CROP YIELD PREDICTION USING MACHINE LEARNING TECHNIQUES

BY

C.SARASWATHI

(17PCS018)

Project Report Submitted

In Partial fulfillment of the requirements for the award of

Master's Degree in Computer Science

Department of Computer Science

Avinashilingam Institute for Home Science and Higher Education for Women, (Deemed to be University)

Coimbatore - 641043

April – 2019

CROP YIELD PREDICTION USING MACHINE LEARNING TECHNIQUES

BY

C.SARASWATHI

(17PCS018)

Project Report Submitted

In Partial fulfillment of the requirements for the award of

Master's Degree in Computer Science

Department of Computer Science

Avinashilingam Institute for Home Science and Higher Education for Women, (Deemed to be University)

Coimbatore - 641043

April-2019

Signature of the Head of the Department **Signature of the Supervisor**

Viva Voce Examination Held on _____

Signature of the Examiners

Acknowledgement

ACKNOWLEDGEMENT

I would like to express my sincere thanks to **God Almighty**, for his constant love and grace that he has showered upon me.

I am very grateful to **Shri, Dr.P.R.KrishnaKumar, Chancellor,** Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for his support and encouragement during the course of my project.

I heartily thank **Dr. (Mrs.) Premavathy Vijayan, M.Sc., M.Ed., Dip.Spl.Edn., M.Phil, Ph.D., Vice Chancellor** Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for providing the facilities to do the project.

I express my humble gratitude to **Dr. (Mrs.) S Kowsalya, M.Sc., M.Phil, Ph.D., Registrar,** Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for providing all the facilities necessary for the project.

I am also thankful to **Dr. (Mrs.) K.Udaya Chandrika M.Sc., M.Phil., Ph.D., Dean,** School of Physical Sciences & Computational Sciences of our university, for granting the facility required.

I wish to place on record my deep sense of gratitude to **Dr. (Mrs.) V.Radha, M.Sc., PGDOR, PGDCA, B.Ed., M.Phil., Ph.D., Professor** and **Head, Department of Computer Science,** Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, for providing all the facilities to complete the project.

I express my honourable thanks to my project coordinator **Dr. (Mrs.) G.Padmavathi, M.Sc., M.Phil., Ph.D., Professor, Department of Computer Science,** for her kind advice and knowledgeable suggestions which helped me to complete my project successfully.

I owe great deal of gratitude to my esteemed guide **Dr. (Mrs.) S.Anitha, M.Sc., M.Phil., Ph.D., Assistant Professor, Department of Computer Science,** for imparting the tremendous assistance and well-timed support for triumph of my project.

Finally, I take pride to thank my parents and those who helped me directly or indirectly for carrying out this work.

Abstract

ABSTRACT

The project entitled as **"CROP YIELD PREDICTION USING MACHINE LEARNING TECHNIQUES"** is developed using Python. An important issue for agricultural planning purposes is the accurate crop yield estimation for the various crops and it involves careful planning. Machine learning techniques are suitable to achieve practical and effective solutions for the crop yield prediction. The prediction of crop yield is based on various factors like temperature, rainfall, weather condition, humidity and season. It is necessary to build a system to better estimate the crop production.

The project predicts the crop yield based on machine learning techniques based regression model. The algorithms such as Linear Regression (LR), K-Nearest Regression (KNR), and Support Vector Regression (SVR) are used to estimate crop yield.

The crop production statistics dataset of India is used for the study. The Indian dataset consists of 246091 records and 7 attributes. The prediction of crop yield is based on the parameters like state name, district name, crop name, crop year, season, area, and production. The dataset includes both numerical and string data.

The performances of the algorithms are evaluated based on the metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). Thus the system will assist in agricultural planning and improve algorithms.

Contents

TABLE OF CONTENT

Introduction

1. INTRODUCTION

1.1 Machine Learning

Machine learning is an application of artificial intelligence (AI) that provides the ability to automatically learn and improve the system from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and uses it to learn for themselves.

The process of learning begins with observations or data, such as examples, direct experience, or instruction, and looks for pattern in data thus it helps to make better decision. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

Machine learning is a data analytics technique that teaches computers to do what comes naturally to humans and animals: learn from experience. Machine learning algorithms use computational methods to “learn” information directly from data without relying on a predetermined equation as a model. The algorithms adaptively improve their performance as the number of samples available for learning increases.

Some of the key machines learning algorithms are

- ✓ Random forests
- ✓ Neural networks
- ✓ Discovery of sequence and associations
- ✓ Decision trees
- ✓ Mapping of nearest neighbour
- ✓ Supporting vector machines
- ✓ Boosting and bagging gradient

Machine learning methods

Supervised Learning

Supervised machine learning builds a model that makes predictions based on evidence in the presence of uncertainty. A supervised learning algorithm takes a known set of input

data and known responses to the data (output) and trains a model to generate reasonable predictions for the response to new data. Supervised learning uses classification and regression techniques to develop predictive models.

- ✓ Classification techniques predict discrete responses

Classification applications include medical imaging, speech recognition, and credit scoring. Commonly used classification algorithms are boosted and bagged decision trees, Naïve Bayes, discriminant analysis, logistic regression, neural networks, Support Vector Machine (SVM),*k*-nearest neighbour.

- ✓ Regression techniques predict continuous responses

Commonly used regression algorithms are linear model, nonlinear model, regularization, step wise regression.

Unsupervised Learning

Unsupervised learning finds hidden patterns or intrinsic structures in data. It is used to draw inferences from datasets consisting of input data without labelled responses. Clustering is the most common unsupervised learning technique. It is used for exploratory data analysis to find hidden patterns or groupings in data. Applications for cluster analysis include gene sequence analysis, market research, and object recognition.

For example, if a cell phone company wants optimize the locations where they build cell phone towers, they can use machine learning to estimate the number of clusters of people relying on their towers. A phone can only talk to one tower at a time, so the team uses clustering algorithms to design the best placement of cell towers to optimize signal reception for groups, or clusters, of their customers.

Commonly used clustering algorithms are k-means and k-Medoids, hierarchical clustering, Gaussian mixture models, hidden Markov models, self-organizing maps, fuzzy c-means clustering, and subtractive clustering. The Figure 1.1 shows the categories of machine learning algorithms.

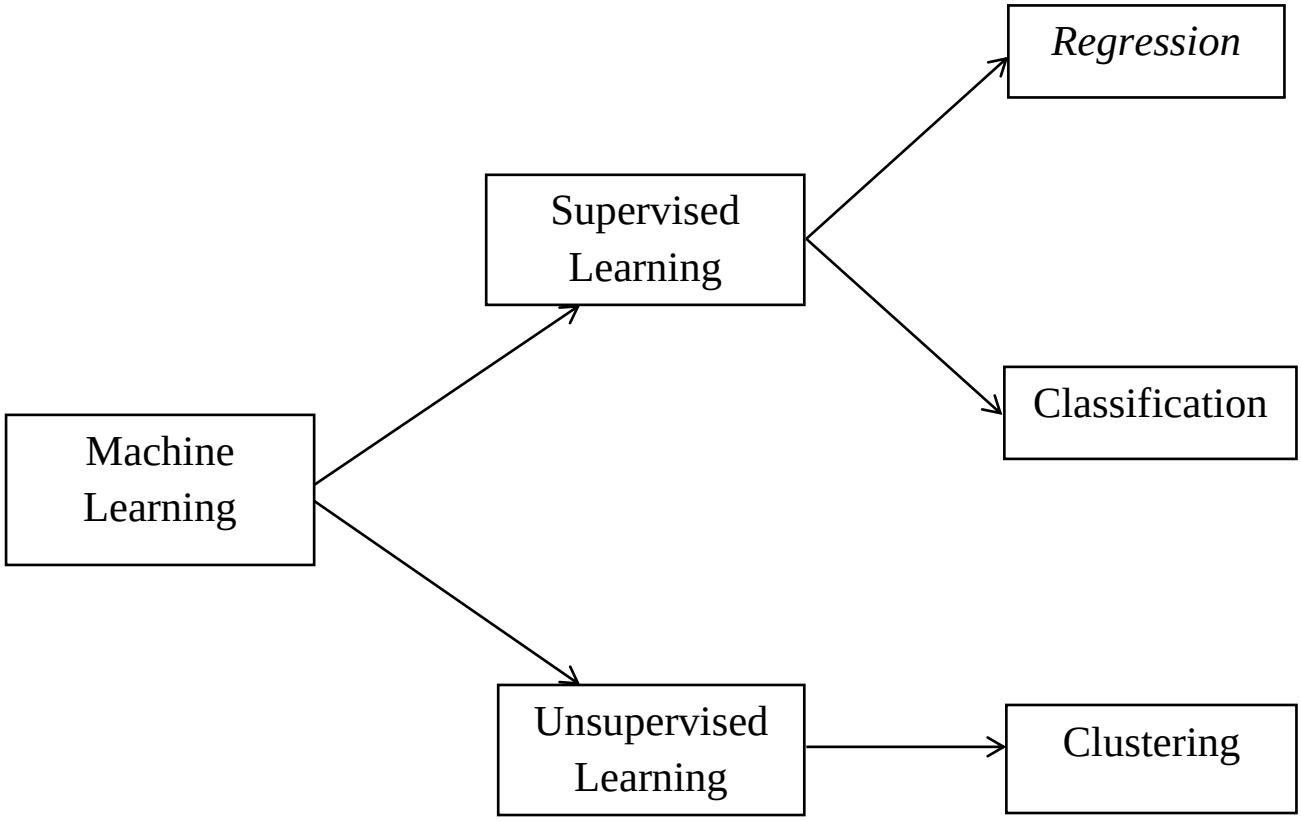


Figure 1.1 Machine Learning Algorithms

Machine learning applications are

- ✓ Image Recognition
- ✓ Speech Recognition
- ✓ Medical Diagnosis
- ✓ Statistical Arbitrage
- ✓ Classification
- ✓ Prediction
- ✓ Regression

1.2 Problem Definition

An important issue for agricultural planning purposes is the accurate yield estimation for various crops and involves careful planning. Machine learning is an essential approach for achieving practical and effective solutions for this problem. There are various factors that may affect crop yield and leads to heavy loss.

