**Methodology:**

The data mining algorithm used for diabetes prediction is Classification. We have used 3 basic classification algorithms to try and determine which one gives the best accuracy. They are J48 classifier, Naïve Bayes Classifier and Lazy IBK classifier.

Before, the data mining model is created and tested, the data is pre-processed. The dataset obtained from Kaggle has 22 attributes and over 2 lakh rows. The data contains 3 classes, diabetic, non-diabetic and pre-diabetic. The dataset contains attributes like blood pressure, cholesterol, BMI, whether the person smokes or not, whether the person had a history of cardiac arrest, physical activity, eating style, sex, age etc. The data is cleaned by removing the attributes like education and financial status which are not useful for prediction of diabetes.

Then, the dataset which is in csv format is converted into. arff format to pre-process and load the data.

**Classification:**

Classification, also known as supervised learning, is a data mining task that maps data into predefined groups and classes. Classification attempts to predict which of a (small) set of classes each individual in a population belongs to. A data mining procedure creates a model that, given a new individual, determines which class that individual belongs to.

WEKA provides several classification models which can be used on our data.

**J48 model:**

The C4.5 algorithm is a classification algorithm that uses information theory to generate decision trees. It is an extension of Ross Quinlan's earlier ID3 (Iterative Dichotomiser 3) algorithm, which is also known as J48 in Weka. C4.5's decision trees are used for classification, and as a result, C4.5 is frequently referred to as a statistical classifier. Accounting for missing values, decision tree pruning, continuous attribute value ranges, rule derivation, and other features are included in the J48 implementation of the C4.5 algorithm. J48 is an open-source Java implementation of the C4.5 algorithm in the WEKA data mining tool.

Using the concept of information entropy, this algorithm constructs decision trees based on a set of training data in the same way that the ID3 algorithm does. The best attribute to split on to achieve the highest classification accuracy is the attribute with the most information. The C4.5 algorithm selects the attribute of the data at each node of the tree that most effectively divides its set of samples into subsets, enriched in one class or the other. The normalized information gain, calculated from the difference in entropy, is the splitting criterion. To make the decision, the attribute with the highest normalized information gain is chosen. The C4.5 algorithm then recurses on the partitioned sub lists using a divide-and-conquer strategy to generate a greedy decision tree.

**Naïve Bayes Model:**

It is a classification technique based on Bayes' Theorem and the assumption of predictor independence. A Naive Bayes classifier, in simple terms, assumes that the presence of one feature in a class is unrelated to the presence of any other feature. The Naive Bayes model is simple to construct and is especially useful for very large data sets. In addition to its simplicity, Naive Bayes has been shown to outperform even the most sophisticated classification methods.

