# **Data Mining**

## **i**

Data mining is a process of extracting valuable and useful information by looking and searching through a large number of data sets. It identifies the hidden patterns of the data sets based on the likely outcomes of the given data and helps to establish relationships between data to solve and predict the solution of future problems. Data mining is used in marketing, mathematics, research and medical areas, cybersecurity, fraud detection, etc. The more the number of data sets is, the more powerful will be the data mining system (Hughes, 2020).

Data mining can help the hospital industries with disease prediction, advise doctors in clinical decisions, predictions in treatment, etc. Some appropriate data mining applications for hospitals are listed below.

* **Treatment effectiveness: -** Data mining applications analyses the courses of treatment, negative effects, causes, symptoms to identify the best effective treatment for an illness. This also helps to address some side effects of treatments.
* **Healthcare management: -** High risks patients, chronic sickness states can be addressed and tracked by data mining applications. This also provides the best treatment schemes, helps to decrease hospital admissions and claims.
* **Customer relationship management: -** Data mining can help to establish a better customer relationship and maintain it as much good as possible in the coming days by tracking the calls, billing systems, ambulatory care, etc.
* **Fraud and abuse prevention: -** Data mining can detect fraud and abuses. This will help to reduce them by identifying the wrong and inappropriate prescriptions, unusual medical claims patterns, and fraud insurance (n.d., 2021).

## **ii**

A data mining system for the hospital can be created by going through the data mining life cycle which consists of four main stages and those are explained below.

1. **Problem definition: -** This is very important and might be a very hard stage of the data mining life cycle. First of all, the hospital should have a clear idea about “what is the actual problem cases and what is needed to be done?”. A clear understanding of the problem context will help to gather only the valuable data but not the unnecessary data. Therefore, there should not be left any doubt about the problems before moving to the next stage.
2. **Data gathering & preparation: -** It will be easier for the hospital to identify the right set of data that will help to answer pertinent questions after defining the problems. Once the collection of required data is done, then the data will be purified by cleaning and removing duplicates, unnecessary, outliers, and missing values to build a more accurate data model. This will help to increase the accuracy level of prediction.
3. **Model Building and Evaluation: -** Depending upon the data sets and type of analysis, the data mining technique might vary while creating the model. Association rules, Neural networks, Decision trees, KNN are the most common techniques. After building the data model, this will be used to predict pertinent questions for the new data sets of the hospital. This will provide some results which might be almost 100% accurate depending upon the created model.
4. **Use of knowledge: -** After getting useful, novel, understandable, and valid information, this knowledge can be used to make changes, implement improvement in their service, start new strategies, achieve the intended objectives for the hospital, provide better treatment for the patients, clarify a patient illness stage and level. Or it can be reused to make a better version of the data mining system for more accurate prediction (Education, 2021).

## **Iii**

The accuracy of a data mining system depends on how well a model has been created. The larger number of data sets will help create a better and powerful data model and the accuracy of the data mining system will also increase. While predicting the outcome of a problem, it uses a tree-like visualization derived from the data model. When the data sets are in large numbers, the decision tree will be more precise and the accuracy level of the prediction will also increase.

For example, let's take “diabetes.arff” data which has 9 columns and 768 data in it. The decision tree derived from it is shown below with the help of a data mining tool “WEKA”.

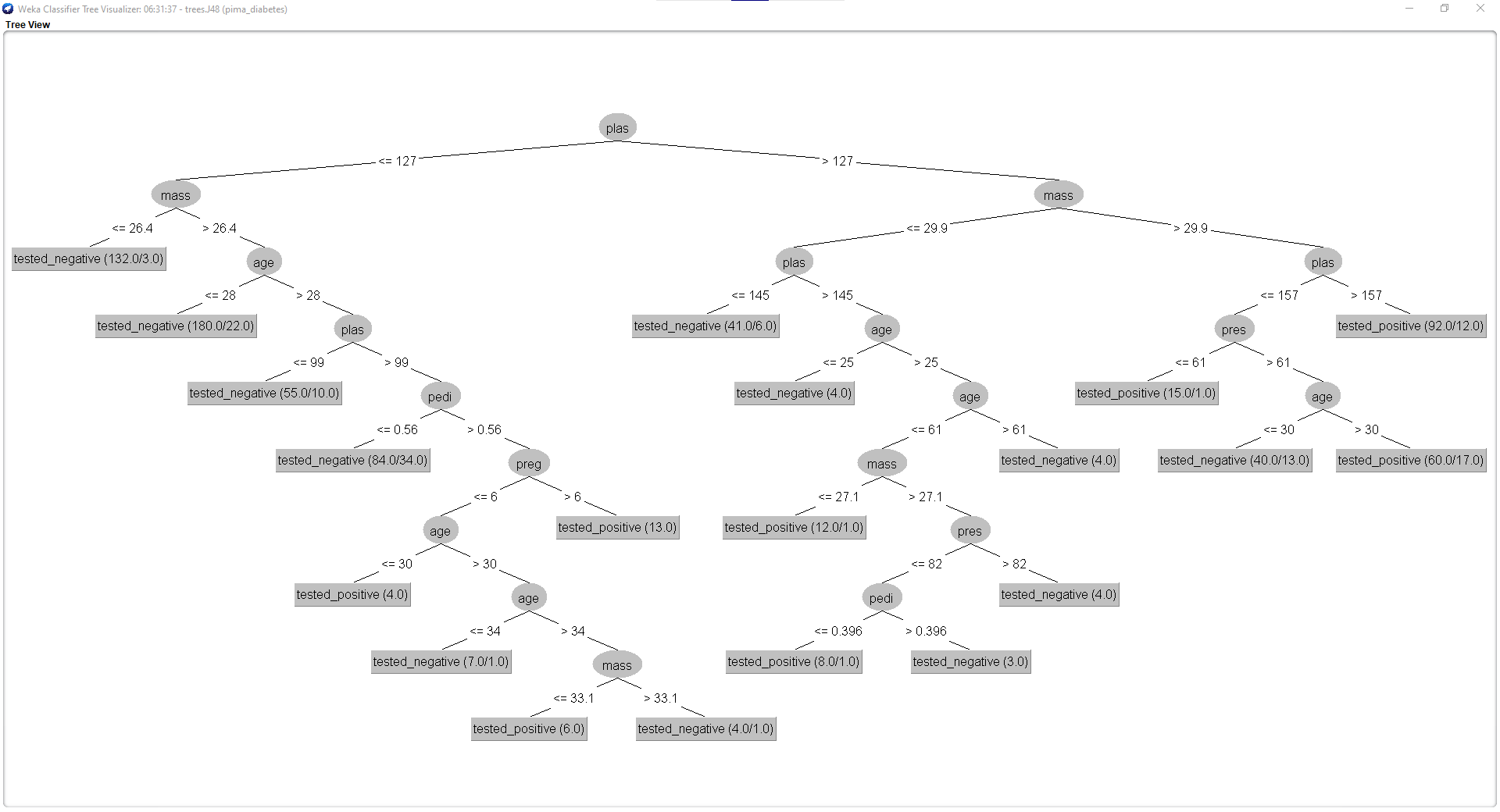


Figure 1: Decision tree of 768 data from diabetes.arff data.

