GALGOTIAS UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

CROP YIELD PREDICTION USING MACHINE LEARNING

A Report for the Evaluation 3 of Project 2

Submitted by

APRAJITA SINGH
16SCSE101508
(1613101171)

In partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

Under the Supervision of

Mr. C Ramesh Kumar
Asst. Professor
APRIL / MAY- 2020

GALGOTIAS
UNIVERSITY
(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

SCHOOL OF COMPUTING AND SCIENCE AND ENGINEERING

Greater Noida, Uttar Pradesh
Winter 2019-2020

BONAFIDE CERTIFICATE

Certified that this project report

"CROP YIELD PREDICYION USING MACHINE LEARNING "

is the bonafide work of **" APRAJITA SINGH(1613101171) "** who carried out the project work under my supervision.

SIGNATURE OF HEAD **SIGNATURE OF SUPERVISOR**

Dr. Munish Sabharwal,
Ph.D.

Professor & Dean
School of Computing Science
& Engineering

Mr. C Ramesh Kumar ,
M.Tech

Assistant Professor
School of Computing Science
& Engineering

TABLE OF CONTENTS

S.No.	Contents	Page No.
1.	Abstract	4
2.	Introduction	5-6
3.	Overall description	7
4.	Purpose of crop yield and Advantages and disadvantages	8
5.	Motivational Scope	9
6.	Related Work	10-13
7.	Other Algorithm	14-15
8.	Proposed Statement	16-17
9.	Proposed Model	18-19
10.	Implementation	20
11.	Testing and Evaluating	21-22
12.	Result and Conclusion	22-27

13.	Model Comparison Based on RMSE Value	28-29
14.	References	30

1. ABSTRACT

Data Mining is emerging research field in crop yield analysis. Yield prediction is a very important issue in agricultural. Any farmer is interested in knowing how much yield he is about to expect. In the past, yield prediction was performed by considering farmer's experience on particular field and crop. The yield prediction is a major issue that remains to be solved based on available data. Data mining techniques are the better choice for this purpose. Different Data Mining techniques are used and evaluated in agriculture for estimating the future year's crop production. This research proposes and implements a system to predict crop yield from previous data. This is achieved by applying association rule mining on agriculture data. This research focuses on creation of a prediction model which may be used to future prediction of crop yield.

Crop yield prediction Crop yields are critically dependent on weather. A growing empirical literature models this relationship in order to project climate change impacts on the sector. We describe an approach to predict yield that uses various regression techniques such as Multiple Linear Regression, Decision Tree Regression, Support Vector Regression and Random Forest Regression for predicting the influence of climatic parameters on the crop yields. In the present study we have collected data related to Maize production of Karnataka from various government sites. This study also presents a comparative study of these regression techniques in predicting maize yield. In this research a sample of environmental factors like annual rainfall, area under cover, temperature will be considered for a period of 10 years. This research can be extended by considering other factors like Minimum Support Price (MSP), Cost Price Index (CPI), Wholesale Price Index (WPI) etc. and their relationship with crop yield.

2. INTRODUCTION

Yield prediction benefits the farmers in reducing their losses and to get best prices for their Crops. The objective of this work is to analyze the Environmental parameters like Area under Cultivation (AUC), Annual Rainfall (AR) and Temperature that influences the yield of crop and to establish a relationship among these parameters. With the impact of climate change in India, Majority of the agricultural crops are being badly affected. Predicting the crop yield well ahead of its harvest would help the policy makers and farmers for taking appropriate measures for marketing and storage.

Such predictions will also help the associated industries for planning the logistics of their business. In the present study we have collected data related to crop production of Karnataka from various government sites and analyzed various regression techniques such as Multiple Linear Regression, Decision Tree Regression, Support Vector Regression, and Random Forest Regression for predicting the influence of climatic parameters on the crop yields. These techniques can provide a insight like which climatic parameter is more important and Significant on the crop yields of selected crops in selected districts of Karnataka state and will also be helpful in predicting the crop yields. In this research, Regression Analysis (RA) is used to analyze the environmental factors and their infliction on crop yield. RA is a multivariate analysis technique which analyzes the factors groups them into explanatory and response variables and helps to obtain a decision. A sample of environmental factors like AR, AUC, and FPI will be considered for a period of 10 years. Linear Regression (LR) is used to establish relationship between explanatory variables (AR, AUC, and FPI) and the crop yield as response variable. R2 value clearly shows that yield is mainly dependent on AR. AUC and FPI are the other two factors influencing the crop yield. This research can be extended by considering other factors like Minimum Support Price (MSP), Cost Price Index (CPI), Wholesale Price Index (WPI) etc. and their relationship Index (WPI) etc. and their relationship with crop yield.

2.1 Overall Description

Applications in Agriculture:

Crop yield prediction is an important agricultural problem. Each and Every farmer is always tries to know, how much yield will get from his expectation. In the past, yield prediction was calculated by analyzing farmer's previous experience on a particular crop. The Agricultural yield is primarily depends on weather conditions, pests and planning of harvest operation. Accurate information about history of crop yield is an important thing for making decisions related to agricultural risk management. This research focuses on evolution of a prediction model which may be used to predict crop yield production. The proposed method use data mining technique to predict the crop yield production based on the association rules. There are several applications in the field of agriculture. Some of them are listed below.

2.1.1 Crop Selection and Crop Yield Prediction

To maximize the crop yield, selection of the appropriate crop that will be sown plays a vital role. It depends on various factors like the type of soil and its composition, climate, geography of the region, crop yield, market prices etc. Techniques like Artificial neural networks, K-nearest neighbors and Decision Trees have carved a niche for themselves in the context of crop selection which is based on various factors. Crop selection based on the effect of natural calamities like famines has been done based on machine learning (Washington Okori, 2011). The use of artificial neural networks to choose the crops based on soil and climate has been shown by researchers (Obua, 2011). A plant nutrient management system has been proposed based on machine learning methods to meet the needs of soil, maintain its fertility levels, and hence improve the crop yield (Shivnath Ghosh, 2014). A crop selection method called CSM has been proposed which helps in crop selection based on its yield prediction and other factors (Kumar, 2009).

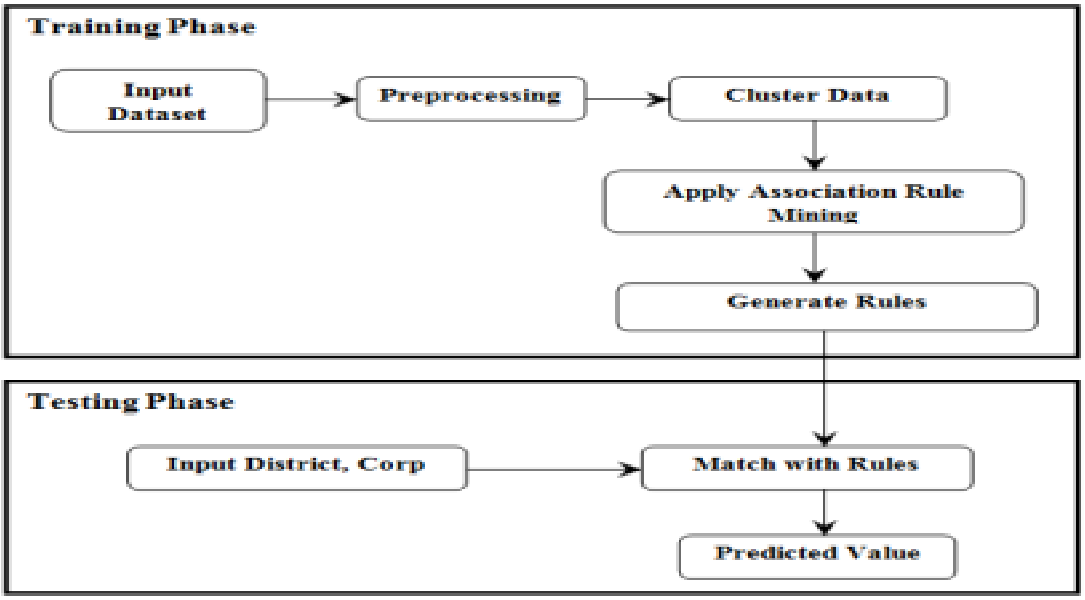


Figure 1.1 Crop Yield Using ML

2.1.2 Weather Forecasting

Indian agriculture mainly relies on seasonal rains for irrigation. Therefore, an accurate forecast of weather can reduce the enormous toil faced by farmers in India including crop selection, watering and harvesting. As the farmers have poor access to the Internet as a result of digital-divide, they have to rely on the little information available regarding weather reports. Up-to-date as well as accurate weather information is still not available as the weather changes dynamically over time. Researchers have been working on improving the accuracy of weather predictions by using a variety of algorithms. Artificial Neural networks have been adopted extensively for this purpose. Likewise, weather prediction based on machine learning technique called Support Vector Machines had been proposed (M.Shashi, 2009). These algorithms have shown better results over the conventional algorithms.

IMPORTANCE OF WEATHER FORECASTING IN
MODERN AGRICULTURE AND FARMING

Figure 1.4 Weather Forecasting

2.1.3 Smart Irrigation System

Farming sector consumes a huge portion of water in India. The levels of ground water are dropping down day-by-day and global warming has resulted in climate changes. The river water for irrigation is a big issue of dispute among many states in India. To combat the scarcity of water, many companies have come up with sensor based technology for smart farming which uses sensors to monitor the water level, nutrient content, weather forecast reports and soil temperature. EDYN Garden sensor is another example (Gupta, 2016). However, the high cost of such devices deters the small land owners and farmers in India to use them. These smart devices are being designed on the principles of machine learning. The nutrient content of soil can also be recorded using the sensors and hence used for supplying fertilizers to the soil using smart irrigation systems. This will also reduce the labor cost in the fields, which is a huge crisis being faced by the Indian farmers these days.

2.2 Purpose of Crop Yield Prediction

Data Mining is widely applied to agricultural issues. Data Mining is used to analyze large data sets and establish useful classifications and patters in the data sets. The overall goal of the Data Mining process is to extract the information from a data set and transform it into understandable structure for further use. This analyzes the crop yield production based on available data. The Data mining technique was used to predict the crop yield for maximizing the crop productivity.

