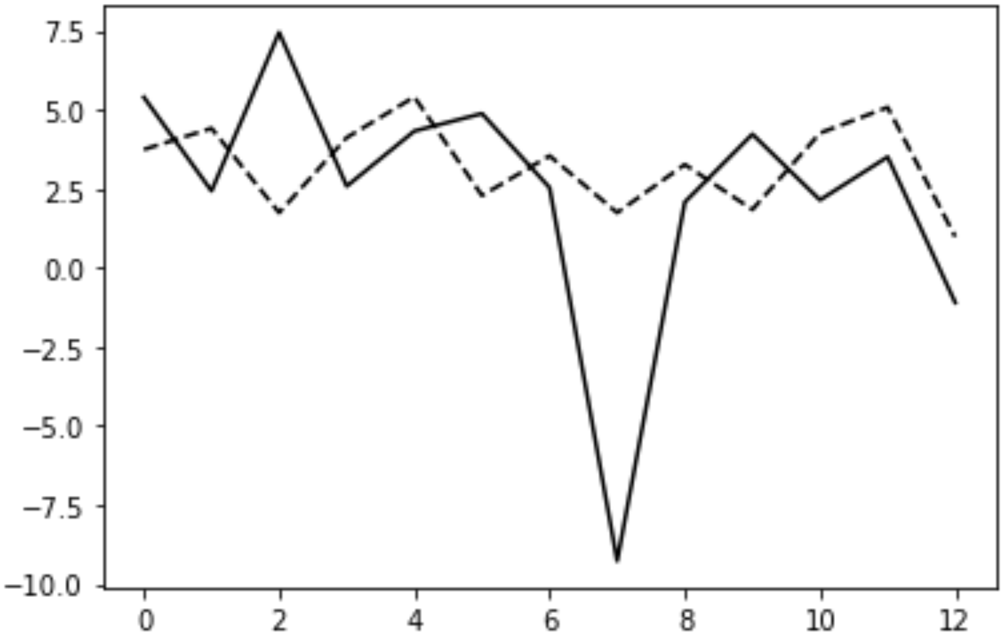
Also the variable model name is given as rf ,the X_train and Y_train values are fitted into the model .



```
In [19]: # Use the forest's predict method on the test data
         predictions = rf.predict(X_test)
         print(predictions)

         [3.73894601 4.4080543  1.74059877 4.11504581 5.40075948 2.28370549
          3.53879827 1.74059877 3.26425098 1.83896724 4.25861985 5.07162289
          0.98211048]

In [23]: print(mean_squared_error(Y_test,predictions))

         14.617568001252998
```

Finally the final random forest mean squared error value is observed as 14.61 ,which again is a very bad model

The plot for the observed and Predicted values for the random forest algorithm is given below



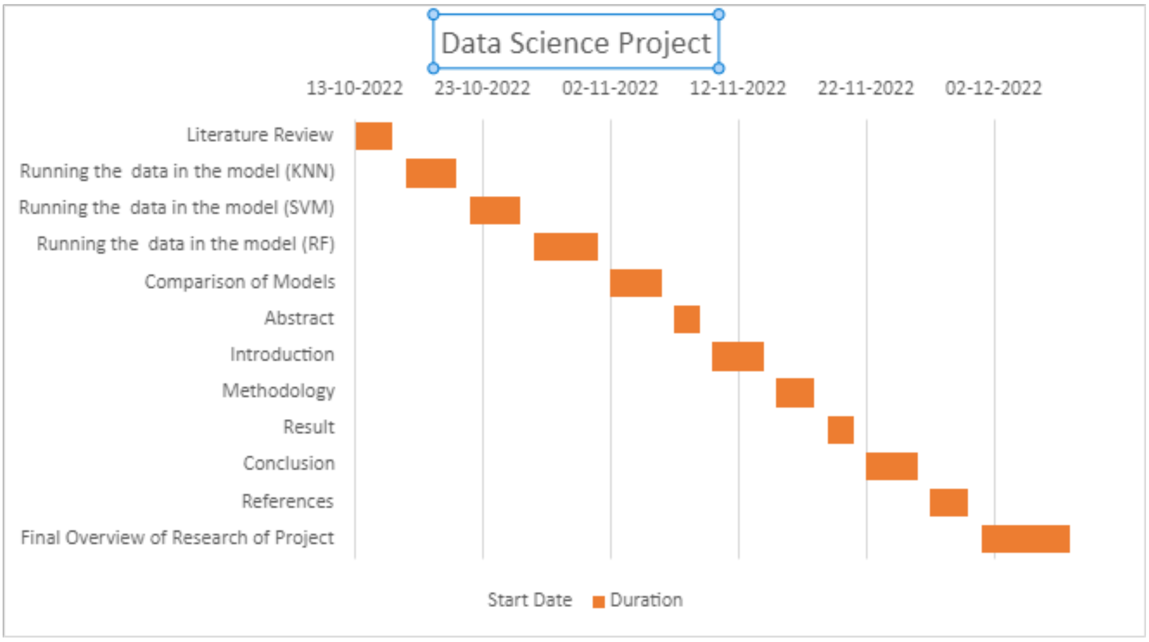
The predicted values are similar in certain places ,but certain data points follow the observed data points.

6. Project Management

I continue to utilise university resources, including the library's collection of textbooks, online essay-structure guidelines, and textbooks for critical analysis. I've started my literature review and have been reading through textbooks on geology, materials science, and glazes. After talking with my tutor, I realised that putting glazes in the context of the social history of material science would be better handled in a written dissertation than in a technical report, which I had originally considered writing for my dissertation. I have done the thorough study for significant books and am developing crucial ideas to solve issues that are raised in the studies.

The project started in September, we had a meeting, and we attended an online meeting. My supervisor leads me on how to start the project and how to analyse the data. In the first week covered a literature review. To take seven research papers and read them. I learned how to do a project on a research paper. The second week covers the machine learning algorithm which is used in this project. I have watched tutors for guidance to develop the dissertation. The dataset was run in the models. To start a comparison of models in the third week and cover the Abstract and Introduction. In the fourth week start the methodology and result and In the fifth week cover the Conclusion and References. In the last week of the period covering the final overview of the project.

6.1 Project Schedule



7. Conclusions

A country's economic size and health throughout time are gauged by its gross domestic product (GDP) (usually one quarter or one year). In addition, it is used to compare the size of various economies through time. To determine GDP, the Office for National Statistics Opens in a separate window (ONS) gathers information from hundreds of UK companies every three months. GDP may be calculated in three distinct ways, further complicating issues. You may calculate it by totaling up all the goods and services produced (the "output"), everyone's income, and the total amount spent nationwide. When the GDP increases, the economy expands because people spend more and businesses could flourish. Because consumers spend more as the GDP increases, the economy expands.
GDP also does not reveal anything about how income is distributed throughout the population. Growth might imply that everyone gets a better deal or that the richest sector gets more wealthy. In actuality, it is frequently somewhere in the middle.

Taking population size variations into account is also crucial. The average income per person will decrease if the UK GDP grows by 2% next year while population increases by 4%.

Then there are things that boost GDP but have no positive effects on the nation. One instance is war (a lot of money is spent, so GDP goes up). Or there would be a substantial rise if a sizable piece of the Amazon rainforest were cleared out in a single week.
According to data from the World Bank, the United Kingdom's GDP was 3186.86 trillion US dollars in 2021. 2.38 percent of the global GDP is made up by the United Kingdom.

When a GDP falls for 2 quarters continuously, that’s what a recession.This project's aim was to predict GDP for future preparation. In this paper we use 3 machine learning algorithm for the prediction of GDP ,those were K nearest neighbour (KNN) , Support vector Machine Regressor and Random Forest. All the 3 models performed very badly giving a similar mean squared error of 14 . One of the reasons why the model performed badly was there that there were not enough data available.Also a lot of factors depend on the GDP which was not taken care into. Machine learning can be really helpful to predict the GDP but with the right parameters

8. Future Work

A better dataset can be taken to analyse the GDP of UK. Also we can other regressor model to analyse the data. We can use other factors of GDP.

9. Bibliography and References

[1] Fernando, J. (2022). *Gross Domestic Product - GDP*. Investopedia. https://www.investopedia.com/terms/g/gdp.asp

[2] Countryeconomy. (2019, November 25). *United Kingdom (UK) GDP - Gross Domestic Product 2019*. Countryeconomy.com. https://countryeconomy.com/gdp/uk

[3] World Bank. (2010). *GDP growth (annual %) - United Kingdom | Data*. Worldbank.org. https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG?locations=GB

[4] An in-depth guide to supervised machine learning classification. (n.d.). Built In. https://builtin.com/data-science/supervised-machine-learning-classification

[5] Nyman, R., & Ormerod, P. (2016). *Predicting Economic Recessions Using Machine Learning Algorithms*. https://www.paulormerod.com/wp-content/uploads/2012/06/Random-Forest-23-Dec-2016.pdf

[6] Richardson, A., Van, T., Mulder, F., & Vehbi, T. (n.d.). *IFC -Bank Indonesia International Workshop and Seminar on “Big Data for Central Bank Policies / Building Pathways for Policy Making with Big Data” Nowcasting New Zealand GDP using machine learning algorithms 1*. https://www.bis.org/ifc/publ/ifcb50_15.pdf

[7] Jiang, Z. (2022). Prediction and Industrial Structure Analysis of Local GDP Economy Based on Machine Learning. *Mathematical Problems in Engineering*, *2022*, e7089914. https://doi.org/10.1155/2022/7089914

[8] Agu, S., Onu, F., Ezemagu, U., & Oden, D. (2022). Predicting gross domestic product to macroeconomic indicators. *Intelligent Systems with Applications*, *14*, 200082. https://doi.org/10.1016/j.iswa.2022.200082

[9] Jiang, Zhiqiang. (2022). Prediction and Industrial Structure Analysis of Local GDP Economy Based on Machine Learning. Mathematical Problems in Engineering. 2022. 1-9. 10.1155/2022/7089914.

