Predictive Analysis On Agricultural Data Using Pandas Framework

Mrs.Suvitha Vani.P[1], Priyadharshini.R[2], Priyadarshini.G[3], Shobana Devi.k[4]
[1]Assistant Professor, Dept of Computer Science and Engineering
[2, 3, 4]Dept of Computer Science and Engineering
[1, 2, 3, 4] Sri Shakthi Institute of Engineering and Technology, Coimbatore

Abstract- *Agricultural crop production depends on various factors such as climate,geography and economy.Several factors have different impacts on agriculture, which can be quantified using appropriate statistical methodologies.Applying such methodologies and techniqueson historical data yield of crops,it is possible to obtain information or Knowledge which can be helpful for formers and government organization to make better decision and policies which lead to increased production.Developing better techniques to predict crop productivity in different climatic condition can assist farmers and other stakeholders in better decision making in terms of agronomy and crop choice.*

Keywords-Predict,over-fitting,modelling, classification,Pruning.

I. INTRODUCTION

Data Mining is also known as Knowledge Discovery in Database(KDD).It is a powerful technology and helps us in extraction of hidden predictive information from large database and converts large dataset to understandable structure.Here large dataset involves the methods at the intersection of machine learning, statistics, and database systems.The prediction algorithms such as Linear Regression,SVM and Decision Tree are used for the prediction of agricultural data for smaller datasets.

II. EXISTING SYSTEM AND PROPOSED SYSTEM

In Machine Learning Algorithm,prediction can be done accurately only for the large datasets.In the proposed system,we can build efficient predictive model to find best solution among Machine Learning Algorithm for Smaller Datasets.

III. LEARNING PANDAS FRAMEWORK AND DATA COLLECTION

Pandas is open source,Python Package well suited for the Tabular data for heterogenous-typed column(such as Excel),any other form of observational/statistical dataset.It is fast,flexible and expressive data structure to work with relational or labelled data.

IV. LINEAR REGRESSION

It is a linear approach for modelling the relationship between scalar dependent variable Y and independent variable X.It does a set of predictor variable to do a good job in predicting the outcomes.The simplest form of regression equation with one depenndent and one independent variable is defined from the formula,

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

Y=a+b*X

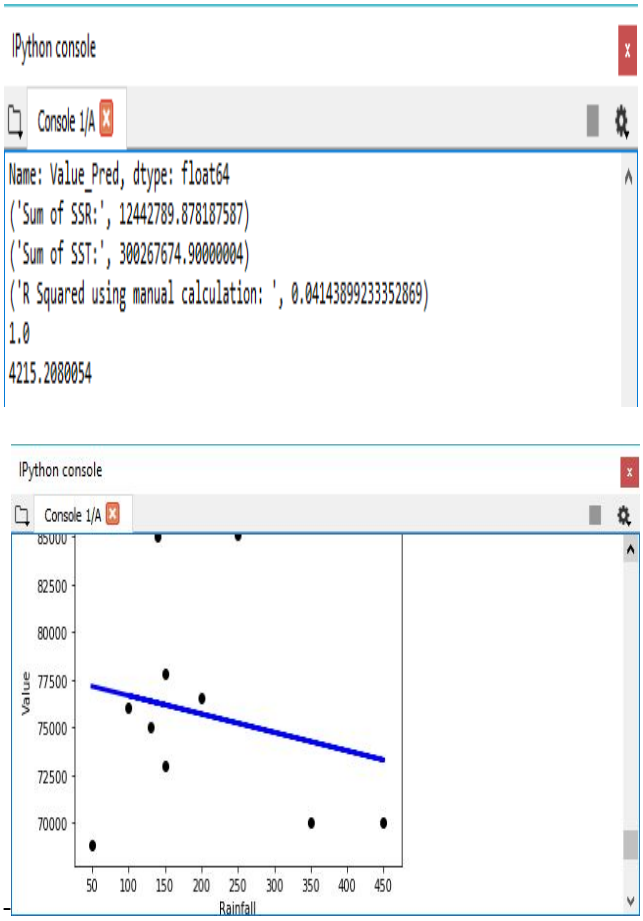
Where,
X-independent variable
Y-dependent variable
a-constant
b-regression coefficient
The above equation is to find the intercept.