Advantages:

- Cost effective method
- Optimize water use
- Sustain high-yielding
- High quality crop production
- Utilization of Resources Efficiently
- Increase Data Collection

Disadvantages:

1. Security
2. Privacy
3. Complexity

Figure 1.1 Optimized Crop Yield

2.3 Motivations and Scope

Motivations:

Different kind of problems faced by the farmers motivated me for the recommended system that is: the Indian farming is on the hitch because of the limited technical know-how of the best and efficient agricultural practices and moreover they are still dependent on conventional methods of agriculture that leads to lesser productivity of crops. So by using upcoming technology the productivity of crops can be maximized at minimal cost. This also reduces burden of taking up of heavy loans on farmers which they have incurred on themselves in order to sustain their livings or to get good yields of their crops. Apart from these issues scarcity of resources also adds up in their problem causing hindrance or stopping framers from cultivating and hence Indian economy is also additionally getting influenced to large extent as most of the fruitful lands of the nation are being destroyed that forms the vital part of GDP.
So through this framework I am presenting solution for this issue by introducing an automated and systematic farming strategies that enable farmers to cultivate in a productive way also with limited resources and greater yield which is assured and efficient.

Scope:

Update farmers with the new technology and to avoid manual labor.

- Reduce wastage of water and enhance productivity of crops by providing them ideal Condition.

- Meet the difficulties such as severe weather conditions and advancing climate and change environmental consequences resulting from intensive farming practices.

Figure 1.2 Optimized Predictors

3. Literature survey

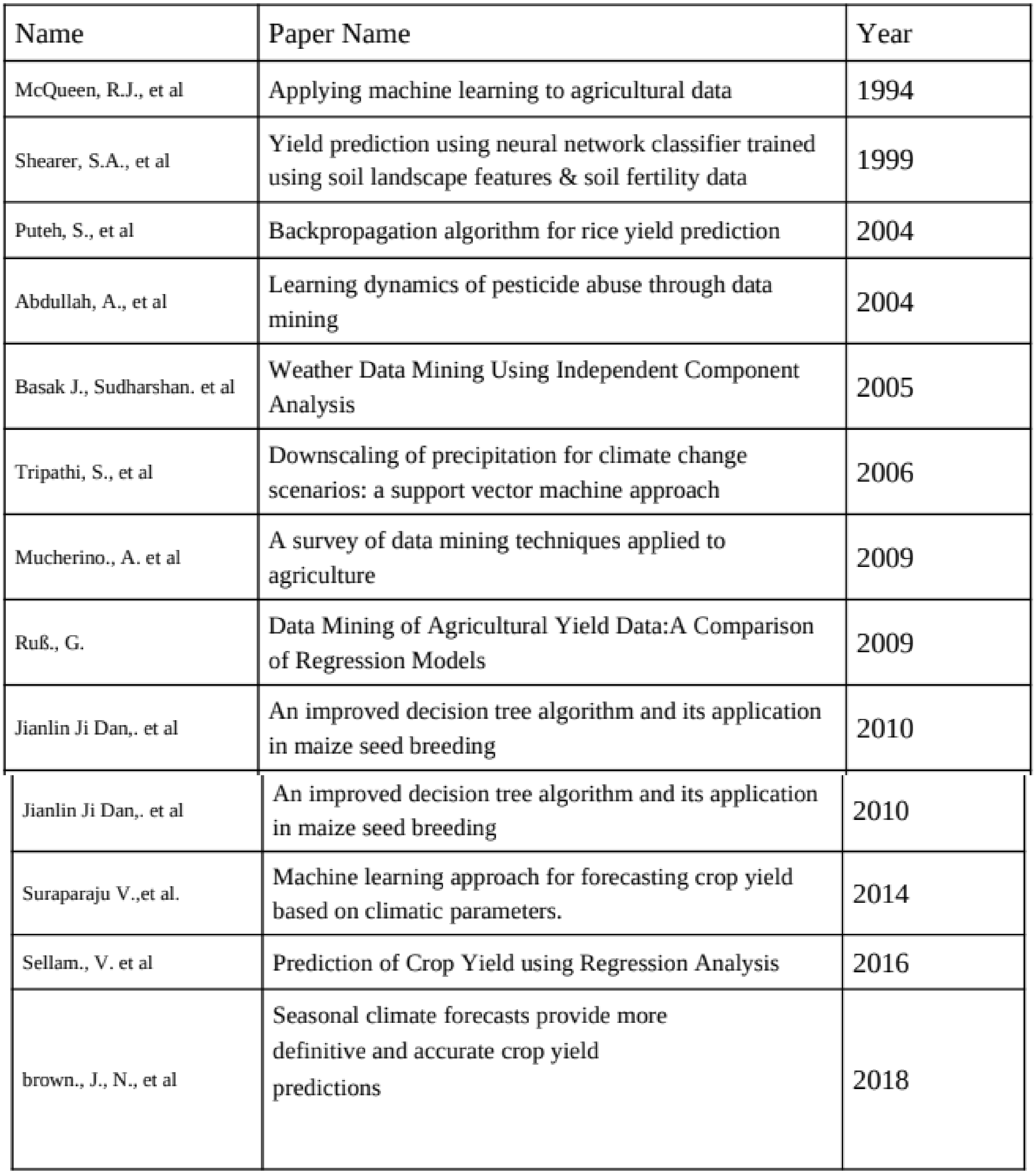
This chapter will provide a detailed overview of the complete methodology and design about the Study. It provides an overview of the data being used for the research study. It outlines the methods and tools taken for implementation of the research study. Finally, it discussed the Techniques used to address the research question in this context and the various methods to evaluate those Techniques.

There are different forecasting methodologies developed and evaluated by the researchers all Over the world in the field of Data Mining on agricultural data or associated sciences. We Can Trace back the History of this Young and Emerging field into 90's. Four scientist of Waikato University in New Zealand Published A Paper Named, "Applying Machine Learning on Agricultural Data [1]" on 1994. Where they introduce a System, WEKA, which allows user to Access a variety of machine learning techniques for the purposes of experimentation and Comparison using real world database. This software system is widely used till then to today. From that time till 2007, this field saw a exponential growth in research using Data Mining techniques such as Artificial Neural Network (ANN), K-nearest neighbor (KNN) etc. To classify soil fertility and predicting site-specific crop yield. In this paper [2] Author Examines the Idea of precision farming using Satellite data and normalized difference vegetation Index (NDVI) to predict crop yield with the help of Multivariate Regression and neural Network. After then this field diverges on different direction on various parameter. Such as in the Paper [5] used Multiple Linear Regression (MLR) technique for crop Analysis. Crop analysis is The science to associate different weather and non-weather parameter with the yield of the crop. This study [3] Used Support Vector Machine (SVM) to analyze different possible change in weather scenario which is important in the era of rapid Climate change. Naive Bayes Data Mining Technique is used to classify soils that analyze large soil profile experimental datasets. Decision tree algorithm in data mining is used for predicting soil fertility.[4] K-nearest(KNN) Algorithm [5] is used in simulating daily precipitations and other weather variable. Decision tree and instance based learning (IB3) also used in that study of four years period data of Temperature, relative humidity and rainfall. In this paper[6] author use cloud cover days, Rainfall And temperature variable to forecast crop yield with average of 85% accuracy [7].

Figure 1.3 Crop Yield Prediction

This table contains Review of the Paper Published on Data mining Approach in Agricultural:

.



Name	Paper Name	Year
McQueen, R.J., et al	Applying machine learning to agricultural data	1994
Shearer, S.A., et al	Yield prediction using neural network classifier trained using soil landscape features & soil fertility data	1999
Puteh, S., et al	Backpropagation algorithm for rice yield prediction	2004
Abdullah, A., et al	Learning dynamics of pesticide abuse through data mining	2004
Basak J., Sudharshan. et al	Weather Data Mining Using Independent Component Analysis	2005
Tripathi, S., et al	Downscaling of precipitation for climate change scenarios: a support vector machine approach	2006
Mucherino., A. et al	A survey of data mining techniques applied to agriculture	2009
Ruß., G.	Data Mining of Agricultural Yield Data:A Comparison of Regression Models	2009
Jianlin Ji Dan,. et al	An improved decision tree algorithm and its application in maize seed breeding	2010
Jianlin Ji Dan,. et al	An improved decision tree algorithm and its application in maize seed breeding	2010
Suraparaju V.,et al.	Machine learning approach for forecasting crop yield based on climatic parameters.	2014
Sellam., V. et al	Prediction of Crop Yield using Regression Analysis	2016
brown., J., N., et al	Seasonal climate forecasts provide more definitive and accurate crop yield predictions	2018

Supervised Learning

In supervised learning, we are given a data set and already know what our Correct output should look like, having the idea that there is a relationship between the input and the output. Alternatively, supervised learning is a learning in which we teach or train the machine using data which is well labeled that means some data is already tagged with correct answer. After that, machine is provided with new set of data so that supervised learning algorithm analyses the training

Data and produces a correct outcome from labeled data. Supervised learning problems are categorized into "regression" and "Classification" problems. In a regression problem, we are trying to predict Results within a continuous output, meaning that we are trying to map input variables to some continuous function. In a classification problem, we are instead trying to predict results in a discrete output. In other words, we are trying to map input variables into discrete categories.

For Instance:

Given data about the size of houses on the real estate market, try to predict their price. Price as a function of size is a continuous output, so this is a regression problem. We could turn this example into a classification problem by instead making our output about whether the house "sells for more or less than the asking price." Here we are classifying the houses based on price into two discrete Categories.

Classification of Supervised Learning

Supervised learning classified into two categories of algorithms:

- Classification: A classification problem is when the output variable is a category, it gives discrete output.
- Regression: A regression problem is when the output variable is a real value, such as “dollars” or “weight”. It produces continuous output.

Example:

1. Regression - Given a picture of a person, we have to predict their age on the basis of the given picture. Regression predicts a numerical value based on previous observed data.

eg: House Price Prediction, Stock Price Prediction, Height-Weight Prediction.