After removing some data, there are only 21 data left in the data set. Then the decision tree derived from it is shown below.

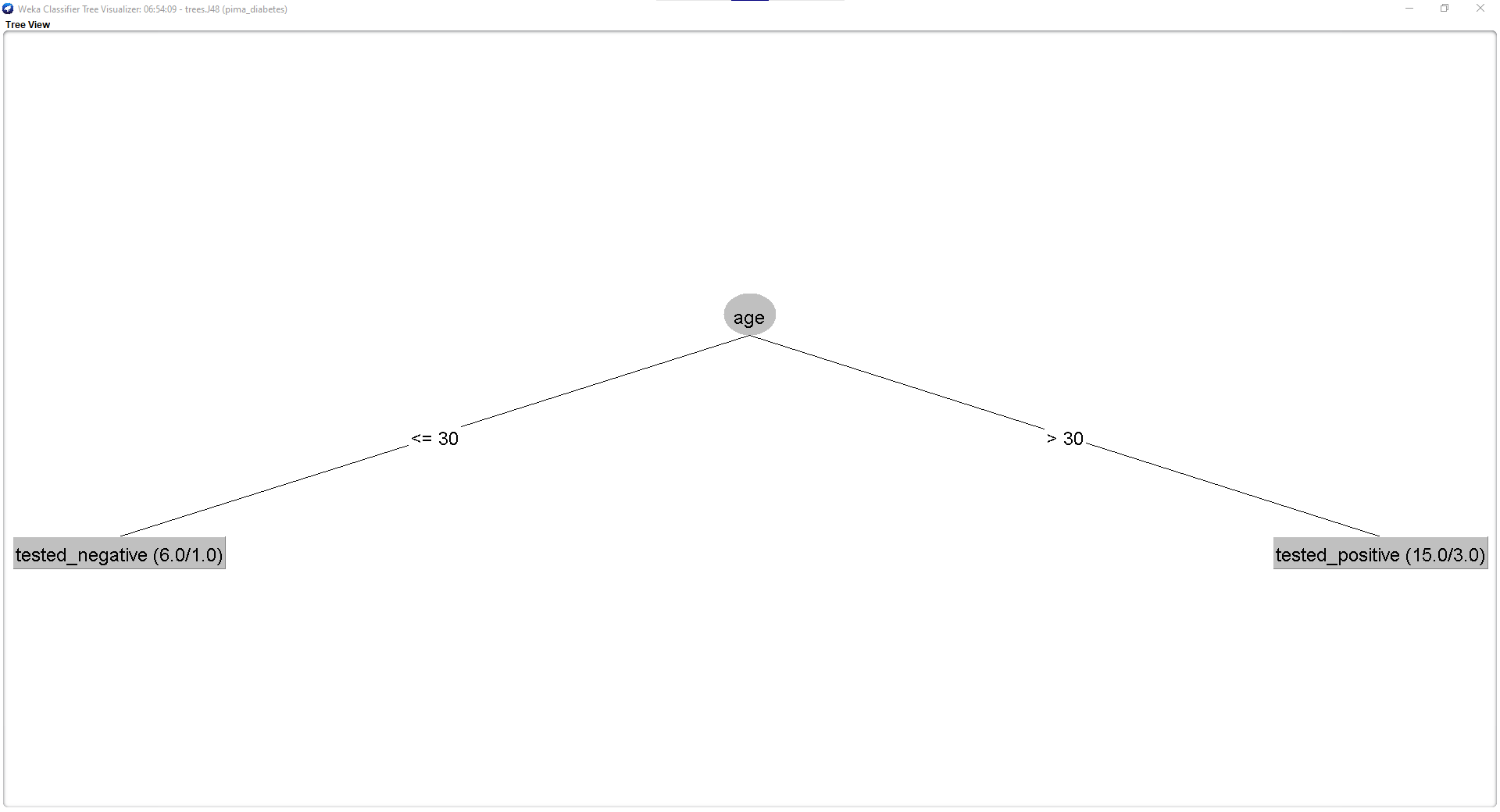


Figure 2: Decision tree of 21 data from diabetes.arff data.

The decision tree from figure 1 is way more precise than that from figure 2 and the predicted value from figure 2 will not as much as accurate as from figure 1. Therefore, the small amount of data collected by the hospital will give a prediction if a person has diabetes or not but the result might not be 100% accurate. To have a very accurate and rigid decision, the hospital needs way more amounts of data than this.

## **iv**

The “diabetes.arff” data has been loaded in weka from the file explorer which is shown below.

