CLASS: TE CMPN A ROLL NO. : 19

NAME: REBECCA DIAS PID: 182027

# EXPERIMENT : 8A

# Classification Algorithm using WEKA

**Aim: -** To Implement all the Classification algorithm using WEKA (Naïve Bayes, ID3, Regression).

**S/W Requirement: -** WEKA Tool

# Theory: -

## 1.Application of all different Classification algorithm: Naïve Bayes, ID3, Regression:

* Naïve Bayes: As this algorithm is fast and efficient, you can use it to make real-time predictions.
* This algorithm is popular for multi-class predictions. You can find the probability of multiple target classes easily by using this algorithm.
* Email services (like Gmail) use this algorithm to figure out whether an email is a spam or not. This algorithm is excellent for spam filtering.
* Its assumption of feature independence, and its effectiveness in solving multi-class problems, makes it perfect for performing Sentiment Analysis. Sentiment Analysis refers to the identification of positive or negative sentiments of a target group (customers, audience, etc.)
* Collaborative Filtering and the Naive Bayes algorithm work together to build recommendation systems. These systems use data mining and machine learning to predict if the user would like a particular resource or not.

ID3:

The ID3 algorithm is a decision tree classification algorithm based on information entropy.

Regression:

1. Predictive Analytics:

Predictive analytics i.e. forecasting future opportunities and risks is the most prominent application of regression analysis in business. Demand analysis, for instance, predicts the number of items which a consumer will probably purchase. However, demand is not the only dependent variable when it comes to business.

1. Operation Efficiency:

Regression models can also be used to optimize business processes. A factory manager, for example, can create a statistical model to understand the impact of oven temperature on the shelf life of the cookies baked in those ovens.

1. Supporting Decisions:

Businesses today are overloaded with data on finances, operations and customer purchases. Increasingly, executives are now leaning on data analytics to make informed business decisions that have statistical significance, thus eliminating the intuition and gut feel. RA can bring a scientific angle to the management of any businesses.

## 2.Advantage and Disadvantage of all different Classification Algorithms:

* Naïve Bayes:

**Advantages:**

1. This algorithm works very fast and can easily predict the class of a test dataset.
2. You can use it to solve multi-class prediction problems as it’s quite

useful with them.

1. Naive Bayes classifier performs better than other models with less training data if the assumption of independence of features holds.

**Disadvantages:**

1. If your test data set has a categorical variable of a category that wasn’t present in the training data set, the Naive Bayes model will assign it zero probability and won’t be able to make any predictions in this regard. This phenomenon is called ‘Zero Frequency,’ and you’ll have to use a smoothing technique to solve this problem.
2. This algorithm is also notorious as a lousy estimator. So, you shouldn’t take the probability outputs of ‘predict\_proba’ too seriously.

* ID3:

**Advantages:**

1. Assuming that the space contains all decision trees, the search space is complete.
2. Good robustness and not affected by noise.
3. The instances with missing attribute values can be trained.

**Disadvantages:**

1. ID3 only considers sub-type features, but does not consider continuous features, such as length and density are continuous values, which cannot be used in ID3. This greatly limits the use of ID3.
2. The ID3 algorithm does not consider missing values.
3. The problem of overfitting is not considered.

* Regression:

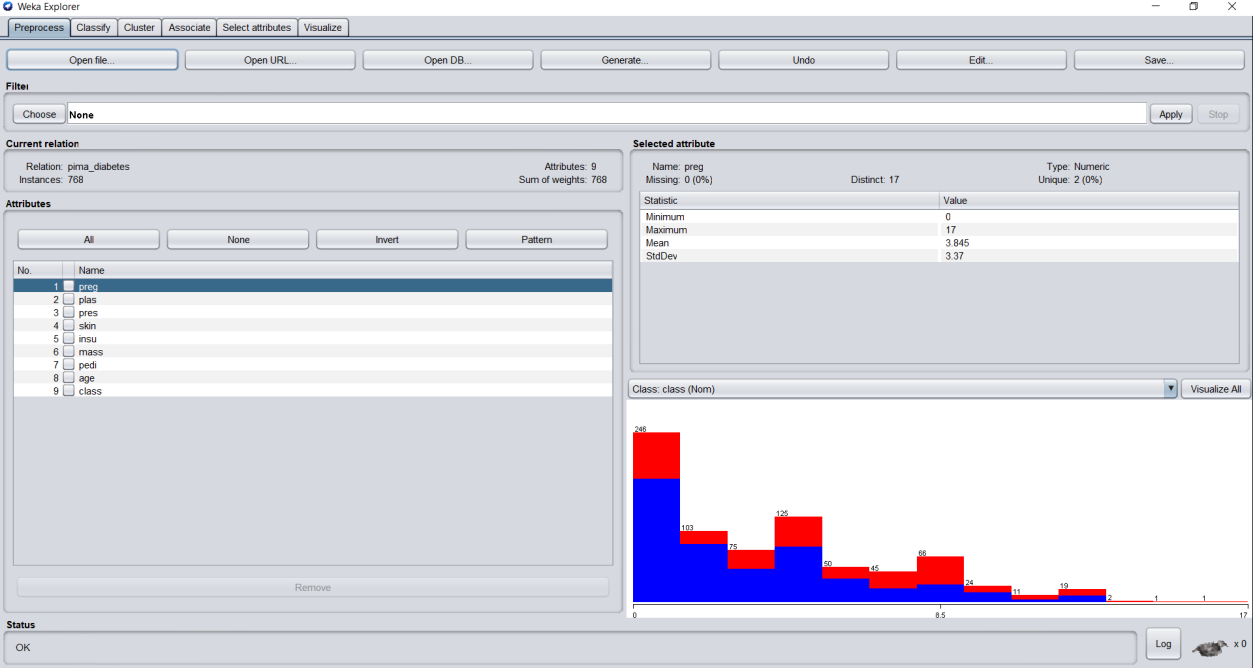
**Advantages:**

1. Linear Regression is a very simple algorithm that can be implemented very easily to give satisfactory results.
2. Linear regression fits linearly seperable datasets almost perfectly and is often used to find the nature of the relationship between variables.
3. Regularization is a technique that can be easily implemented and is capable of effectively reducing the complexity of a function so as to reduce the risk of overfitting.