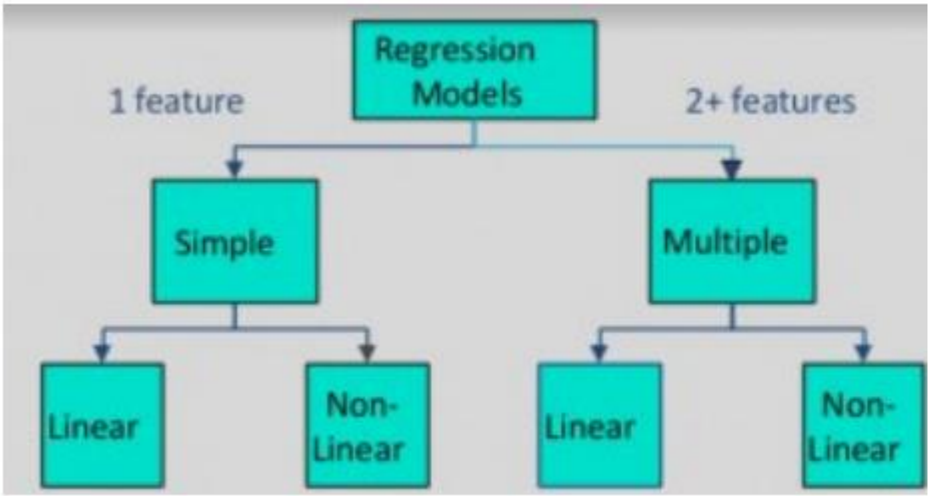
2. Classification - Given a patient with a tumor, we have to predict whether the tumor is malignant or benign. Classification predicts the a category the data belongs to.

eg: Spam Detection, Churn Prediction, Sentiment Analysis, Dog Breed Detection.

2.3 Regression Analysis

Unlike traditional statistical methods, ML does not make assumptions about the correct structure of the data model, which describes the data. This characteristic is very useful to model complex non-linear behaviors, such as a function for crop yield prediction. ML techniques most successfully applied to Crop Yield Prediction (CYP).Supervised learning algorithm consist of a target /outcome variable (or dependent variable) which is to be predicted from a given set of predictors (independent variables). Using these set of variables, we generate a function that map inputs to desired outputs. The training process continues until the model achieves a desired level of accuracy on the training data. A regression problem is when the output variable is a real or continuous value, such as “salary” or “weight”. Many different models can be used; the simplest is the linear regression. It tries to fit data with the best hyper-plane which goes through the points.

Types of Regression:



Simple Linear Regression

It is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables:

- One variable, denoted x, is regarded as the predictor, explanatory, or independent variable.
- he other variable, denoted y, is regarded as the response, outcome or dependent variable.

Simple Linear regression is nothing but a manifestation of the below simple equation.

$y = mx + c$

y is the dependent variable i.e. the variable that needs to be estimated and predicted.

x is the independent variable i.e. the variable that is controllable. It is the input.

m is the slope. It determines what will be the angle of the line. It is the parameter denoted as β.

c is the intercept. A constant that determines the value of y when x is 0. Linear regression models are not perfect. It tries to approximate the relationship between dependent and independent variables in a straight line. Approximation leads to errors. Some errors are inherent in the nature of the problem.

The same equation of a line can be re-written as:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

β0 and β1 are two unknown constants that represent the intercept and slope. They are the parameters. ε is the error term.

Multiple Linear Regression:

In real application scenarios there is typically more than one independent or explanatory variable, so in order to handle the problem of omitted bias, multiple regressions are applied in which there are ‘n’ independent variables. Linear regression with multiple variables is also known as "Multivariate linear regression". It is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the linear relationship between the explanatory (independent) variables and response (dependent) variable. In essence, multiple regressions is the extension of ordinary least-squares (OLS) regression that involves more than one explanatory variable.

The Formula for Multiple Linear Regression is:

hθ(x)=θ0+θ1x1+θ2x2+θ3x3+⋯+θnxn + ε
Where, for i=n observations:
h =dependent variable
x =explanatory variables
θ0 =y-intercept (constant term)
θn =slope coefficients for each explanatory variable
ε =the model’s error term (also known as the residuals)
In the least-squares model, the best-fitting line for the observed data is calculated by minimizing the sum of the squares of the vertical deviations from each data point to the line (if a point lies on the fitted line exactly, then its vertical deviation is 0).

Because the deviations are first squared, then summed, there are no cancellations between positive and negative values. The least-squares estimates b0, b1, ... bp are usually computed by statistical software.

Examples:

- The selling price of a house can depend on the desirability of the location, the number of bedrooms, the number of bathrooms, the yea the house was built, the square footage of the lot and a number of other factors.
- The height of a child can depend on the height of the mother, the height of the father, nutrition, and environmental factors.

KEY TAKEAWAYS:

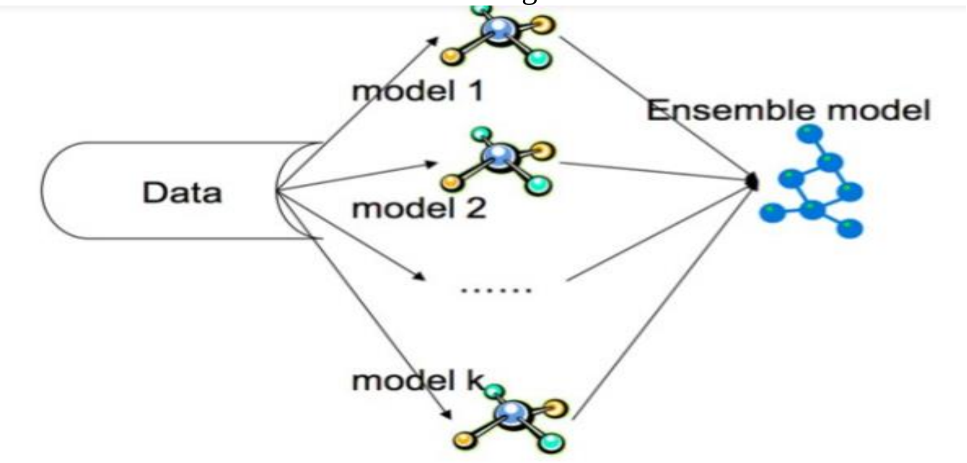
- Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable.
- Multiple regressions is an extension of linear (OLS) regression that uses just one explanatory variable.
- MLR is used extensively in econometrics and financial inference.

Decision Tree Regression

Decision tree builds regression or classification models in the form of a tree Structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy), each representing values for the attribute tested. Leaf node (e.g., Hours Played) represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

Random Forest Regression:

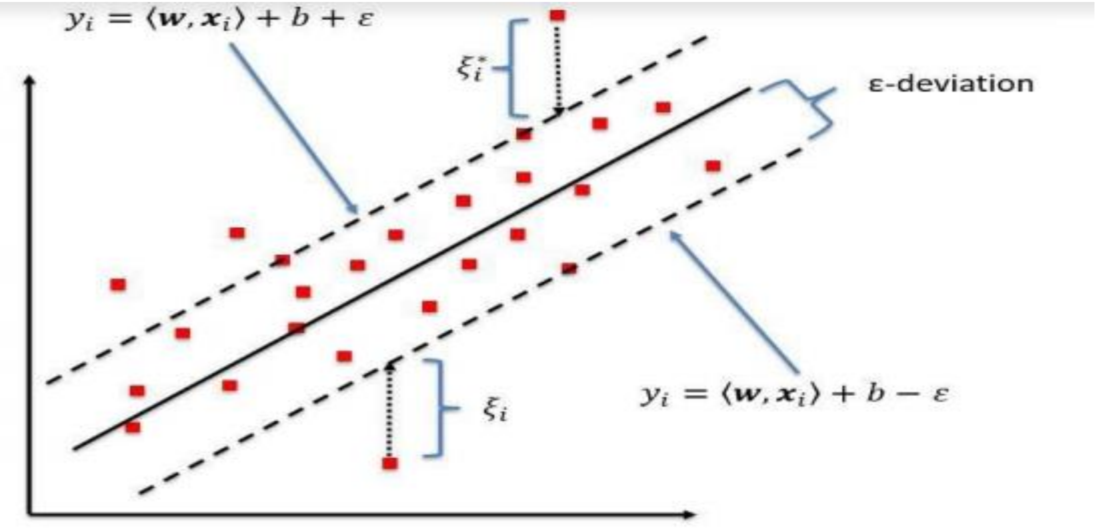
A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap Aggregation, commonly known as bagging. The random forest model is a type of additive model that makes predictions by combining decisions from a sequence of base models. The Random Forest is one of the most effective machine learning models for predictive analytics, making it an industrial workhorse for machine learning.



Support Vector Regression
It is a supervised machine learning algorithm which can be used for regression challenges.

The terms that we are going to be using frequently are:
1. **Kernel**: The function used to map a lower dimensional data into a higher dimensional data.
2. **Hyper Plane**: In SVM this is basically the separation line between the data classes. Although in SVR we are going to define it as the line that will help us predict the continuous value or target value
3. **Boundary line**: In SVM there are two lines other than Hyper Plane which creates a margin . The support vectors can be on the Boundary lines or outside it. This boundary line separates the two classes. In SVR the concept is same.
4. **Support vectors**: This is the data points which are closest to the boundary. The distance of the points is minimum or least.

In simple regression we try to minimize the error rate. While in SVR we try to fit the error within a certain threshold. Support Vector Machine can also be used as a regression method, maintaining all the main features that characterize the algorithm (maximal margin). The Support Vector Regression (SVR) uses the same principles as the SVM for classification, with only a few minor differences. First of all, because output is a real number it becomes very difficult to predict the information
at hand, which has infinite possibilities. In the case of regression, a margin of tolerance (epsilon) is set in approximation to the SVM which would have already requested from the problem.