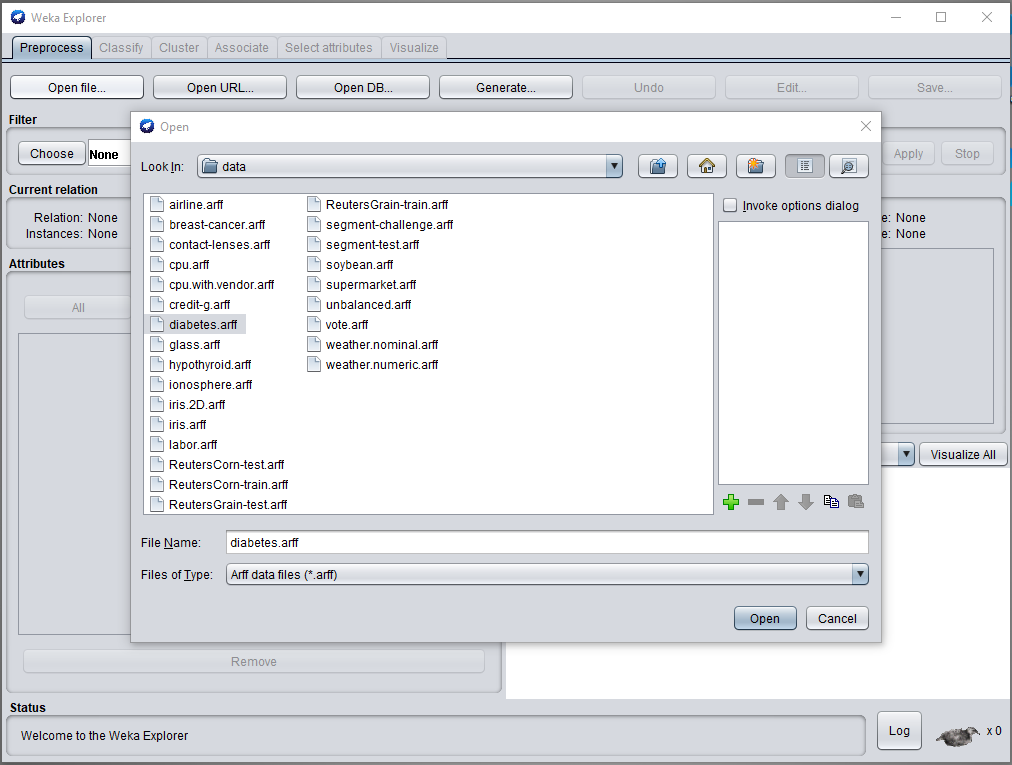


Figure 3: Importing diabetes.arff in WEKA

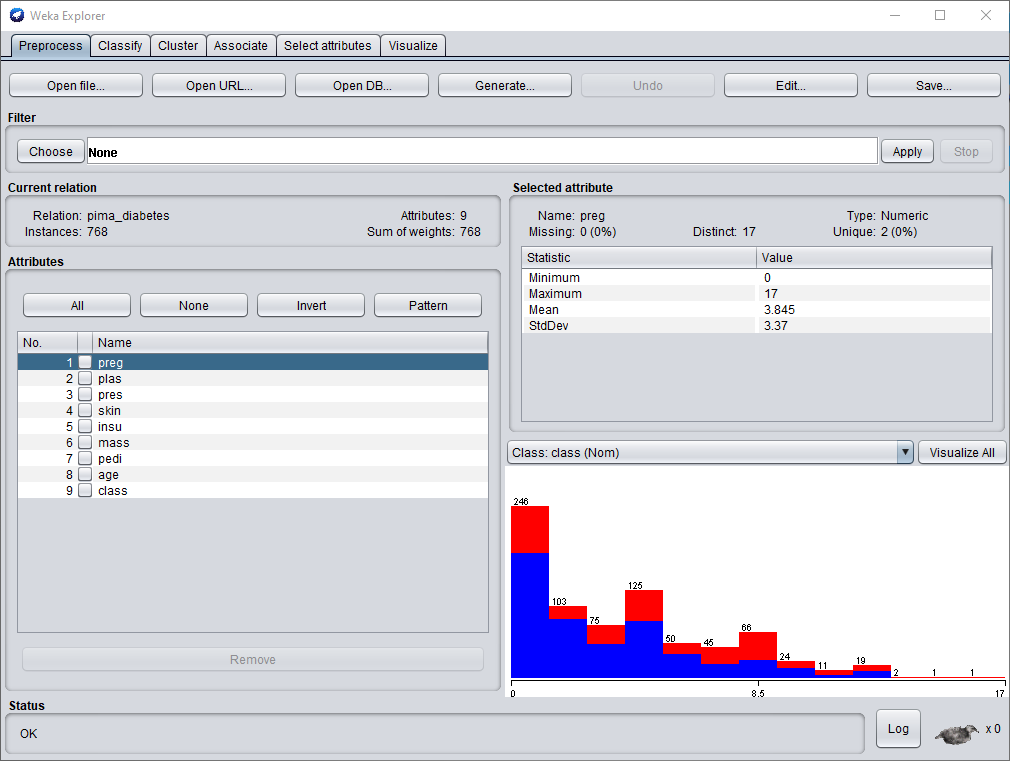


Figure 4: diabetes.arff data in WEKA

The data model for diabetes.arff data has been created which is shown below.