[10] Vyas, Nikhil & Patel, Jay & Vala, Darshit & Patel, Devansh & Patel, Rohit. (2021). MACHINE LEARNING BASED GENERIC GDP ANALYSIS AND PREDICTION SYSTEM. 10.51319/2456-0774.2021.5.0012.

[11] Angelin Gladston, Arjun Sharmaa I., Bagirathan S. S. K. G., "Regression Approach for GDP Prediction Using Multiple Features From Macro-Economic Data", *International Journal of Software Science and Computational Intelligence*, vol.14, no.1, pp.1, 2022.

[12] Machine Learning Random Forest Algorithm - Javatpoint. (n.d.). Www.javatpoint.com. h

Magazzino, C., Mele, M., & Schneider, N. (2021). A machine learning approach on the relationship among solar and wind energy production, coal consumption, GDP, and CO2 emissions. *Renewable Energy*, *167*, 99-115. https://doi.org/10.1016/j.renene.2020.11.050

[13] Richardson, A., van Florenstein Mulder, T., & Vehbi, T. (2021). Nowcasting GDP using machine-learning algorithms: A real-time assessment. *International Journal of Forecasting*, *37*(2), 941-948. https://doi.org/10.1016/j.ijforecast.2020.10.005

[14] Nyman, R., & Ormerod, P. (2017). Predicting Economic Recessions Using Machine Learning Algorithms. *arXiv*. https://doi.org/10.48550/arXiv.1701.01428

[15] Perkel, J. M. (2018). Why Jupyter is data scientists' computational notebook of choice. *Nature*, *563*(7729), 145–146. https://doi.org/10.1038/d41586-018-07196-1

[16] Perkel, J. M. (2018). Why Jupyter is data scientists' computational notebook of choice. *Nature*, *563*(7729), 145–146. https://doi.org/10.1038/d41586-018-07196-1

[17] Team, G. L. (2021, May 28). *What is Numpy in Python - Everything you Need to Know About*. GreatLearning Blog: Free Resources What Matters to Shape Your Career! https://www.mygreatlearning.com/blog/python-numpy-tutorial/18

[18] What Is Pandas in Python? Everything You Need to Know. (n.d.). ActiveState. https://www.activestate.com/resources/quick-reads/what-is-pandas-in-python-everything-you-need-to-know/

[19]Data Preprocessing: Definition, Key Steps and Concepts. (n.d.). SearchDataManagement. https://www.techtarget.com/searchdatamanagement/definition/data-preprocessing

[20] Engineering Statistics Handbook. (2019). *7.1.6. What are outliers in the data?* Nist.gov. https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm

[21] What is Exploratory Data Analysis? (n.d.). Www.ibm.com. https://www.ibm.com/uk-en/cloud/learn/exploratory-data-analysis

[22] Wikipedia Contributors. (2019, August 15). *Double-precision floating-point format*. Wikipedia; Wikimedia Foundation. https://en.wikipedia.org/wiki/Double-precision_floating-point_format

[23] Installing and getting started — seaborn 0.11.1 documentation. (n.d.). Seaborn.pydata.org. https://seaborn.pydata.org/installing.html

[24] SL, S. (2020, July 2). *PAIRPLOT VISUALIZATION*. Analytics Vidhya. https://medium.com/analytics-vidhya/pairplot-visualization-16325cd725e6#:~:text=Pairplot%20visualizes%20given%20data%20to

[25] A Guide on Splitting Datasets With Train_test_split Function. (n.d.). Www.bitdegree.org. https://www.bitdegree.org/learn/train-test-split

[26] What is the accuracy_score function in Sklearn? (n.d.). Educative: Interactive Courses for Software Developers. Retrieved December 10, 2022, from https://www.educative.io/answers/what-is-the-accuracyscore-function-in-sklearn

[27] Mean Squared Error: Definition and Example. (n.d.). Statistics How To. https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/mean-squared-error/

[28] Bank of England. (2019, January 10). *What is GDP?* Bankofengland.co.uk. https://www.bankofengland.co.uk/knowledgebank/what-is-gdp

[29] Saadah, Siti & Wibowo, Muhammad. (2020). Prediction of Gross Domestic Product (GDP) in Indonesia Using Deep Learning Algorithm. 32-36. 10.1109/ISRITI51436.2020.9315519.

[30] Van Rossum, G., & Drake Jr, F. (1995). Python tutorial (vol. 620). CWI Report CS-R9526, Amsterdam, Netherlands, msekce. karlin. mff. cuni. cz/˜ halas/IT/tutorial. pdf.

[31] Rolon-Merette, Damiem & Ross, Matt & Rolon-Merette, Thaddé & Church, Kinsey. (2020). Introduction to Anaconda and Python: Installation and setup. The Quantitative Methods for Psychology. 16. S3-S11. 10.20982/tqmp.16.5.S003.

[32] Python Numpy Tutorial (with Jupyter and Colab). (n.d.). Cs231n.github.io. https://cs231n.github.io/python-numpy-tutorial/

[33] McCann, S. (2022, May 19). *Fears are growing that the UK will enter a recession after interest rates rise to 9%*. Www.nationalworld.com.

https://www.nationalworld.com/lifestyle/money/when-was-the-last-recession-uk-how-long-what-does-mean-recover-3700204

[34] United Kingdom GDP | 1960-2019 Data | 2020-2022 Forecast | Historical | Chart | News. (n.d.). Tradingeconomics.com. https://tradingeconomics.com/united-kingdom/gdp#:~:text=GDP%20in%20the%20United%20Kingdom

[35] FocusEconomics. (n.d.). *United Kingdom GDP - UK Economy Forecast & Outlook*. FocusEconomics | Economic Forecasts from the World's Leading Economists. https://www.focus-economics.com/country-indicator/united-kingdom/gdp

[36] nigeledwardsceramics. (2019, February 27). *DISSERTATION PROPOSAL PROJECT MANAGEMENT LOGWC 11th and 18th February*. Nigel Edwards Learning Journal.

[37] Why is Data Preprocessing required? Explain the different steps involved in Data Preprocessing. (n.d.). Www.ques10.com. Retrieved December 10, 2022, from https://www.ques10.com/p/9224/why-is-data-preprocessing-required-explain-the-dif/?

[38] Walstrum, T. (2017). Can the CFSBC Activity Index Nowcast U.S. Real GDP Growth? *Chicago Fed Letter*. https://econpapers.repec.org/article/fipfedhle/00067.htm

[39] Sargeant, A. (2016). *THE INFLUENCE OF WEATHER AND ICE ON FERRY OPERATIONS: MODELLING PRESENT-DAY EFFECTS TO PREDICT FUTURE TRENDS*. https://dalspace.library.dal.ca/bitstream/handle/10222/72105/Sargeant-Andrew-MASc-IENG-August-2016.pdf?isAllowed=y&sequence=3

[40] Pandey, Y. N., Rastogi, A., Kainkaryam, S., Bhattacharya, S., & Saputelli, L. (2020). *Machine Learning in the Oil and Gas Industry*. Apress. https://doi.org/10.1007/978-1-4842-6094-4

[41] Using Python Visuals in Power BI. (n.d.). AbsentData. Retrieved December 10, 2022, from https://absentdata.com/how-to-user-python-and-power-bi/

[42] A, P. R., Praneash, G. P., Rashmika, T., & Natarajan, A. (2022, August 1). Forest Fire Detection using Computer Vision. IEEE Xplore.

[43] Raut, J., Sharma, Y., & Shinde Head, V. (2020). *European Journal of Molecular & Clinical Medicine PERFORMANCE EVALUATION OF VARIOUS SUPERVISED MACHINE LEARNING ALGORITHMS FOR DIABETES PREDICTION*. https://ejmcm.com/article_7221_98356dd221c7772117e1690b6544f29a.pdf

[44] Larsson, M., Megyesi, B., Schulman, R., & Johansson, M. (2006). *Development of a Speech-Driven Automatic Telephone Service Retrieving pronunciation and spelling of names*. https://cl.lingfil.uu.se/exarb/arch/2006_larsson2.pdf

[45]Phishing Detection Using Machine Learning - ProQuest. (n.d.). Www.proquest.com. Retrieved December 10, 2022, from https://www.proquest.com/docview/2622582696

[46] Magazzino, C., Mele, M., & Schneider, N. (2020). *A Machine Learning Approach on the Relationship Among Solar and Wind Energy Production, Coal Consumption, GDP, and CO2 Emissions*. Papers.ssrn.com. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3745072

[47] Correcting Code conventions (47c8420c) · Commits · Earth Sciences / hermesv3_gr · GitLab. (n.d.). GitLab. Retrieved December 10, 2022, from https://earth.bsc.es/gitlab/es/hermesv3_gr/-/commit/47c8420c83c07f0443d6300466402a28b9f9fd4a?view=parallel

[48] Wikipedia Contributors. (2019, April 9). Economy of the United Kingdom. Wikipedia; Wikimedia Foundation. https://en.wikipedia.org/wiki/Economy_of_the_United_Kingdom

[49] Kurian, B., & Liyanapathirana, R. (2019). Machine Learning Techniques for Structural Health Monitoring. *Lecture Notes in Mechanical Engineering*, 3–24. https://doi.org/10.1007/978-981-13-8331-1_1