**Disadvantages:**

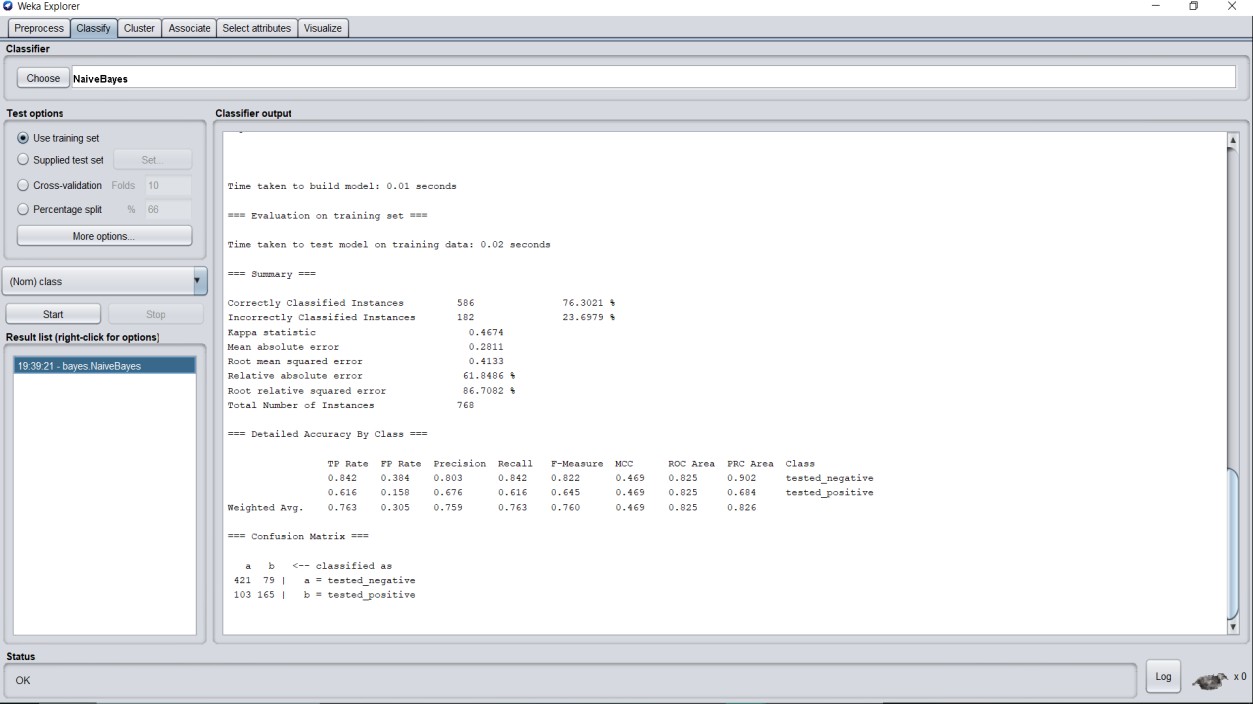
1. Underfitting: A sitiuation that arises when a machine learning model fails to capture the data properly.This typically occurs when the hypothesis function cannot fit the data well.
2. Outliers of a data set are anomalies or extreme values that deviate from the other data points of the distribution.Data outliers can damage the performance of a machine learning model drastically and can often lead to models with low accuracy.
3. Outliers can have a very big impact on linear regression's performance and hence they must be dealt with appropriately before linear regression is applied on the dataset.

# Implementation: -



## **OUTPUT: -**

1. **Naïve Bayes:**



**Classifier Output:**

=== Run information ===

Scheme: weka.classifiers.bayes.NaiveBayes

Relation: pima\_diabetes Instances: 768

Attributes: 9

preg plas pres skin insu mass pedi age class

Test mode: evaluate on training data

=== Classifier model (full training set) ===

Naive Bayes Classifier

Class

Attribute tested\_negative tested\_positive (0.65) (0.35)

===============================================

preg

|  |  |  |
| --- | --- | --- |
| mean | 3.4234 | 4.9795 |
| std. dev. | 3.0166 | 3.6827 |
| weight sum | 500 | 268 |
| precision | 1.0625 | 1.0625 |

plas

|  |  |  |
| --- | --- | --- |
| mean | 109.9541 | 141.2581 |
| std. dev. | 26.1114 | 31.8728 |
| weight sum | 500 | 268 |
| precision | 1.4741 | 1.4741 |

pres

|  |  |  |
| --- | --- | --- |
| mean | 68.1397 | 70.718 |
| std. dev. | 17.9834 | 21.4094 |
| weight sum | 500 | 268 |
| precision | 2.6522 | 2.6522 |

skin

|  |  |  |
| --- | --- | --- |
| mean | 19.8356 | 22.2824 |
| std. dev. | 14.8974 | 17.6992 |
| weight sum | 500 | 268 |
| precision | 1.98 | 1.98 |

insu

|  |  |  |
| --- | --- | --- |
| mean | 68.8507 | 100.2812 |
| std. dev. | 98.828 | 138.4883 |
| weight sum | 500 | 268 |
| precision | 4.573 | 4.573 |

mass

|  |  |  |
| --- | --- | --- |
| mean | 30.3009 | 35.1475 |
| std. dev. | 7.6833 | 7.2537 |
| weight sum | 500 | 268 |
| precision | 0.2717 | 0.2717 |

pedi

|  |  |  |
| --- | --- | --- |
| mean | 0.4297 | 0.5504 |
| std. dev. | 0.2986 | 0.3715 |
| weight sum | 500 | 268 |
| precision | 0.0045 | 0.0045 |

age

|  |  |  |
| --- | --- | --- |
| mean | 31.2494 | 37.0808 |
| std. dev. | 11.6059 | 10.9146 |
| weight sum | 500 | 268 |
| precision | 1.1765 | 1.1765 |