**Lazy IBK model:**

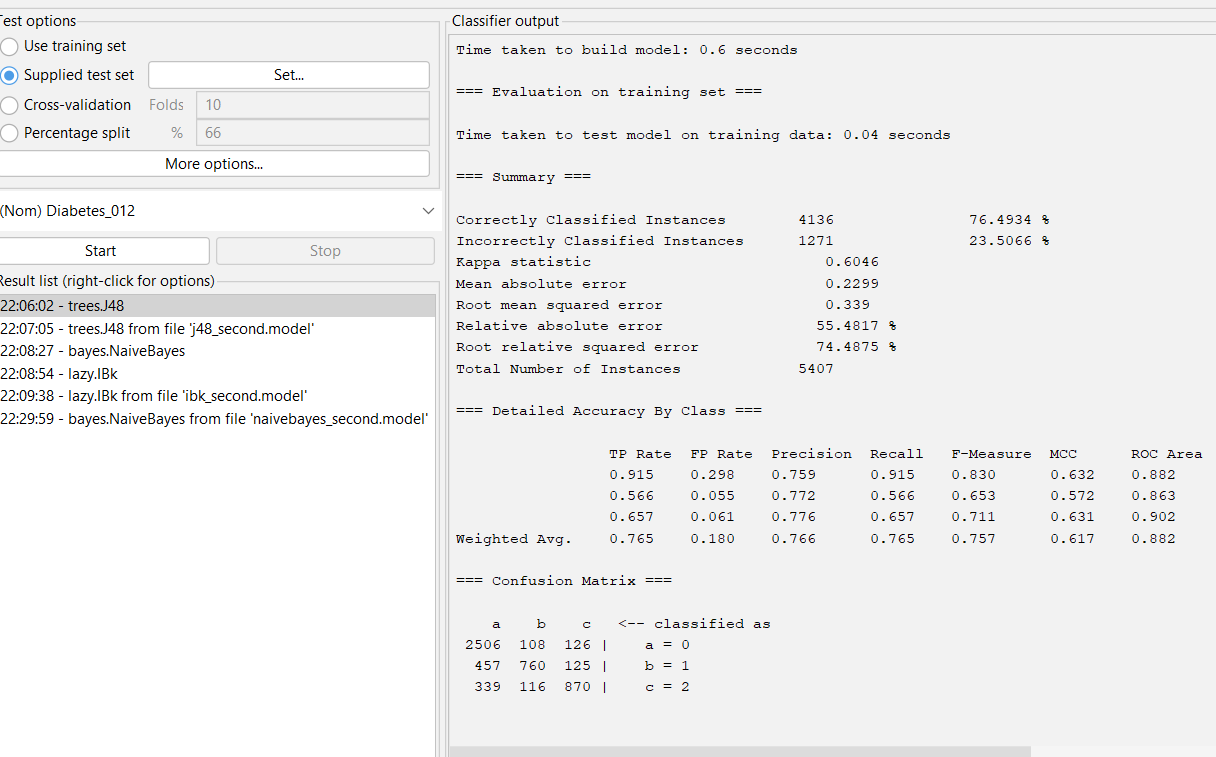
IBK is a classifier that uses the same distance metric as k-nearest-neighbor. The number of nearest neighbors can be explicitly specified in the object editor or determined automatically using leave-one-out cross-validation with a focus to an upper limit specified by the value. IBK is a classifier for closest neighbors. A variety of search algorithms can be used to accelerate the task of locating the nearest neighbors. The linear search is the default, but you can also use KD-trees, ball trees.”

Evaluation and Results:

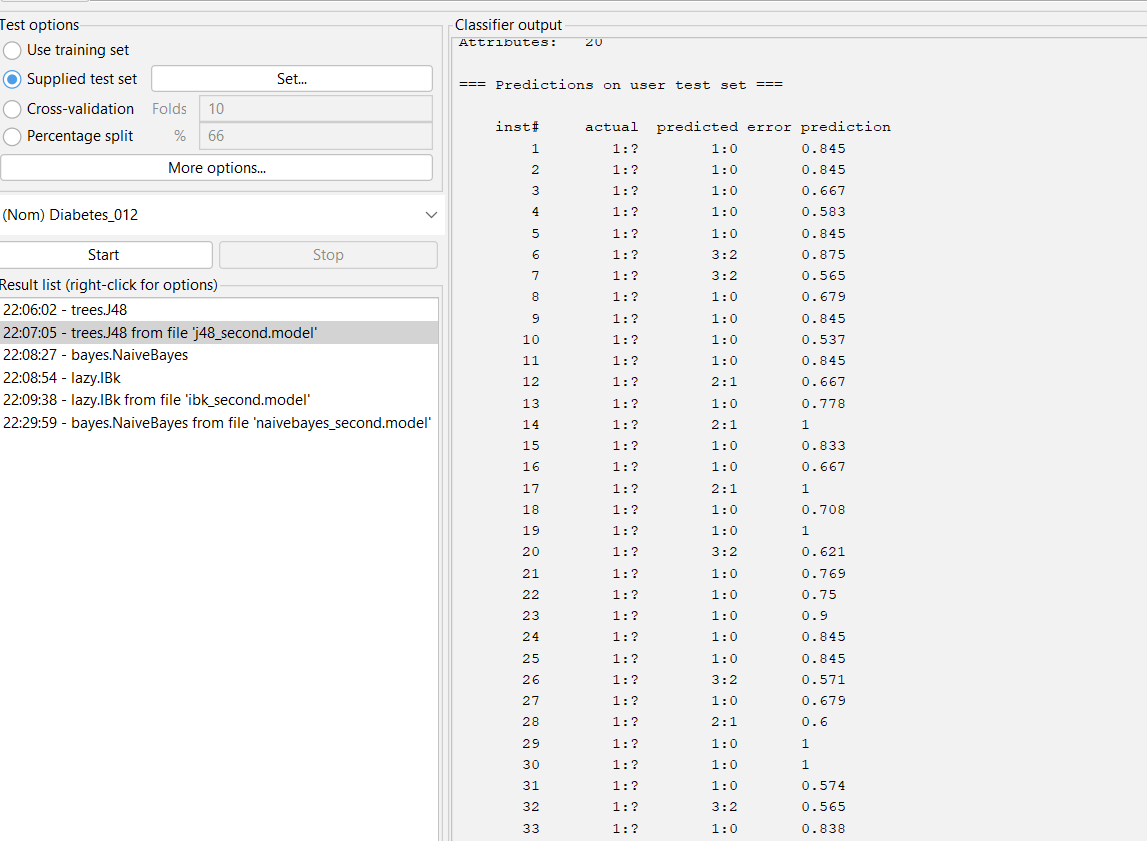
1. **J48 Model:**

**Training:**

Accuracy: 76.49%

****

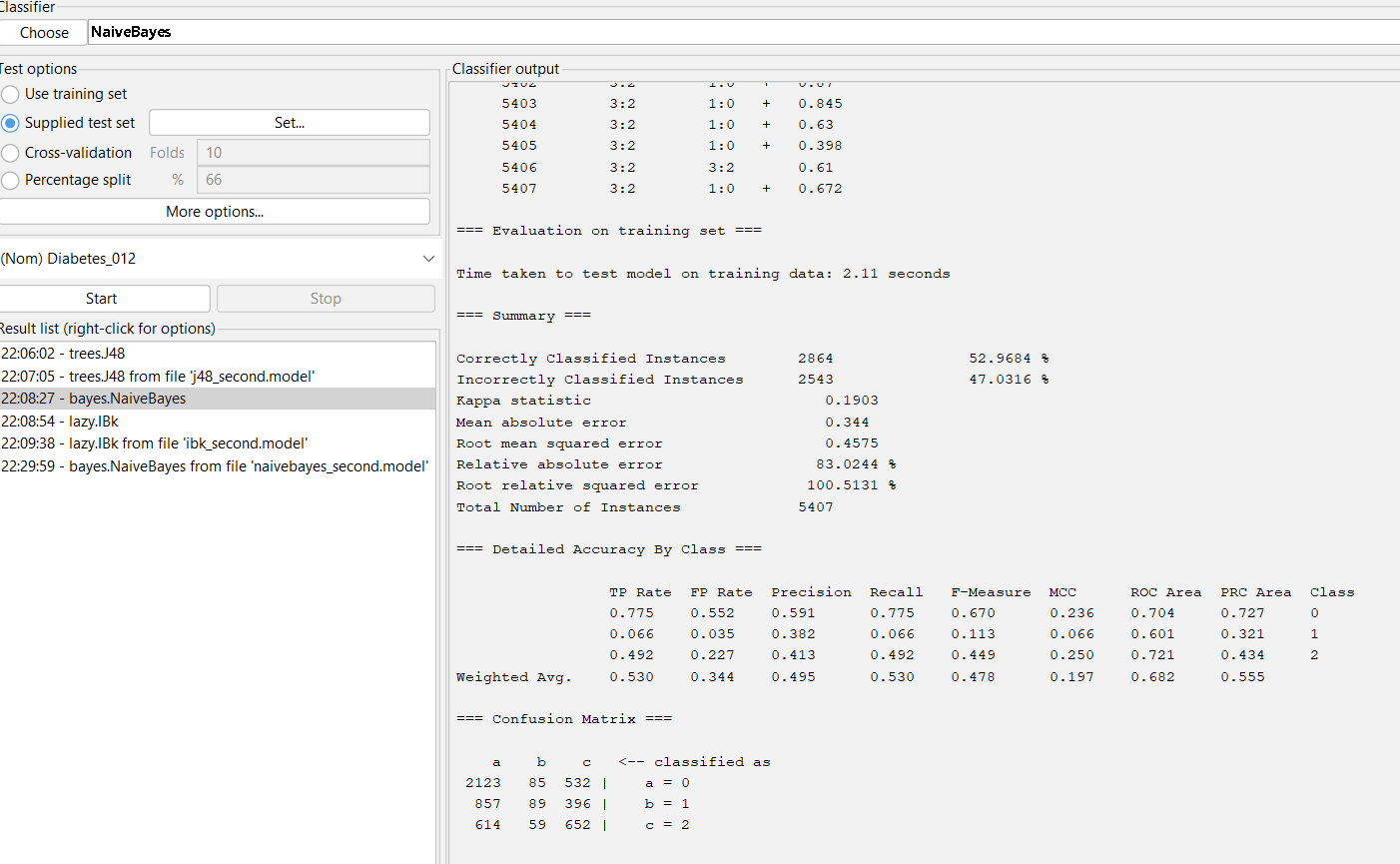