[50] How to split the Dataset With scikit-learn's train_test_split() Function. (2022, June 24). GeeksforGeeks. https://origin.geeksforgeeks.org/how-to-split-the-dataset-with-scikit-learns-train_test_split-function/

[51] Verdhan, V. (2020). *Supervised Learning with Python*. Apress. https://doi.org/10.1007/978-1-4842-6156-9

[52] Ghag, Y., Upadhyay, D., Gadkari, S., Pandit, S., Rathod, S., & Student. (2019). *CSEIT19485 | A Survey on Smart Mirror*. *4*, 2456–3307. https://doi.org/10.32628/IJSRCSEIT

[53]United Kingdom GDP - 2022 Data - 2023 Forecast - 1960-2021 Historical - Chart - News. (n.d.). Tradingeconomics.com. Retrieved December 10, 2022, from https://cdn.tradingeconomics.com/united-kingdom/gdp

[54] python. (n.d.). Educational Research Techniques. Retrieved December 10, 2022, from https://educationalresearchtechniques.com/tag/python/

[55] (2022). Coursehero.com. https://www.coursehero.com/file/1944050/prc/

[56] Agu, S. C., Onu, F. U., Ezemagu, U. K., & Oden, D. (2022). Predicting gross domestic product to macroeconomic indicators. *Intelligent Systems with Applications*, *14*, 200082. https://doaj.org/article/494ea9086c394d23a418adcc28868976

[56] Vyas, N., Patel, J., Vala, D., Patel, D., & Patel, R. (n.d.). *MACHINE LEARNING BASED GENERIC GDP ANALYSIS AND PREDICTION SYSTEM Student*. https://doi.org/10.51319/2456-0774.2021.5.0012

[57] janeiro 2017. (n.d.). Data Mining / Machine Learning / Data Analysis. Retrieved December 10, 2022, from https://mineracaodedados.wordpress.com/2017/01/

[58] Muchisha, D., Tamara, N., Dwi Muchisha, N., Andriansyah, Agus, M., & Soleh. (2020). *Munich Personal RePEc Archive Nowcasting Indonesia's GDP Growth Using Machine Learning Algorithms Nowcasting Indonesia's GDP Growth Using Machine Learning Algorithms*. https://mpra.ub.uni-muenchen.de/105235/1/MPRA_paper_105235.pdf

[59] Nyman, R., & Ormerod, P. (2020). *Understanding the Great Recession Using Machine Learning Algorithms*. http://export.arxiv.org/pdf/2001.02115

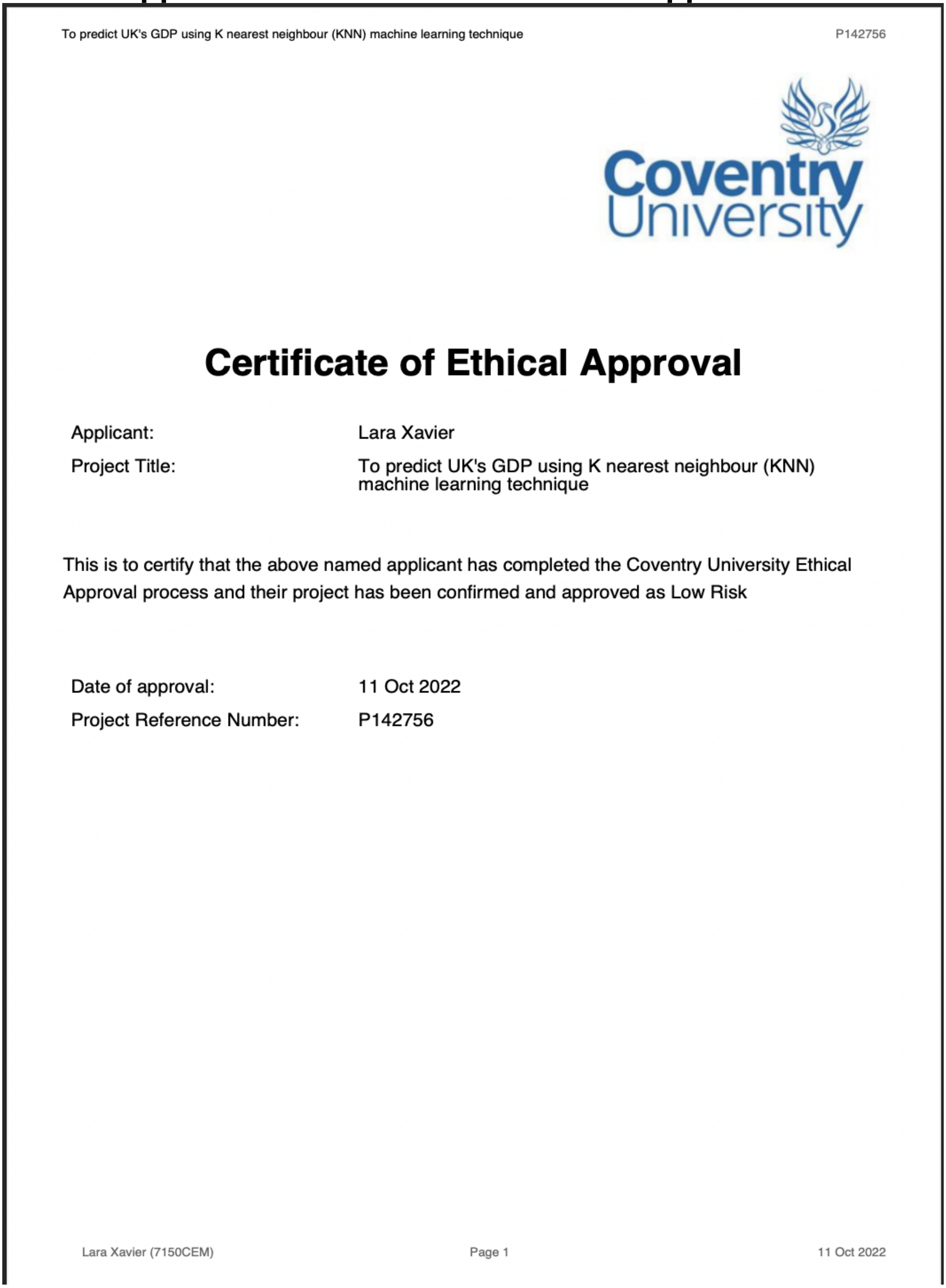
10. Appendix A – Project Specification

```
import numpy as np
import pandas as pd
import pandas as pd

df=pd.read_csv('ukdata.csv')
df1 = df[['Year','GDP']]
df1.dropna()
df1.drop(df1.index[[61,66]], inplace=True)

df1.drop(df1.index[[63,64]], inplace=True)

df1.drop(df1.index[[61,62]], inplace=True)

print(df1[df1['GDP'].isnull()])
df1
df1.shape
#Finding null values in the dataset
df1.isnull().sum()
df.dtypes
import matplotlib.pyplot as plt
import seaborn as sns
sns.pairplot(df1)
#Finding the distribution of data

def distplots(col):
    sns.distplot(df[col])
    plt.show()

for i in list(df1.columns)[1:]:
     distplots(i)
#Finding the outliers of data

def boxplots(col):
    sns.boxplot(df[col])
    plt.show()

for i in list(df1.select_dtypes(exclude=['object']).columns)[1:]:
    boxplots(i)
plt.figure(figsize=(20,20))
corr=df1.corr()
sns.heatmap(corr,annot=True)

import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsRegressor
from sklearn.metrics import mean_squared_error
from sklearn.metrics import accuracy_score
import pandas as pd
```

```
df=pd.read_csv('ukdata.csv')
df1 = df[['Year','GDP']]
df1.dropna()
df1.drop(df1.index[[61,66]], inplace=True)

df1.drop(df1.index[[63,64]], inplace=True)

df1.drop(df1.index[[61,62]], inplace=True)

print(df1[df1['GDP'].isnull()])
df1.corr()
X=df1[['Year']]
Y=df1[['GDP']]
X_train ,X_test ,Y_train ,Y_test=train_test_split(X,Y,test_size=0.2,random_state=0)
clf = KNeighborsRegressor(2)
clf.fit(X_train,Y_train)
y_pred = clf.predict(X_test)
print(y_pred)
print((mean_squared_error(Y_test,y_pred)))
#Plotting the observed and predicted data
import matplotlib.pyplot as plt

x_ax =range(len(X_test))
plt.plot(x_ax,Y_test,label='Observed',color='k',linestyle = '-')
plt.plot(x_ax,y_pred,label='Prediction',color='k',linestyle = '--')
# Support Vector Regression
from sklearn.svm import SVR
#Creating and fitting SVR model
ll_svm = SVR().fit(X_train,Y_train)
print(ll_svm.score(X_train,Y_train))
ytrain_pred = ll_svm.predict(X_train)
print(ytrain_pred)
print(mean_squared_error(Y_train,ytrain_pred))
#Prediction on the testing dataset
ytest_pred=ll_svm.predict(X_test)
print(ytest_pred)
print(mean_squared_error(Y_test,ytest_pred))
#Plotting the observed and predicted data
import matplotlib.pyplot as plt

x_ax =range(len(X_test))
plt.plot(x_ax,Y_test,label='Observed',color='k',linestyle = '-')
plt.plot(x_ax,ytest_pred,label='Prediction',color='k',linestyle = '--')
# Import the model we are using
from sklearn.ensemble import RandomForestRegressor
# Instantiate model with 1000 decision trees
rf = RandomForestRegressor(n_estimators = 1000, random_state = 42)
# Train the model on training data
rf.fit(X_train, Y_train);
# Use the forest's predict method on the test data
predictions = rf.predict(X_test)
```

```
print(predictions)
print(mean_squared_error(Y_test,predictions))
#Plotting the observed and predicted data
import matplotlib.pyplot as plt

x_ax =range(len(X_test))
plt.plot(x_ax,Y_test,label='Observed',color='k',linestyle = '-')
plt.plot(x_ax,predictions,label='Prediction',color='k',linestyle = '--')
```

