Classification Models Comparison and Analysis Using Weather Prediction Dataset

Course Name: Data Warehousing and Data Mining

Section- D

Course Instructor- Akinul Islam Jony

Group Members:

NAME	ID
MD. Bakhtiar Azim Niloy	19-41011-2
MD. AHASANUL ISLAM	19-40396-1
ASRAFUL ISLAM	19-40166-1
WALID BIN WAHID BADHAN	19-40845-2

1. Introduction: (Project Overview)

The goal of this project is to predict weather condition of a year based on date to measure precipitation, temperature minimum and temperature maximum. Three different classification models were used to predict the outcome and comparison between the models were done.

Classification in machine learning and statistics is a supervised learning approach in which the computer program learns from the data given to it and makes new observations or classifications. Here, 3 classification models were used - Naïve Bayes, k-nearest neighbors (KNN) and Decision Tree. These are some popular classification algorithms that can be used for classifications.

Decision tree gives better classification accuracy than Naïve bayes and KNN. The percentage split test method gave slightly more accuracy than cross-validation 10 folds. Also the F_measure analysis of the three models.

2. Dataset:

The dataset is collected from Kaggle. Here is the link of the dataset: [WEATHER PREDICTION | Kaggle](Kaggle).

This datasets represent the prediction weather of one year 2012.

There is a total of 1460 instances and 6 attributes.

Fig-2: Dataset

3. Model Development:

Model used in this project are Naïve bayes, KNN and Decision Tree. Two types test option were applied on the dataset. First one is Cross-validation, second one is percentage split. 66% of instances were taken for the trainset and 34% for the test set.

To create these three models, a data analysis program called "WEKA" is used. To begin, the dataset is uploaded and reviewed to ensure that all instances are valid. We chose the classify option since this task is classification-based. “<=50” attribute is selected as the class attribute and for test options two methods were used for all the selected models. One is called "Percentage Split (Train 66%, Test 34%)," while the other is called cross validation (10 folds). After training and testing the dataset, we obtained a statistical report on the model's prediction accuracy.

3.1. Naïve Bayes:

It is a classification strategy that is based on Bayes' Theorem and makes the assumption of predictor independence. To put it simply, a Naive Bayes classifier believes that the existence of a specific feature in a class has no correlation with the presence of any other feature.