The prediction of crop yield is based on various factors like temperature, rainfall, weather condition, humidity and season. It is necessary to build a system to better estimate crop production.

System Study

2. SYSTEM STUDY

This section gives the details about the hardware and software requirements

2.1 HARDWARE SPECIFICATION

The following are minimum hardware requirements for this project

- ✓ Hard Disk : 285 GB
- ✓ Monitor : 21’ Color with VGI card support
- ✓ RAM : 2.00 GB
- ✓ Processor : Intel(R) Pentium(R)
- ✓ Processor speed : CPU A 1018@2.10GHz
- ✓ System Type : 64-bit Operating System

2.2 SOFTWARE SPECIFICATION

The following are minimum software requirements for this project

- ✓ Operating System : Windows 8
- ✓ Programming Language : Python
- ✓ Framework : Anaconda
- ✓ IDE : Spyder
- ✓ Back End : MS Excel

2.3 ABOUT THE SOFTWARE

2.3.1 Python Programming

Python Programming Language is a high-level and interpreted programming language which was created by Guido Van Rossum in 1989. It was first released in 1991, which results in a great general purpose language capable of creating anything from desktop software to web applications and frameworks.

Python is a high-level dynamic programming language. It is quite easy to learn and provides powerful typing. Python code has a very ‘natural’ style to it, in that it is easy to read and understand. Python programming language runs on any platform, ranging from Windows to Linux to Macintosh, Solaris etc. Python can easily be used for small, large, online and

offline projects. The best options for utilizing Python are web development, simple scripting and data analysis

Features of Python

- ✓ Cross platform
- ✓ Expressive language
- ✓ Interpreted language
- ✓ Free and Open Source
- ✓ Object oriented
- ✓ Extensible
- ✓ Large standard library

Advantages of Python

- ✓ Presence of third-party modules
- ✓ Extensive support libraries (NumPy for numerical calculations, Pandas for data analytics etc.)
- ✓ Open source and community development
- ✓ User-friendly data structures
- ✓ Dynamically typed language(No need to mention data type based on value assigned, it takes data type)

Applications of Python

- ✓ GUI based desktop applications (Games, Scientific Applications)
- ✓ Web frameworks and applications
- ✓ Enterprise and Business applications
- ✓ Python Applications in Education
- ✓ Software Development Application

2.3.2 Spyder

Spyder, the scientific python development environment, is a free Integrated Development Environment (IDE) that is included with anaconda. Spyder is a powerful scientific environment written in Python, for Python, and designed by and for scientists, engineers and data analysts. It offers a unique combination of the advanced editing, analysis, debugging, and profiling functionality of a comprehensive development tool with the data exploration, interactive execution, deep inspection, and beautiful visualization capabilities of a scientific package.

The spyder was created and developed by Pierre Raybaut in 2009, since 2012 spyder has been maintained and continuously improved by a team of scientific python developers and the community.

Spyder is extensible with first- and third-party plugins, includes support for interactive tools for data inspection and embeds Python-specific code quality assurance and introspection instruments, such as Pyflakes, Pylint and Rope. It is available cross-platform through Anaconda, on Windows, on macOS through MacPorts, and on major Linux distributions such as Arch Linux, Debian, Fedora, Gentoo Linux, open SUSE and Ubuntu.

Features of Spyder

- ✓ Easy to read
- ✓ Expressive
- ✓ Free and Open-Source
- ✓ High- Level
- ✓ Portable
- ✓ Interpreted
- ✓ Object-Oriented
- ✓ Extensible

2.3.3 MS-Excel

Microsoft Excel is a software program that allows users to organize, format and calculate data with formulas using a spreadsheet system. This software as part of the Microsoft office suite and is compatible with other applications in the office suite.

Excel is a commercial spreadsheet application produced and distributed by Microsoft for Microsoft windows and Mac OS X. It features the ability to perform basic calculations, use graphing tools, create pivot tables and create macro programming language.

Excel has the same basic features as every spreadsheet, which use a collection of cells arranged into rows and columns to organize data manipulation. They also display data as charts, histograms and line graphs.

Features of MS-Excel

- ✓ Multi-Threading Recalculation (MTR) for commonly used functions
- ✓ Improved pivot tables
- ✓ More condition formatting options
- ✓ Additional image editing capabilities
- ✓ Ability to preview before pasting
- ✓ Ability to customize the Ribbon
- ✓ Many new formulas, most highly specialized to improve accuracy

Advantages of MS-Excel

- ✓ Analyzing and storing data
- ✓ Excel tools make your work easier
- ✓ Data recovery and spreadsheet
- ✓ Mathematical formulas of MS Excel make things easier
- ✓ Keeps data combined at one location
- ✓ Helps businessmen in developing future strategy

Disadvantages of MS-Excel

- ✓ Excel is vulnerable to change
- ✓ Excel is difficult to troubleshoot or test
- ✓ Excel is obstructive to regulatory compliance

Methodology

3. METHODOLOGY

3.1 Overview of the Project

In this project three machine learning algorithms are used to predict crop production and they are as follows.

- ✓ Linear Regression
- ✓ K-Nearest Regression
- ✓ Support Vector Regression

Linear Regression is a machine learning algorithm based on supervised learning. A linear regression is a statistical model that analyzes the relationship between a response variable (often called y) and one or more variables and their interactions (often called x or explanatory variables).

There are two types of linear regression.

- ✓ Simple Linear Regression
- ✓ Multiple Linear Regression

Simple Linear Regression is characterized by one independent variable. And, Multiple Linear Regression (as the name suggests) is characterized by multiple (more than 1) independent variables.

K-nearest regression is a very simple, easy to understand, versatile and one of the topmost machine learning algorithms. KNN used in the variety of applications such as finance, healthcare, political science, handwriting detection, image recognition and video recognition. In credit ratings, financial institutes will predict the credit rating of customers. KNN algorithm used for both classification and regression problems.

In KNN, K is the number of nearest neighbors. The number of neighbors is the core deciding factor. K is generally an odd number if the number of classes is 2. When K=1, then the algorithm is known as the nearest neighbor algorithm. This is the simplest case. Suppose P1 is the point, for which label needs to predict.

Support vector machine is considered to be a classification approach, but it can be employed in both classification and regression problems. It can easily handle multiple continuous and categorical variables. SVR constructs a hyperplane in multidimensional space to separate different classes. SVR generates optimal hyperplane in an iterative manner, which is used to minimize an error. The core idea of SVR is to find a maximum marginal hyperplane that best divides the dataset into classes.

The SVM algorithm is implemented in practice using a kernel. A kernel transforms an input data space into the required form. SVM uses a technique called the kernel trick. Here, the kernel takes a low-dimensional input space and transforms it into a higher dimensional space. In other words, you can say that it converts non separable problem to separable problems by adding more dimension to it.

The performances of three algorithms are analyzed using the metrics. Such as

- ✓ Mean Absolute Error
- ✓ Mean Squared Error
- ✓ Root Mean Squared Error

The Mean Absolute Error (or MAE) is the sum of the absolute differences between predictions and actual values. It gives an idea of how wrong the predictions were. A value of 0 indicates no error or perfect predictions.

The Mean Squared Error (or MSE) is much like the mean absolute error in that it provides a gross idea of the magnitude of error. A mean squared error of zero indicates perfect skill, or no error.

The square root of the mean squared error converts the units back to the original units of the output variable and can be meaningful for description and presentation. This is called the Root Mean Squared Error (or RMSE). This metric too is inverted so that the results are increasing.

3.2 Existing System

In existing system the performance was analyzed using only two machine learning algorithms such as linear regression and support vector regression, and involves single performance metric mean absolute error (MAE). It considers limited number of crops and involves minimum numbers of features.

Drawbacks of the Existing System

- ✓ Limited numbers of crops are considered
- ✓ Minimum numbers of features are considered
- ✓ Area of cultivation is not considered
- ✓ Produces poor results

3.3 Proposed System

It is necessary to build a system that better predicts the crop estimation. The important factor that affects the crop yield can also be added as additional features. For example geographical area of the field plays an important role in estimating the crop yield. Regression based algorithm k-nearest regression is also considered for prediction. Additionally the Mean Squared Error (MSE)and Root Mean Squared Error (RMSE) performance metrics are also added.

Advantages of Proposed System

- ✓ The system is expandable to add more number of records or attributes, thus it generates new significant rules
- ✓ Handles large dataset
- ✓ Additional performance metrics are added
- ✓ Error rate is minimized

3.4 About the Dataset

The crop production dataset is collected from the kaggle.com website and it contains 246091 instances with 7 attributes. The dataset contains 33 states with crop yield production details. The dataset contains both numerical values and string values. Table 3.1 shows the description of the dataset.

Table 3.1 Dataset Description

S.no	Attributes	Type
1	State name	String
2	District name	String
3	Crop name	String
4	Season	String
5	Crop year	Numerical
6	Area	Numerical
7	Production	Numerical

3.5 Modules

This project includes five major modules. Such as,

- Data collection
- Data preprocessing
- Splitting the data into training and testing
- Machine Learning Techniques
- Performance Metrics

3.5.1.Data Collection

The crop production statics dataset of India is used for the study. The crop production dataset is downloaded fromkaggle.com and loaded into spyder. The crop production dataset contains 246091 instances with 7 attributes. Such as state name, district name, crop name, year, season, area, production. (Refer Figure 8.2.1)

3.5.2 Data Preprocessing

Data preprocessing is an important technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a non-evitable steps for resolving such issues. Data preprocessing prepares raw data for further processing. (Refer Figure 8.2.2).

Steps during preprocessing:

- ✓ Data Cleaning
    - ✓ Removal null values
    - ✓ Scaling the data

3.5.3 Splitting the data into training and testing

The dataset is divides into training and testing. A common strategy is to take all available labeled data, and split it into training and evaluation subsets, usually with a ratio of 70 or 80 percent for training and 20 or 30 percent for evaluation. In our study machine

learning algorithms uses the first 70 percent of the input data in the order it appears in the source data for the training data source and the remaining 30 percent of the data for the evaluation. (Refer Figure 8.2.3)

3.5.4 Machine Learning techniques

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and learn it and updates the knowledge.