11. Appendix B – Certificate of Ethics Approval



To predict UK's GDP using K nearest neighbour (KNN) machine learning technique P142756

Certificate of Ethical Approval

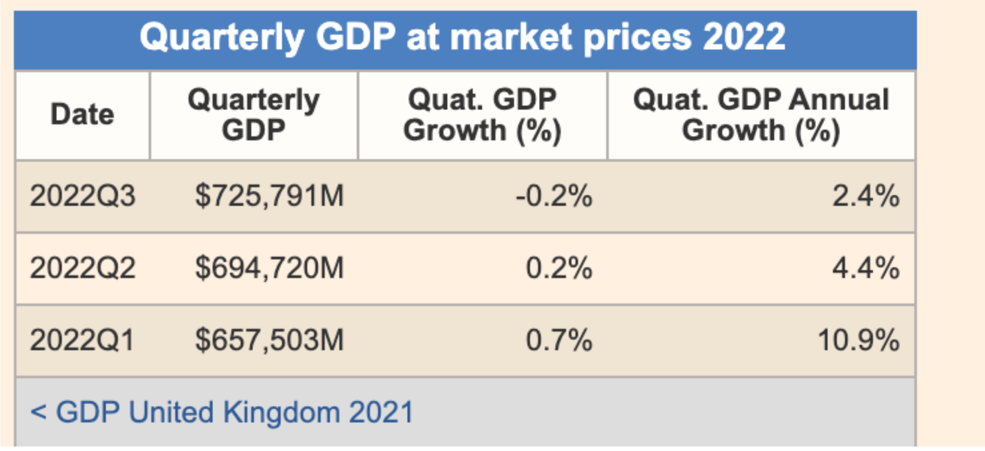
Applicant:	Lara Xavier
Project Title:	To predict UK's GDP using K nearest neighbour (KNN) machine learning technique

This is to certify that the above named applicant has completed the Coventry University Ethical Approval process and their project has been confirmed and approved as Low Risk

Date of approval:	11 Oct 2022
Project Reference Number:	P142756

Lara Xavier (7150CEM) Page 1 11 Oct 2022

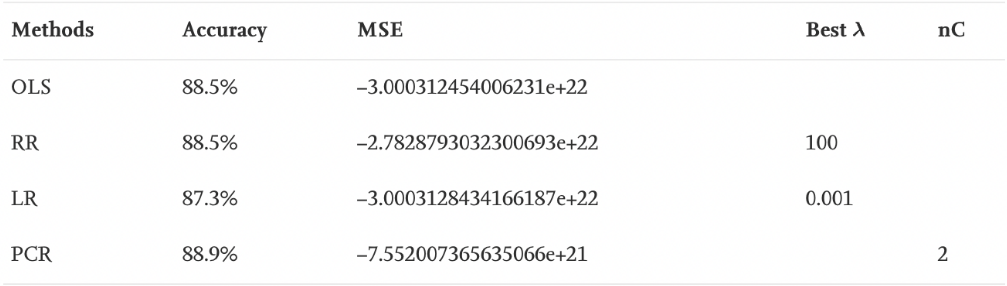
12. Appendix-C Images



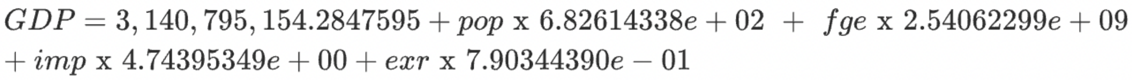
Quarterly GDP at market prices 2022

Date	Quarterly GDP	Quat. GDP Growth (%)	Quat. GDP Annual Growth (%)
2022Q3	$725,791M	-0.2%	2.4%
2022Q2	$694,720M	0.2%	4.4%
2022Q1	$657,503M	0.7%	10.9%

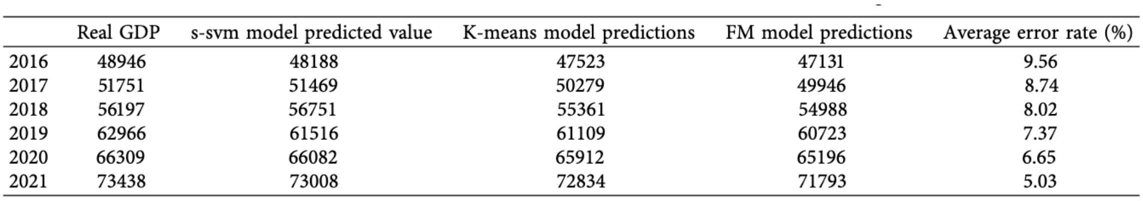
< GDP United Kingdom 2021



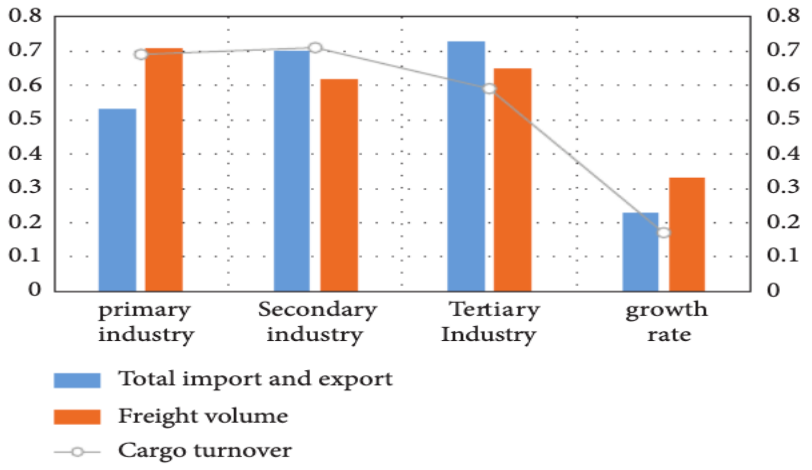
Methods	Accuracy	MSE	Best λ	nC
OLS	88.5%	−3.000312454006231e+22		
RR	88.5%	−2.7828793032300693e+22	100	
LR	87.3%	−3.0003128434166187e+22	0.001	
PCR	88.9%	−7.552007365635066e+21		2

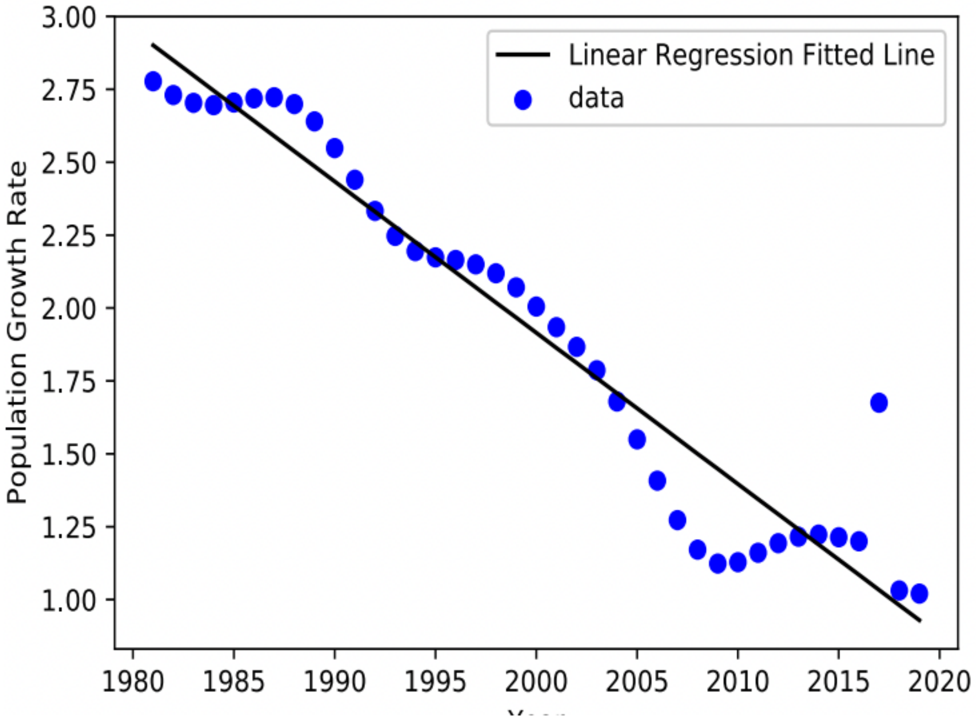


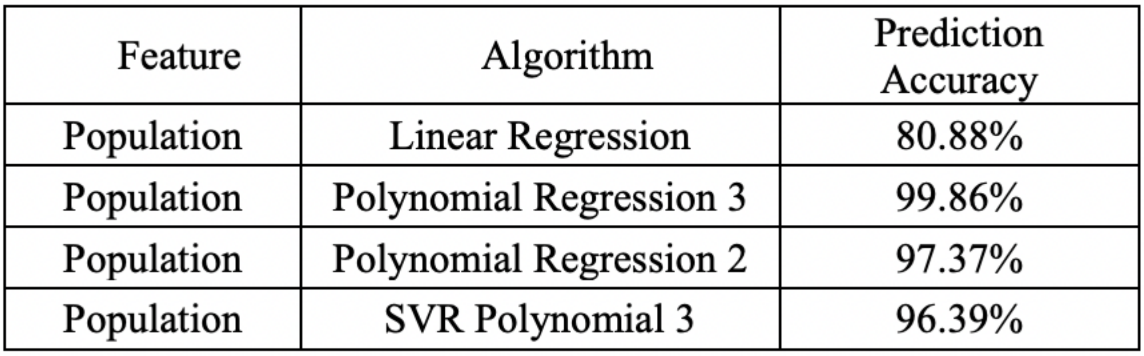
$$GDP = 3,140,795,154.2847595 + pop \text{ x } 6.82614338e+02 \; + \; fge \text{ x } 2.54062299e+09 + imp \text{ x } 4.74395349e+00 + exr \text{ x } 7.90344390e-01$$