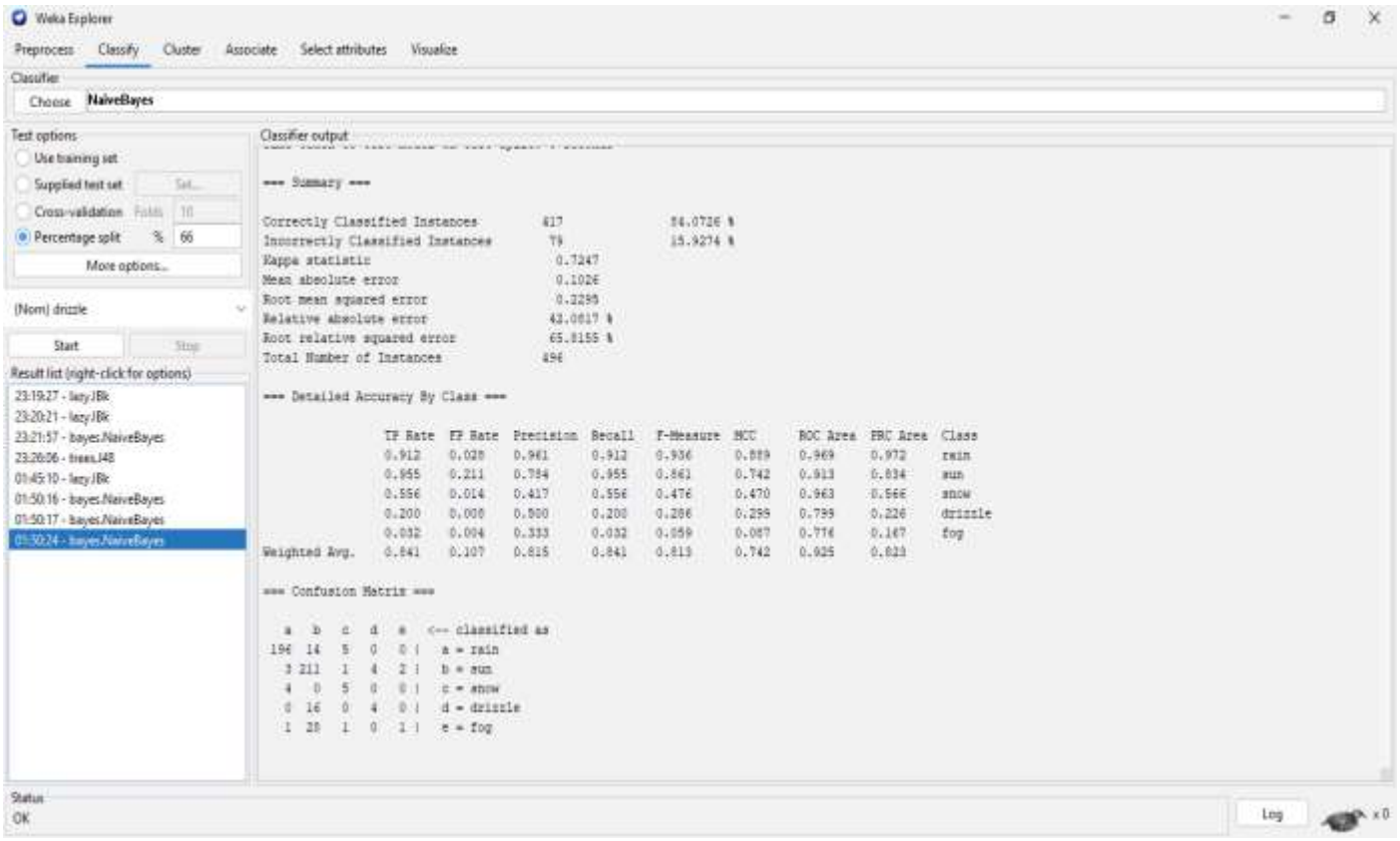


fig-3.1: Naïve Bayes (Percentage split)