The project involves the following three machine learning algorithms and identifies better algorithm for the crop production dataset.

- ✓ Linear Regression
- ✓ K-Nearest Neighbors
- ✓ Support Vector Regression

Linear Regression

Linear Regression is a machine learning algorithm based on supervised learning. A linear regression is a statistical model that analyzes the relationship between a response variable (often called y) and one or more variables and their interactions (often called x or explanatory variables).

This best fit line is known as regression line and represented by a linear equation

$$Y = a * X + b.$$

In this equation:

- ✓ Y – Dependent Variable
- ✓ a – Slope
- ✓ X – Independent variable
- ✓ b – Intercept

There are two types of linear regression

- ✓ Simple Linear Regression
- ✓ Multiple Linear Regression

Simple Linear Regression is characterized by one independent variable. And, Multiple Linear Regression (as the name suggests) is characterized by multiple (more than 1) independent variables. Figure 3.1 shows the example of linear regression algorithm

Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables. One variable denoted x is regarded as an independent variable and other one denoted y is regarded as a dependent variable. It is assumed that the two variables are linearly related.

Multiple linear regressions basically describe how a single response variable Y depends linearly on a number of predictor variables.(Refer Figure 8.2.4)

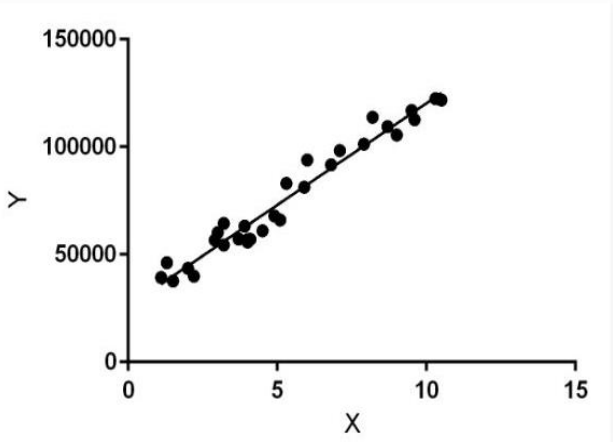


Figure: 3.1 Examples of linear regression

Advantages

- ✓ Linear regression is an extremely simple method.
- ✓ It is very easy and intuitive to use and understand.
- ✓ To find the nature of the relationship between the two variables.

Disadvantage

- ✓ Linear regression models relationships between dependent and independent variables that are linear and there is a straight-line relationship between them which is incorrect.
- ✓ Linear regression is very sensitive to the anomalies in the data (or outliers).
- ✓ A number of parameters than the number of samples available then the model start to model the noise rather than the relationship between the variables.

K-Nearest Regression

Regression based k-nearest neighbors algorithm is a very simple, easy to understand, versatile and one of the topmost machine learning algorithms. KNN used in the variety of applications such as finance, healthcare, political science, handwriting detection, image recognition and video recognition. In credit ratings, financial institutes will predict the credit rating of customers. KNN algorithm used for both classification and regression problems.

KNN is a non-parametric and lazy learning algorithm. Non-parametric means there is no assumption for underlying data distribution. In other words, the model structure determined from the dataset. This will be very helpful in practice where most of the real world datasets do not follow mathematical theoretical assumptions. Lazy algorithm means it does not need any training data points for model generation. All training data used in the testing phase. This makes training faster and testing phase slower and costlier. Costly testing phase means time and memory. In the worst case, KNN needs more time to scan all data points and scanning all data points will require more memory for storing training data. (Refer Figure 8.2.5)

In KNN, K is the number of nearest neighbors. The number of neighbors is the core deciding factor. K is generally an odd number if the number of classes is 2. When K=1, then the algorithm is known as the nearest neighbor algorithm. This is the simplest case. Suppose P1 is the point, for which label needs to predict. First, you find the one closest point to P1 and then the label of the nearest point assigned to P1. Figure 3.2 shows the example of K-Nearest Neighbors algorithm works.

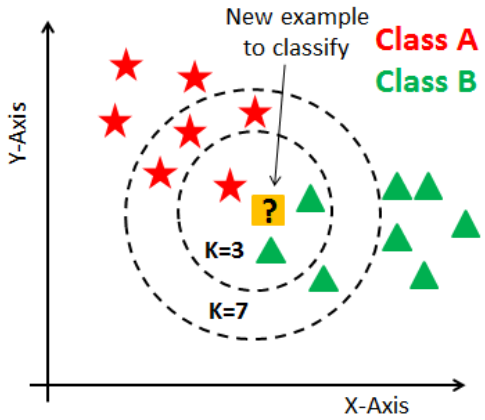


Figure: 3.2 KNN works

We can implement a KNN model by following the below steps

- ✓ Load the data
- ✓ Initialize the value of k
- ✓ For getting the predicted class, iterate from 1 to total number of training data points
- ✓ Calculate the distance between test data and each row of training data.
- ✓ Sort the calculated distances in ascending order based on distance values
- ✓ Get top k rows from the sorted array
- ✓ Get the most frequent class of these rows
- ✓ Return the predicted class

Advantages

- ✓ Robust noisy training data
- ✓ Effective if the training data is large

Disadvantages

- ✓ To determine value of parameter K (number of nearest neighbors)
- ✓ Distance based learning is not clear which type of distance to use and which attribute to use to produce the best result.

- ✓ Computation cost is quite high because we need to compute distance of each query instance to all training samples.

Support Vector Regression

Support vector machine is considered to be a classification approach, but it can be employed in both classification and regression problems. It can easily handle multiple continuous and categorical variables. SVR constructs a hyperplane in multidimensional space to separate different classes. SVR generates optimal hyperplane in an iterative manner, which is used to minimize an error. The core idea of SVR is to find a maximum marginal hyperplane that best divides the dataset into classes. Figure 3.3 shows the example of Support Vector Regression.

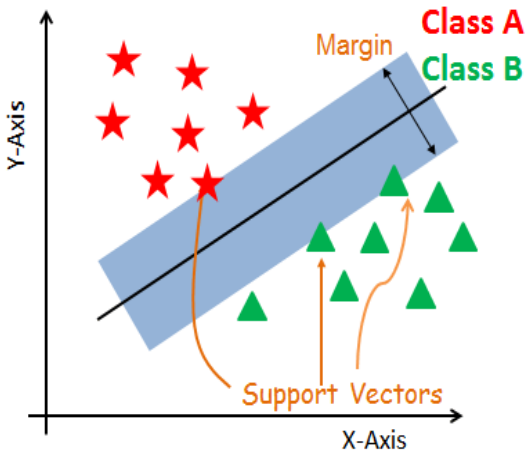


Figure: 3.3 Support Vector Regression

Support Vectors

Support vectors are the data points, which are closest to the hyperplane. These points will define the separating line better by calculating margins. These points are more relevant to the construction of the classifier.

Hyperplane

A hyperplane is a decision plane which separates between a set of objects having different class memberships

Margin

A margin is a gap between the two lines on the closest class points. This is calculated as the perpendicular distance from the line to support vectors or closest points. If the margin is larger in between the classes, then it is considered a good margin, a smaller margin is a bad margin.

The main objective is to segregate the given dataset in the best possible way. The distance between the either nearest points is known as the margin. The objective is to select a hyperplane with the maximum possible margin between support vectors in the given dataset. SVM searches for the maximum marginal hyperplane in the following steps:

- ✓ Generate hyperplanes which segregates the classes in the best way. Left-hand side figure showing three hyperplanes black, blue and orange. Here, the blue and orange have higher classification error, but the black is separating the two classes correctly.
- ✓ Select the right hyperplane with the maximum segregation from the either nearest data points as shown in the right-hand side figure. Figure 3.4 show the how does SVM works.

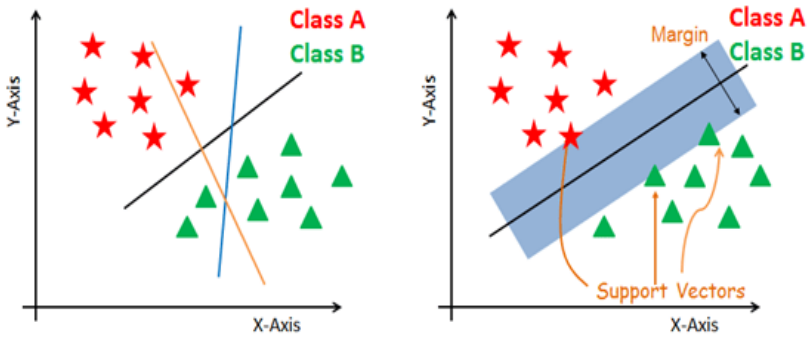


Figure 3.4 SVM works

Advantages

SVM Classifiers offer good accuracy and perform faster prediction compared to Naïve Bayes algorithm. They also use less memory because they use a subset of training points in the decision phase. SVM works well with a clear margin of separation and with high dimensional space.

Disadvantages

SVR is not suitable for large datasets because of its high training time and it also takes more time in training when compared to other algorithms. It works poorly with overlapping classes and is also sensitive to the type of kernel used. (Refer Figure 8.2.6)

3.5.5 Performance Metrics

The next step after implementing a machine learning algorithm is to find out how effective is the model based on metric for the given dataset. Different performance metrics are used to evaluate the machine learning algorithms. If the machine learning model is trying to predict, then RMSE (root mean squared error), MSE (mean squared error) and MAE (mean absolute error) can be used to calculate the efficiency of the model. The efficiency of the machine learning algorithms are evaluated using three performance metrics namely mean absolute error , mean squared error, and root mean squared error.