$y_i = \langle \mathbf{w}, \mathbf{x}_i \rangle + b + \varepsilon$
ξ_i^*
ε-deviation
ξ_i
$y_i = \langle \mathbf{w}, \mathbf{x}_i \rangle + b - \varepsilon$

3 . Problem Statement

This is the project from the motivation of the farmers working in the farm lands are solely dependent on the rains and bore wells for irrigation of their land. In recent times, the farmers have been using irrigation technique through the manual control in which the farmers irrigate the land at regular intervals by turning the water-pump ON/OFF when required. Moreover, for the power indication they are glowing a single bulb between any one of phase and neutral, meanwhile when there is any phase deduction occurs in other phases, the farmer cannot know their supply is low. If they Switch ON any of the motor, there will be the sudden defuse in motor circuit. They may have to travel so far for SWITCHING ON/OFF the motor. They may be suffering from hot Sun, rain and night time too. After reaching their farm, they found that there is no power, so they quietly disappointed to it!! Is there any solution for it???

- The small scale farmers are not able to frequently analyze their farm condition leading too crop failures.
- The irrigation of the farm may become irregular a tedious with manual irrigation procedures.
- Farmers are not able to monitor and control their farms remotely.
- Yield of crops gets reduced if perfect conditions for the crops are not maintained.

4. Proposed Model

4.1 Introduction

This chapter will provide a detailed overview of the complete methodology and design about the study. It provides an overview of the data being used for the research study. It outlines the methods and tools taken for implementation of the research study. Finally, it discussed the techniques used to address the research question in this context and the various methods to evaluate those Techniques.

4.2 Data

All the datasets used in the research were sourced from the openly accessible records of the Indian Government. This was sourced for the years 1990 to 2002 for different seasons like Kharif and Rabi of maize production in Karnataka. From the vast initial dataset, only a limited number of important factors which have the highest impact on agricultural yield were selected for the present research.

The parameters selected for the present study listed below.

- **Rainfall (mm)**: The total amount of precipitation for Kharif and Rabi season of each year of every district.
- **Average Temperature (degree Celsius)**: Crop production will definitely have an impact due to maximum temperature for each year of every district was considered for the present research.
- **Crop Production(x)**: The crop cultivated area in Hectares and production in tones for Kharif and Rabi seasons for ever y year in each selected district of Karnataka state was considered for the present research.
- **Precipitation**: Perception data for every year in each selected district of Karnataka was considered for accurate yield.

Historical weather data is taken from the gridded surface meteorological dataset (METDATA) of Karnataka state. Variables are observed daily and include minimum and maximum air temperature and relative humidity, precipitation, incoming shortwave radiation (sunlight), and average wind, speed. An overview of the variables used is presented in table 1.

We note that all variables that are included parametrically are also included nonparametrically, with the exception of the quadratic terms in time and total precipitation. This allows for a parametric 'main effect,' while also allowing these variables to form nonlinear combinations with other input data, which could be useful if the effects of these variables depend partially on the levels of other variables.

When training the model, we convert all nonparametric covariates into a matrix of their principal components, retaining those that comprise 95% of the variance of the data. The prediction of crop is dependent on numerous factors such as Soil Nutrients, weather and past crop production in order to predict the crop accurately. All these factors are location reliant and thus the location of user is taken as an input to the system.

The selection of input variables and their combinations is an important and complex task.

The following combinations were initially chosen expecting the combinations will give an insight into variable selection for yield predictions.

4.3 Data Preparation

After a detailed and careful examination of the original dataset, a number of data manipulation steps were carried out to prepare the data for investigation and predictive modeling. Each of the dependent and independent variables is analyzed.
Non-relevant variables were discarded and removed from the dataset. The probability distributions for the features were examined using histograms and density curves to understand the variance and outliers.

Data transformations are performed if the distributions are highly skewed from normal to meet the underlying statistical assumptions. Scatter plots are used to visualize the relationship between feature variables and response variables. The significance and strength of the relationships are determined using the correlation coefficients and p values. Similarly, the relationships between features are also examined. After a detailed investigation into the semantics of Karnataka state Agricultural datasets and existential diversities, a number of data manipulation tasks were performed to prepare the data for this research. For regression modeling, the categorical features are recoded to continuous features and new features were created. The data was partitioned into train and test sets for predictive modeling.

4.4 Data Exploration and Visualization

A crop can be cultivable only if apropos conditions are met. These include extensive parameters allied to soil and weather. These constraints are compared and the apt crops are ascertained.
Multiple Linear Regression is used by the system to predict the crop. The prediction is based on past production data of crops i.e.: identifying the tangible weather and soil parameters and comparing it with current conditions which will predict the crop more accurately and in a practical manner

4.5 Model Selection:

A regression analysis is a powerful statistical tool that allows for establishing relationships and characterization within data. In brief, a regression analysis is used for:

A statistical description of variables.

Estimation of a response variable provided a given setoff input variables.
To determine the risk factors which can influence the response variable?
The empirical model will be of multivariate linear regression.
Production = w0 + w1 (Feature 1) + w2 (Feature 2) + w3 (Feature 3) + ...+ wn (Feature n) + e

Linear regression will allow us to estimate the effect of each variable on production. The coefficient of variables will provide us respective impact on the response variable. Linear models are by far the most widely used technique on the subject and have provided successful results (Gerhart, 1988; Rumberger, 1993; Scholz, 1996).
The predictive regression models are build using Multiple regression models are built using various techniques. All the models are compared based on their accuracy using Root mean squared Error.

4.5 Model Evaluation:

In order to establish relationships, it is very important to critically evaluate a regression model structure. The aptness of a regression model is critical to derive effective inference from the model. A regression model is susceptible to misguided inference if the underlying assumptions are not met. There a number of methods available to perform diagnostics and evaluation of regression models - in one of the studies by (Alf, 1984; Lommele and Sturgis, 1974), the author discusses a few standard criteria to evaluate and diagnose regression models. The various evaluation and diagnostics measures used for the purpose of this study are discussed below.

Goodness-of-fit:

A goodness-of-fit determines how well the selected model fits the underlying data. One of the widely adopted measures for determining the goodness-of-fit is the R- squared coefficient of determination. The coefficient is calculated as the square of the correlation between observed response values and predicted response values.

$$R^2=\sum(\breve{Y}i-\bar{Y})^2/\sum(Yi-\bar{Y})^2'$$

Equation - Formula to calculate R- squared measure

R-squared has the useful property that its scale is intuitive: it ranges from zero to one, with zero indicating that the proposed model does not improve prediction over the mean model, and one indicating perfect prediction. Improvement in the regression model results in proportional increases in R- squared. The R 2 (R-squared) value has been used to analyze the variance explained towards the response variable (Production) by the input features within a model

4.5 Root Mean Squared Error RMSE):

RMSE is the most popular evaluation metric used in regression problems. It follows an assumption that error is unbiased and follows a normal distribution. Here are the key points to consider on RMSE:

1. The power of ‘square root’ empowers this metric to show large number deviations.
2. The ‘squared’ nature of this metric helps to deliver more robust results which prevents cancelling the positive and negative error values. In other words, this metric aptly displays the plausible magnitude of error term.
3. It avoids the use of absolute error values which is highly undesirable in mathematical calculations.

4. When we have more samples, reconstructing the error distribution using RMSE is considered to be more reliable.
5. RMSE is highly affected by outlier values. Hence, make sure you’ve removed outliers from your data set prior to using this metric.
6. As compared to mean absolute error, RMSE gives higher weightage and punishes large errors.

RMSE metric is given by:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

Where, N is Total Number of Observations.

Data Split:

In data mining applications, the source dataset is generally split into two or three parts for multiple purposes. The train set is used to train a predictive model and then a test set is used to measure performance on unseen data. The test is used to measure the accuracy of the model. Sometimes a third set, a validation set, is used for the optimization of models. A 80/20 (80% for Training and 20% hold out the sample as Test set) split will be used for this study.

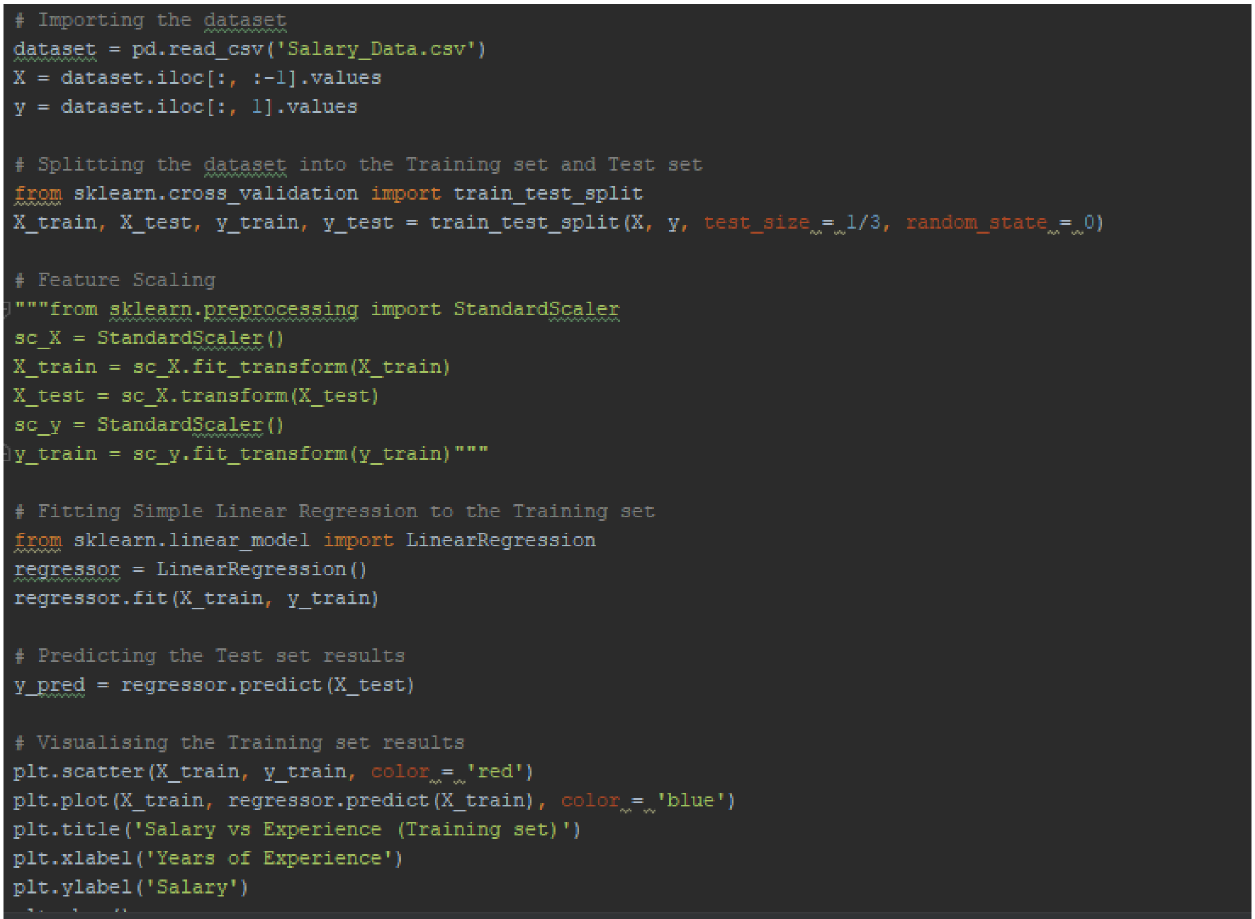
4.10 IMPLEMENTATION:

In order to build an accurate yield prediction model, multiple regression models are created using feature selection and regularization techniques. The regression models are built using R Statistical Package. All the models are compared based on the Root Mean Square Error (RMSE) on the test set. The dataset was split into (80:20) as training and test set. In this study, the below steps are used to build models.