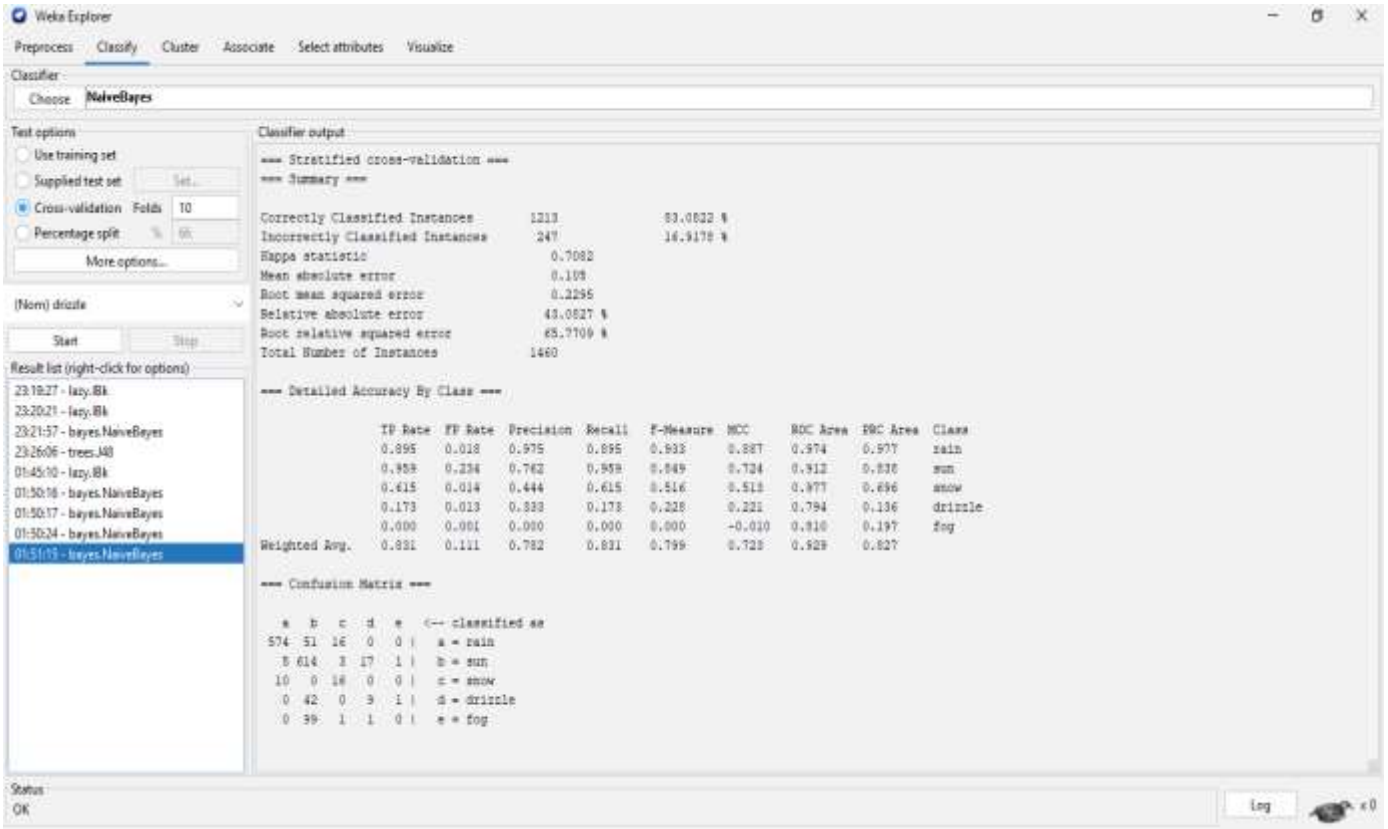


fig-3.2: Naïve Bayes (Cross validation)

3.2. KNN:

Classification KNN is a nearest-neighbor classification model in which both the