Mean Absolute Error (MAE)

Mean Absolute Error is the average of the difference between the original values and the predicted values. It gives us the measure of how far the predictions were from the actual output. It will not provide direction of the error i.e. whether we are under predicting the data or over predicting the data. As with the mean absolute error of zero indicates no error. The Figure 8.2.7, Figure, 8.2.8, Figure.8.2.9 shows the Mean Absolute Error of three machine learning algorithms.

$$\text{MAE} = \frac{1}{n}\sum|y - \hat{y}|$$

$\frac{1}{n}$ - Divide by the total number of data points

$\sum$ - Sum of

y - Actual output values

$\hat{y}$ - Predicted output values

$|y - \hat{y}|$- The absolute values of the residual

The picture below is a graphical description of the MAE. The green line represents our model's predictions, and the blue points represent our data.

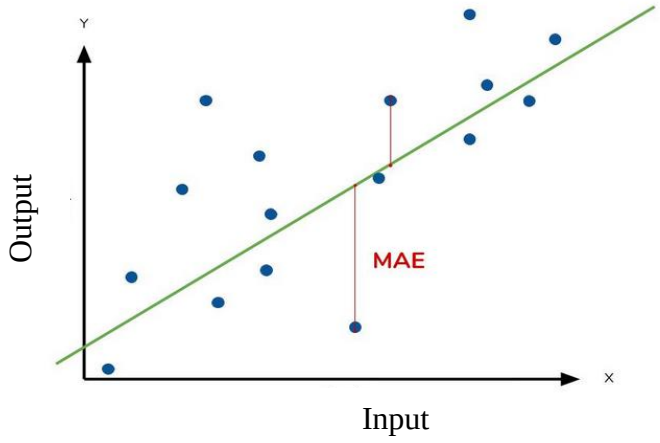


Figure 3.5 Mean Absolute Error

Mean Squared Error (MSE)

Mean Squared Error (MSE) is quite similar to mean absolute error, the only difference being that MSE takes the average of the squared of the difference between the original values and the predicted values. As, we take square of the error, the effect of larger errors become more pronounced then smaller error, hence the model can now focus more on the larger errors. As with the mean squared error of zero indicates no error. The Figure 8.2.7, Figure, 8.2.8, Figure.8.2.9 shows the Mean Squared Error of three machine learning algorithms.

$$\text{MSE} = \frac{1}{n}\sum(y - \hat{y})^2$$

$\frac{1}{n}$ - Divide by the total number of data points

$\sum$ - Sum of

y - Actual output values

$\hat{y}$ - Predicted output values

$(y - \hat{y})^2$- The square of the difference between actual and predicted

The following picture graphically demonstrates what an individual residual in the MSE might look like

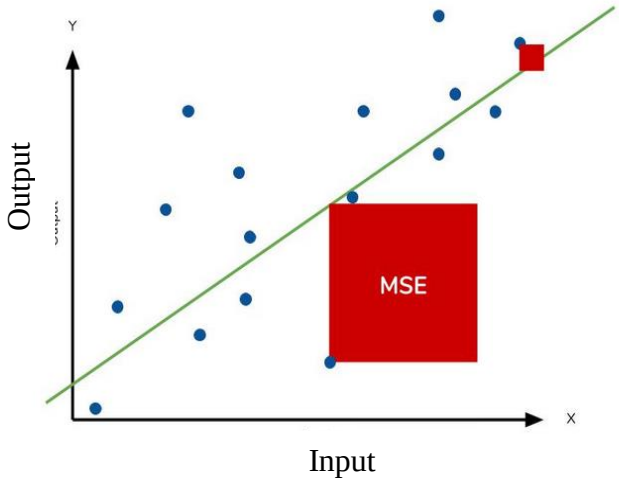


Figure 3.6 Mean Squared Error

Root Mean Squared Error (RMSE)

The most commonly used metric for regression tasks is RMSE (Root Mean Square Error). This is defined as the square root of the average squared distance between the actual score and the predicted score. Root Mean Squared Error can be transformed back into the original units of the predictions by taking the square root of the mean squared error score. This is called the root mean squared error, or RMSE. As with the mean squared error, an RMSE of zero indicates no error. The Figure 8.2.7, Figure, 8.2.8, Figure.8.2.9 shows the Root Mean Squared Error of three machine learning algorithms.

$$RMSE=\sqrt{MSE}$$

Experimental Results and Discussion

4. EXPERIMENTAL RESULTS AND DISCUSSION

The crop production dataset is collected from the kaggle.com website and it contains 246091 instances with 7 attributes namely state name, district name, crop name, season, year, area, production. The dataset contains 33 states with crop yield production details. Machine learning algorithms are used predict crop production namely linear regression, k-nearest regression, support vector regression. Different performance metrics are used to evaluate the machine learning algorithms such as mean absolute error, mean squared error, and root mean squared error.

Mean Absolute Error is the average of the difference between the original values and the predicted values. It gives us the measure of how far the predictions were from the actual output.

Mean Squared Error (MSE) is quite similar to mean absolute error, the only difference being that MSE takes the average of the squared of the difference between the original values and the predicted values. As, we take square of the error, the effect of larger errors become more pronounced then smaller error, hence the model can now focus more on the larger errors.

Root Mean Squared Error can be transformed back into the original units of the predictions by taking the square root of the mean squared error score. This is called the root mean squared error, or RMSE. As with the mean squared error, an RMSE of zero indicates no error.

The following Table 4.1 shows the performance of the three machine learning algorithms in terms of MAE, MSE, and RMSE.

Table 4.1 Comparisons of three machine learning algorithms

Machine Learning Techniques	Performance Metrics		
	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	Root Mean Squared Error(RMSE)
Linear Regression	0.0855	0.8397	0.9163
K-Nearest Regression	***0.0157***	***0.2244***	***0.4737***
Support Vector Regression	0.0640	0.8242	0.9078

Generally error refers to the difference between the actual value and predicted values for given instance. The error value will be a non-negative integer and ranges between zero and one, the error would be zero for a perfect model. The error value close to zero represents a better prediction

The comparison of these three performance metrics shows that k-nearest regression performs better with minimum error rate. The result shows that k-nearest regression algorithm performs better with low MAE, MSE, and RMSE values when compared to other two machine learning algorithms linear regression and support vector regression. Ref (Figure 8.2.10)

Conclusion

5. CONCLUSION

The project evaluates the performance of regression based models to predict crop production. Three machine learning algorithms namely linear regression, k-nearest regression, and support vector regression are applied on the crop production dataset. The performances of the algorithms are analyzed based on three metrics, such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). As the error decreases it shows the increased performance of the system. The result shows that k-nearest regression algorithm performs better with low MAE, MSE, and RMSE values when compared to other two machine learning algorithms linear regression and support vector regression.

Scope for Future Enhancement

6. SCOPE FOR FUTURE ENHANCEMENT

The work can be future enhanced by including the other machine learning algorithms namely logistic regression, decision tree, random forest algorithm etc. It can also be future enhanced by adding performance metrics such as relative absolute error, relative mean squared error, relative squared error, and root relative squared error. The performance of the algorithms can be future improved by considering additional features for the crop dataset.

Bibliography

7. BIBLIOGRAPHY

Journal References

- ✓ International Journal of Advance Engineering and Research Development Special Issue on Recent Trends in Data Engineering Volume 4, Special Issue 5, Dec.-2017 ,Crop Prediction System using Machine Learning.
- ✓ International Research Journal of Engineering and Technology (IRJET) Volume: 05 Issue: 06 | June-2018, EFFICIENT CROP YIELD PREDICTION USING MACHINE LEARNING ALGORITHMS.
- ✓ International Research Journal of Engineering and Technology (IRJET) Volume: 05 Issue: 02 | Feb-2018, Prediction of Crop Yield using Machine Learning.
- ✓ International Journal of Innovations & Advancement in Computer Science IJIACS ISSN 2347 – 8616 Volume 7, Issue 4 April 2018, Application of Data Mining Techniques for Prediction of Crop Production in India.
- ✓ International Journal on Engineering Technology and Sciences – IJETSVolume III, Issue III, March- 2016, CROP YIELD PREDICTION

