IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 4, April 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 4, April 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1004047|

Comparative Analysis of Machine Learning Algorithm for classification Using Orange Tool

Pratibha Jaisingh[1], Dr. R.K.Dhuware[2]

Department of Mathematics and Computer Science, RD University, Jabalpur, Madhya Pradesh, India[1]

Assistant Professor, D.B.Science College, Gondia, (Maharashtra), India[2]

ABSTRACT*:* Machine Learning plays a significant role in the areas of Artificial Intelligence where a computer can be trained using a data and get into self-learning mode without explicitly programming it to do so. Classification uses machine learning algorithms for recognizing, understanding and predicting the class. There are many tools available for performing machine learning techniques. Orange is open-source powerful platform built on Python that is extensively used for machine learning. This paper analyzes five classification algorithms using Orange tool and compare their performance according to various factors like Classification Accuracy and Precision Score. Some conclusion drawn from the difference between actual and predicted value will help us to determine which classifier works best. [1]

KEYWORDS: Orange Tool, Logistic Regression, Random Forest, Naïve Bayes, K-nearest neighbor, Neural Network.

I. INTRODUCTION

Machine Learning uses Statistical algorithm to predict patterns in massive data. It gives the ability to computer system to automatically learn and improve without explicitly programming them. In Supervised classification, Computer system are trained to build a model using "Training data" to make future predictions and when they are fed with "Test data", they learn, grow, change and develop by themselves. The target variable in a dataset is independent feature that is whose values are to be modeled and predicted by other variable. Classification algorithm uses historic data to learn patterns and uncover the relationship between other features and target. Cross Validation is resampling statistical method that is used to estimate the accuracy of the model. In K-fold cross validation, the dataset is divided into k subsets and holdout method is used k times. In each of the k iterations, out of k subsets, one subset is used as Test data and other 'k-1' is used as training data. Stratified K-fold cross validation is slight variation of k-fold cross validation in which each fold contains approximately the same percentage of samples of each target class as the complete set. Orange is code-free GUI based platform which includes drag and drop features for preprocessing, data visualization exploration and modelling. [2]

II. EXPERIMENT

A. Dataset

The dataset used in the paper is taken from Kaggle. The dataset is about Mobile price classification which is used to predict mobile price range depending upon many feature like RAM, Internal storage, battery power etc.

B. Machine Learning Algorithms :

The classification algorithm used in the paper are: Random Forest, logistics regression, Naïve Bayes, Decision Tree and K-Nearest Neighbor.

i. Logistics Regression: It is supervised machine learning technique which is used for predicting the output of categorical dependent variable depending upon the value of independent variables. It is much similar to linear regression algorithm except that linear regression output a continues number values and logistics regression transforms its output using the logistics sigmoid function to return a probabilistic value which can then be mapped to two or more discrete values.[3]

ii. Decision Tree: It is supervised machine learning technique in which continuously split the data according to certain conditions in the form of tree representation. The nodes in the tree can be decision nodes or leaf nodes. Decision nodes are where data is split and decisions are made. The final outcomes of the decision are leaf nodes.[4]

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 4, April 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1004047|

iii. Random Forest: It is also a supervised machine learning technique which uses the concept of ensemble learning. It creates multiple decision tress on different subsets of a dataset, merges them to find stable and more accurate predictive accuracy of the dataset.[5]

iv. K-Nearest Neighbor: It is supervised machine learning technique which is based on feature similarity. It predicts the class of new data depending upon its similarity to the available classes. It calculates the similarity in terms of how the new object is located physically close to existing objects and then it assign the object to class where distance is least.[6]

v. Naïve Bayes: It is supervised machine learning algorithm which uses Bayes theorem with the assumption of independence between predictors. It is a probabilistic classifier which predicts the class based on the probability of the object.[7]

C. Parameters for first set of evaluation :

The regularization for logistic regression is set to LASSO (L1) regularization. The decision tree is binary induced with minimum 2 instances in leaves, minimum instances in leaves is 2 and maximal depth to 100.The random forest has 10 trees with split subset not more than 5.The number of neighbor in KNN is set to 5 and Euclidean distance metric with distance based weight metric is used.

D. Parameters for second set of evaluation

The regularization for logistic regression is set to RIDGE (L2) regularization. The decision tree is binary induced with minimum 2 instances in leaves, minimum instances in leaves is 2 with maximal depth to 200. The random forest has 20 trees with split subset not more than 5. The number of neighbor in KNN is set to 5 and Manhattan distance metric with uniform weight metric is used.