distance metric and the number of nearest neighbors can be altered. It is also called Lazy in Weka.

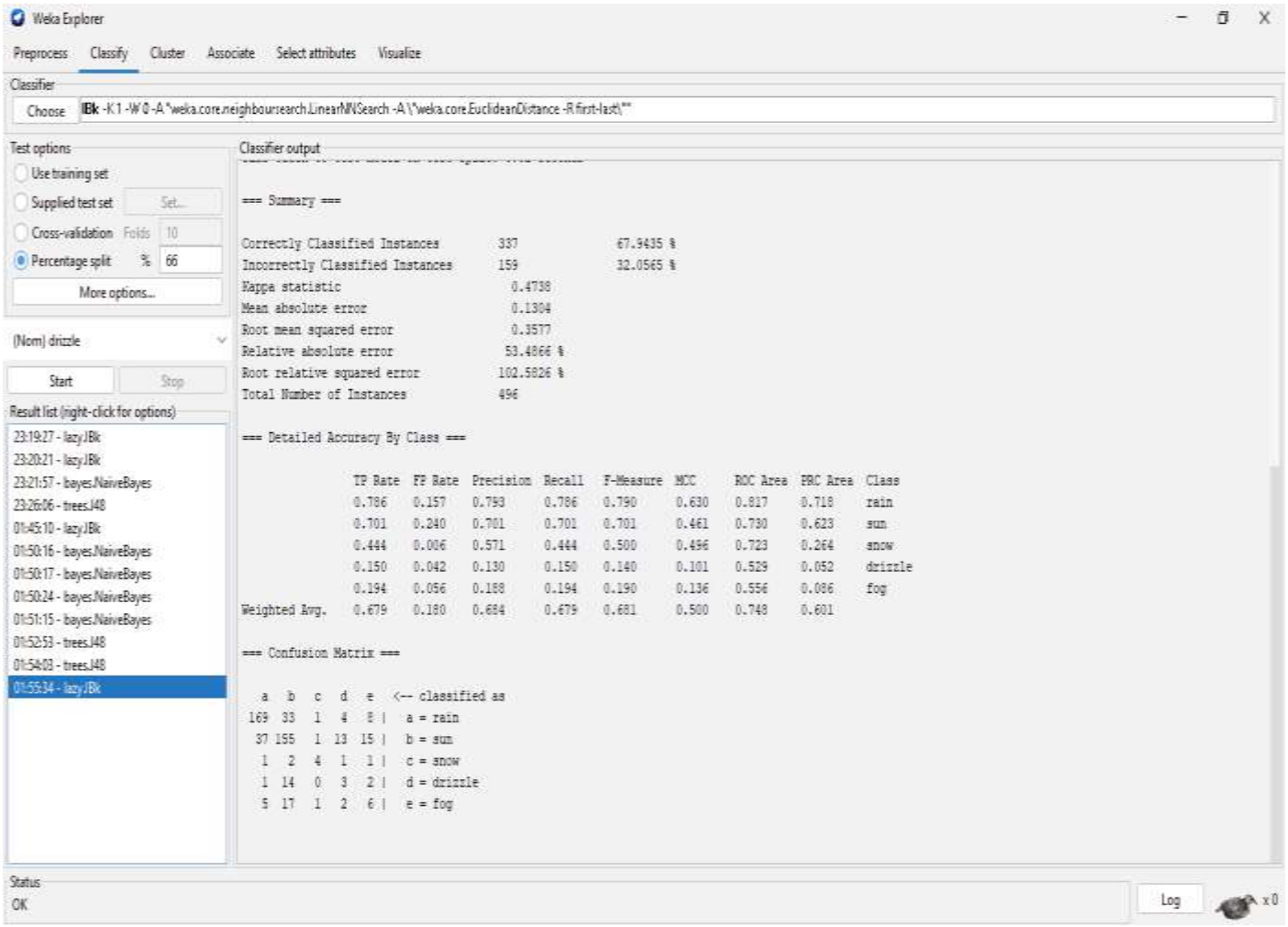


Fig-3.3: KNN (Percentage split)