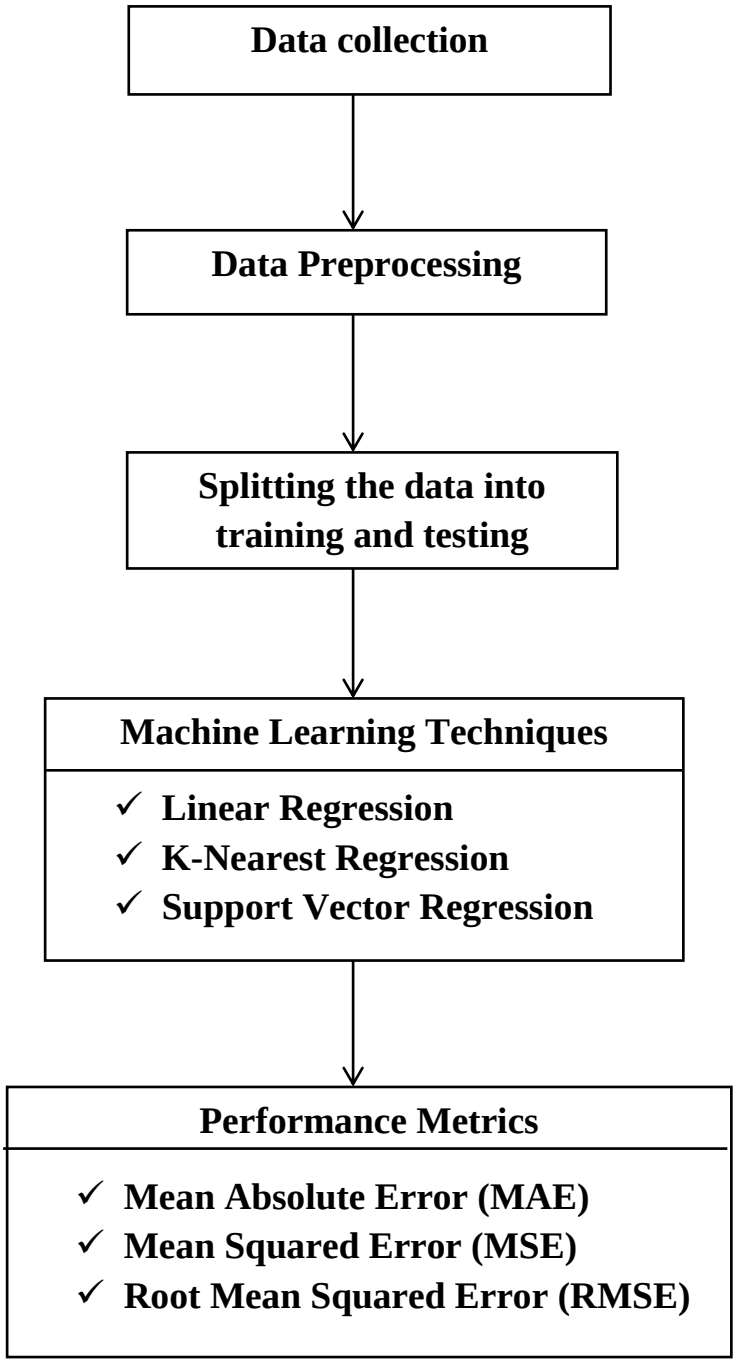
Website References

- ✓ https://www.researchgate.net/publication/263368516_Predictive_ability_of_machine_learning methods_for_massive_crop_yield_prediction
- ✓ https://www.python-course.eu/k_nearest_neighbor_classifier.php
- ✓ https://datascienceplus.com/linear-regression-in-python-predict-the-bay-areas-home-prices/
- ✓ https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python
- ✓ https://datascienceplus.com/linear-regression-in-python-predict-the-bay-areas-home-prices/
- ✓ http://www.datacamp.com/community/tutorials/naive-bayes-scikit-learn
- ✓ https://www.kaggle.com/abhiseklewan/crop-production-statistics-from-2000-in-india

Appendix

8. APPENDIX

8.1 Workflow diagram



8.2 Screen shots

Data Collection

sf - DataFrame

Index	State_Name	District_Name	Crop_Year	Season	Crop	Area	Production
207	Andhra Pradesh	ANANTAPUR	1997	Kharif	Dry chillies	3700	7100
208	Andhra Pradesh	ANANTAPUR	1997	Kharif	Groundnut	650800	228400
209	Andhra Pradesh	ANANTAPUR	1997	Kharif	Horse-gram	3300	1000
210	Andhra Pradesh	ANANTAPUR	1997	Kharif	Jowar	10100	10200
211	Andhra Pradesh	ANANTAPUR	1997	Kharif	Korra	2200	700
212	Andhra Pradesh	ANANTAPUR	1997	Kharif	Maize	2800	4900
213	Andhra Pradesh	ANANTAPUR	1997	Kharif	Moong(Green Gram)	1300	500
214	Andhra Pradesh	ANANTAPUR	1997	Kharif	Other Kharif pulses	800	100
215	Andhra Pradesh	ANANTAPUR	1997	Kharif	Ragi	6700	11800
216	Andhra Pradesh	ANANTAPUR	1997	Kharif	Rice	35600	75400
217	Andhra Pradesh	ANANTAPUR	1997	Kharif	Sugarcane	700	72900
218	Andhra Pradesh	ANANTAPUR	1997	Kharif	Sunflower	35900	11100
219	Andhra Pradesh	ANANTAPUR	1997	Kharif	Tobacco	200	200
220	Andhra Pradesh	ANANTAPUR	1997	Rabi	Dry chillies	100	100
221	Andhra Pradesh	ANANTAPUR	1997	Rabi	Gram	28000	14800
222	Andhra Pradesh	ANANTAPUR	1997	Rabi	Groundnut	20200	21700
223	Andhra Pradesh	ANANTAPUR	1997	Rabi	Horse-gram	600	200
224	Andhra Pradesh	ANANTAPUR	1997	Rabi	Jowar	18800	9400
225	Andhra Pradesh	ANANTAPUR	1997	Rabi	Korra	100	100
226	Andhra Pradesh	ANANTAPUR	1997	Rabi	Maize	600	2400
227	Andhra Pradesh	ANANTAPUR	1997	Rabi	Ragi	600	1000