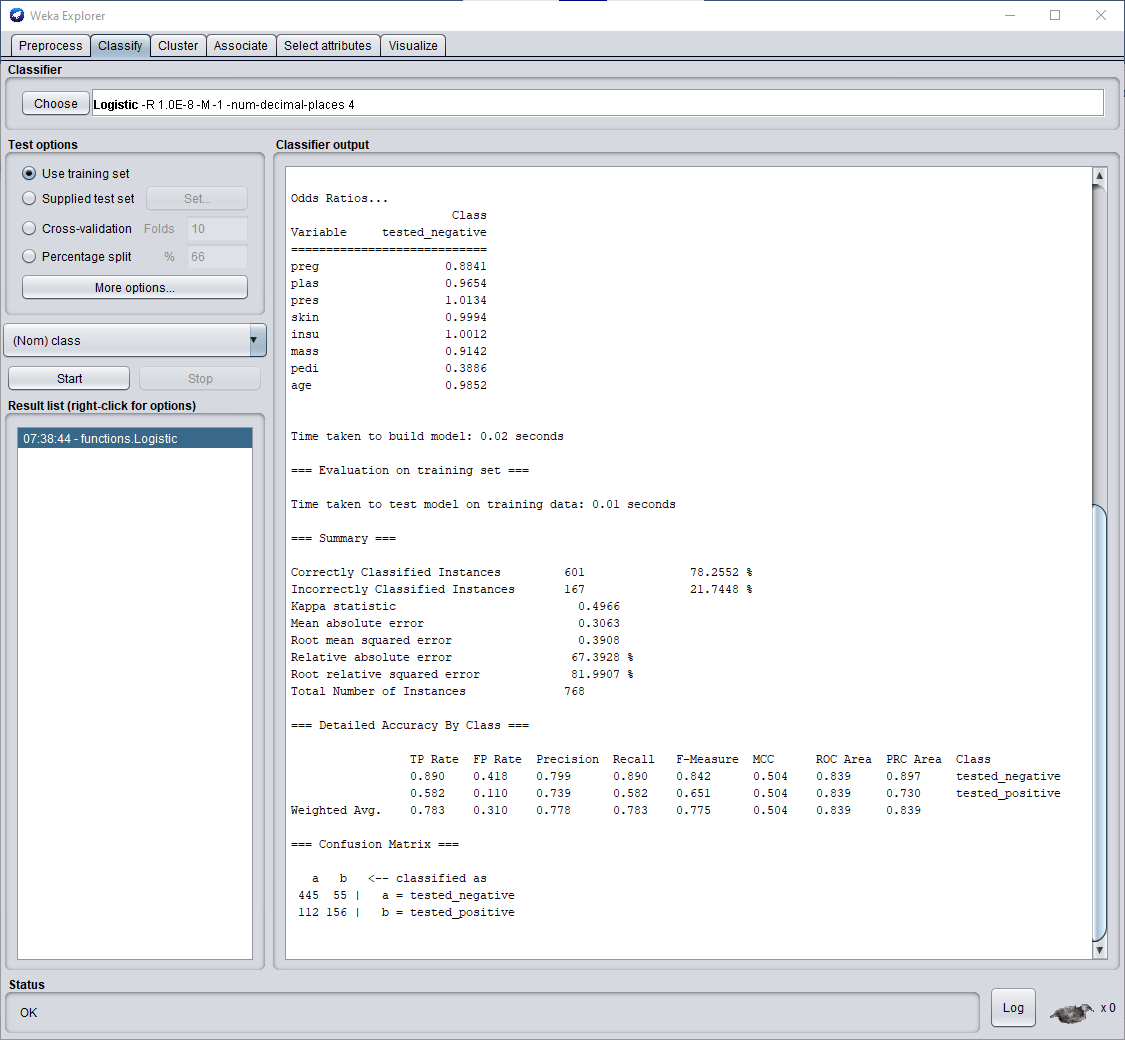


Figure 5: Data model creation for diabetes.arff

The decision tree of diabetes.arff data has been created which is shown below.