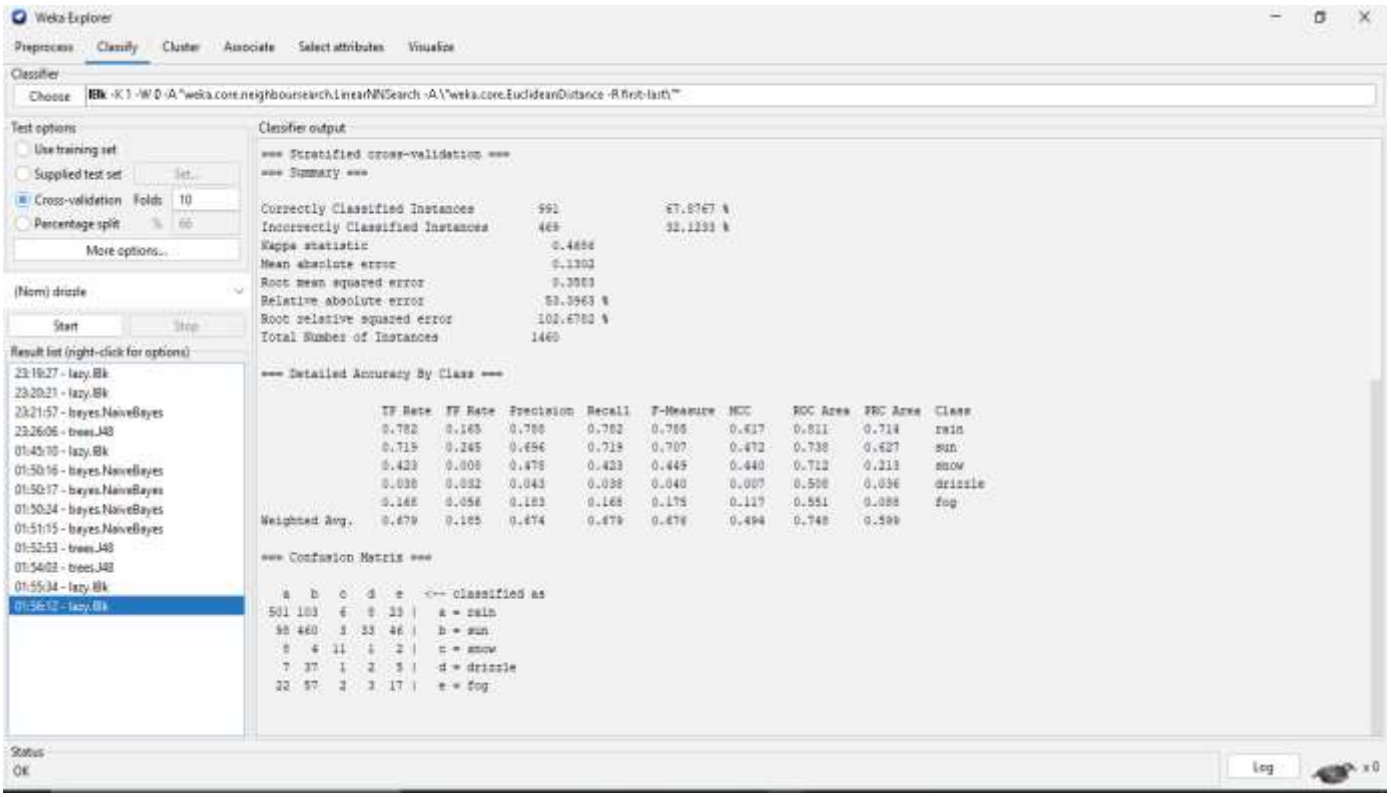


Fig-3.4: KNN (Cross validation)

3.3. Decision Tree:

Decision Trees (DTs) are a kind of non-parametric supervised learning algorithm that is often used for classification and regression. The objective is to develop a model that accurately predicts the value of a target variable by inferring basic decision rules from data attributes.