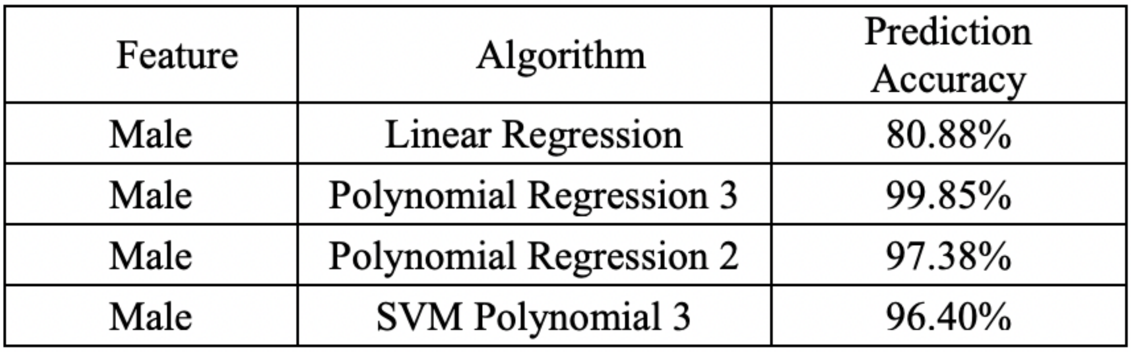
	Real GDP	s-svm model predicted value	K-means model predictions	FM model predictions	Average error rate (%)
2016	48946	48188	47523	47131	9.56
2017	51751	51469	50279	49946	8.74
2018	56197	56751	55361	54988	8.02
2019	62966	61516	61109	60723	7.37
2020	66309	66082	65912	65196	6.65
2021	73438	73008	72834	71793	5.03



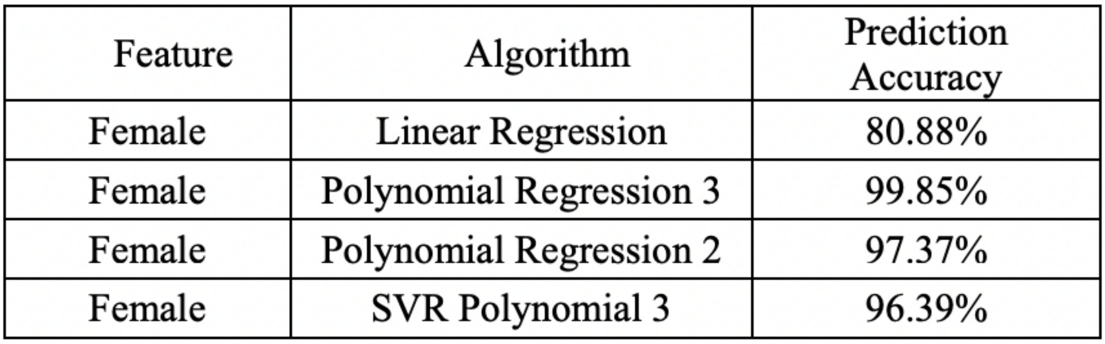




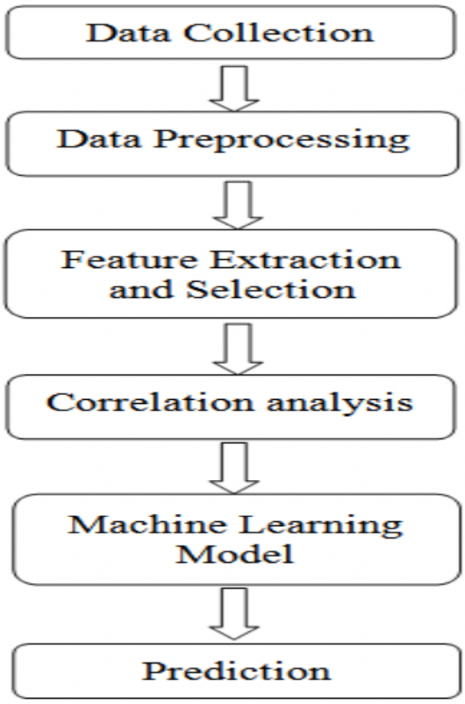
Feature	Algorithm	Prediction Accuracy
Population	Linear Regression	80.88%
Population	Polynomial Regression 3	99.86%
Population	Polynomial Regression 2	97.37%
Population	SVR Polynomial 3	96.39%



Feature	Algorithm	Prediction Accuracy
Male	Linear Regression	80.88%
Male	Polynomial Regression 3	99.85%
Male	Polynomial Regression 2	97.38%
Male	SVM Polynomial 3	96.40%



Feature	Algorithm	Prediction Accuracy
Female	Linear Regression	80.88%
Female	Polynomial Regression 3	99.85%
Female	Polynomial Regression 2	97.37%
Female	SVR Polynomial 3	96.39%