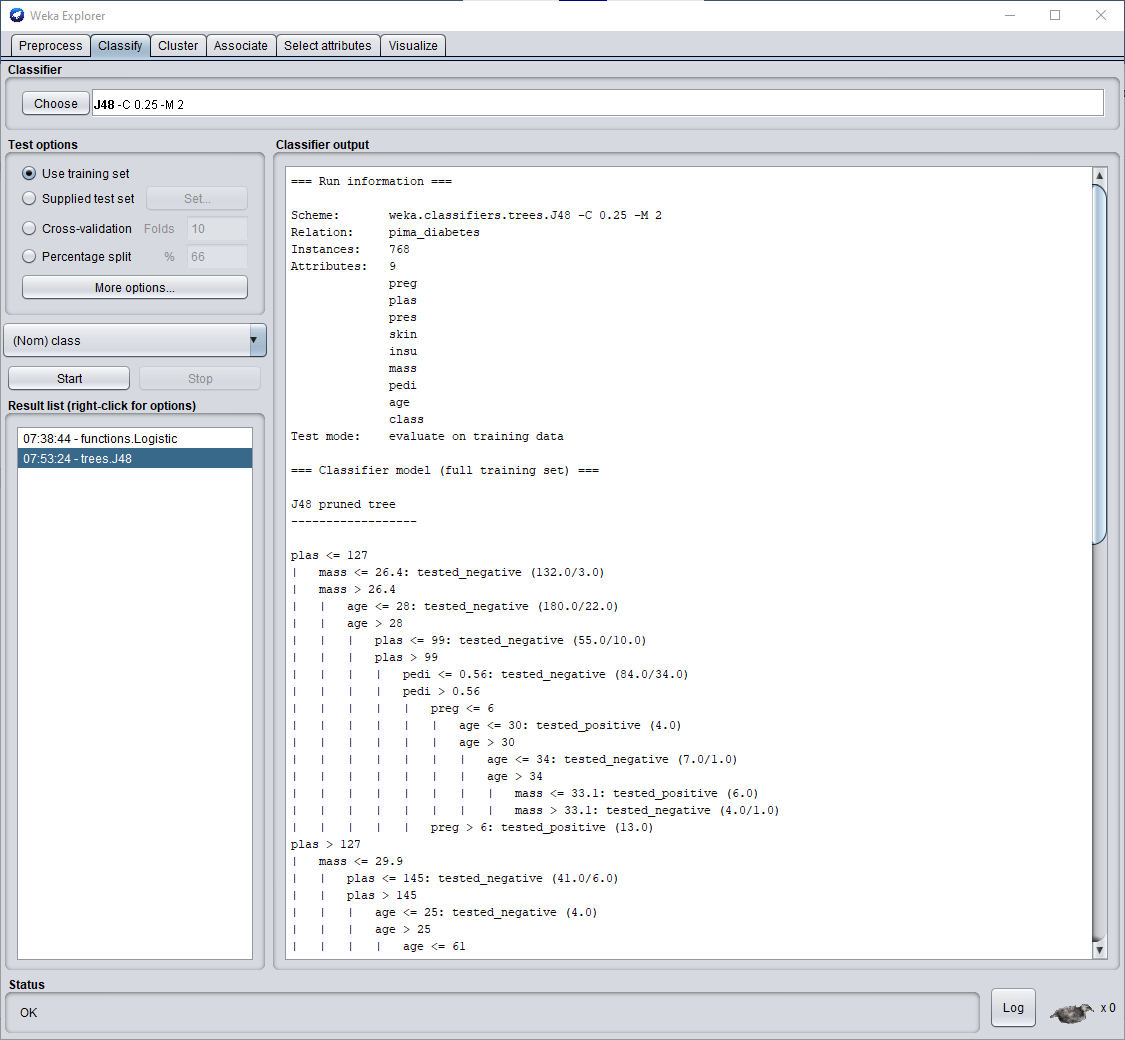


Figure 6: Creating decision tree for diabetes.arff data

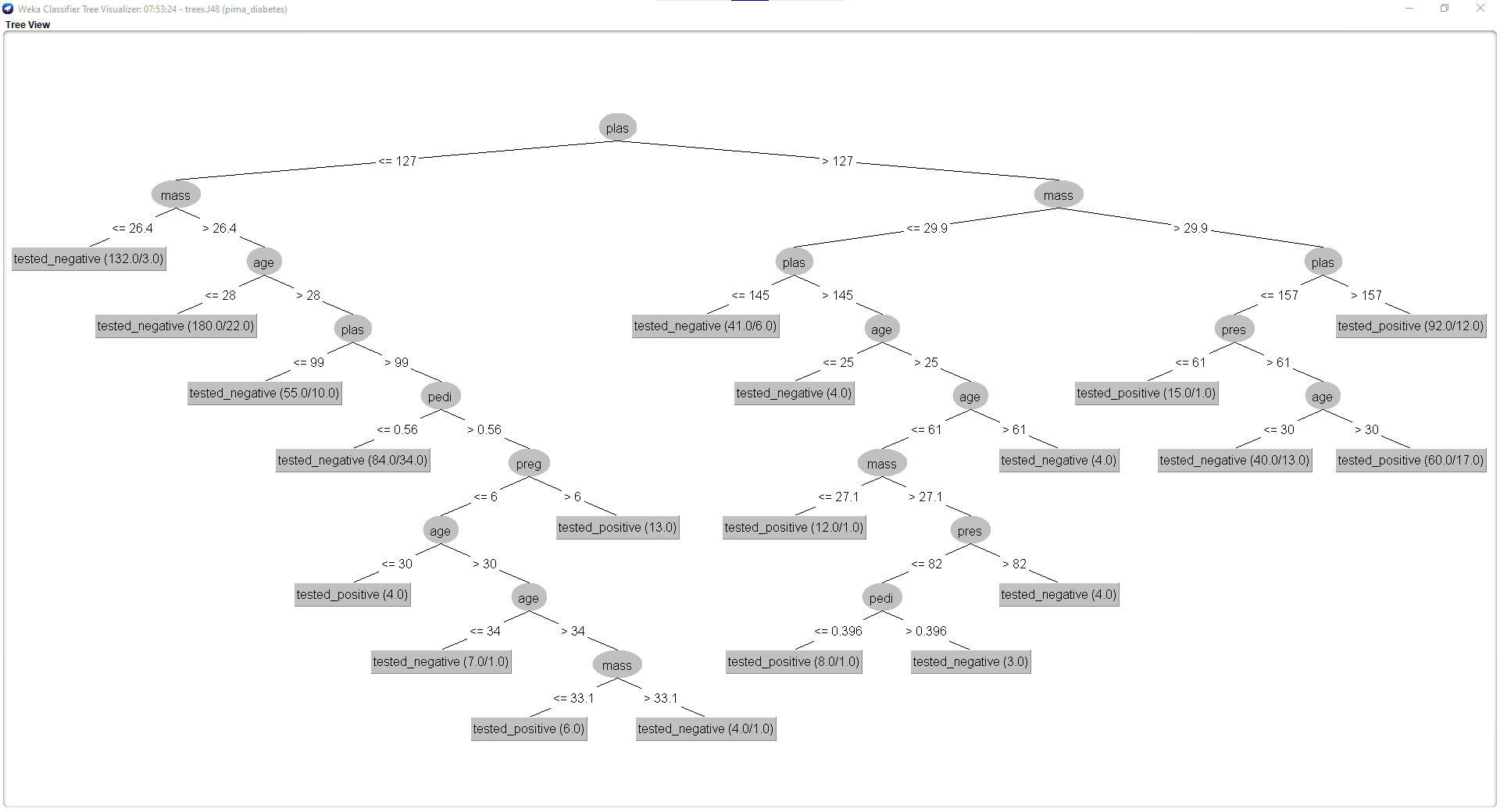


Figure 7: Decision tree of diabetes.arff data

The data from diabetes.arff has been edited, only 10 data are left in it and the predicted value of the data are replaced by “?” which is shown below.