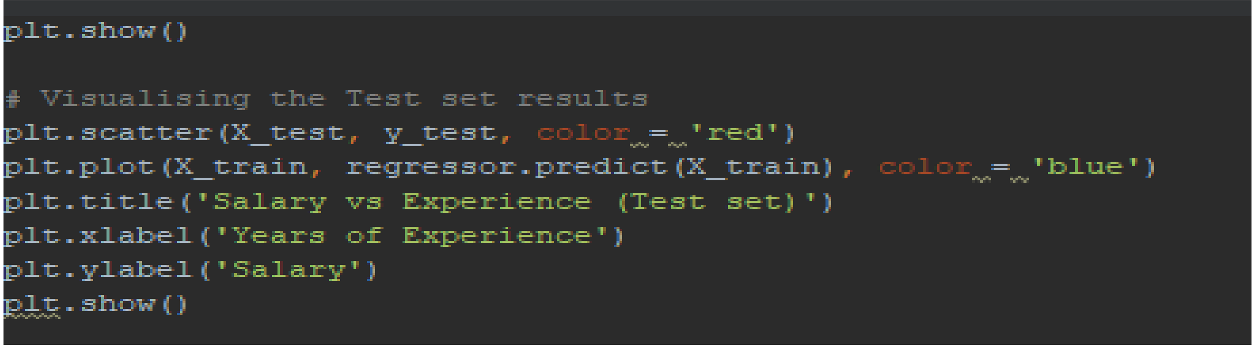
Steps	Description

1.	Import Data and split (80:20) as Training and Test Set
2.	Train the model on the training set.
3.	Select Best Model using Feature Selection/ Regularization
4.	Apply parameter engineering to improve performance
5.	Validate Model and Calculate RMSE

SIMPLE LINEAR REGRESSION

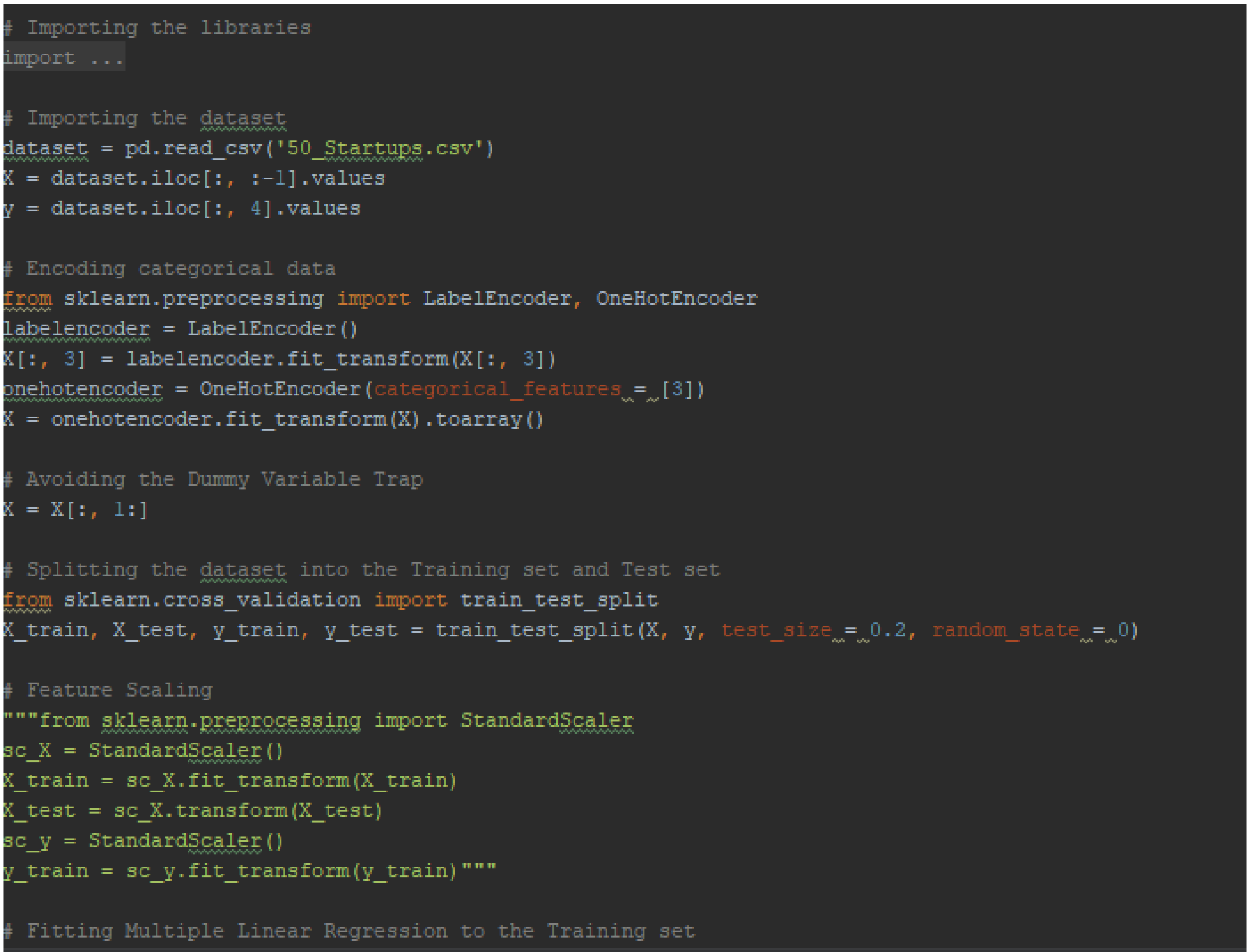


```
# Importing the dataset
dataset = pd.read_csv('Salary_Data.csv')
X = dataset.iloc[:, :-1].values
y = dataset.iloc[:, 1].values

# Splitting the dataset into the Training set and Test set
from sklearn.cross_validation import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 1/3, random_state = 0)

# Feature Scaling
"""from sklearn.preprocessing import StandardScaler
sc_X = StandardScaler()
X_train = sc_X.fit_transform(X_train)
X_test = sc_X.transform(X_test)
sc_y = StandardScaler()
y_train = sc_y.fit_transform(y_train)"""

# Fitting Simple Linear Regression to the Training set
from sklearn.linear_model import LinearRegression
regressor = LinearRegression()
regressor.fit(X_train, y_train)

# Predicting the Test set results
y_pred = regressor.predict(X_test)

# Visualising the Training set results
plt.scatter(X_train, y_train, color = 'red')
plt.plot(X_train, regressor.predict(X_train), color = 'blue')
plt.title('Salary vs Experience (Training set)')
plt.xlabel('Years of Experience')
plt.ylabel('Salary')
```

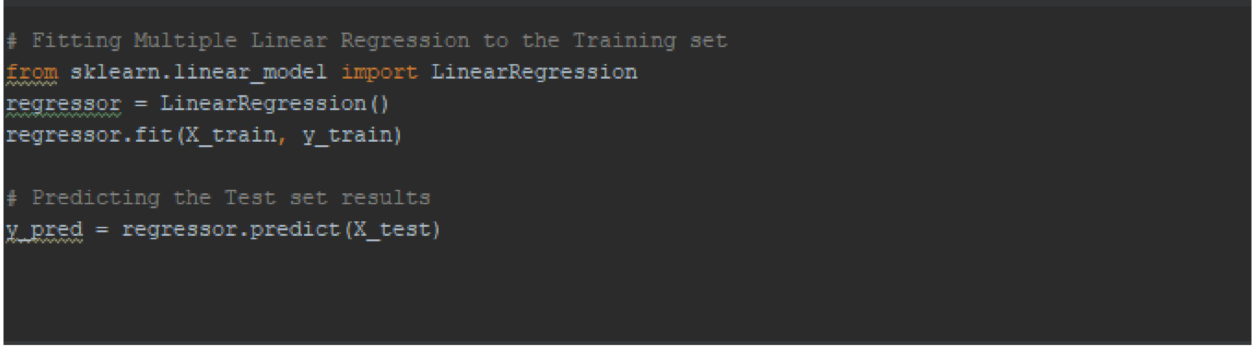


```
plt.show()

# Visualising the Test set results
plt.scatter(X_test, y_test, color = 'red')
plt.plot(X_train, regressor.predict(X_train), color = 'blue')
plt.title('Salary vs Experience (Test set)')
plt.xlabel('Years of Experience')
plt.ylabel('Salary')
plt.show()
```

MULTIPLE LINEAR REGRESSIONS:

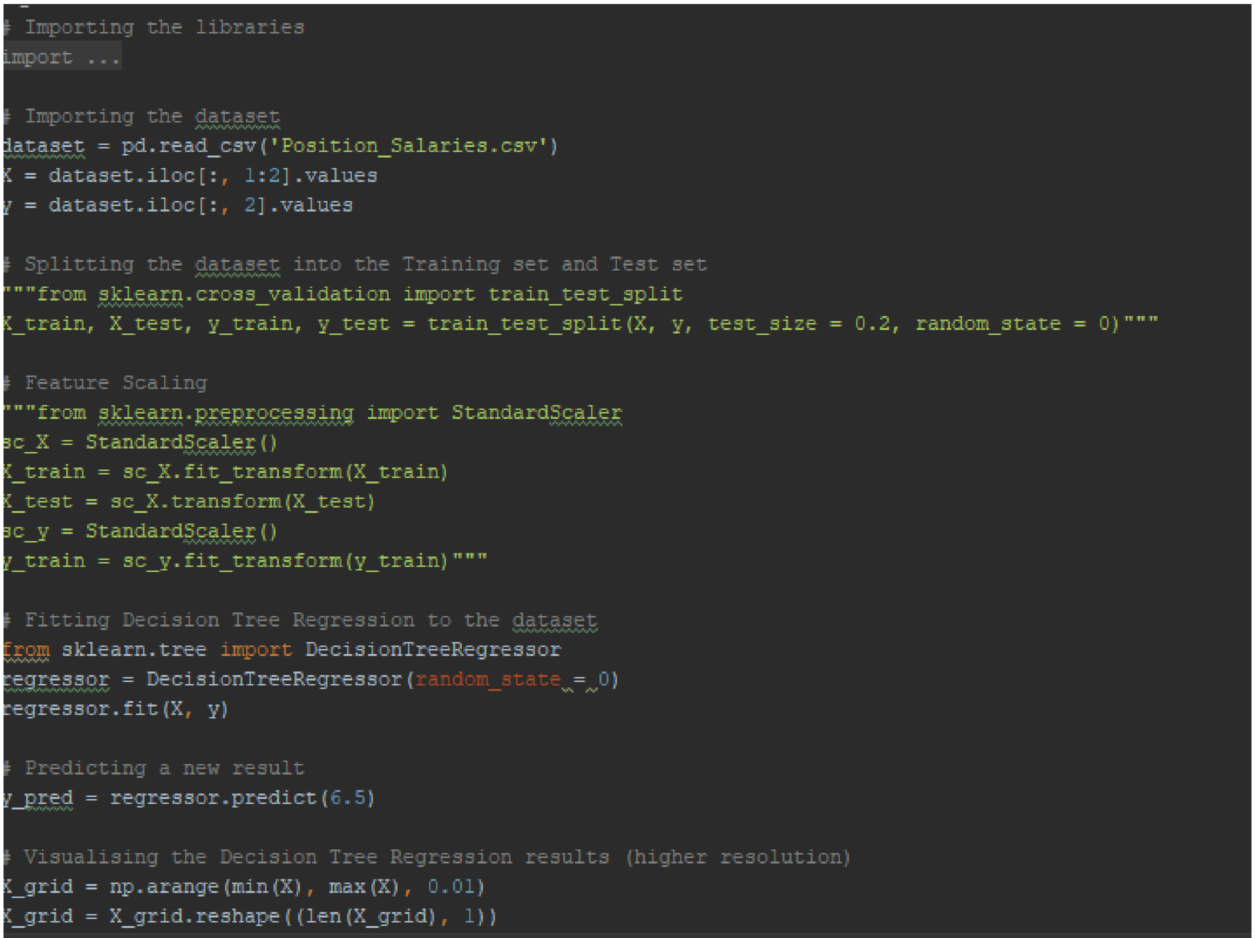


```
# Importing the libraries
import ...

# Importing the dataset
dataset = pd.read_csv('50_Startups.csv')
X = dataset.iloc[:, :-1].values
y = dataset.iloc[:, 4].values

# Encoding categorical data
from sklearn.preprocessing import LabelEncoder, OneHotEncoder
labelencoder = LabelEncoder()
X[:, 3] = labelencoder.fit_transform(X[:, 3])
onehotencoder = OneHotEncoder(categorical_features = [3])
X = onehotencoder.fit_transform(X).toarray()

# Avoiding the Dummy Variable Trap
X = X[:, 1:]

# Splitting the dataset into the Training set and Test set
from sklearn.cross_validation import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)

# Feature Scaling
"""from sklearn.preprocessing import StandardScaler
sc_X = StandardScaler()
X_train = sc_X.fit_transform(X_train)
X_test = sc_X.transform(X_test)
sc_y = StandardScaler()
y_train = sc_y.fit_transform(y_train)"""

# Fitting Multiple Linear Regression to the Training set
```

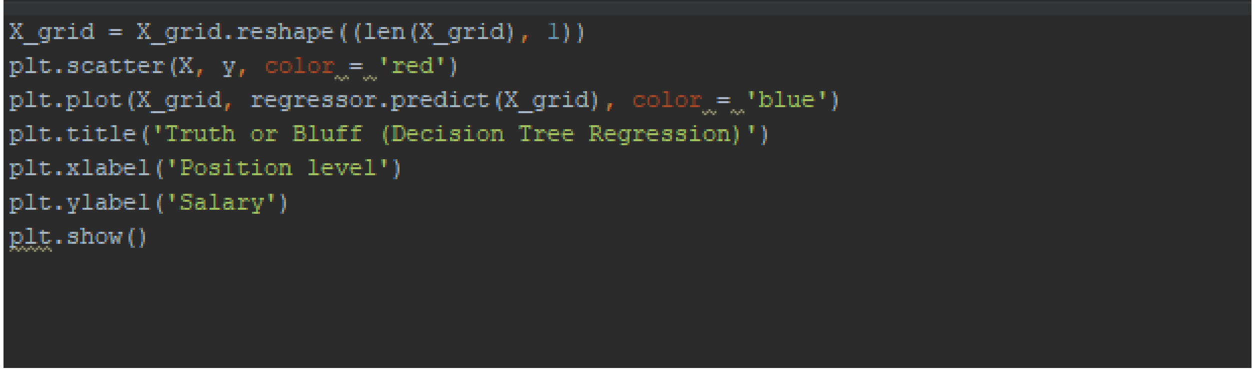


```
# Fitting Multiple Linear Regression to the Training set
from sklearn.linear_model import LinearRegression
regressor = LinearRegression()
regressor.fit(X_train, y_train)

# Predicting the Test set results
y_pred = regressor.predict(X_test)
```

DESCISION TREE REGRESSION :

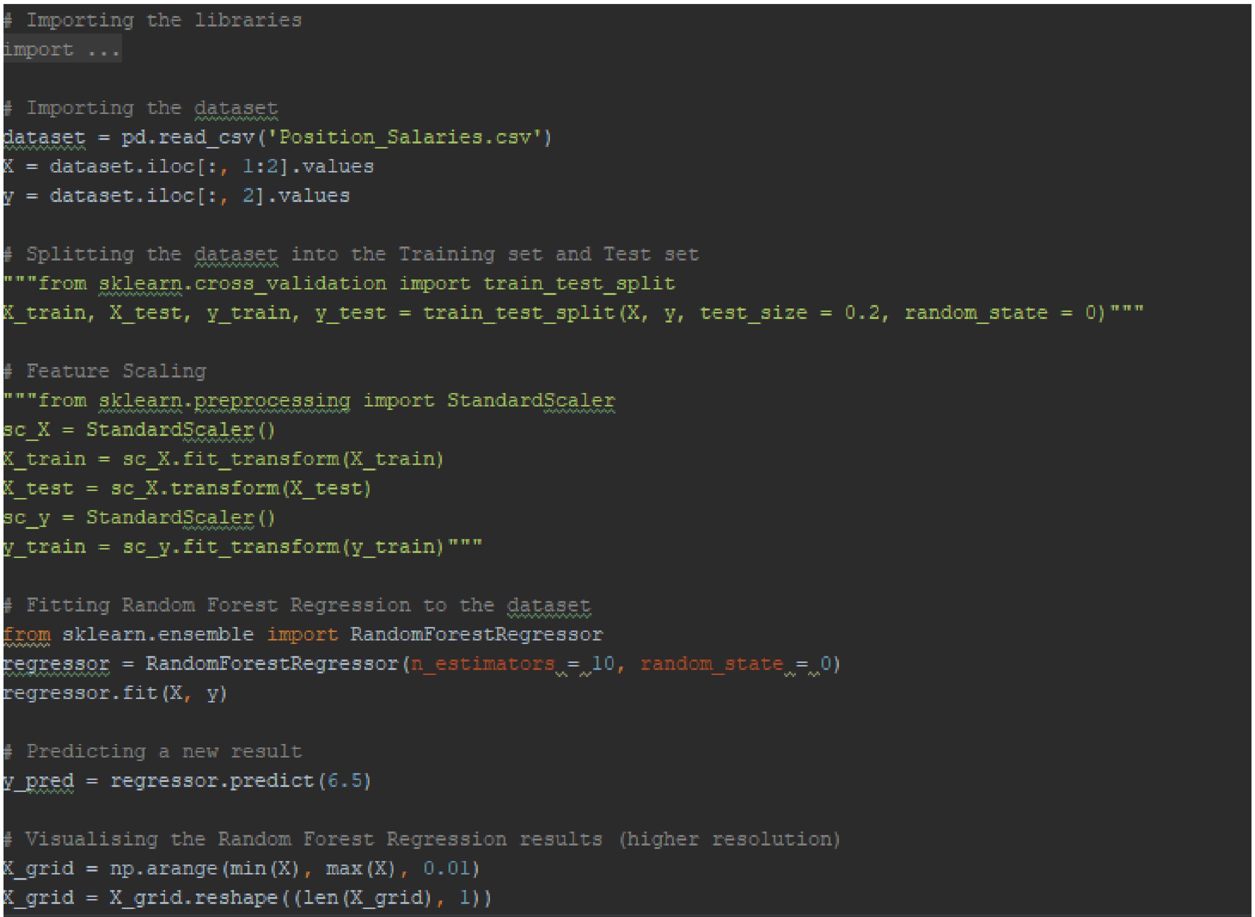


```
# Importing the libraries
import ...

# Importing the dataset
dataset = pd.read_csv('Position_Salaries.csv')
X = dataset.iloc[:, 1:2].values
y = dataset.iloc[:, 2].values

# Splitting the dataset into the Training set and Test set
"""from sklearn.cross_validation import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)"""

# Feature Scaling
"""from sklearn.preprocessing import StandardScaler
sc_X = StandardScaler()
X_train = sc_X.fit_transform(X_train)
X_test = sc_X.transform(X_test)
sc_y = StandardScaler()
y_train = sc_y.fit_transform(y_train)"""

# Fitting Decision Tree Regression to the dataset
from sklearn.tree import DecisionTreeRegressor
regressor = DecisionTreeRegressor(random_state = 0)
regressor.fit(X, y)

# Predicting a new result
y_pred = regressor.predict(6.5)

# Visualising the Decision Tree Regression results (higher resolution)
X_grid = np.arange(min(X), max(X), 0.01)
X_grid = X_grid.reshape((len(X_grid), 1))
```