Format Resize Background color Column min/max

Figure 8.2.1 Load the Dataset

Data Preprocessing

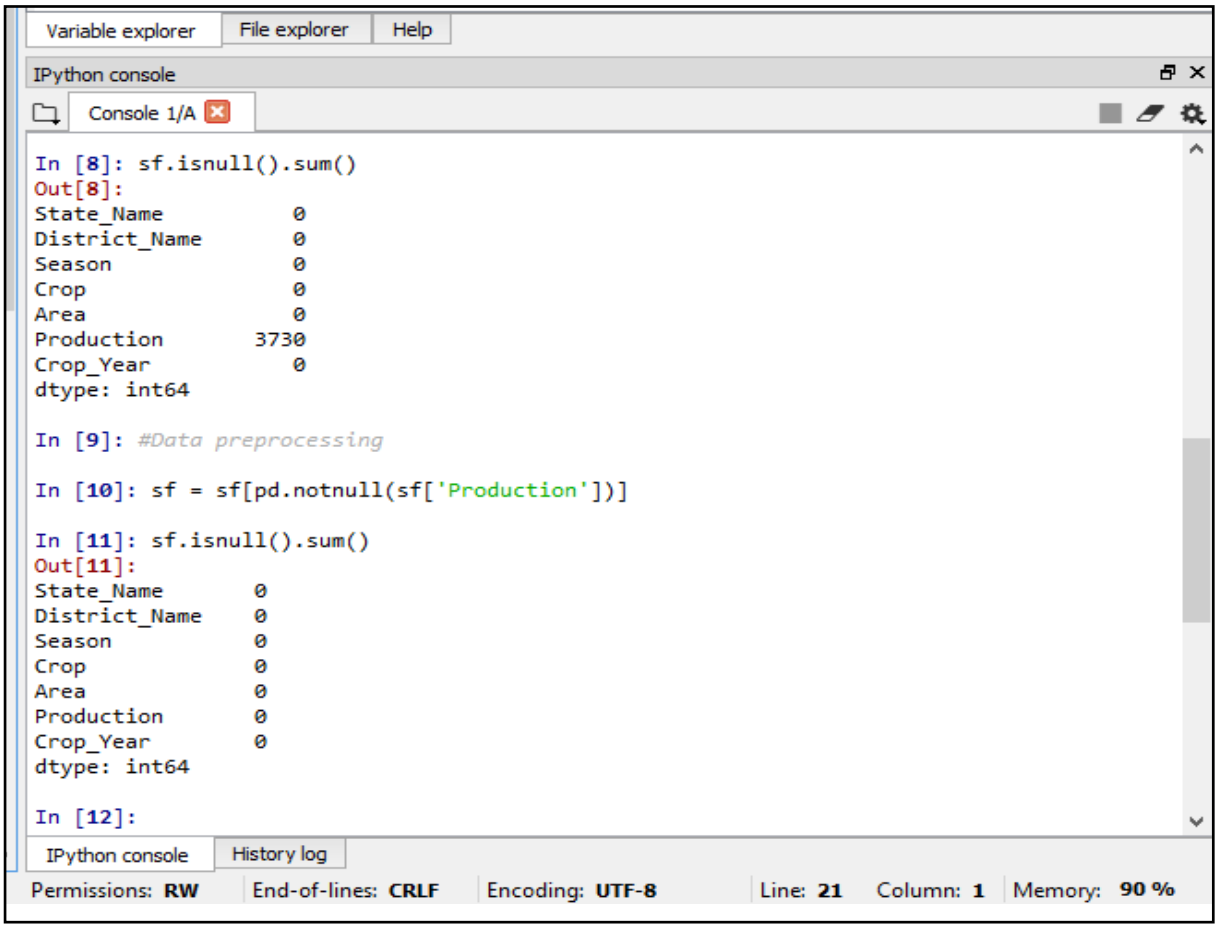


Figure 8.2.2 Data Preprocessing

Splitting the data into training and testing

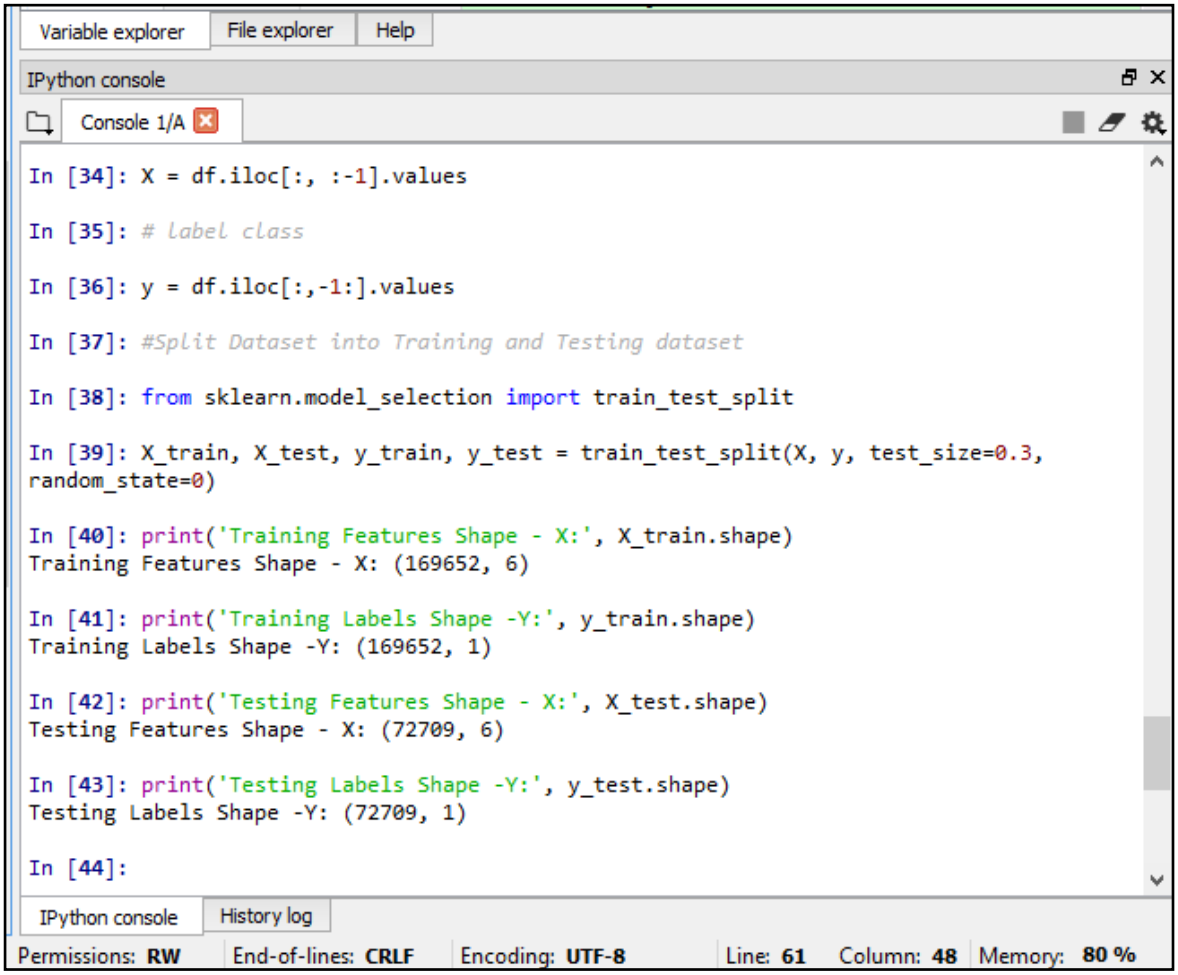


Figure 8.2.3 Splitting the data into training and testing