Data Collection

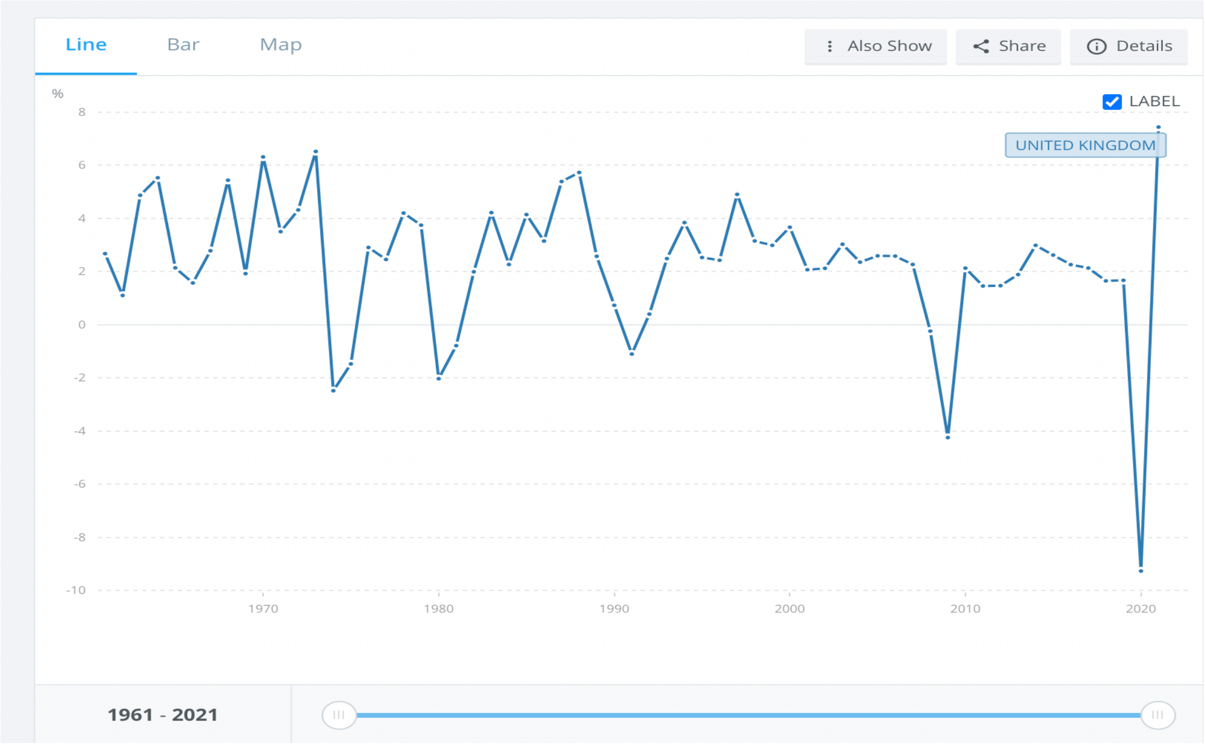
Data Preprocessing

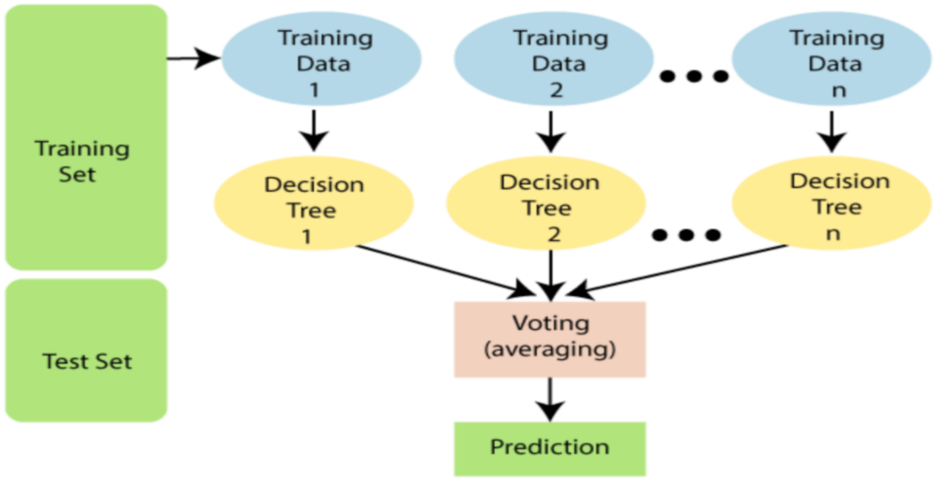
Feature Extraction and Selection

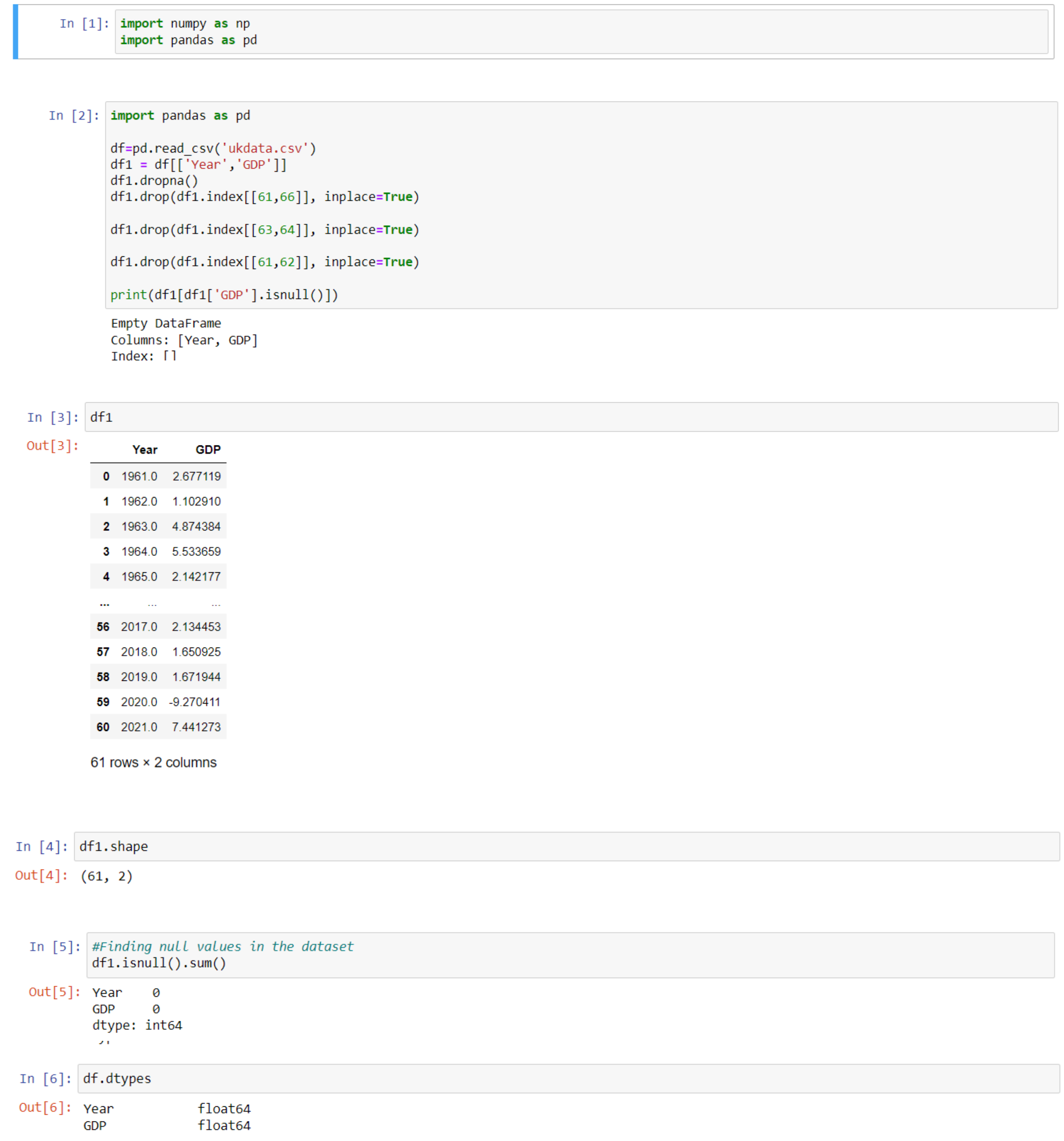
Correlation analysis

Machine Learning Model

Prediction







```
In [1]: import numpy as np
        import pandas as pd
```

```
In [2]: import pandas as pd

        df=pd.read_csv('ukdata.csv')
        df1 = df[['Year','GDP']]
        df1.dropna()
        df1.drop(df1.index[[61,66]], inplace=True)

        df1.drop(df1.index[[63,64]], inplace=True)

        df1.drop(df1.index[[61,62]], inplace=True)

        print(df1[df1['GDP'].isnull()])

        Empty DataFrame
        Columns: [Year, GDP]
        Index: []
```

```
In [3]: df1
```

Out[3]:

	Year	GDP
0	1961.0	2.677119
1	1962.0	1.102910
2	1963.0	4.874384
3	1964.0	5.533659
4	1965.0	2.142177
...	...	...
56	2017.0	2.134453
57	2018.0	1.650925
58	2019.0	1.671944
59	2020.0	-9.270411
60	2021.0	7.441273

61 rows × 2 columns

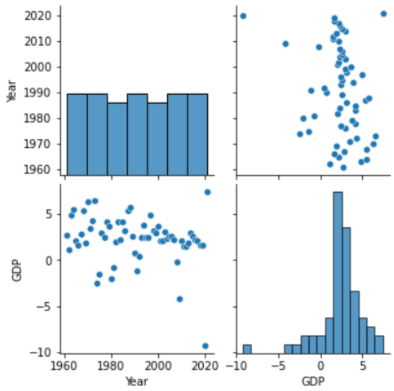
```
In [4]: df1.shape

Out[4]: (61, 2)
```

```
In [5]: #Finding null values in the dataset
        df1.isnull().sum()

Out[5]: Year    0
        GDP     0
        dtype: int64
```

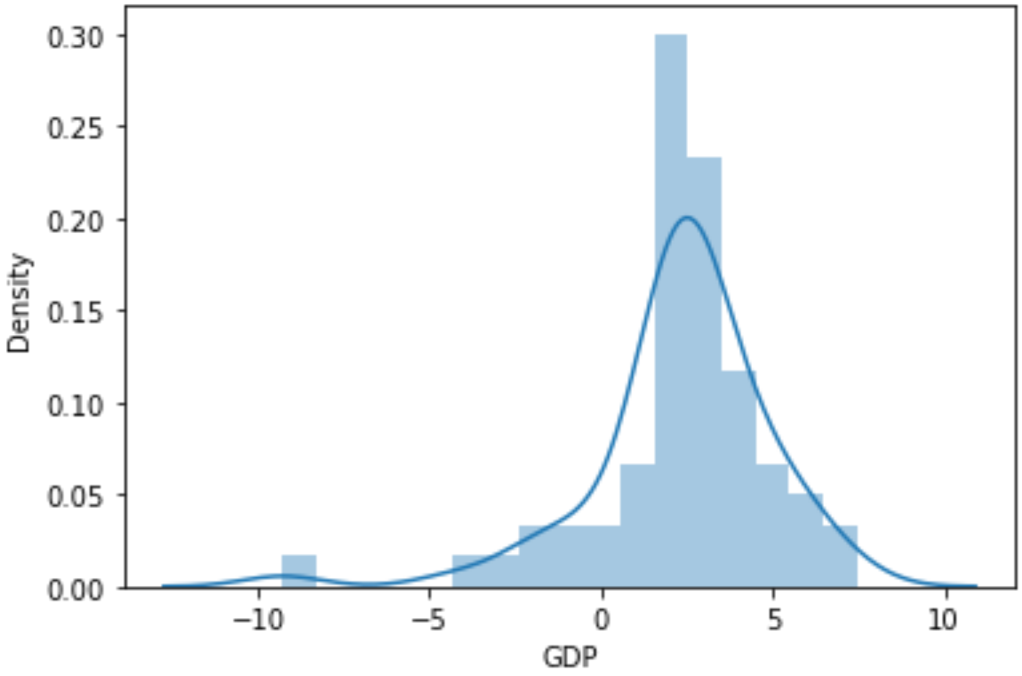
```
In [6]: df.dtypes

Out[6]: Year            float64
        GDP             float64
```

```
In [8]: import matplotlib.pyplot as plt
        import seaborn as sns
```