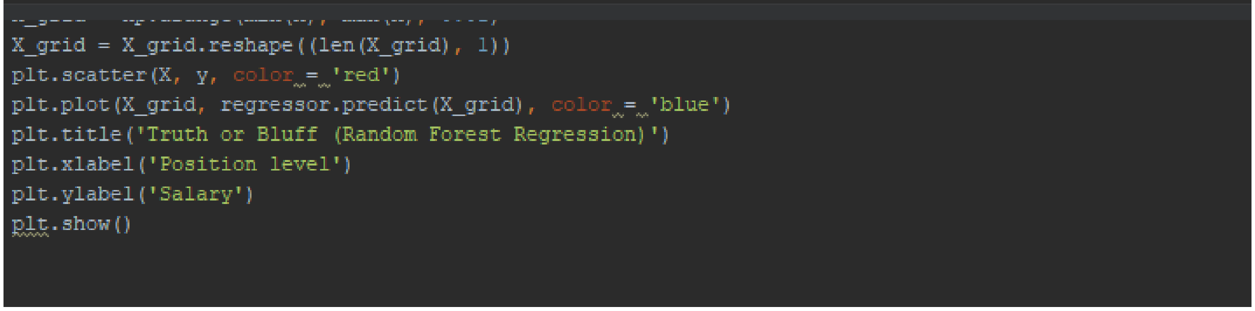


```
X_grid = X_grid.reshape((len(X_grid), 1))
plt.scatter(X, y, color = 'red')
plt.plot(X_grid, regressor.predict(X_grid), color = 'blue')
plt.title('Truth or Bluff (Decision Tree Regression)')
plt.xlabel('Position level')
plt.ylabel('Salary')
plt.show()
```

RANDOM FORSET REGRESSION:

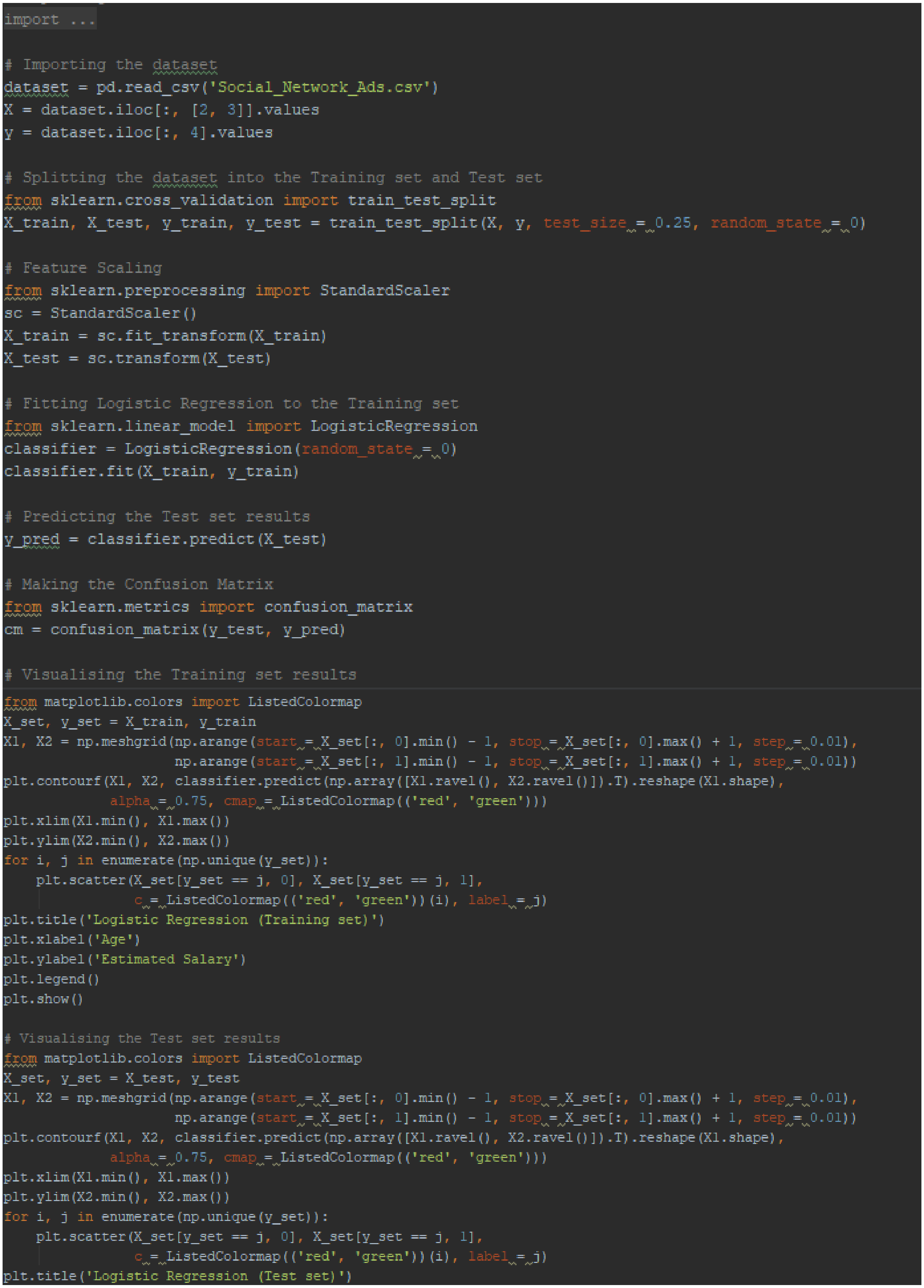


```
# Importing the libraries
import ...

# Importing the dataset
dataset = pd.read_csv('Position_Salaries.csv')
X = dataset.iloc[:, 1:2].values
y = dataset.iloc[:, 2].values

# Splitting the dataset into the Training set and Test set
"""from sklearn.cross_validation import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)"""

# Feature Scaling
"""from sklearn.preprocessing import StandardScaler
sc_X = StandardScaler()
X_train = sc_X.fit_transform(X_train)
X_test = sc_X.transform(X_test)
sc_y = StandardScaler()
y_train = sc_y.fit_transform(y_train)"""

# Fitting Random Forest Regression to the dataset
from sklearn.ensemble import RandomForestRegressor
regressor = RandomForestRegressor(n_estimators = 10, random_state = 0)
regressor.fit(X, y)

# Predicting a new result
y_pred = regressor.predict(6.5)

# Visualising the Random Forest Regression results (higher resolution)
X_grid = np.arange(min(X), max(X), 0.01)
X_grid = X_grid.reshape((len(X_grid), 1))
```



```
X_grid = X_grid.reshape((len(X_grid), 1))
plt.scatter(X, y, color = 'red')
plt.plot(X_grid, regressor.predict(X_grid), color = 'blue')
plt.title('Truth or Bluff (Random Forest Regression)')
plt.xlabel('Position level')
plt.ylabel('Salary')
plt.show()
```

LOGISTIC REGRESSION:



```
import ...

# Importing the dataset
dataset = pd.read_csv('Social_Network_Ads.csv')
X = dataset.iloc[:, [2, 3]].values
y = dataset.iloc[:, 4].values

# Splitting the dataset into the Training set and Test set
from sklearn.cross_validation import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.25, random_state = 0)

# Feature Scaling
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)

# Fitting Logistic Regression to the Training set
from sklearn.linear_model import LogisticRegression
classifier = LogisticRegression(random_state = 0)
classifier.fit(X_train, y_train)

# Predicting the Test set results
y_pred = classifier.predict(X_test)

# Making the Confusion Matrix
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, y_pred)

# Visualising the Training set results
from matplotlib.colors import ListedColormap
X_set, y_set = X_train, y_train
X1, X2 = np.meshgrid(np.arange(start = X_set[:, 0].min() - 1, stop = X_set[:, 0].max() + 1, step = 0.01),
                     np.arange(start = X_set[:, 1].min() - 1, stop = X_set[:, 1].max() + 1, step = 0.01))
plt.contourf(X1, X2, classifier.predict(np.array([X1.ravel(), X2.ravel()]).T).reshape(X1.shape),
             alpha = 0.75, cmap = ListedColormap(('red', 'green')))
plt.xlim(X1.min(), X1.max())
plt.ylim(X2.min(), X2.max())
for i, j in enumerate(np.unique(y_set)):
    plt.scatter(X_set[y_set == j, 0], X_set[y_set == j, 1],
                c = ListedColormap(('red', 'green'))(i), label = j)
plt.title('Logistic Regression (Training set)')
plt.xlabel('Age')
plt.ylabel('Estimated Salary')
plt.legend()
plt.show()

# Visualising the Test set results
from matplotlib.colors import ListedColormap
X_set, y_set = X_test, y_test
X1, X2 = np.meshgrid(np.arange(start = X_set[:, 0].min() - 1, stop = X_set[:, 0].max() + 1, step = 0.01),
                     np.arange(start = X_set[:, 1].min() - 1, stop = X_set[:, 1].max() + 1, step = 0.01))
plt.contourf(X1, X2, classifier.predict(np.array([X1.ravel(), X2.ravel()]).T).reshape(X1.shape),
             alpha = 0.75, cmap = ListedColormap(('red', 'green')))
plt.xlim(X1.min(), X1.max())
plt.ylim(X2.min(), X2.max())
for i, j in enumerate(np.unique(y_set)):
    plt.scatter(X_set[y_set == j, 0], X_set[y_set == j, 1],
                c = ListedColormap(('red', 'green'))(i), label = j)
plt.title('Logistic Regression (Test set)')
```

```
plt.xlabel('Age')
plt.ylabel('Estimated Salary')
plt.legend()
plt.show()
```