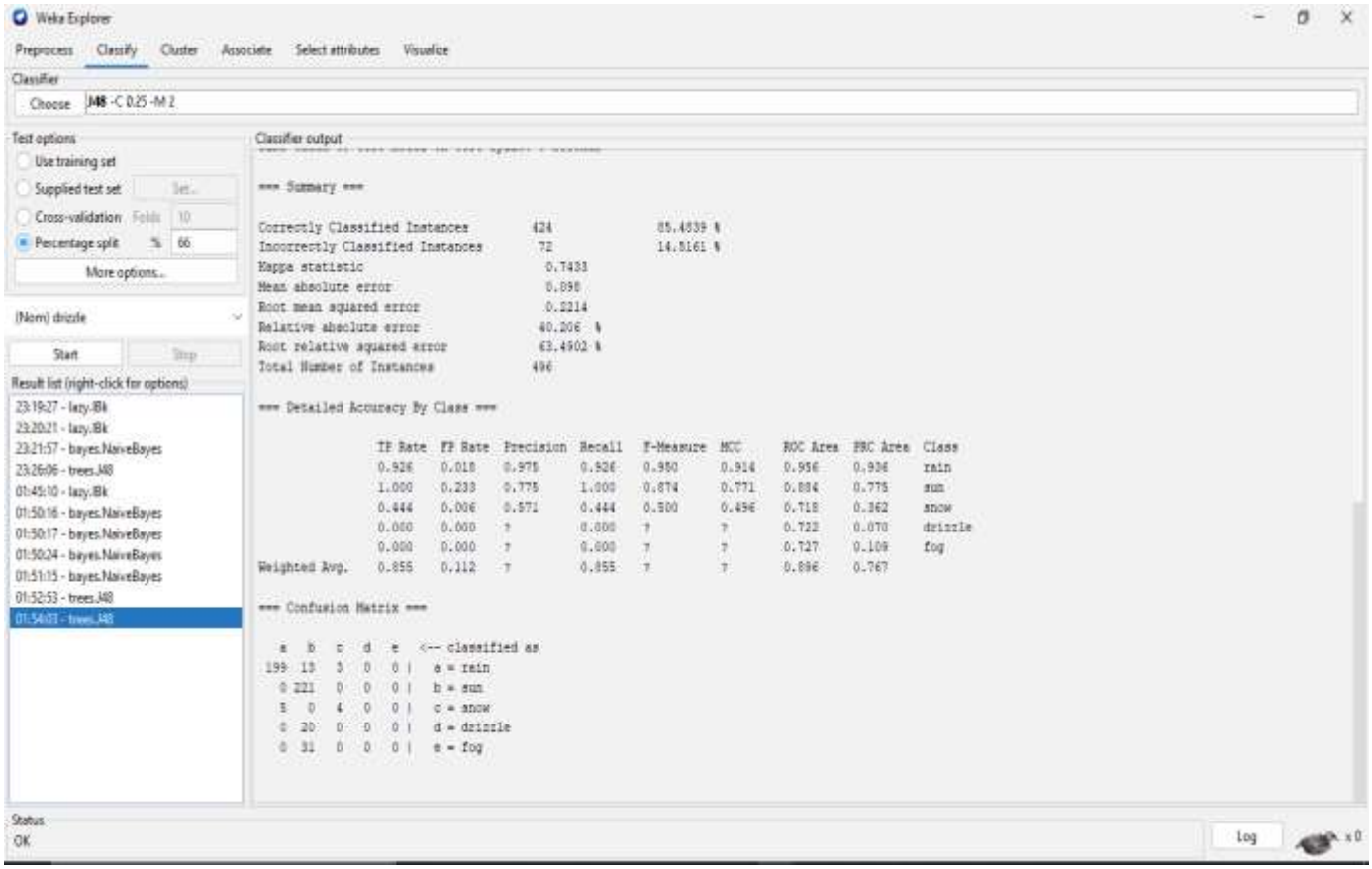
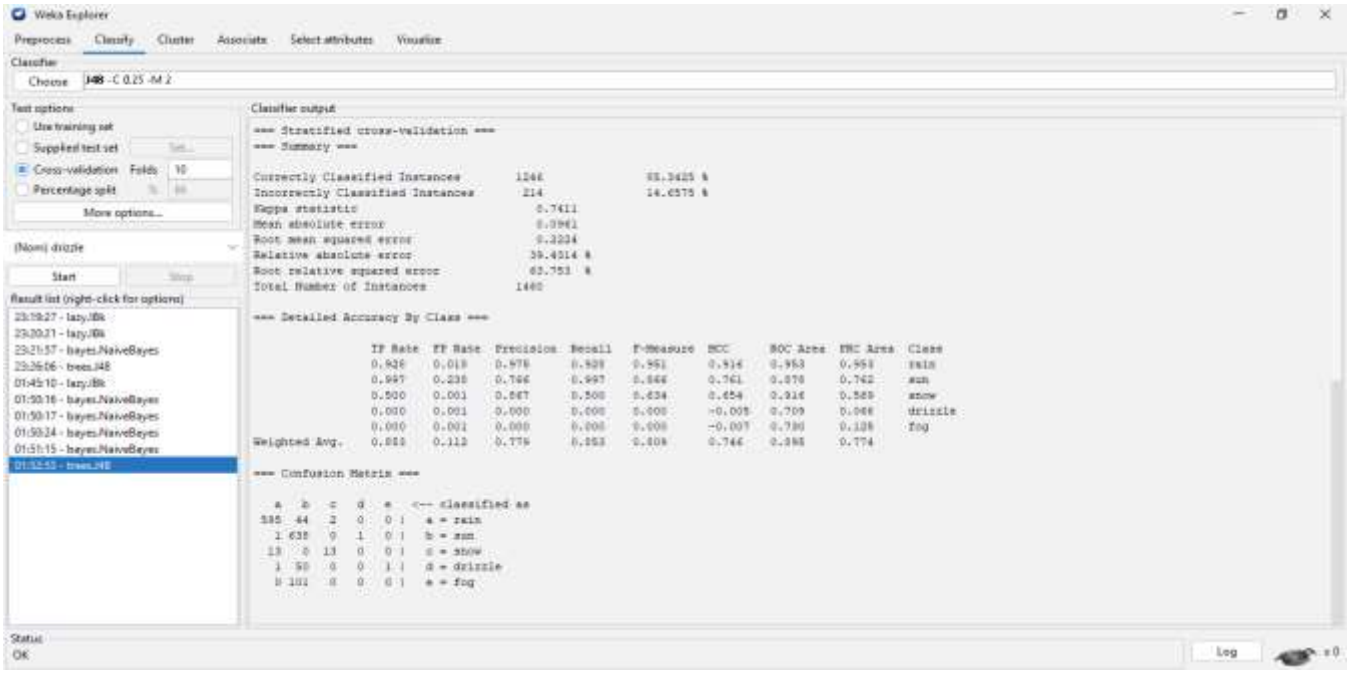