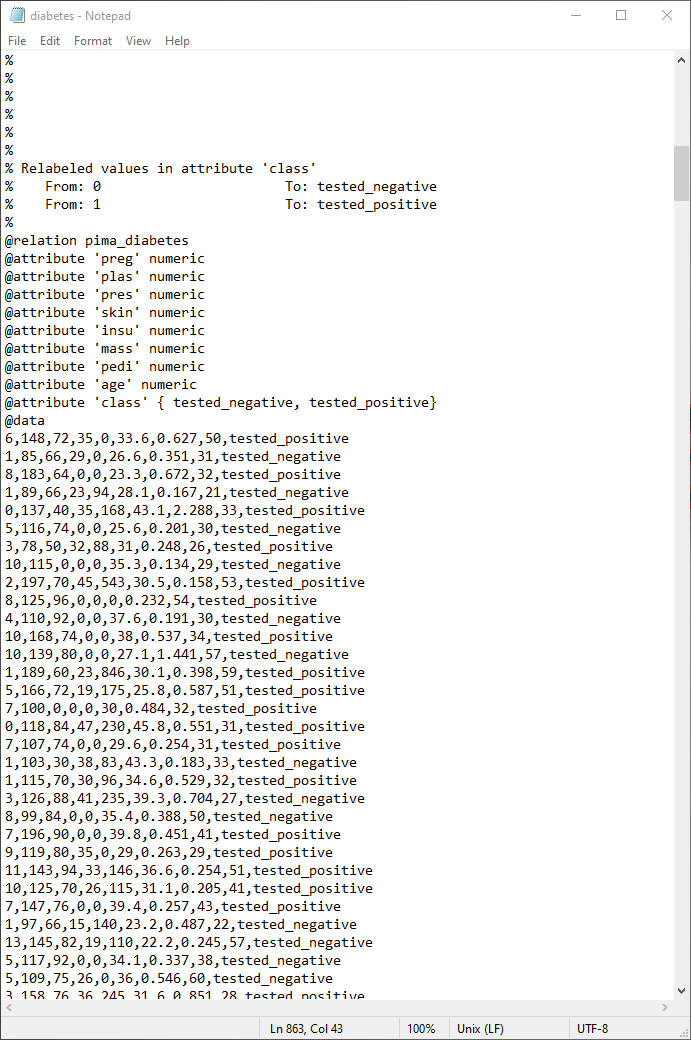


Figure 8: diabetes.arff before editing

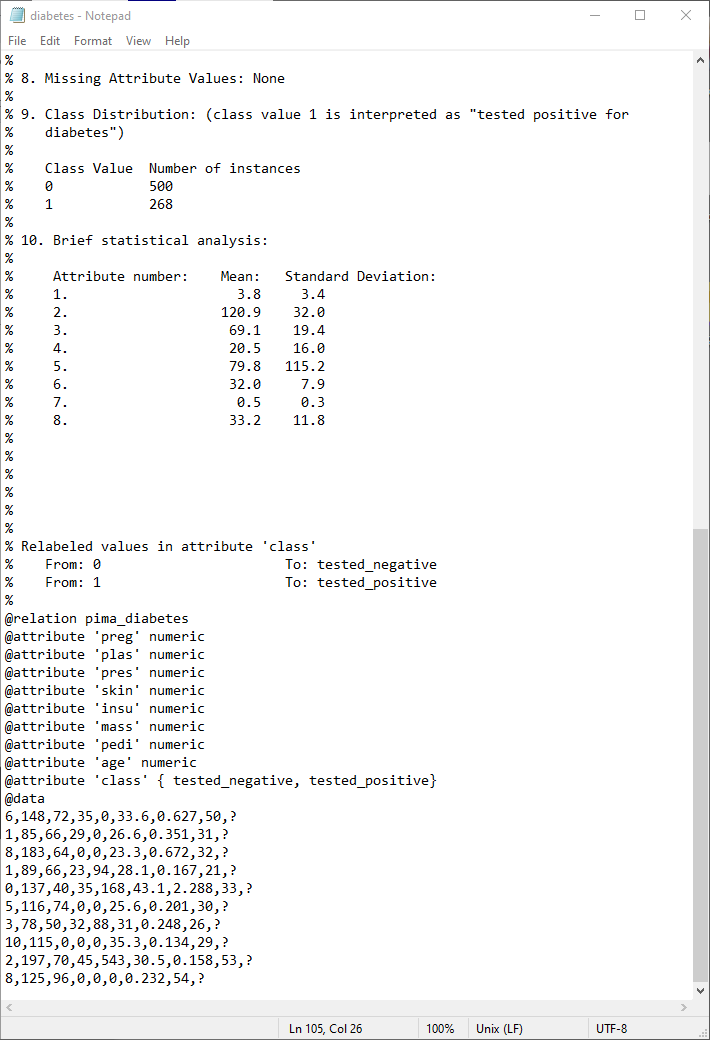


Figure 9: diabetes.arff after editing

Then the edited data has been tested with the help of the created model. This is shown below.