E. Workflow of the model in the Orange Tool

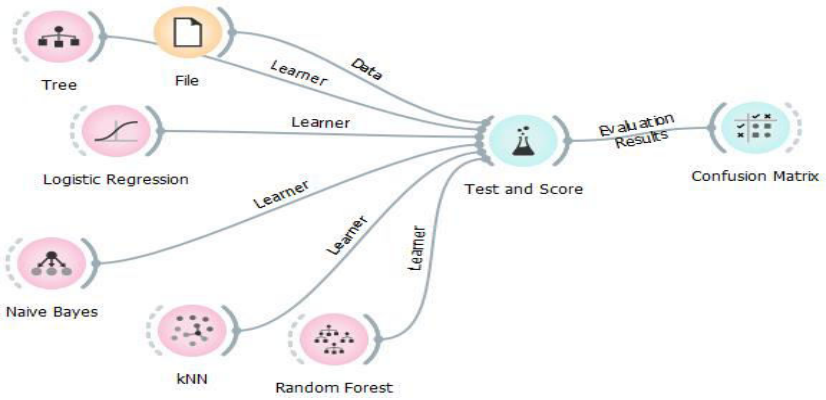

Fig1. Workflow of the model in the Orange tool

F. Evaluation

The Naïve Bayes is used for both algorithms without any parameters changes. The cross validation with 10 folds is used predict the classification Accuracy and precision score of said five algorithms.

III. EVALUATION RESULTS

The said five algorithms are compared using five attributes which are AUC, CA, F score, precision and Recall. All the attributes are computed automatically by the Orange tool.

AUC or Area under Curve is the measure of ability of the classifier to correctly predict the classes in the data. The higher the AUC, the better the performance of the model in distinguishing between positive and negative classes. CA or Classification Accuracy is ratio of number of correct predictions to the total number of predictions. Precision is the number of positive classes that actually belongs to the positive classes. Recall or sensitivity is the number of positive

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 4, April 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1004047|

class predictions made out of all positive examples in the dataset. F-score is way of combining Recall and precision and it is harmonic mean of Precision and Recall of the model.

Table 1 shows the comparison result using parameters in first set of evaluation and its graph is depicted in fig.2. The two main attribute used for comparison of the algorithm are Classification Accuracy and Precision. Fig.3. shows the bar chart of both the parameters for the five algorithms.

Model	AUC	CA	F score	Precision	Recall
Logistics Regression	0.961	0.862	0.859	0.857	0.862
Decision Tree	0.920	0.868	0.867	0.867	0.868
Random Forest	0.948	0.811	0.810	0.810	0.811
KNN	0.991	0.921	0.921	0.921	0.921
Naïve Bayes	0.922	0.764	0.764	0.765	0.764

Table 1. First Set of Evaluations

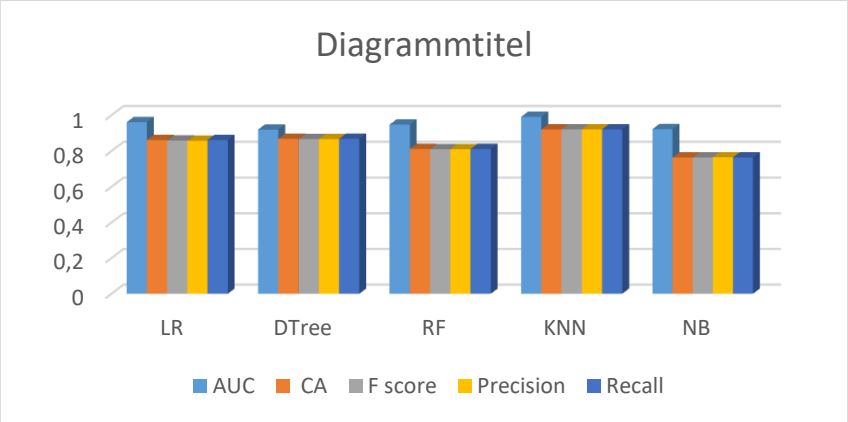

Fig.2. Graph of comparison of Logistics Regression, Decision Tree, Random Forest, KNN and Naïve Bayes using first set of evaluation parameters in terms of Area Under Curve, Classification Accuracy, F-score, Precision and Recall.