Applying Machine Learning Algorithms

(I)Linear Regression Algorithm

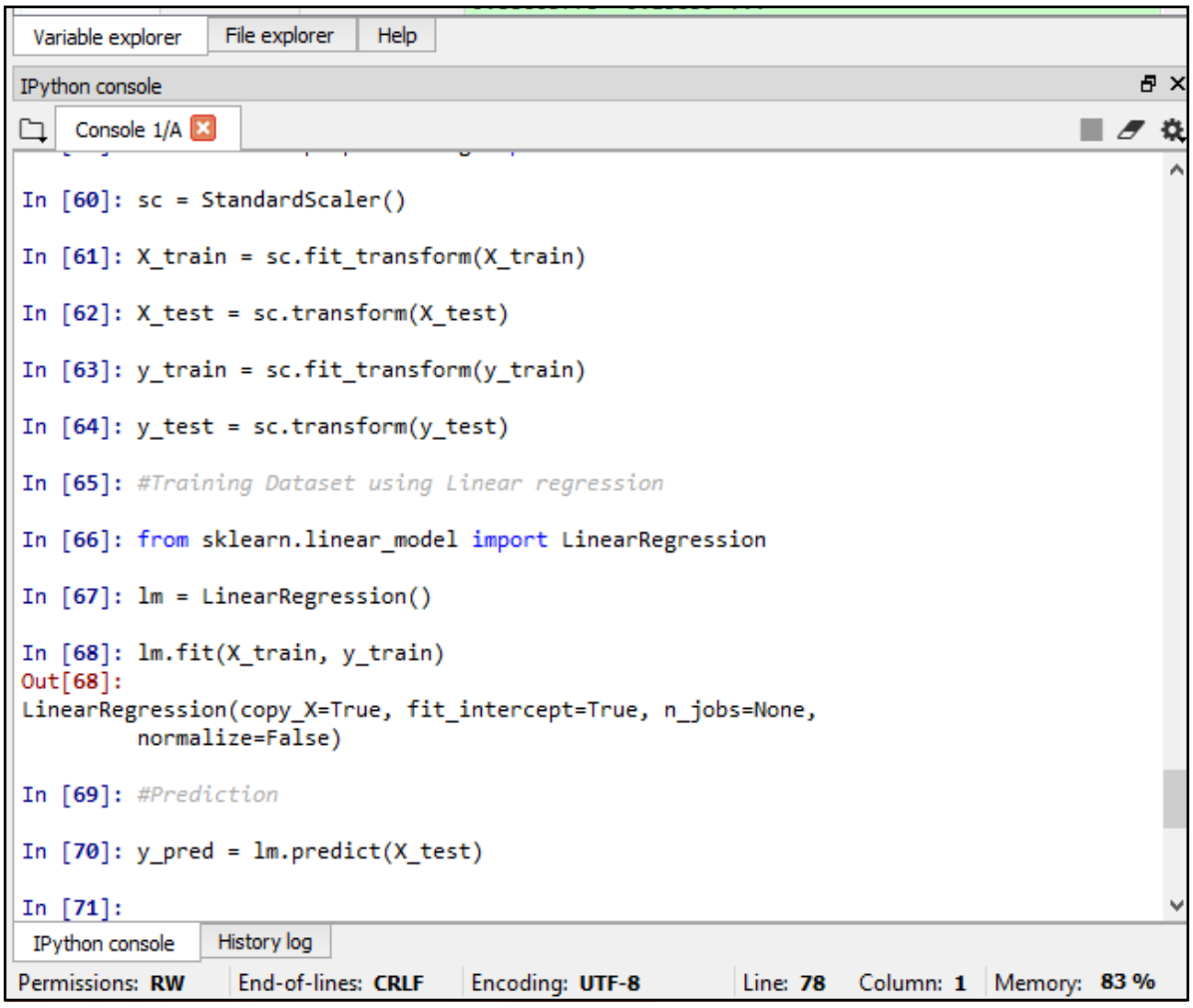


Figure 8.2.4 Linear Regression

(II)K-Nearest Regression Algorithm

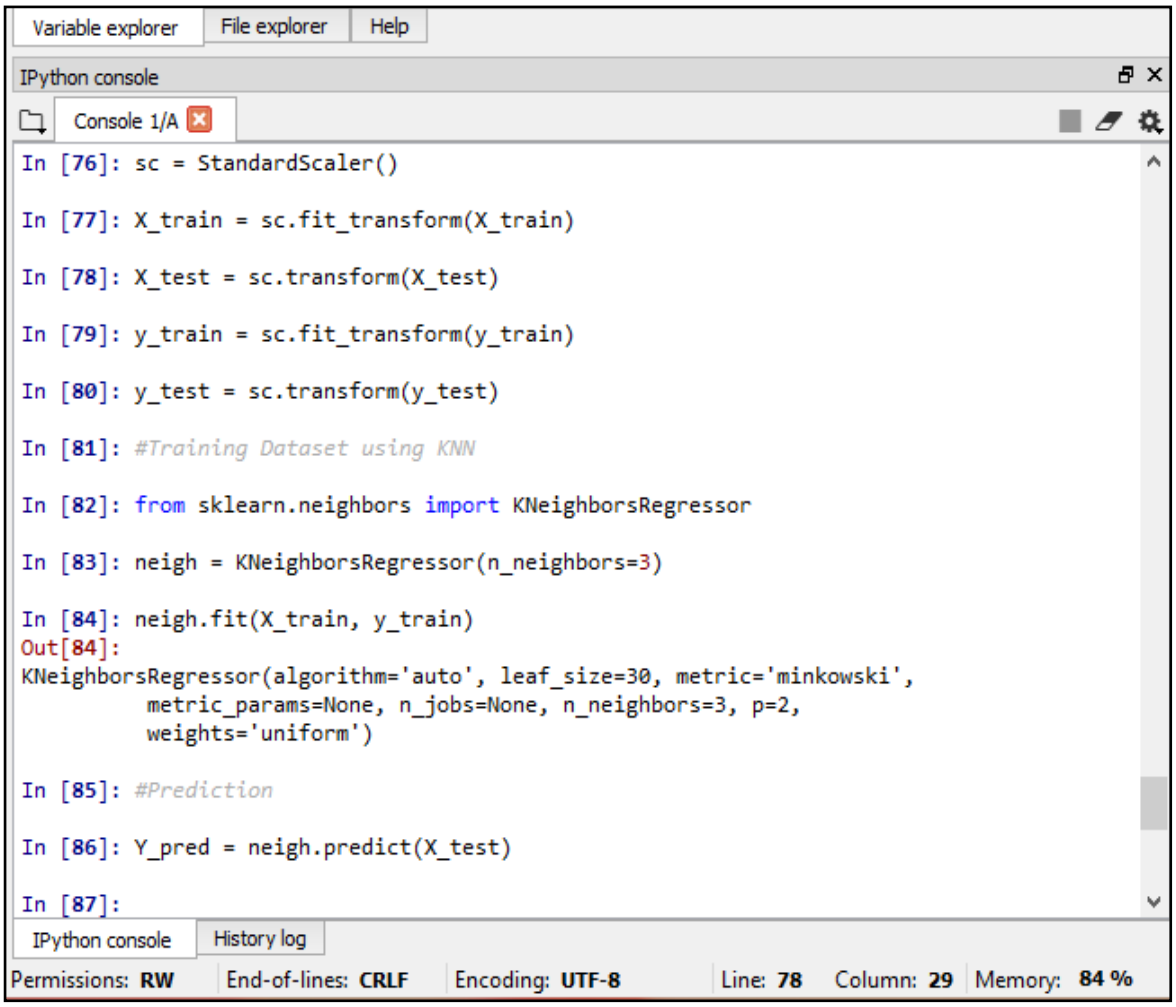
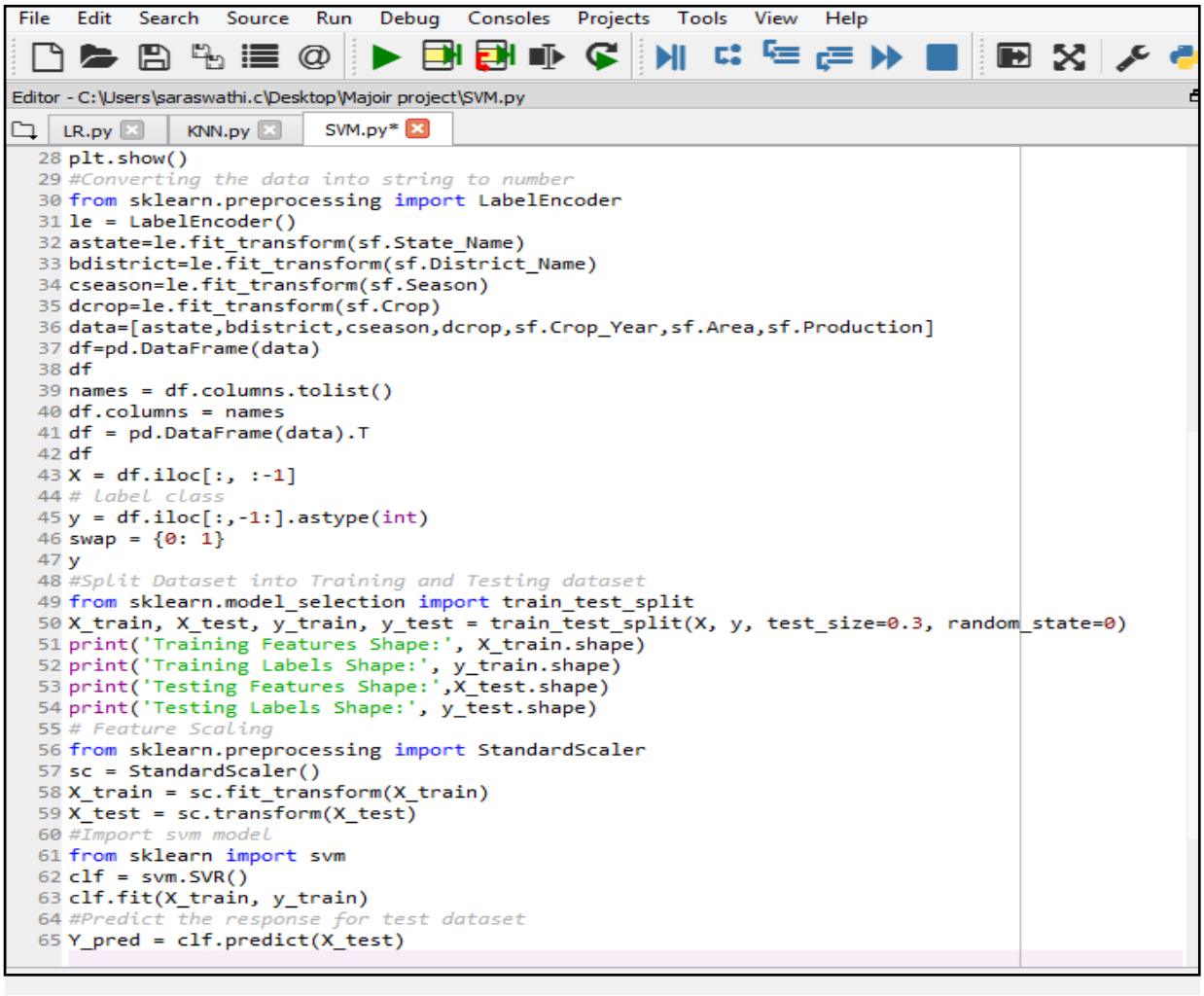