fig-3.5: Decision Tree (Percentage split)



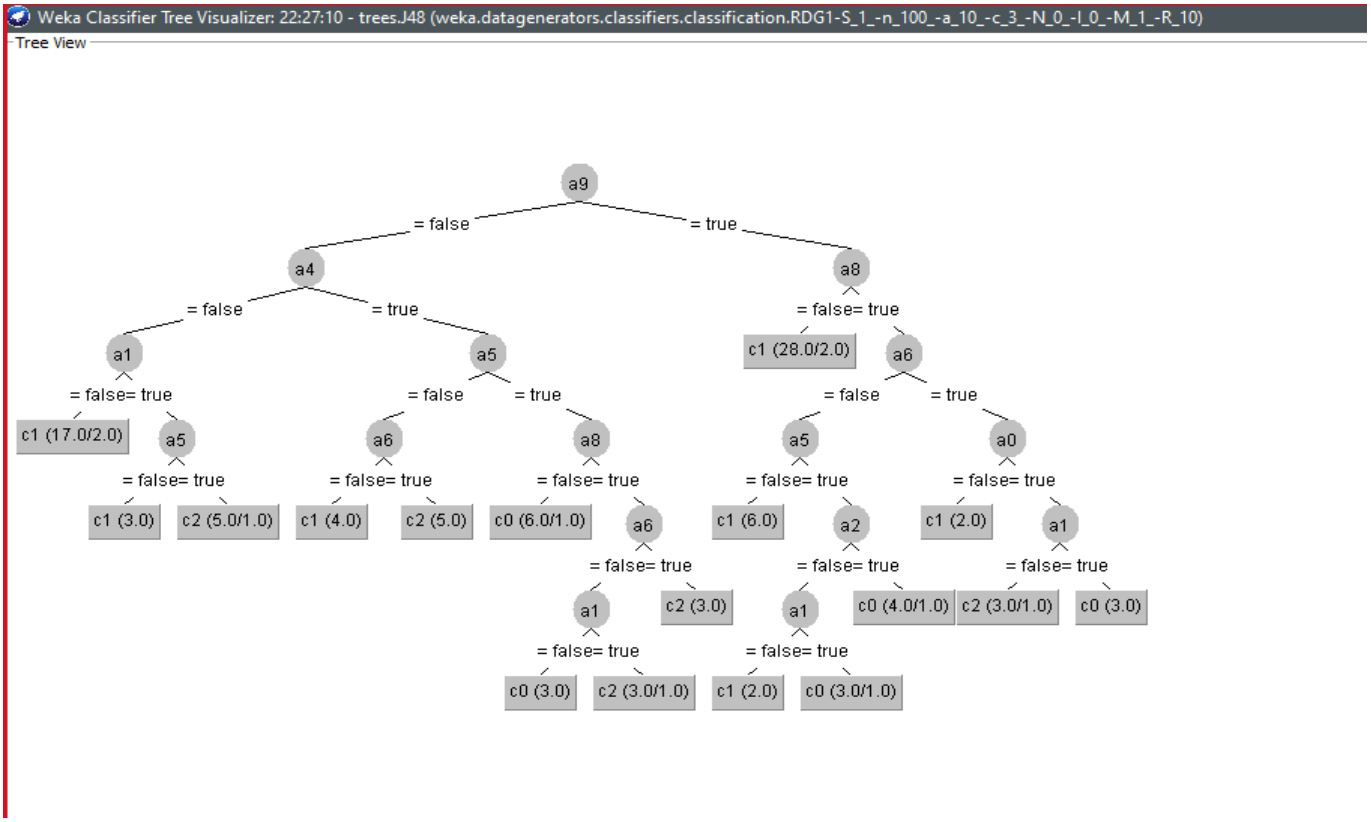


fig-3.6: Decision Tree (Cross validation)

4. Discussion & Conclusion:

This paper presents a study of weather prediction dataset using a total of three classification models (Naïve Bayes, KNN and Decision Tree) and comparison was done to find out which model is more efficient. These models are based on classification supervised learning approaches. The dataset on the weather predict which was collected from Kaggle was used here. The dataset contains 1460 instances and 6 attributes in which the target attribute was <=50. As the correctly classified instances percentage value was higher in case of percentage split, the comparison analysis is done on the percentage split. From the analysis it is seen that the decision tree shows better F_measure value (0.95) than the other two models. In the F_measure value there is a v sign beside decision tree and star (*) sign beside KNN which indicates that decision tree performed better than the other two models and the performance of KNN was poorer. That is why Decision Tree is better than Naïve Bayes and KNN in case of this weather prediction dataset. Also show the confusion matrix of the 3 classes.