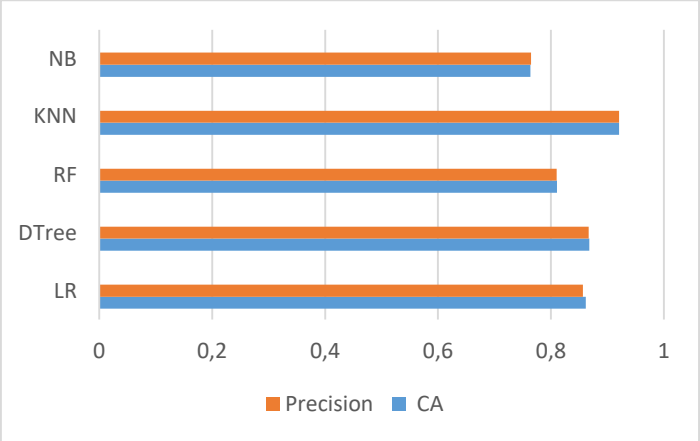

Fig 3. Bar Chart of Classification Accuracy and Precision for comparison of five algorithm

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 4, April 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1004047|

Table 2 shows the comparison result using parameters in Second set of evaluation and its graph is depicted in fig.4. Fig.5. shows the bar chart of Classification Accuracy and Precision for five algorithms

Classification Accuracy cannot predict the accuracy of the classifier model well. We need confusion matrix which is N×N matrix for evaluating the performance of the classification model. It compares the actual predicted values with those predicted by the machine learning model. Fig 6-fig 10 shows the confusion matrix for the five algorithms for both the set of evaluation

Model	AUC	CA	F score	Precision	Recall
Logistics Regression	0.880	0.634	0.632	0.632	0.634
Decision Tree	0.916	0.862	0.862	0.862	0.862
Random Forest	0.965	0.846	0.845	0.845	0.846
KNN	0.991	0.922	0.923	0.923	0.922
Naïve Bayes	0.922	0.764	0.764	0.765	0.764

Table 2 second set of evaluation

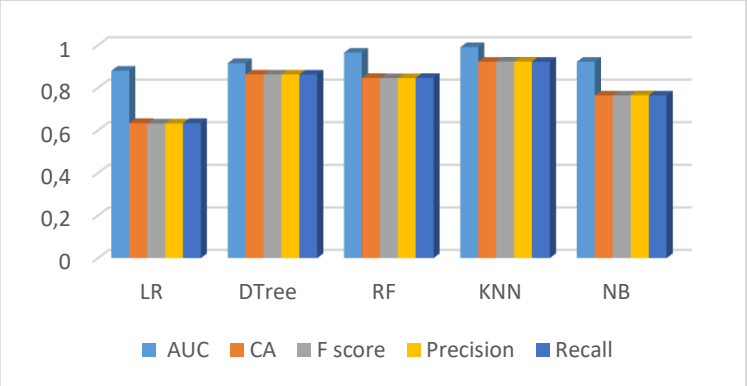

Fig.4. Graph of comparison of Logistics Regression, Decision Tree, Random Forest, KNN and Naïve Bayes using second set of evaluation parameters in terms of Area Under Curve, Classification Accuracy, F-score, Precision and Recall

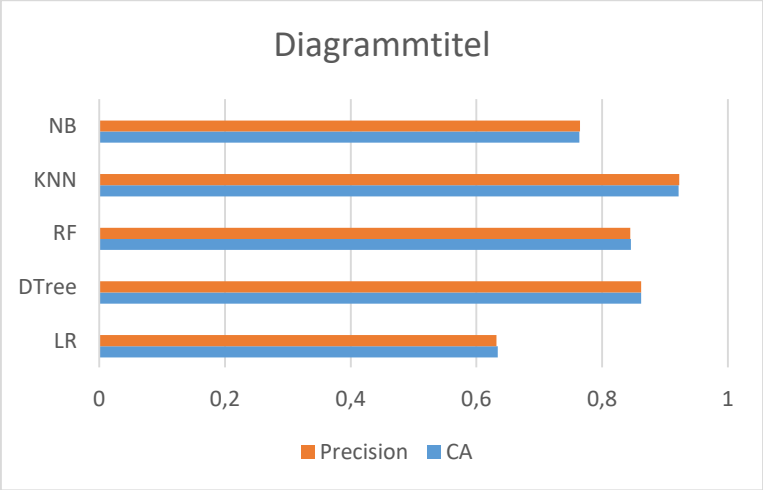

Fig 5. Bar Chart of Classification Accuracy and Precision for comparison of five algorithm

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 4, April 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1004047|