LEAST-SQUARES REGRESSION:

The most common method for fitting a regression line is the method of least-square.It calculates for the observed data by minimizing the sum of squares of vertical deviation from each data point to line.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\,n\Sigma x^2 - (\Sigma x)^2\,][\,n\Sigma y^2 - (\Sigma y)^2\,]}}$$

The value of r is calculated is cor-relation coefficient.The square of cor-realtion coefficient is coefficient of determination.The diagrammatic representation of the Linear Regression are as follows:



V. SUPPORT VECTOR MACHINE

A discriminative classifier formally defined by seperating hyperplane is called SVM. In Machine learning algorithm,SVM are supervised learning algorithm model associated learning algorithms that analyze data used for regression and classification analysis.Mostly used in classification problem,where we can plot every dataset in n-dimensional space with the value of being featured every other value of particular co-ordinate.Here we perform classification by finding hyper-plane that helps to differentiate two classes.

Working:

1. It runs entirely inside the database, such that guarantees about the consistency as well as the security of the data can be given.
2. It does use as little main memory as needed for an efficient implementation. In particular, it does not duplicate the complete example set in its memory space.
3. It uses standard interfaces to access the database, so that it is database independent.
4. The evaluation of the decision function on new examples is as easy as possible.
5. It is as efficient as possible.

The objective of SVM is to find optimal seperating hyperplane by which the margin of training data maximizes,

- It correctly classifies the training data
- It is the one which will generalize better with unseen data.

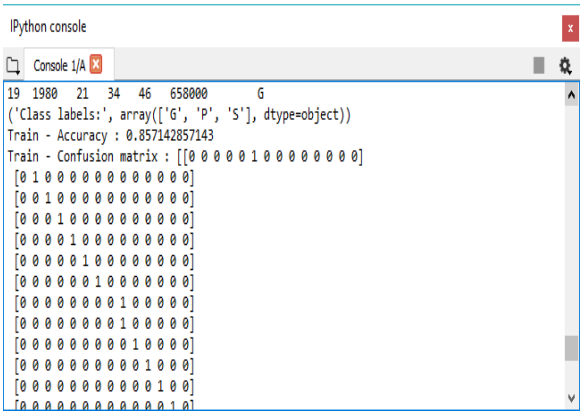
A set of objects having different class memberships is seperated by a decision plane.the concept of decision planes that define decision boundaries are based by SVM.A good hyperplane separation is achieved by which has the largest distance to the nearest training-data point of any class,in general lower the classifier of generalization errorthe larger the margin. The initial form of SVMs is a binary classifier where the output of learned function is either positive or negative. A multiclass classification can be implemented by combining multiple binary classifiers using pairwise coupling method. . A linear classifier separates them with an hyperplane.Two groups of data and separating hyperplanes that are lines in a two-dimensional space. There are many linear classifiers that correctly classify (or divide) the two groups of data. In order to achieve maximum differentiate between the two classes, SVM picks the hyperplane which has the largest margin. The margin is the summation of the shortest distance from the separating hyperplane to the nearest data point of both categories. “Unseen” or testing data points are correctly classified by the hyperplane. The kernel trick is helpful by allowing the absence of the exact formulation of mapping function which could cause the issue of curse of dimensionality. This makes a linear classification in the new space (or the feature space) equivalent to nonlinear classification in the original space (or the input space). By mapping input vectors to a higher dimensional space (or feature space) where a maximal separating hyperplane is constructed.

Advantages

It is regularisation parameter,makes user to avoid over fitting.It is effective in high dimensional spaces. It works well with clear margin of separation.

Disadvantages

When the dataset is large it does not perform well.When there is noise in the data the algorithm does not perform well.The problem of over-fitting from optimising the parameters to model selection in a way the SVM moves. Kernel models are quite sensitive to over-fitting the model selection criterion.



VI. DECISION TREE

In Decision Tree, each non-leaf node denotes a test on attributes,and each leaf node holds the class label which is tree like structure or graph.The topmost node is the root node.The objective of decision tree is to predicts value of a target variable based on several input variable by creating models.Mostly used in classification problems because it is a type of supervised learning algorithm. Both continous and categorical model will work for input and output variable. we will split the sample into two or more homogenous sets(sub-population) are differentiator/splitter in input variable.Classification tree or Regression tree are more descriptive name for tree model.In data mining,decision will not not describe decision but it describes the data.The resulting classification tree will be an input for decision making.There are many specific algorithm for decision tree,

- ID3 (Iterative Dichotomiser 3)
- C4.5 (successor of ID3)
- CART (Classification And Regression Tree)
- CHAID (CHi-squared Automatic Interaction Detector). Performs multi-level splits when computing classification trees.
- MARS: extends decision trees to better handle numerical data.

The associated class label is unknown is given a tuple X, the attribute values of the tuple are tested against decision tree. A path is traced from the root to a leaf node, which holds the class prediction for that tuple. Construction of decision tree classifiers is useful for decision tree because it does not require any domain knowledge. It can handle hi-dimensional data. The learning and classification steps of decision tree induction are simple and fast. Their representation of acquired knowledge in tree form is easy to assimilate by users. Decision tree classifiers have good accuracy.