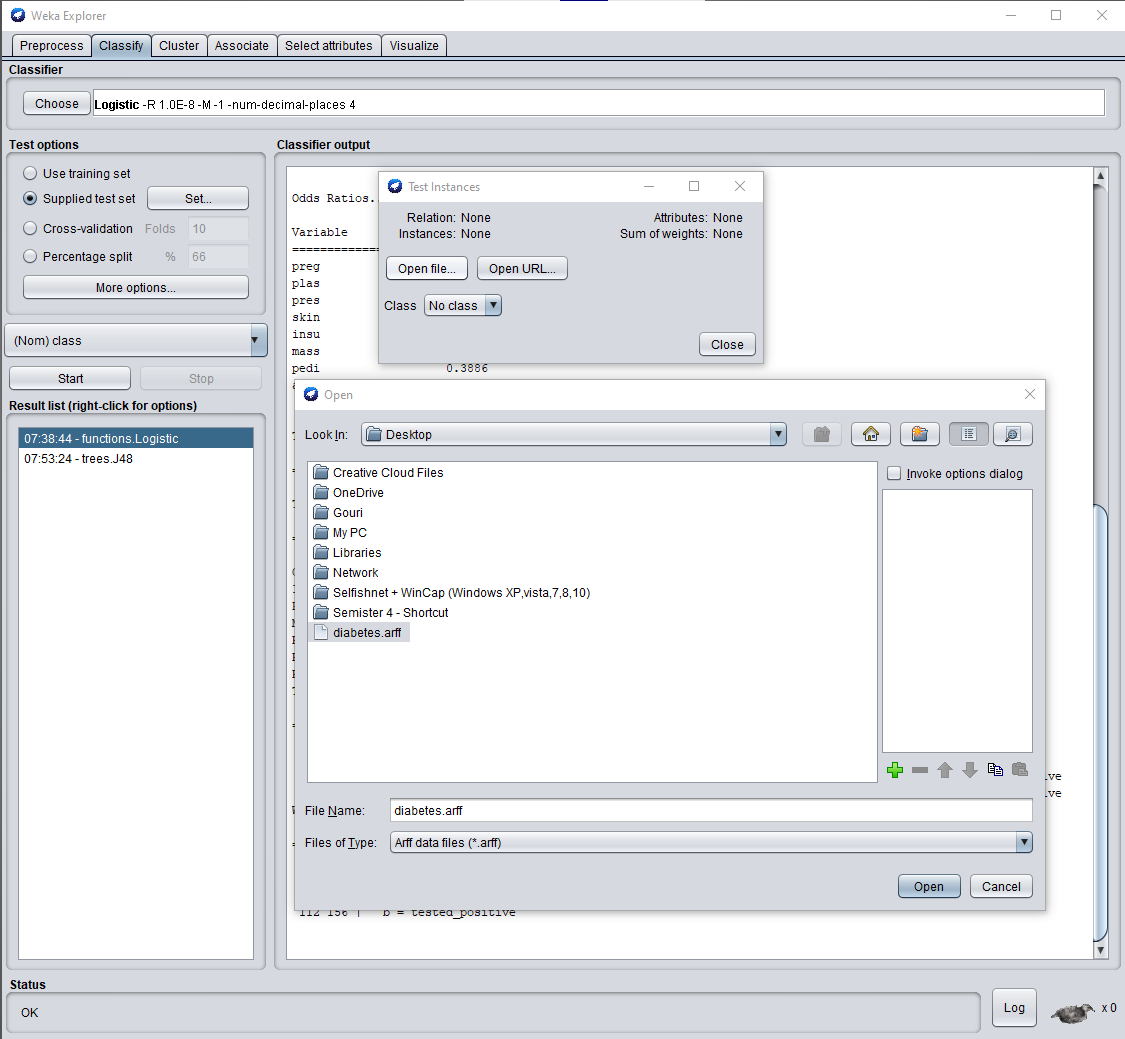


Figure 10: Importing edited diabetes.arff

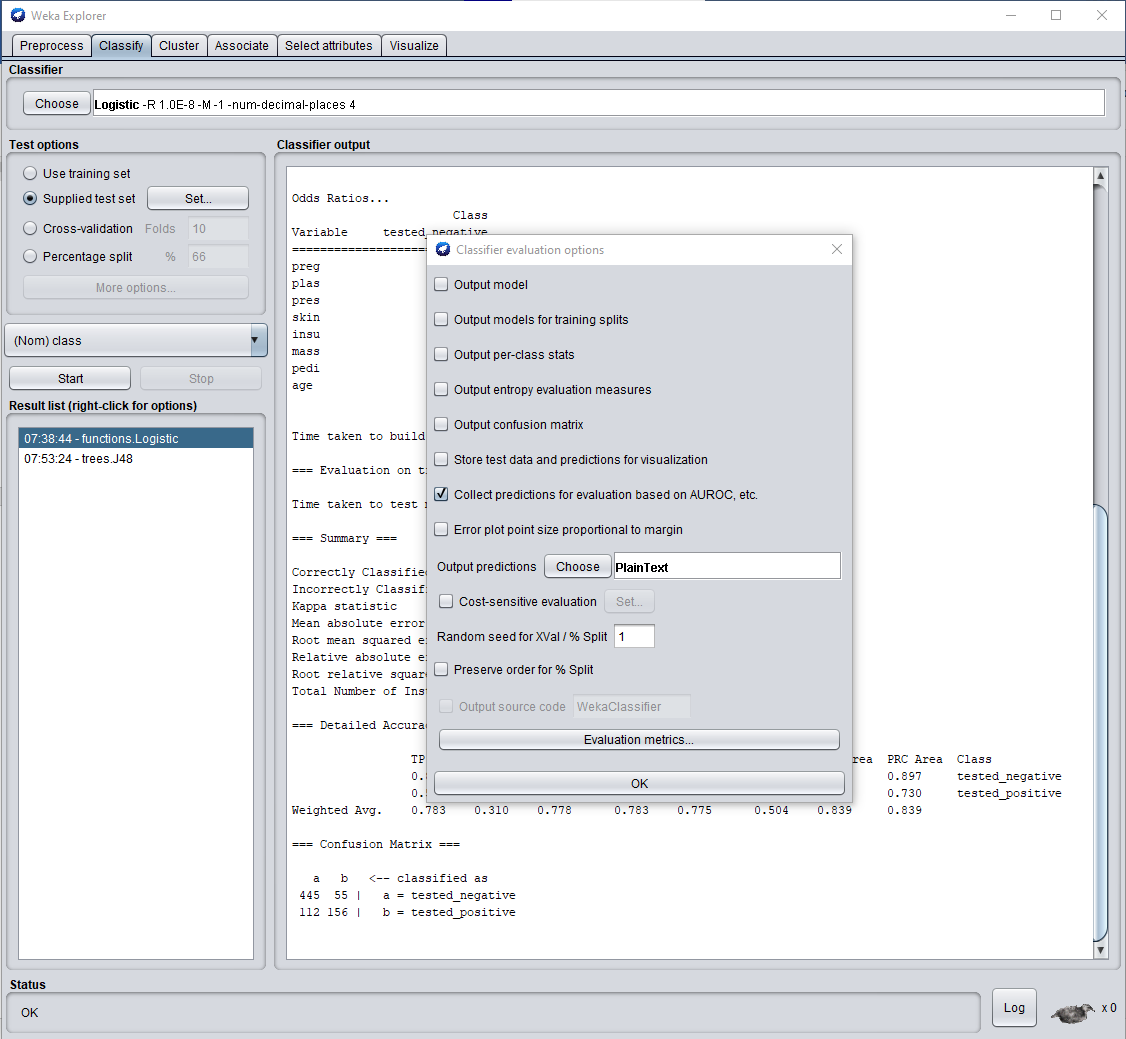


Figure 11: Choosing output options for testing

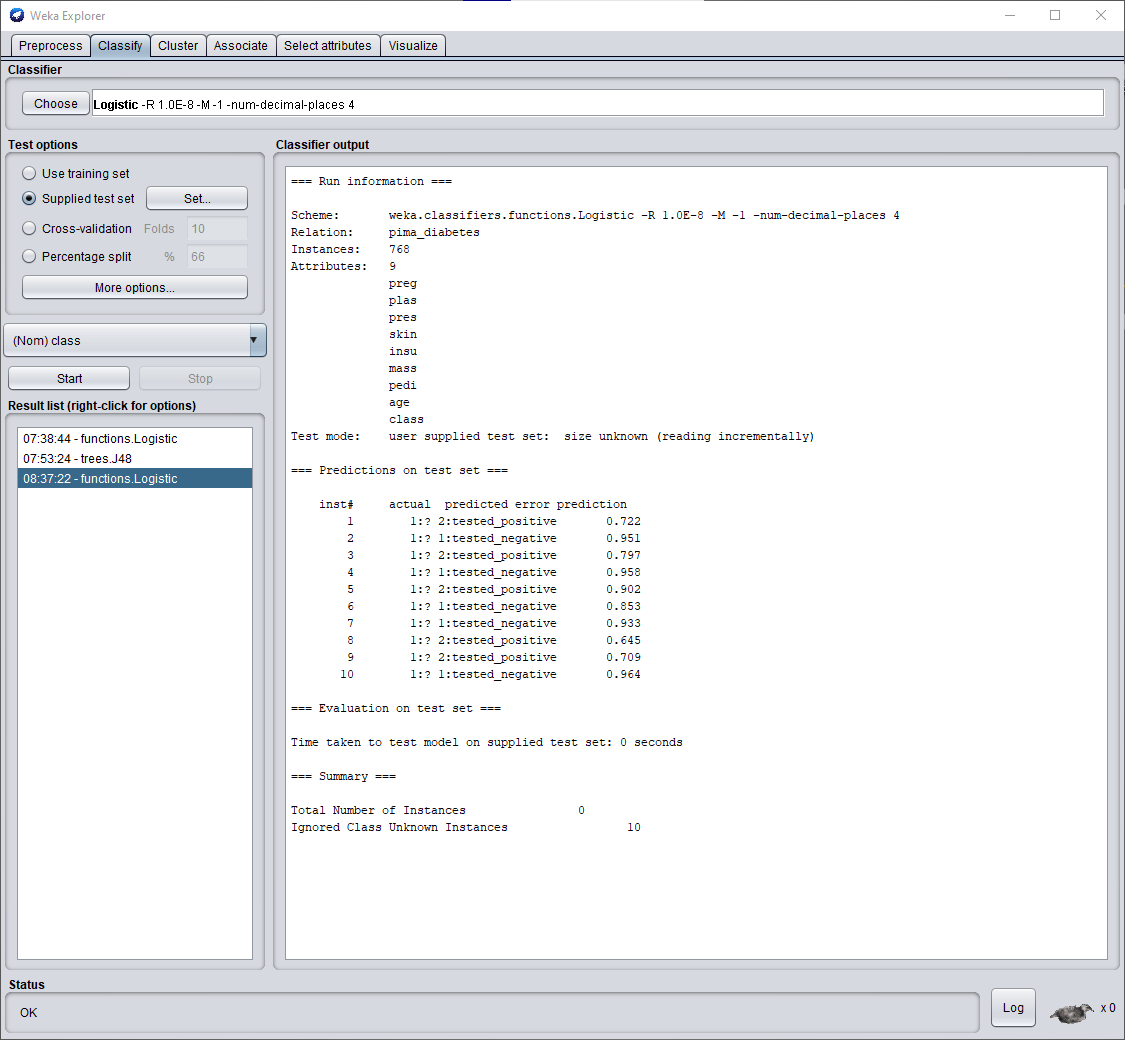


Figure 12: Predicted result value of tested diabetes.arff

The result shown above is not 100% accurate because the new predicted value of no.7, 8, and 10 do not match with the actual value of the data. The remaining predictions are correct.