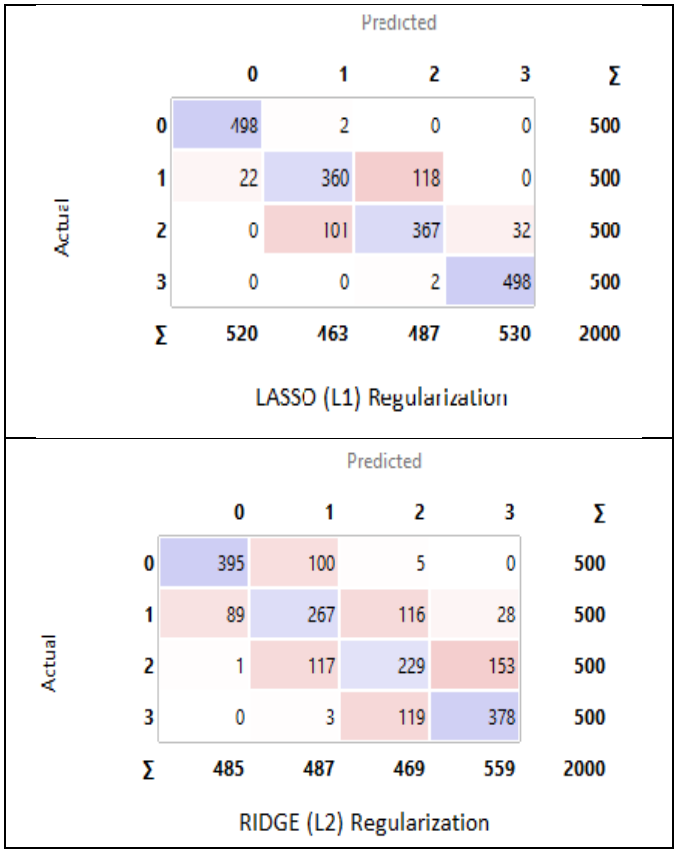

Fig.6. Confusion Matrix for Logistic regression using LASSO (L1) Regularization in the first row and RIDGE (L2) Regularization in the second row

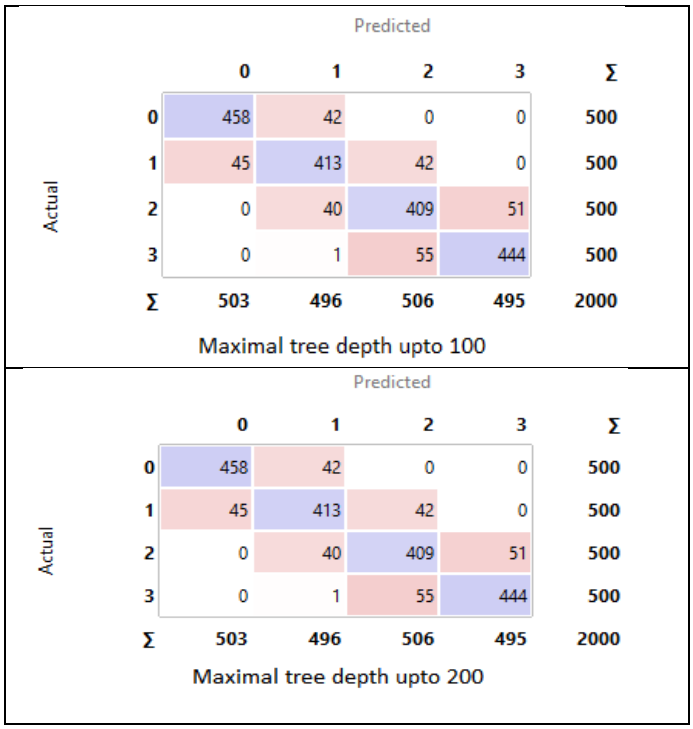

Fig.7. Confusion Matrix for Decision Tree Algorithm with maximal tree depth upto 100 in the first row and 200 trees in the second row

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 4, April 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1004047|