Time taken to build model: 0.01 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.02 seconds

|  |  |  |  |
| --- | --- | --- | --- |
| === Summary === | |  | |
| Correctly Classified Instances | | 586 | 76.3021 % |
| Incorrectly Classified Instances | | 182 | 23.6979 % |
| Kappa statistic | 0.4674 | | |
| Mean absolute error | 0.2811 | | |
| Root mean squared error | 0.4133 | | |
| Relative absolute error | 61.8486 % | | |
| Root relative squared error | 86.7082 % | | |

Total Number of Instances 768

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

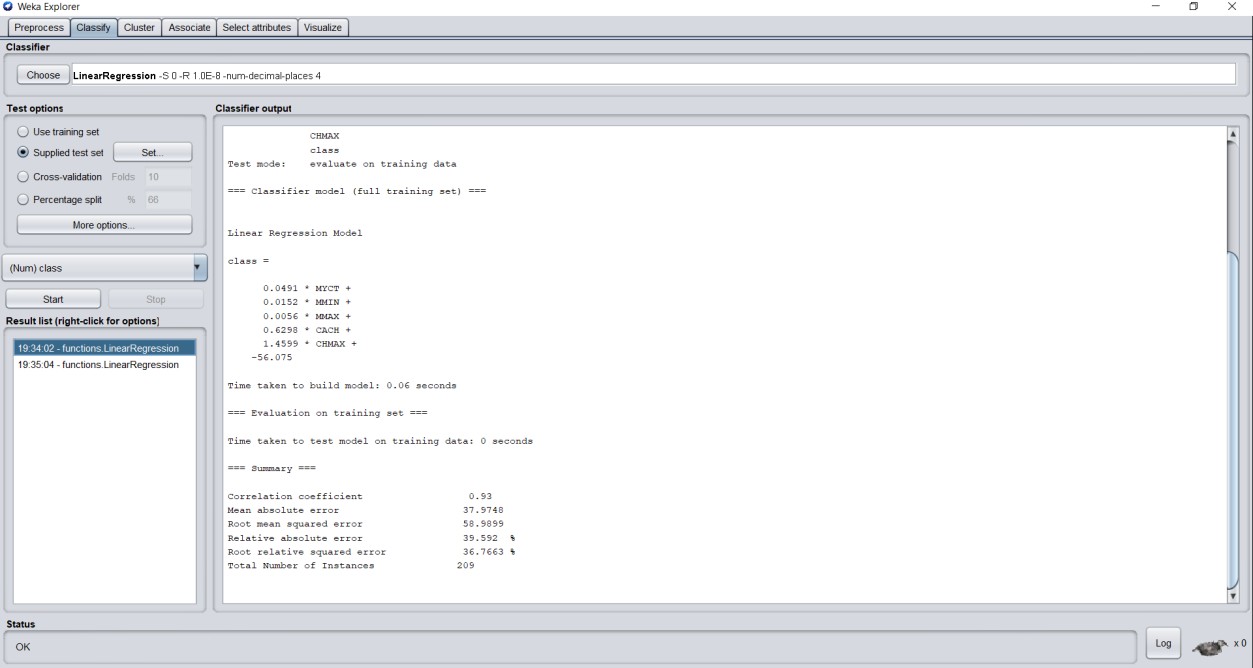
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0.842  tested\_negative | 0.384 | 0.803 | 0.842 | 0.822 | 0.469 | 0.825 | 0.902 |
| 0.616  tested\_positive | 0.158 | 0.676 | 0.616 | 0.645 | 0.469 | 0.825 | 0.684 |
| Weighted Avg. 0.826 | 0.763 | 0.305 | 0.759 | 0.763 | 0.760 | 0.469 | 0.825 |