Two type of Decision tree:

1. Continous Variable Decision Tree:
   The Tree which has continous target variable then it is continous variable decision tree.
2. Categorical Variable Decision Tree:
   The tree which has categorical variable then it is categorical variable decision tree.

Important Terminologies:

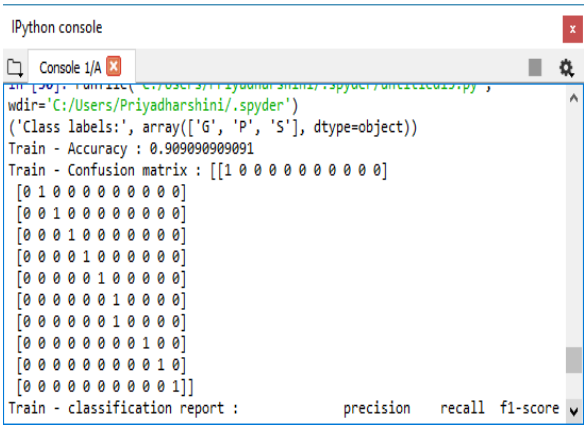
i. Root Node-It represents entire population and future divides into two or homogenous sets.
ii. Splitting: Dividing nodes into two or more sub-nodes.
iii. Decision node: When sub-nodes spilts future into sub-nodes
iv. Leaf/Terminal node: Nodes don't split
v. Pruning: when we remove the sub-nodes of a decision tree.

Advantage:

It is easy to understand for the person who is non-analytical background. It requires less data cleaning when compared to other techniques. Data type is not constrain, it handle both numerical and categorical variables.

Disadvantage:

Overfitting is difficult for decision tree models. Decision tree loses information when working with continous numerical variable, when it categorizes different category.



VII. COMPARISON

In this paper the various classification algorithm like decision tree, linear regression and support vector machine are applied to datasets to predict the accuracy of the algorithm. The accuracy is predicted using the confusion matrix. The accuracy varies according to the algorithm.

Confusion matrix for SVM:

[0 0 0 0 0 0 0 1 0 0]
[0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 0 0 0 0 1]
[0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 0 0 0 0 0]
[0 0 0 1 0 0 0 0 0 0]
[0 0 0 1 0 0 0 0 0 0]
[0 0 0 0 0 0 0 0 0 0]
[0 1 0 0 0 0 0 0 0 0]
[0 0 0 0 0 0 0 0 0 0]

Confusion Matrix for DT:

[0 0 0 0 1 0 0 0 0 0]
[0 0 0 0 0 1 0 0 0 0]
[0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 0 0 0 0 1]
[0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 0 0 0 0 0]
[0 0 0 0 0 1 0 0 0 0]
[0 0 0 0 1 0 0 0 0 0]
[0 0 0 0 0 0 0 0 0 0]

Algorithm	Accuracy
SVM	86
Decision tree	91
Linear Regression	43

VIII. CONCLUSION

The decision tree is better than SVM.A decision tree is a very typical example of an algorithm for learning from training examples. Here values of source data attributes are used to construct some decision function, evaluate model parameters, and so on.The algorithm itself that automatically builds a prediction model based on source data.Any assumptions of linearity in the data is not required by Decision tree. Simply we can say nonlinear relationships between parameters do not affect tree performance.

REFERENCES

[1] A Brief survey of Data Mining Techniques Applied to Agricultural Data Hetal Patel Research Scholar Charusat, Changa Dharmendra Patel Assistant Professor Charusat, Chang
[2] Data Mining Techniques Used In Agriculture Dr. Devesh Katiyar, Department of Computer Science, Dr. Vinodani Katiyar, Department of Information Technoloyg, Dr. Shakuntala Misra National Rehabilitation University, Lucknow.
[3] A Survey on Data Mining Techniques in Agriculture M.C.S.Geetha Assistant Professor, Dept. of Computer Applications, Kumaraguru College of Technology, Coimbatore, India.
[4] ANALYSIS AND PREDICTION IN AGRICULTURAL DATA USING DATA MINING TECHNIQUES ,Vinayak A. Bharadi,Prachi P. Abhyankar,Ravina S. Patil, Sonal S. Patade,Tejaswini U. Nate,Anaya ,M. Joshi, Information Technology, Finolex Academy of Management and Technology, India.
[5] Analysis of Agricultural Data Using Big Data Analytics K.Ravisankar,K.Sidhardha,Prabadevi B,Department of Information Technology and Engineering,VIT University, Vellore.