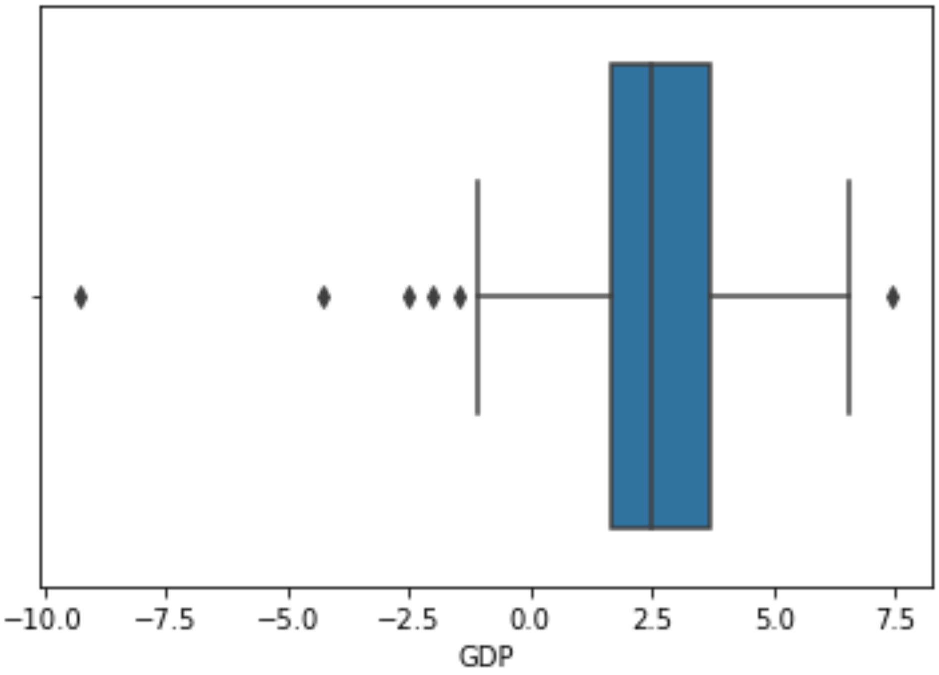
```
In [9]: sns.pairplot(df1)
Out[9]: <seaborn.axisgrid.PairGrid at 0x202d26214f0>
```

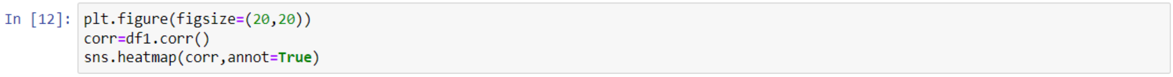


```
In [10]: #Finding the distribution of data

         def distplots(col):
             sns.distplot(df[col])
             plt.show()

         for i in list(df1.columns)[1:]:
              distplots(i)
```

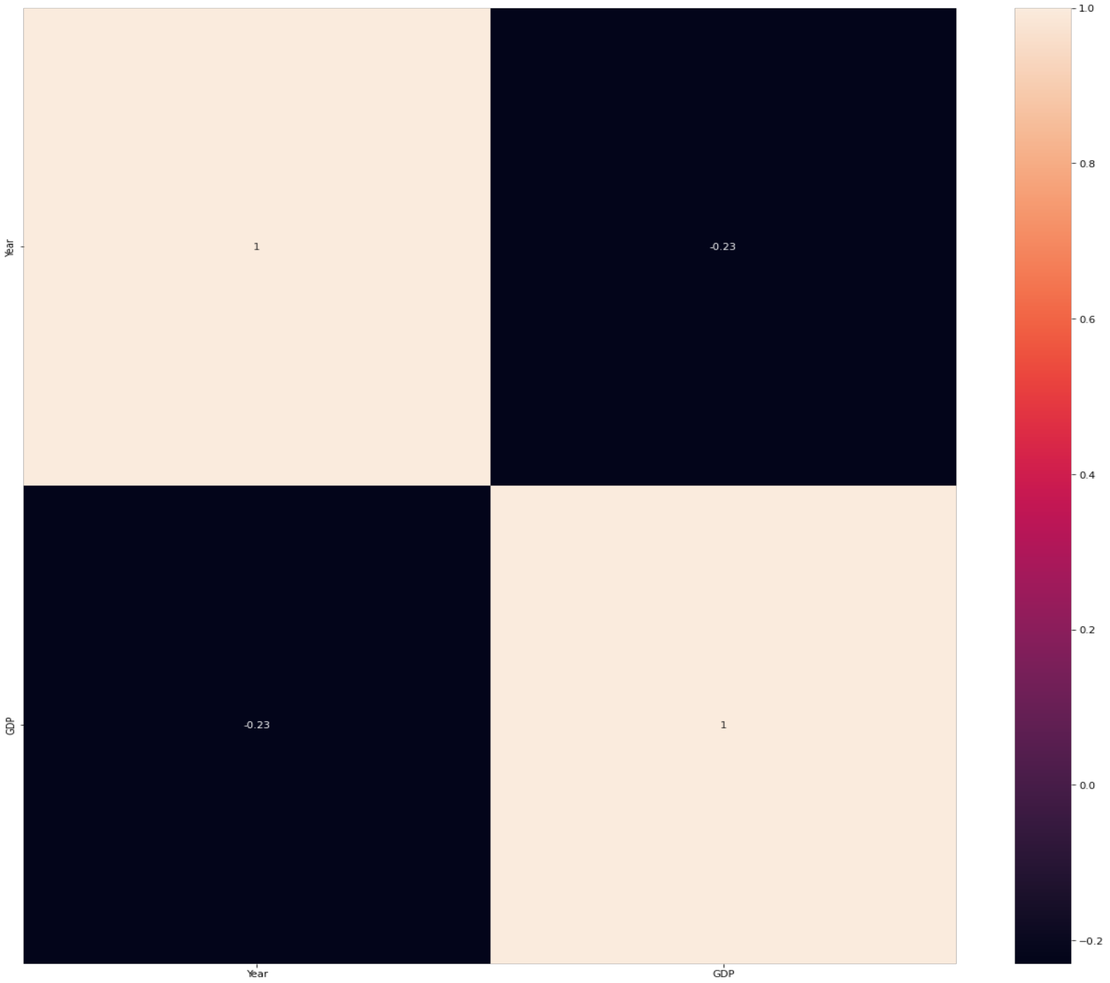


```
In [11]: #Finding the outliers of data

         def boxplots(col):
             sns.boxplot(df[col])
             plt.show()

         for i in list(df1.select_dtypes(exclude=['object']).columns)[1:]:
             boxplots(i)
```





```
In [12]: plt.figure(figsize=(20,20))
         corr=df1.corr()
         sns.heatmap(corr,annot=True)
```



```
In [29]: import pandas as pd
         import numpy as np
         from sklearn.model_selection import train_test_split
         from sklearn.neighbors import KNeighborsRegressor
         from sklearn.metrics import mean_squared_error
         from sklearn.metrics import accuracy_score
```

```
In [31]: df1.corr()
```

Out[31]:

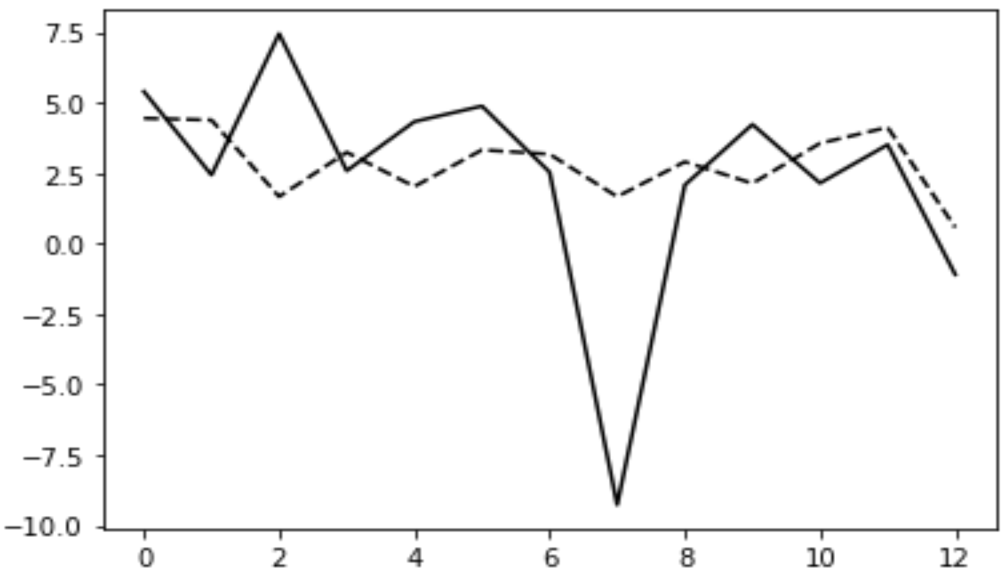
	Year	GDP
Year	1.000000	-0.230051
GDP	-0.230051	1.000000

```
In [4]: X=df1[['Year']]
        Y=df1[['GDP']]
        X_train ,X_test ,Y_train ,Y_test=train_test_split(X,Y,test_size=0.2,random_state=0)
```

```
In [5]: clf = KNeighborsRegressor(2)
        clf.fit(X_train,Y_train)

Out[5]: KNeighborsRegressor(n_neighbors=2)
```

```
In [6]: y_pred = clf.predict(X_test)
        print(y_pred)
        print((mean_squared_error(Y_test,y_pred)))

        [[4.44137713]
         [4.37751745]
         [1.66143485]
         [3.23308462]
         [2.01972222]
         [3.31828452]
         [3.16792008]
         [1.66143485]
         [2.89779594]
         [2.13199825]
         [3.5533799 ]
         [4.12100175]
         [0.5674188 ]]
        13.567739082467769
```

```
In [7]: #Plotting the observed and predicted data
        import matplotlib.pyplot as plt

        x_ax =range(len(X_test))
        plt.plot(x_ax,Y_test,label='Observed',color='k',linestyle = '-')
        plt.plot(x_ax,y_pred,label='Prediction',color='k',linestyle = '--')
```



```
In [8]: # Support Vector Regression
        from sklearn.svm import SVR
```

```
In [9]: #Creating and fitting SVR model
        ll_svm = SVR().fit(X_train,Y_train)
```

```
In [11]: ytrain_pred = ll_svm.predict(X_train)
```

```
In [12]: print(ytrain_pred)
```

```
[2.71973348 2.98026727 2.5749445  2.13761059 2.68023202 2.57877407
 2.96795745 2.03118545 2.8850891  2.64942522 1.91585925 2.55525456
 1.90604142 2.80628717 2.56662299 2.68915893 2.94711803 2.5580821
 1.98981235 2.57534805 1.91132861 2.67050794 2.59353685 2.72393606
 2.92885721 2.5540456  2.42053371 2.20127908 1.93222502 2.88990545
 2.6620186  2.60985904 2.34415122 2.68533865 2.68655507 2.65024657
 2.95753125 2.08063778 2.5945853  2.56225259 2.65787705 2.8413573
 1.90731955 2.68281471 2.64706131 1.95681211 2.2705633  2.49805035]
```

```
In [13]: print(mean_squared_error(Y_train,ytrain_pred))
```

```
4.305635612687594
```