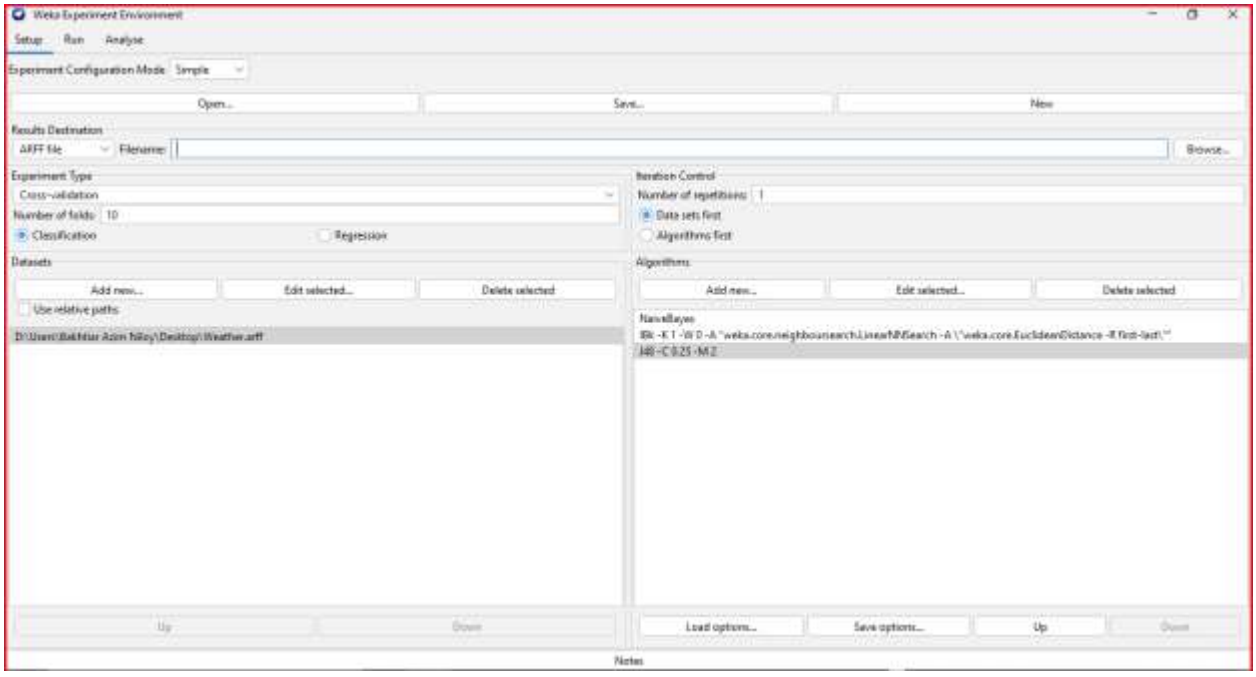


fig-4.1: F_measure analysis

fig-4.2: F_measure analysis

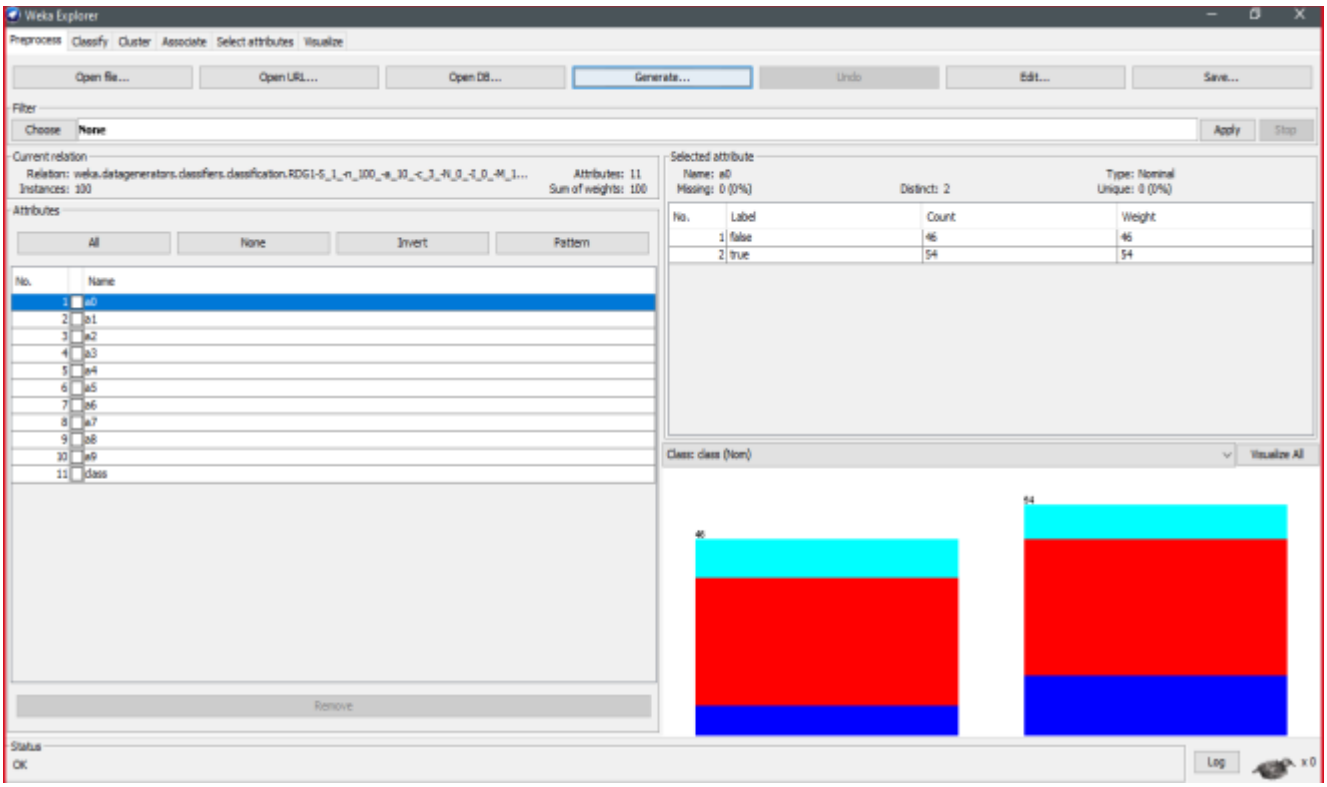


fig-4.3: Confusion Metrix analysis