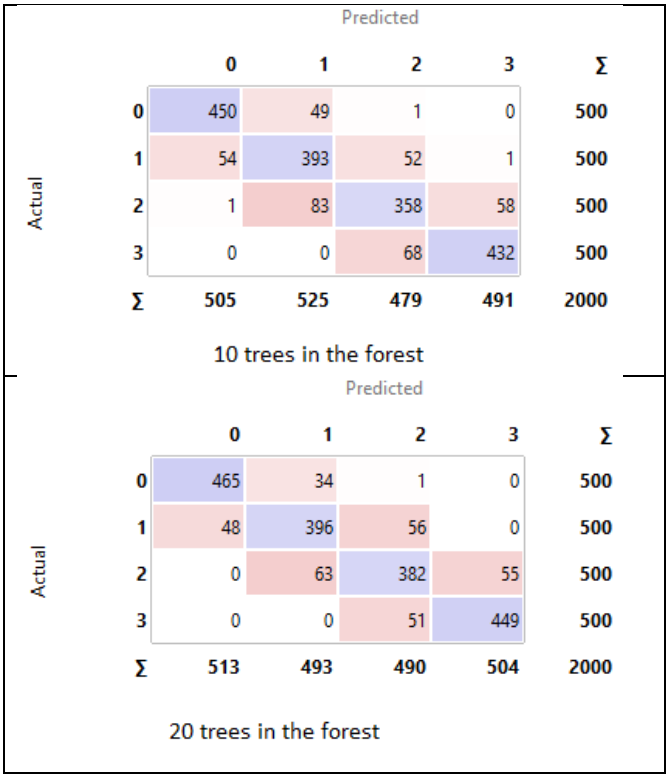

Fig.8. Confusion matrix for Random Forest with 20 trees in the first row and 10 trees in the second row

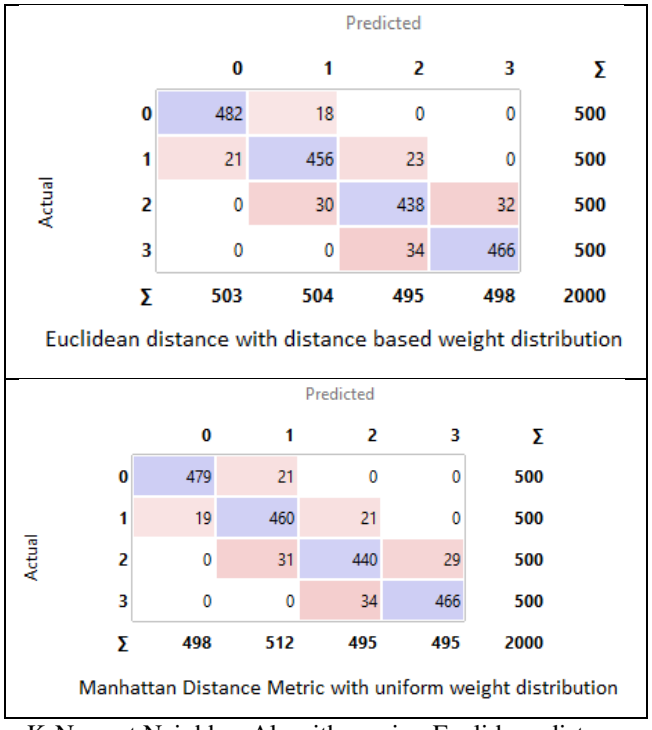

Fig.9. Confusion Matrix for K-Nearest Neighbor Algorithm using Euclidean distance with distance based weight distribution in first row and using Manhattan Distance Metric with uniform weight distribution in second row

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 4, April 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1004047|

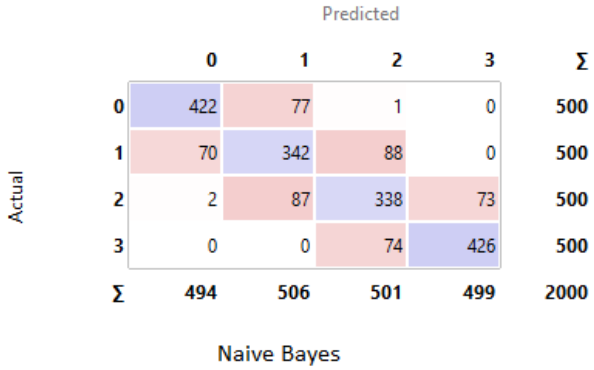

Fig.10. confusion Matrix for Naïve Bayes Algorithm for both the set of evaluations

Confusion matrix can be used to calculate prediction error which is used to measure how well the model can classify objects to the correct class. It is difference between predicted and actual data values. Table.3. shows the prediction error for the first set of evaluation and table.4. shows the prediction error for the second set of evaluation.