**Testing:**

****

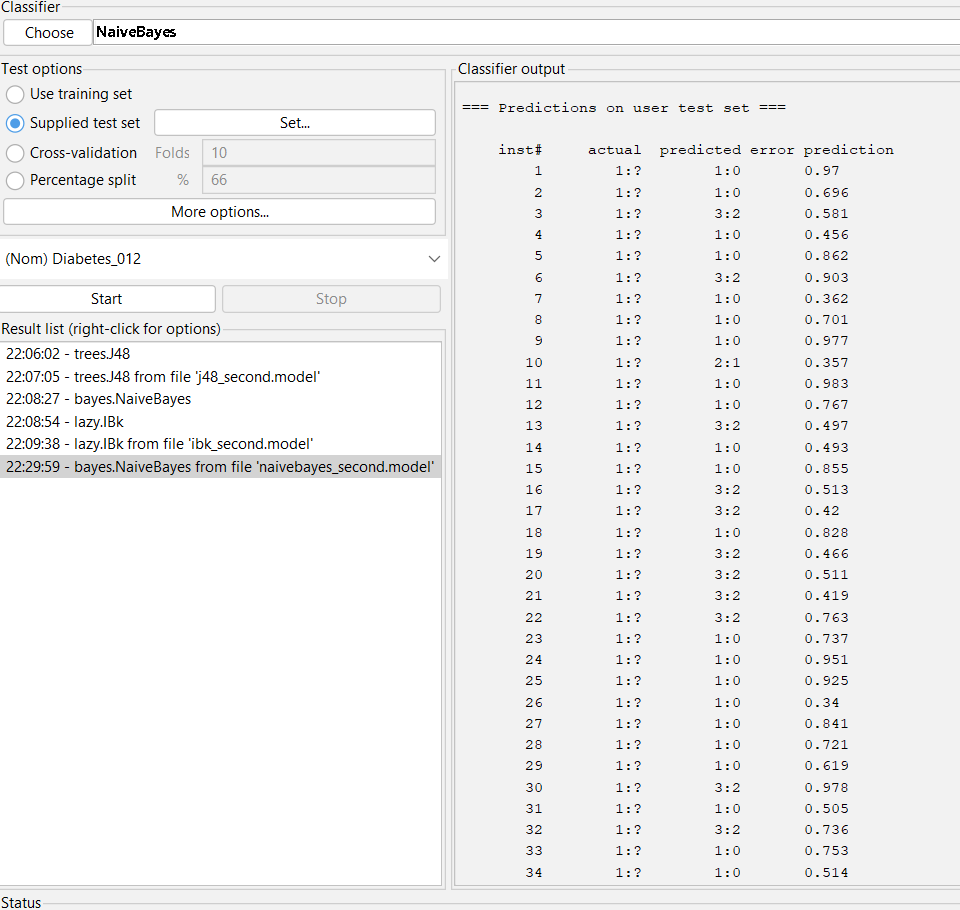
1. **Naïve Bayes model:**

**Training:**

Accuracy: 52.96%