5. RESULT AND DISCUSSIONS:

5.1 Model Comparison Based on RMSE Value:

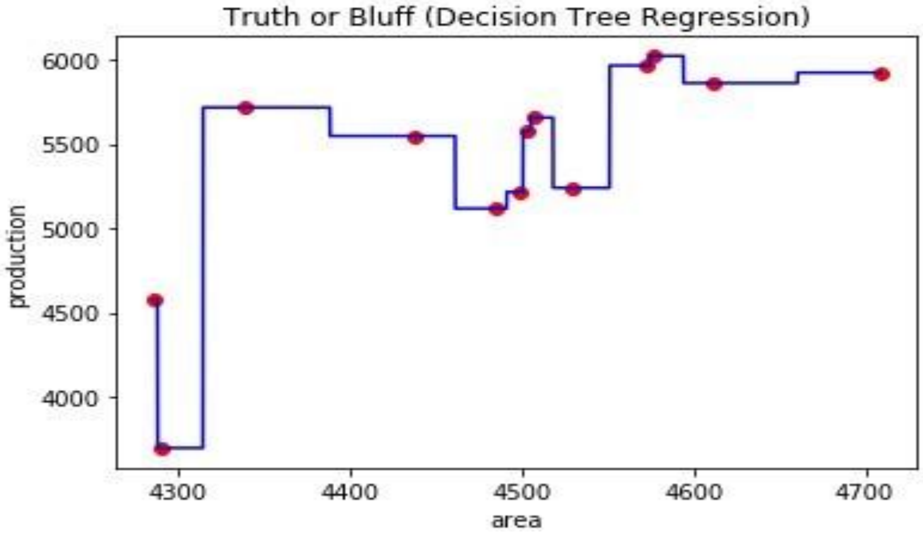
The below table shows the Root Mean Squared error from all the regression Models.

In the predictive modelling Linear Regression, Decision Tree Regression and Random Forest Regression have a very small difference in RMSE. The model Having lower RMSE value is better. Multiple Linear Regression have the least RMSE value and it is the most accurate technique in this case.

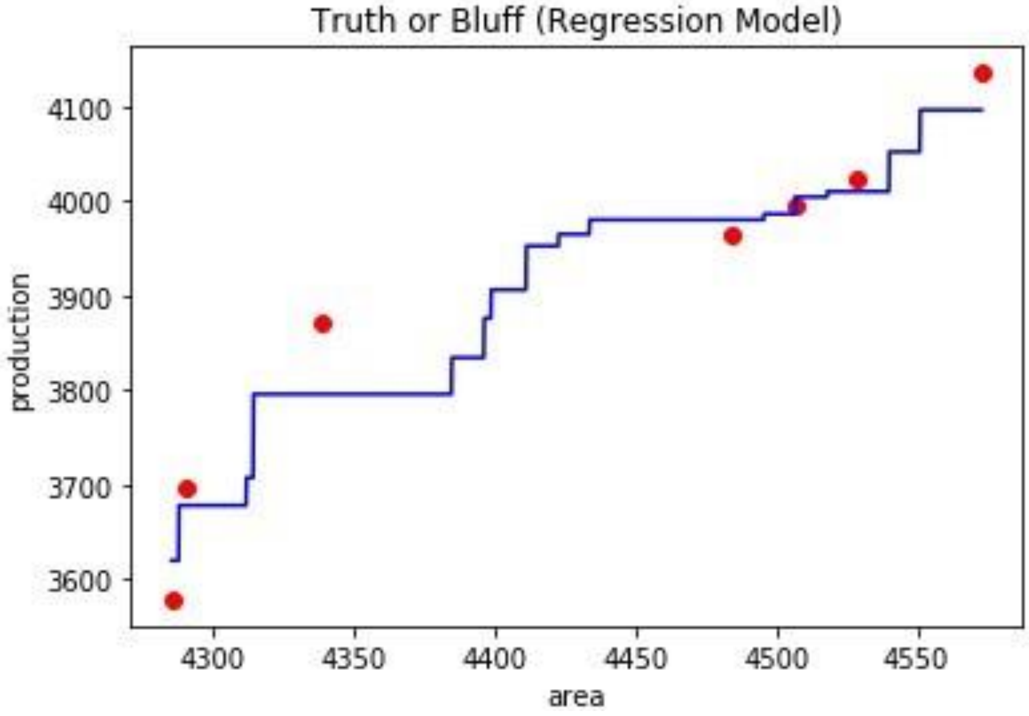
Model RMSE Value

Model	**RMSE Value**
Multiple Linear Regression	692.4732244
Decision Tree Regression	760.5808524
Random Forest Regression	710.5085925
Support Vector Regression	790.2346109

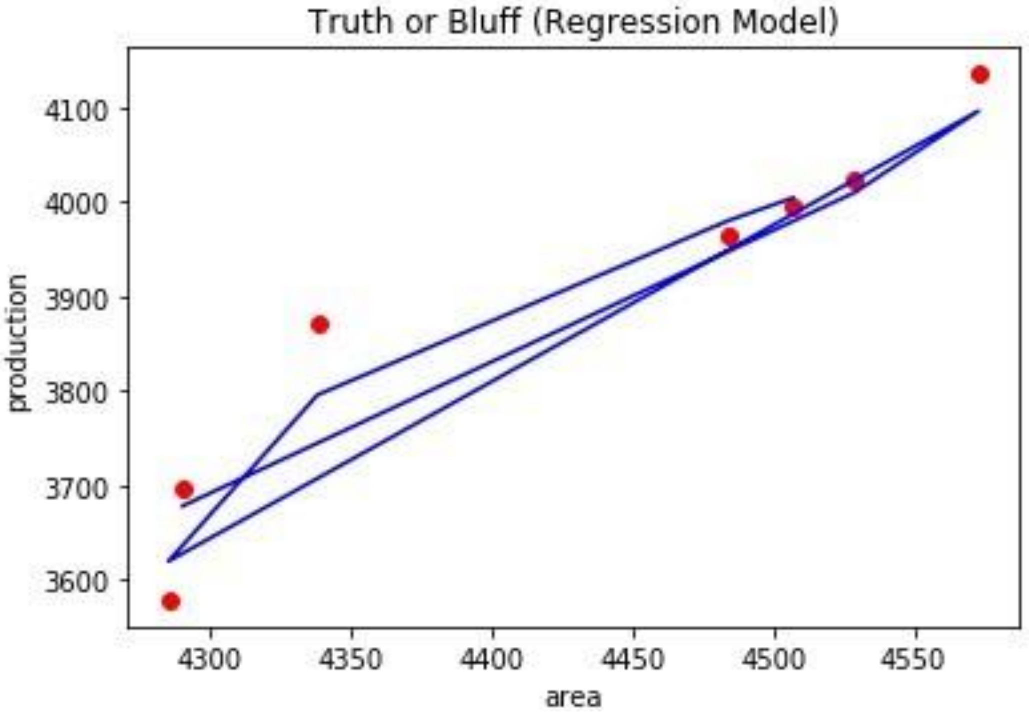
Decision Tree Regression:



Random Forest Regression:



Support Vector Regression



6. CONCLUSION

The present study demonstrated the potential use of regression techniques in predicting the crop yield based on the climatic input parameters. While the performance of the model was tested and a close inspection was performed and residual over fitting kind of problems will be addressed. This work can be extended by considering more factors like Minimum Support Price (MSP), Soil Parameters etc. that affects the yield of a crop and by using various data mining, statistical techniques to analyze the factors influencing the yield. Several evaluations of ML methods applied to CYP have been made in the literature, each one with different researching purposes. Some works measure the ML performance using a particular attribute set.

The evaluated techniques were ranked, from the best to the worst, according to RMSE value. Finally, it is necessary to point out that this work deals only with comparing the predictive accuracy of the above-mentioned techniques. Machine learning techniques are complex, and several factors are related with their performance measuring. Some examples of these factors are the model structure, knowledge representation, implementation cost, missing data handling and training time. Further research will be dedicated to compare these characteristics of ML algorithms and their compatibility with agricultural planning.

In future iterations, we would try to fix these by:

- getting more data,
- engineering additional and/or different features,
- using ensemble techniques by combining the results of different models

7. REFERENCES

[1]Dhanya, C.T. and D. Nagesh Kumar, 2009. Data mining for evolution of association rules for droughts and floods in India using climate inputs. J. of Geo. Phy.Res.114:1-15.

[2]Kannan, M. Prabhakaran S and P. Ramachandran (2011).Rainfall forecasting using data mining technique. International Journal of Engineering and Technology Vol.2 (6), 2010, 397-401.

[3]Li,. A..Estimating crop yield from multi-temporal satellite data using multivariate regression & nn techniques(2007).

[4] P.Vinciya,Dr.A. Valarmathi,"Agricultural Analysis for Next Generation High Tech Farming in Data Mining .

[5] https://www.cse.msu.edu/~tangjili/publication/feature_selection_for_classification.pdf

[6] Rajagopalan, B., & Lall, U. (1999). A k–nearest neighbor simulator for daily precipitation and other weather variables. WATER RESOURCES RESEARCH,35(10), 3089-3101.

[7] https://medium.com/coinmonks/support-vector-regression-or-svr-8eb3acf6d0ff

[8]www.researchgate.netpublication263368516_Predictive_ability_machine_learning_methods

[9] Shevade SK, Keerthi SS, Bhattacharyya C, Murthy KRK, 2000. Improvements to the SMO algorithm for SVM regression. IEEE Transactions on Neural Networks 11(5): 1188-1193.

[10] https://www.investopedia.com/terms/m/mlr.asp

[11] Breiman L, 2001. Statistical modeling: the two cultures (with discussion). Statist Sci 16: 199-231.