Algorithm	Class			
	Class 0	Class 1	Class 2	Class 3
Logistic Regression	2	140	133	2
Decision Tree	42	87	91	56
Random Forest	50	107	142	68
K-Nearest Neighbor	18	44	62	34
Naïve Bayes	78	158	162	74

Table.3. Prediction error for first set of evaluation

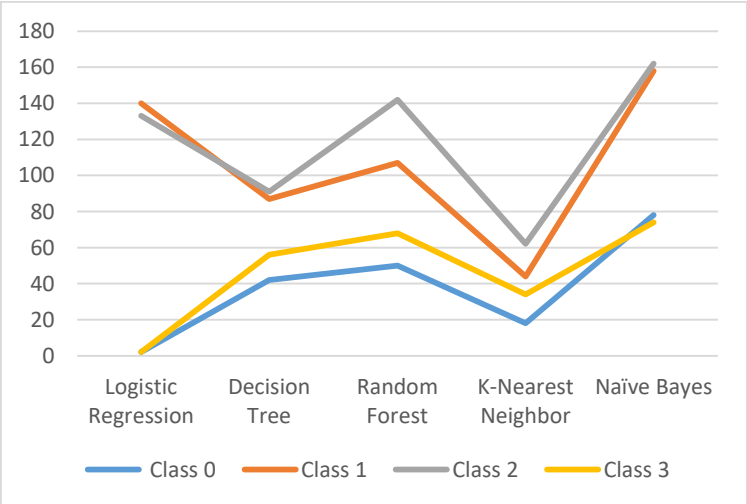

Fig.11. Plot of Prediction error for first set of evaluation

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 4, April 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1004047|

Algorithm	Class			
	Class 0	Class 1	Class 2	Class 3
Logistic Regression	105	233	271	122
Decision Tree	42	87	91	56
Random Forest	35	104	118	51
K-Nearest Neighbor	21	40	60	34
Naïve Bayes	78	158	162	74

Table.4. Prediction error for second set of evaluation

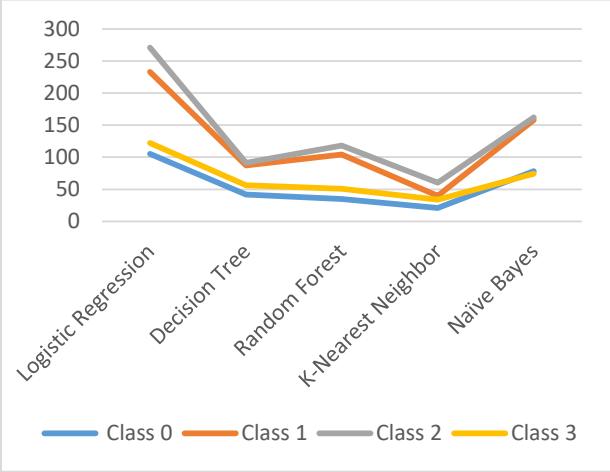

Fig.12. Prediction error plot for second set of evaluation

IV. CONCLUSION

K-Nearest Neighbor provided an accuracy of 92% with precision score of 92% for both the set of evaluation. We can conclude that, among all the five algorithm viz. Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbor and Naïve Bayes, K-nearest Neighbor worked well for the dataset we used in the paper. It provided highest classification Accuracy and Precision Score among others. The prediction error for the K-Nearest Neighbor are less as compared to all other especially for class 0 which can be seen from confusion matrix.

REFERENCES

[1] DevashreeVaishnav and B.Rama Rao, "Comparison of Machine Learning Algorithm and Fruit classification using Orange data Mining Tool", Proceedings of the International Conference on Inventive Computation Technologies (ICICT-2018) IEEE Xplore Part Number: CFP18F70-ART; ISBN:978-1-5386-4985-5

[2] Maqsud S. Kukasvadiya and Dr.Nidhi H. Divech , "Analysis of Data Using Data Mining tool Orange" , International Journal of Engineering Development and Research , 2017 IJEDR | Volume 5, Issue 2 | ISSN: 2321-9939

[3] Harsh H. Patel *, PurviPrajapati ," Study and Analysis of Decision Tree Based Classification Algorithms" , Vol.-6, Issue-10, Oct. 2018 E-ISSN: 2347-2693

[4] EeshaGoel ,Er. Abhilasha , "Random Forest: A Review" , International Journal of Advanced Research in Computer Science and Software Engineering, Volume 7, Issue 1, January 2017 ISSN: 2277 128X

[5] PouriaKaviani, Mrs.SunitaDhotre ," Short Survey on Naive Bayes Algorithm ",*Volume 4, Issue 11, November - 2017* e-ISSN (O): 2348-4470 p-ISSN (P): 2348-6406

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details