=== Confusion Matrix ===

a b <-- classified as

421 79 | a = tested\_negative 103 165 | b = tested\_positive

## **Linear Regression:**



**Classifier Output:**

=== Run information ===

Scheme: weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8 - num-decimal-places 4

Relation: pima\_diabetes Instances: 768

Attributes: 9

preg plas pres skin insu mass pedi age class

Test mode: evaluate on training data

=== Classifier model (full training set) ===

Linear Regression Model

age =

1.6654 \* preg + 0.0573 \* plas + 0.1034 \* pres +

-0.0911 \* skin +

1.4283 \* class=tested\_positive + 14.1346

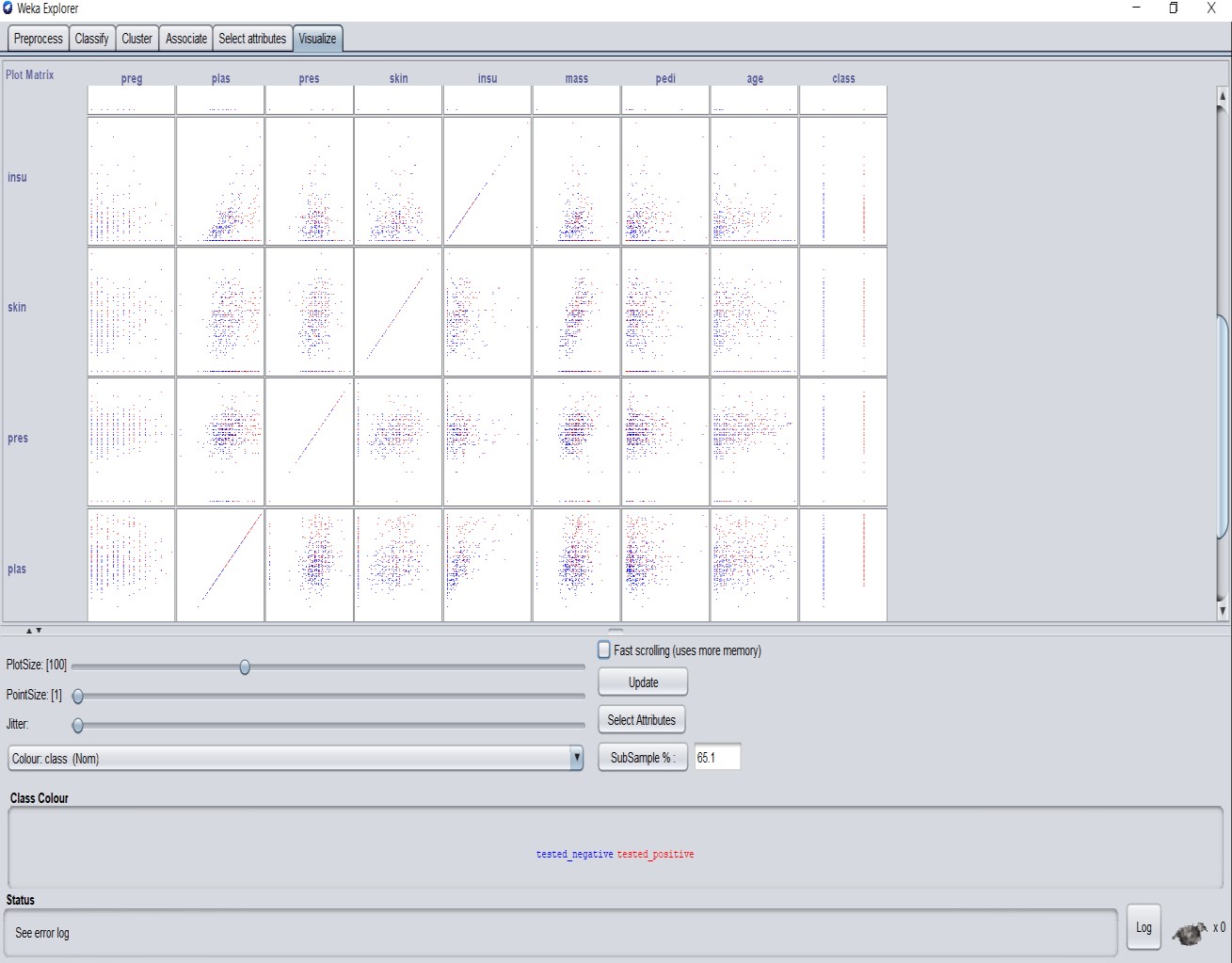
Time taken to build model: 0.02 seconds

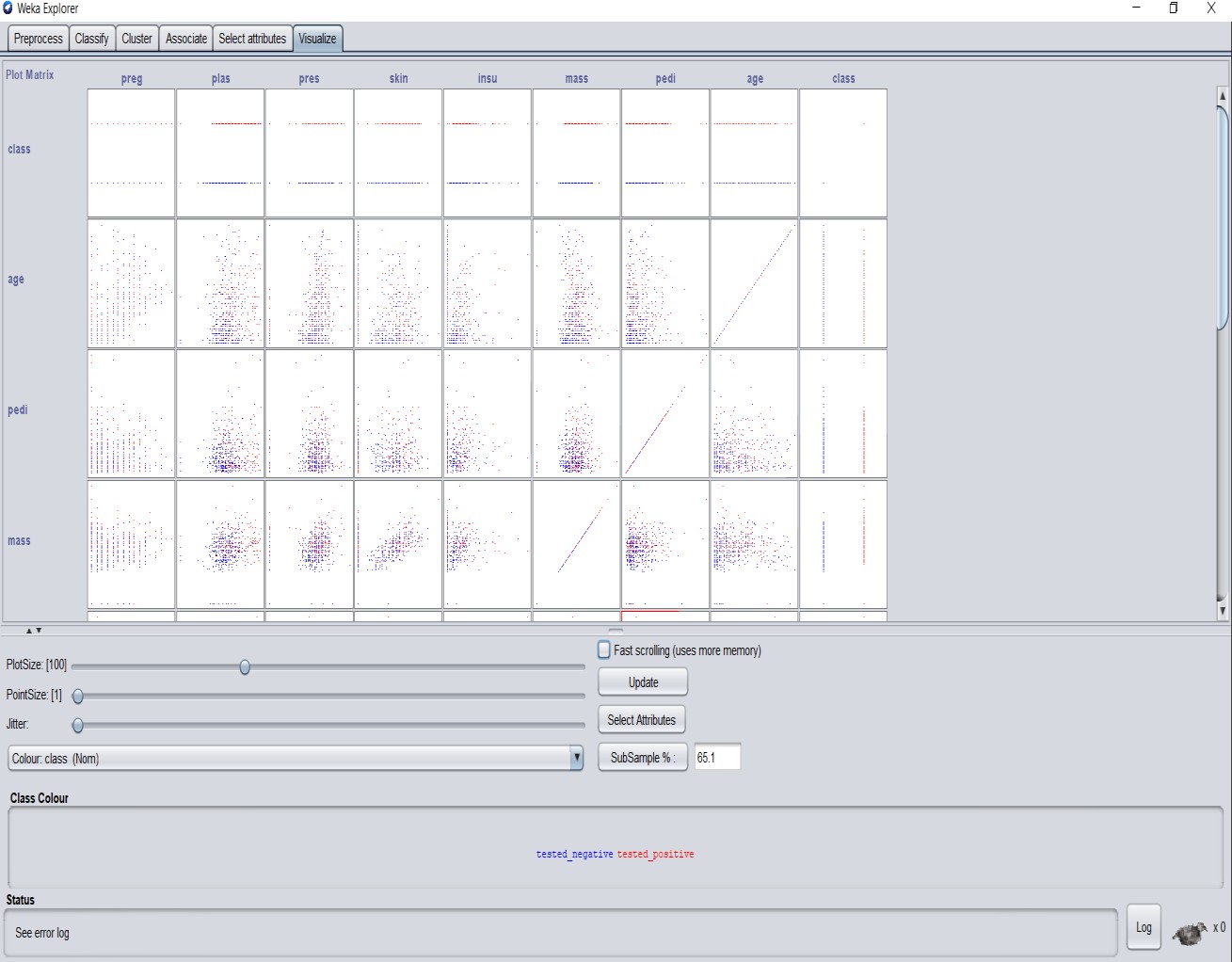
=== Evaluation on training set ===

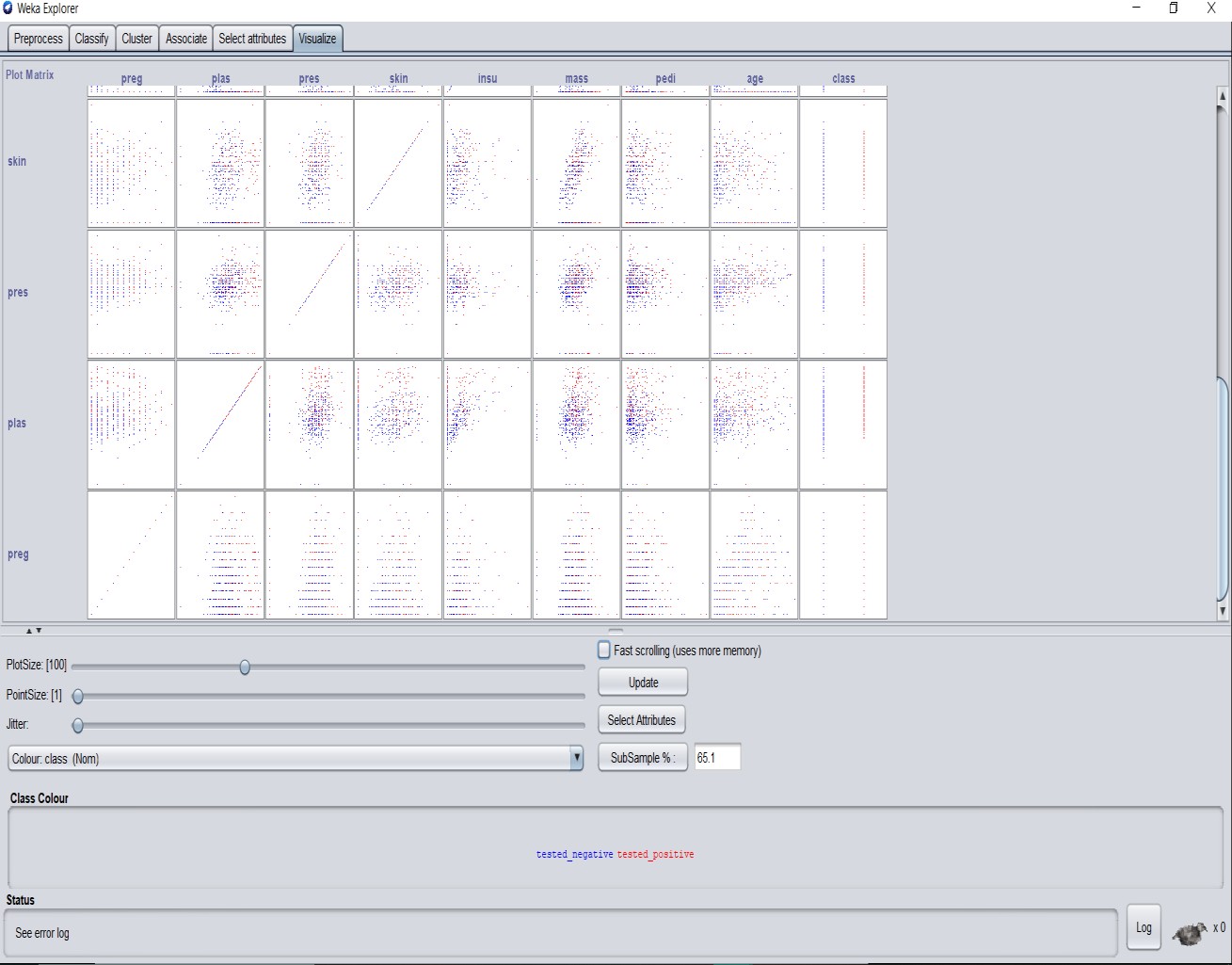
Time taken to test model on training data: 0.01 seconds