Figure 8.2.5 K-Nearest Regression

(III)Support Vector Regression Algorithm




```
plt.show()
#Converting the data into string to number
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
astate=le.fit_transform(sf.State_Name)
bdistrict=le.fit_transform(sf.District_Name)
cseason=le.fit_transform(sf.Season)
dcrop=le.fit_transform(sf.Crop)
data=[astate,bdistrict,cseason,dcrop,sf.Crop_Year,sf.Area,sf.Production]
df=pd.DataFrame(data)
df
names = df.columns.tolist()
df.columns = names
df = pd.DataFrame(data).T
df
X = df.iloc[:, :-1]
# label class
y = df.iloc[:,-1:].astype(int)
swap = {0: 1}
y
#Split Dataset into Training and Testing dataset
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)
print('Training Features Shape:', X_train.shape)
print('Training Labels Shape:', y_train.shape)
print('Testing Features Shape:',X_test.shape)
print('Testing Labels Shape:', y_test.shape)
# Feature Scaling
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
#Import svm model
from sklearn import svm
clf = svm.SVR()
clf.fit(X_train, y_train)
#Predict the response for test dataset
Y_pred = clf.predict(X_test)
```

Figure 8.2.6 Support Vector Regression

Applying Performance Metrics

(I)Linear Regression Algorithm

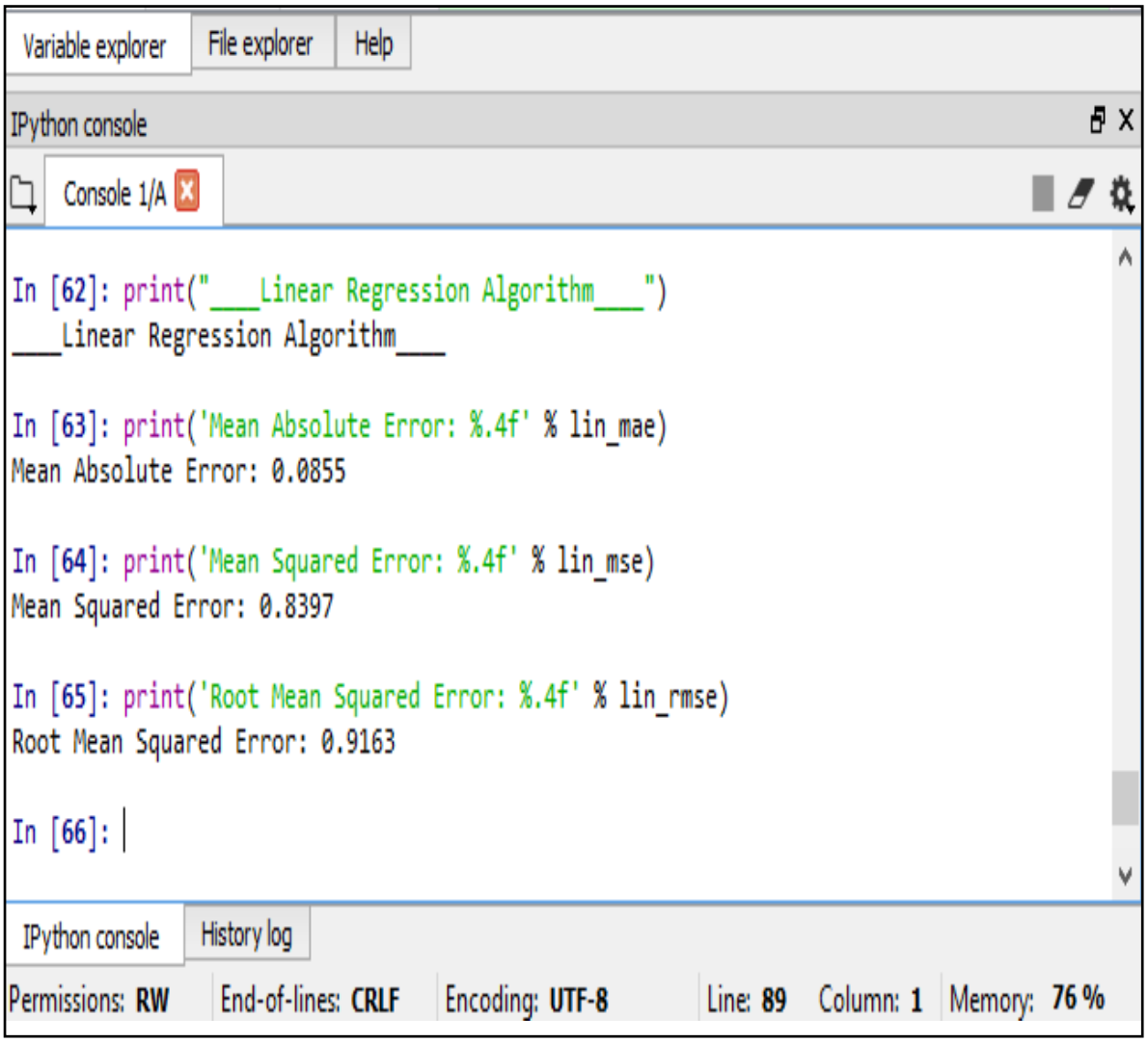


Figure 8.2.7 Performance Metrics of Linear Regression algorithm

(II) K-Nearest Regression Algorithm

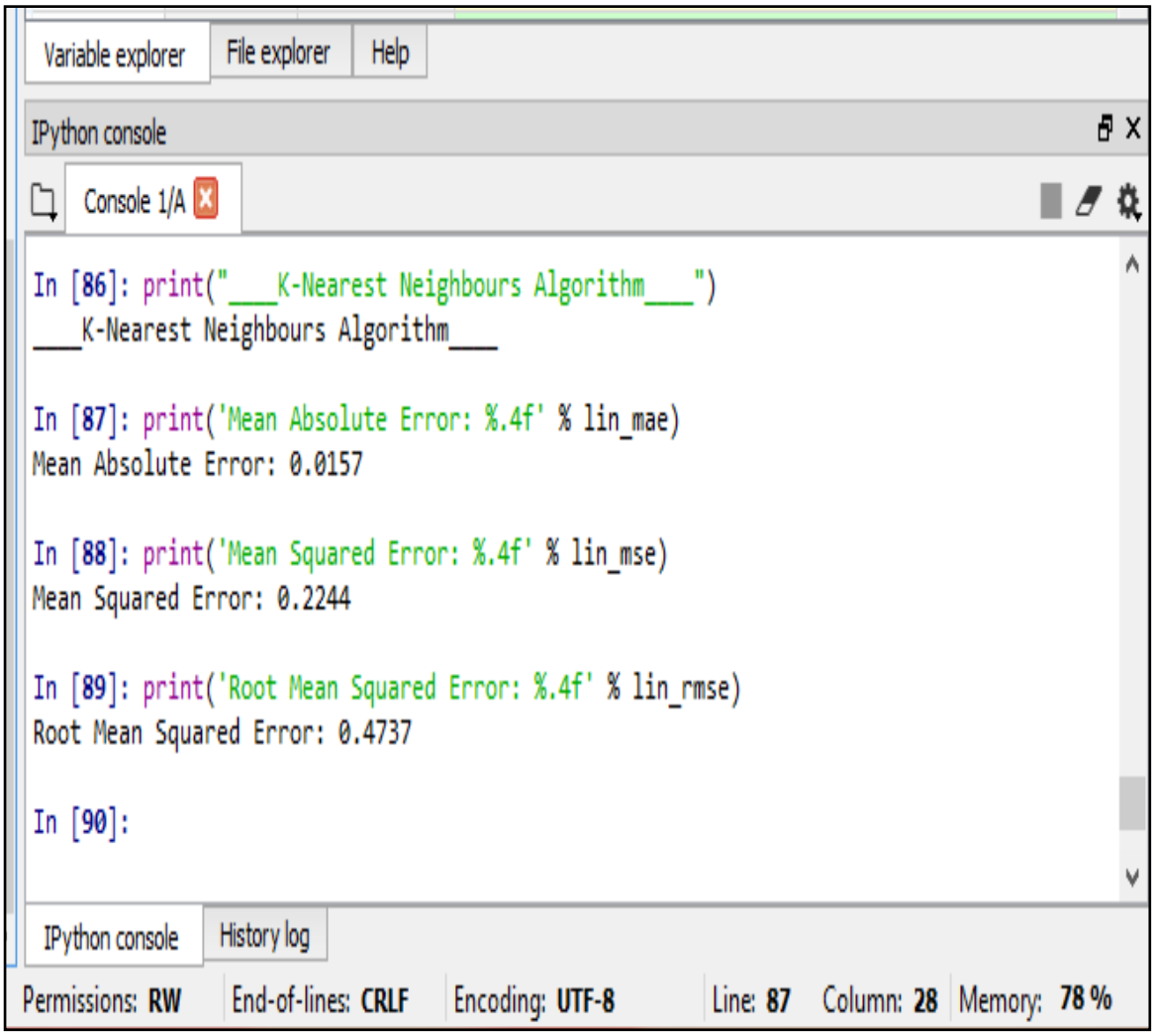


Figure 8.2.8 Performance metrics of K-Nearest Regression Algorithm

(III)Support Vector Regression Algorithm

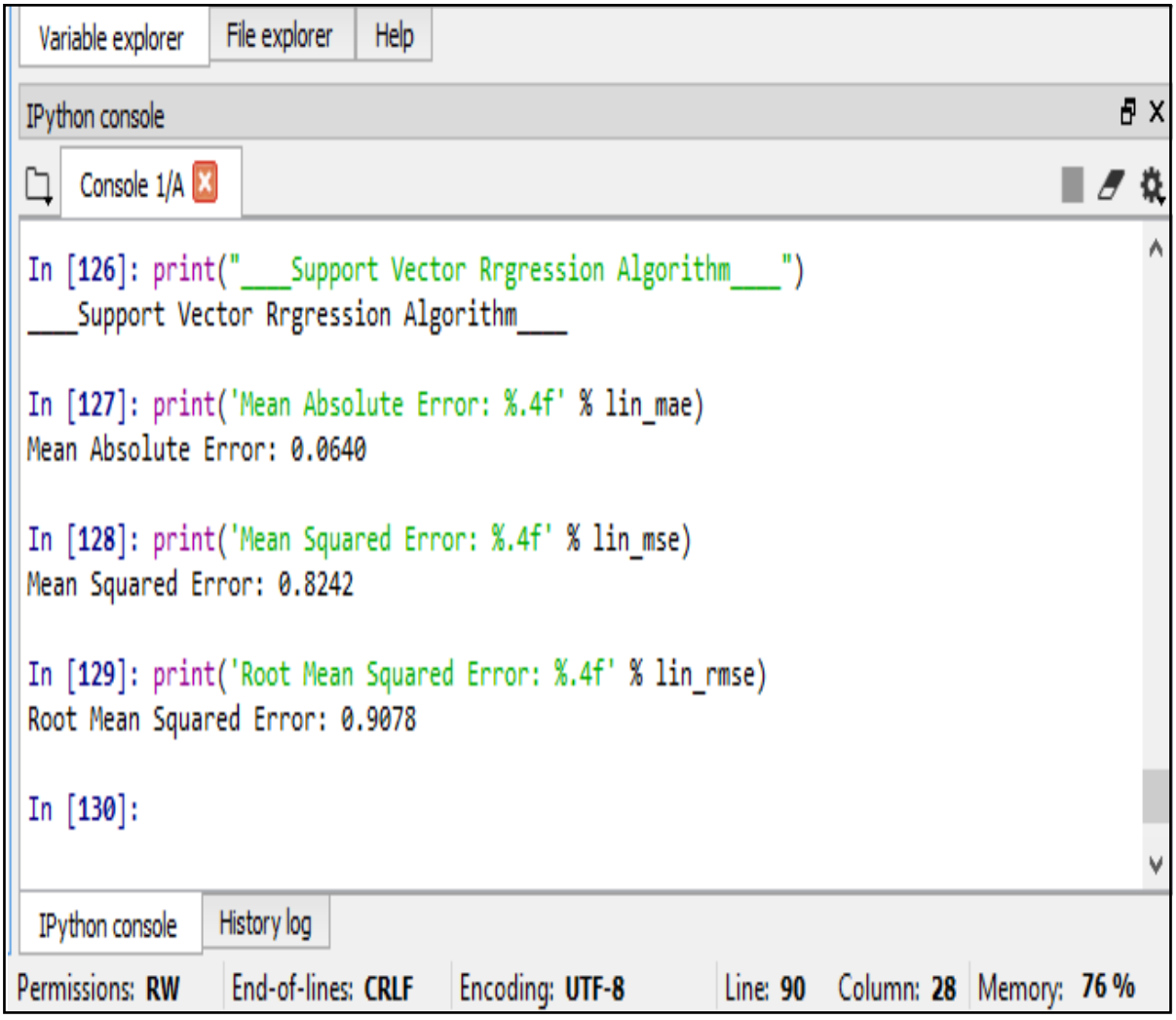


Figure 8.2.9 Performance metrics of Support Vector Regression Algorithm

Comparisons of three algorithms

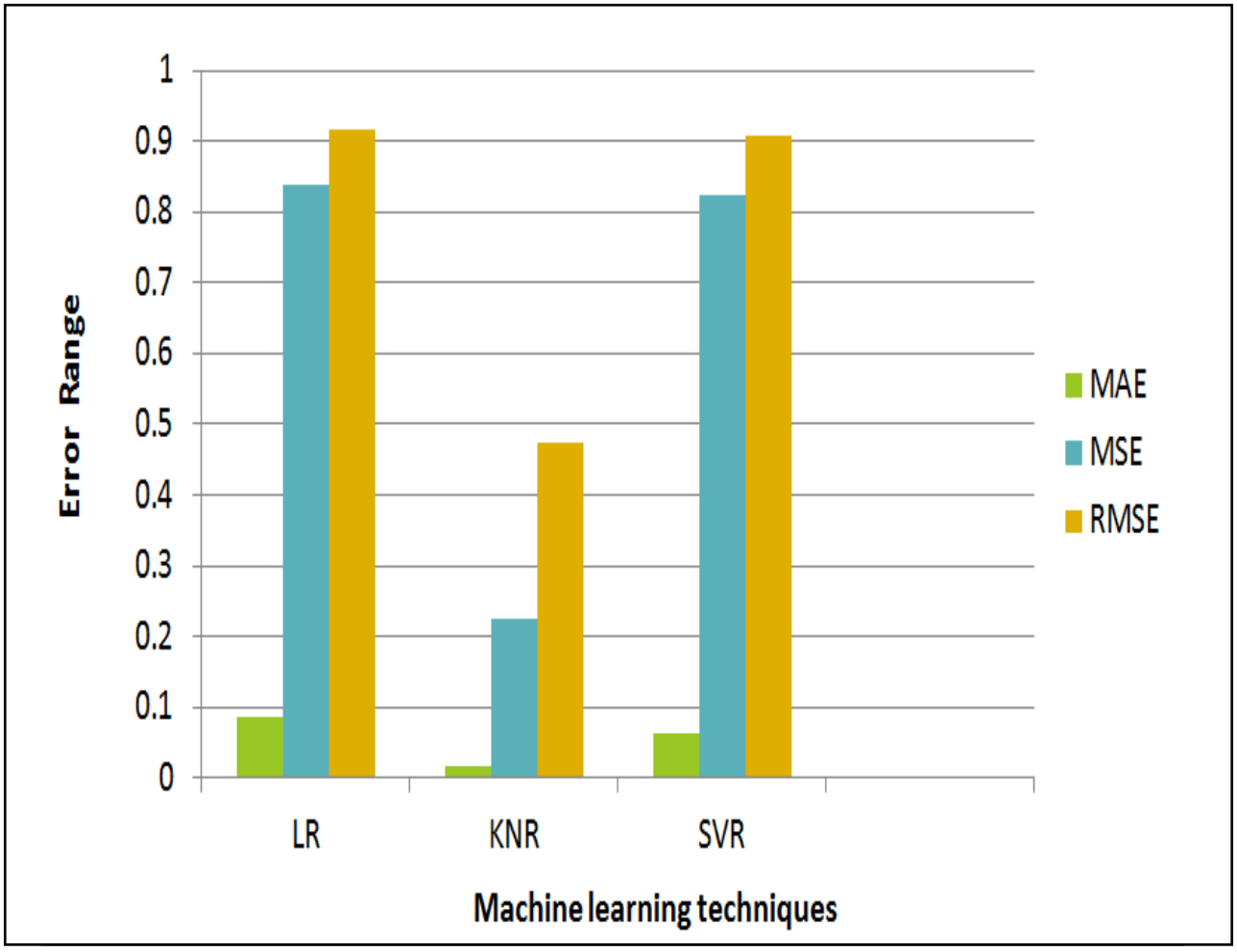


Figure 8.2.10 Comparison of three algorithms Results