Some attributes value from edited diabetes.arrf is again changed and the testing is again done which is shown below.

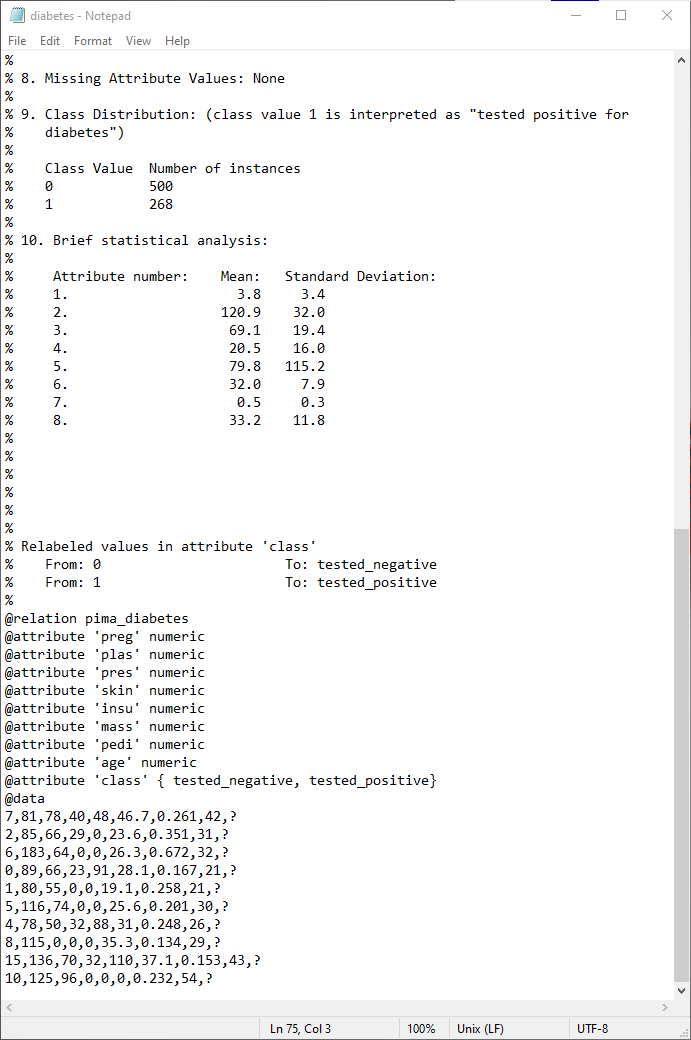


Figure 13: New edited diabetes.arff

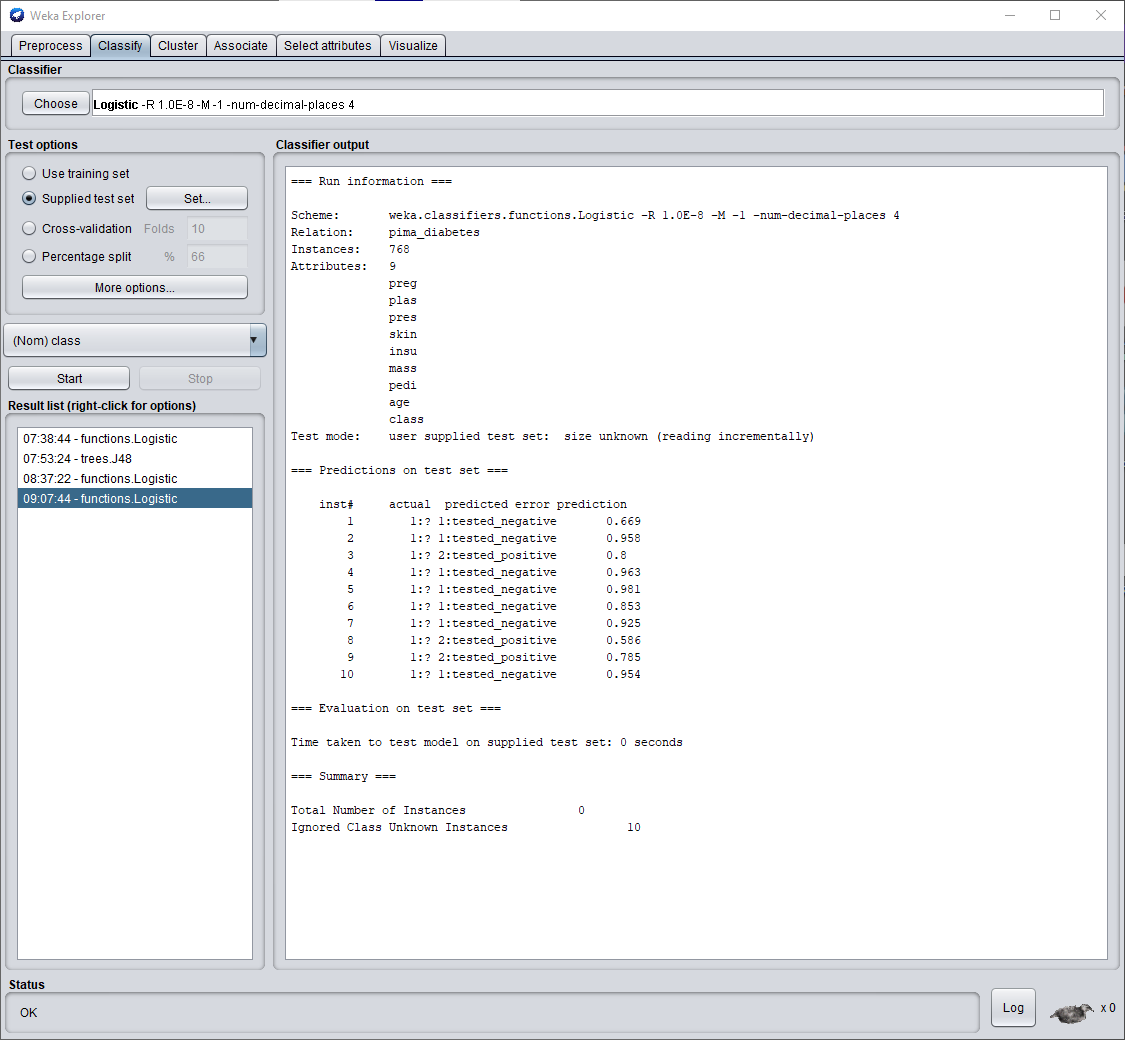


Figure 14: New predicted result of edited diabetes.arff

After going through the same process explained given above, the new prediction has been evaluated which is different from the previously predicted values.