****

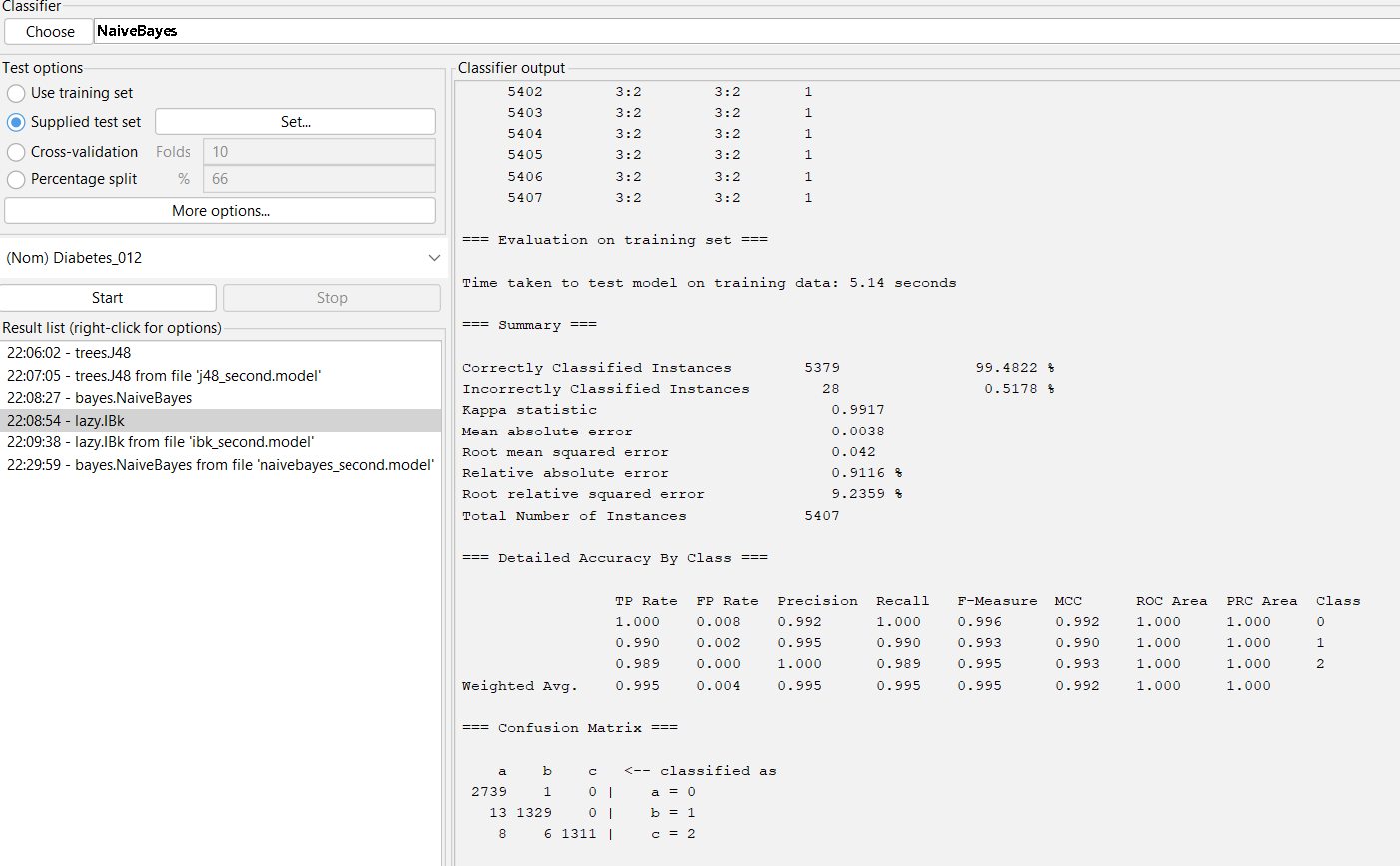
**Testing:**

****

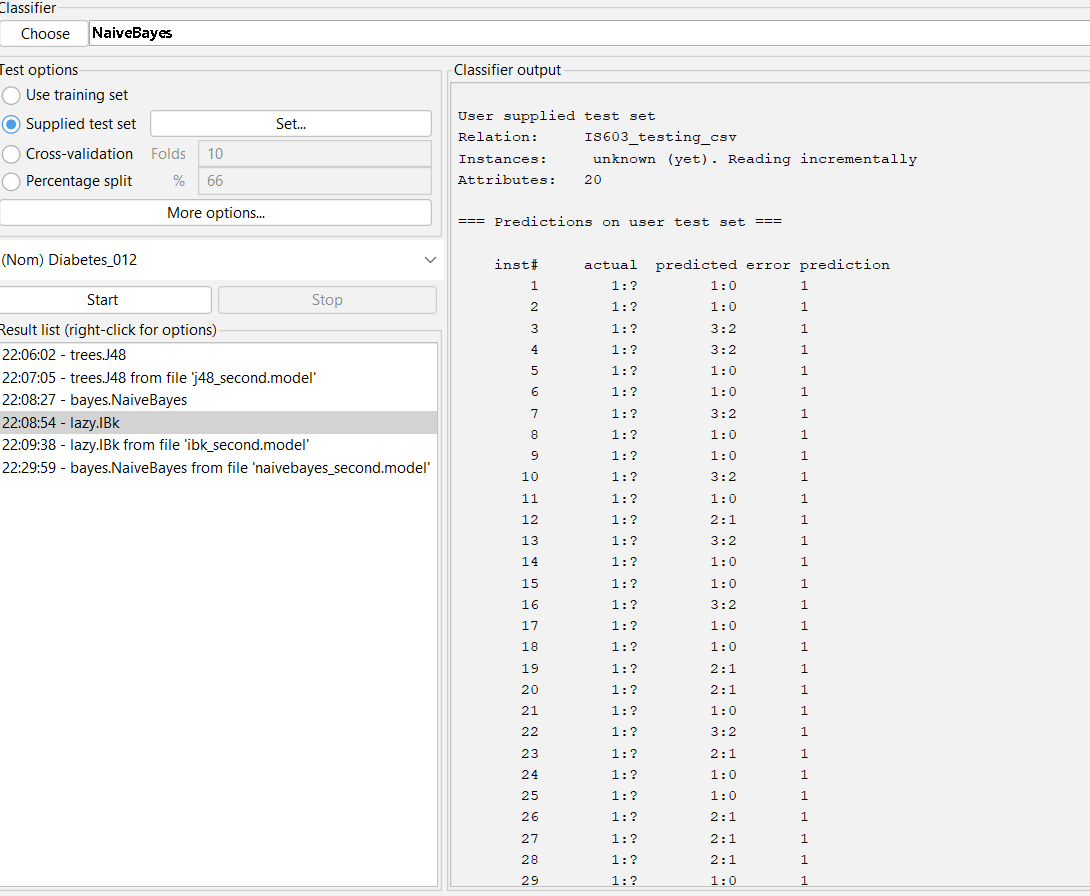
1. **IBK model:**

**Training:**

Accuracy: 99.48%

****

**Testing:**

****

The last IBK model has the highest accuracy with 99.48%.