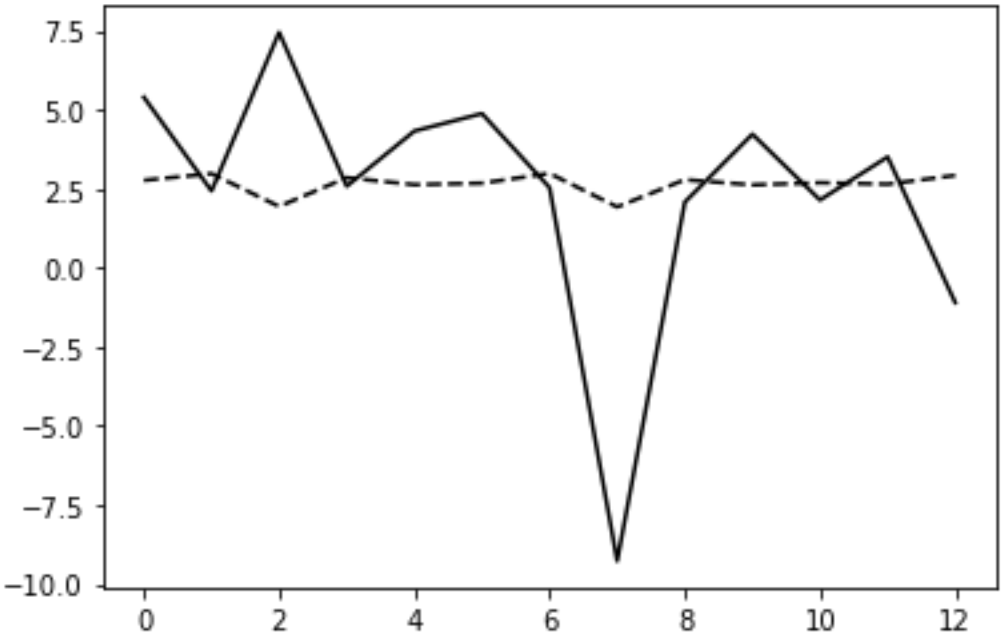
```
In [14]: #Prediction on the testing dataset
         ytest_pred=ll_svm.predict(X_test)
```

```
In [15]: print(ytest_pred)
```

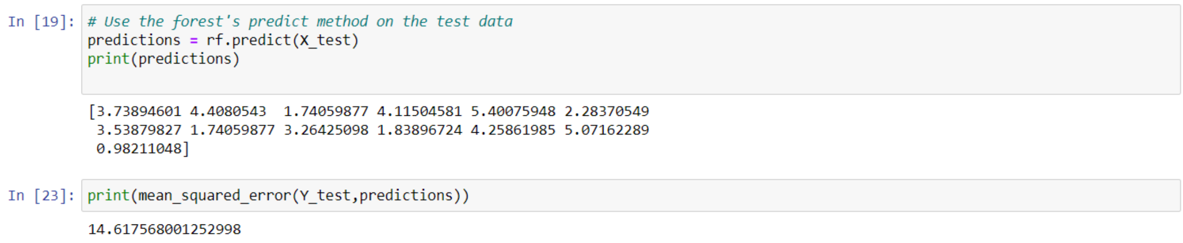
```
[2.76474455 2.97558854 1.9383859  2.84695893 2.62668444 2.67408005
 2.98304473 1.92239172 2.78420939 2.61972666 2.68790252 2.64299987
 2.91900838]
```

```
In [22]: print(mean_squared_error(Y_test,ytest_pred))
```

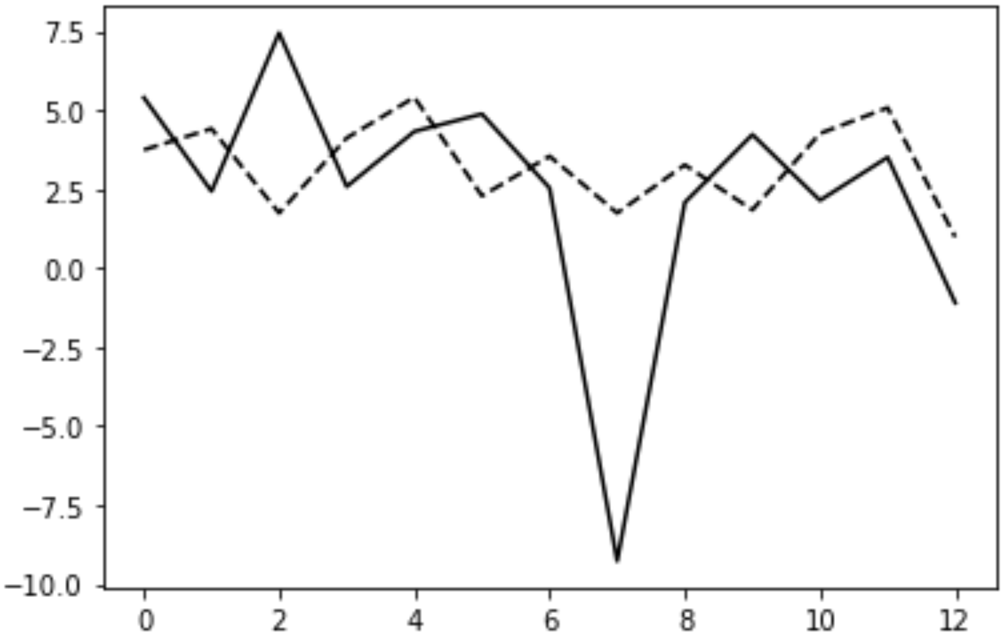
```
14.695987322792138
```

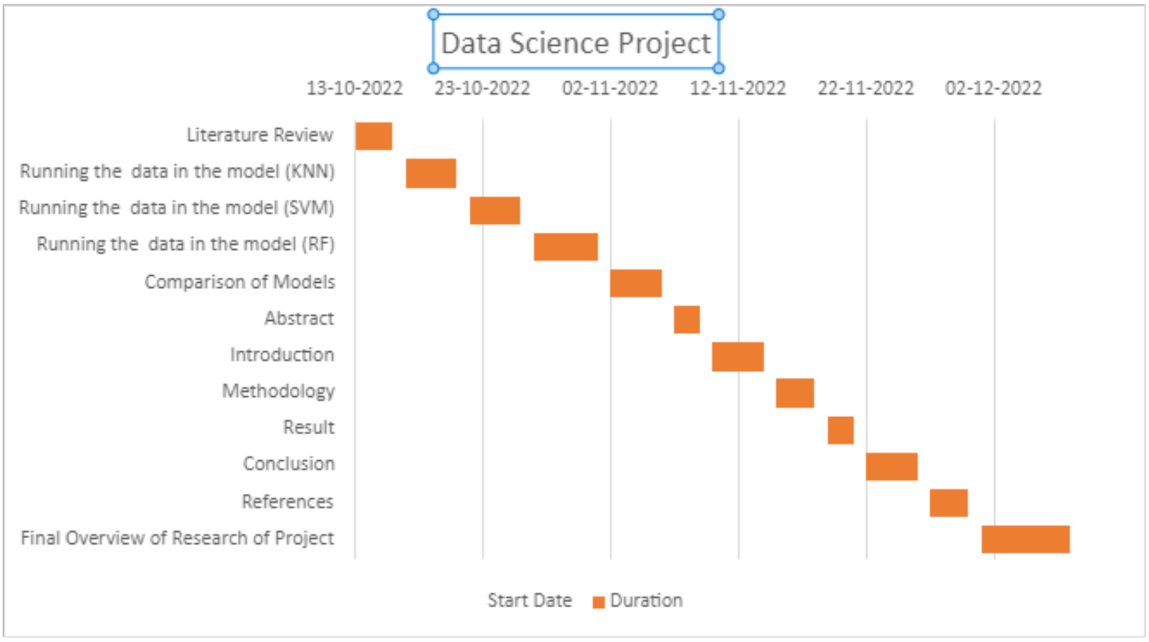


```
In [18]: # Import the model we are using
         from sklearn.ensemble import RandomForestRegressor
         # Instantiate model with 1000 decision trees
         rf = RandomForestRegressor(n_estimators = 1000, random_state = 42)
         # Train the model on training data
         rf.fit(X_train, Y_train);
```



```
In [19]: # Use the forest's predict method on the test data
         predictions = rf.predict(X_test)
         print(predictions)

         [3.73894601 4.4080543  1.74059877 4.11504581 5.40075948 2.28370549
          3.53879827 1.74059877 3.26425098 1.83896724 4.25861985 5.07162289
          0.98211048]

In [23]: print(mean_squared_error(Y_test,predictions))

         14.617568001252998
```


7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
0
2
4
6
8
10
12


Data Science Project
Literature Review
Running the data in the model (KNN)
Running the data in the model (SVM)
Running the data in the model (RF)
Comparison of Models
Abstract
Introduction
Methodology
Result
Conclusion
References
Final Overview of Research of Project
Start Date
Duration