=== Summary ===

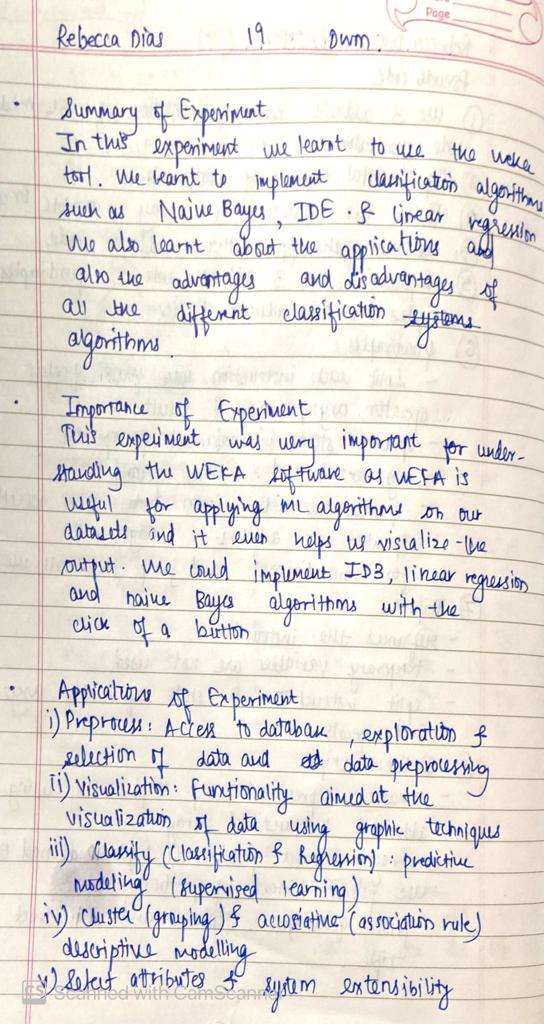
|  |  |
| --- | --- |
| Correlation coefficient | 0.6078 |
| Mean absolute error | 6.7393 |
| Root mean squared error | 9.3323 |
| Relative absolute error | 70.3002 % |
| Root relative squared error | 79.4063 % |
| Total Number of Instances | 768 |
| **Visualization:** |  |







# Conclusion: -

****

# EXPERIMENT NO. 8B: Clustering Algorithm using WEKA

**Aim: -** To Implement all the Clustering algorithm using WEKA (K Means, Agglomerative).

**S/W Requirement: -** WEKA Tool

# Theory: -

## Application of all different Clustering based on types (Partition and Hierarchal):

**Partition:**

* + Academic performance
  + Diagnostic systems
  + Search engines
  + Wireless sensor networks **Hierarchal:**

1. US Senator Clustering through Twitter Can we find the party lines through Twitter?
2. Charting Evolution through Phylogenetic Trees How can we relate different species together?
3. Tracking Viruses through Phylogenetic Trees Can we find where a viral outbreak originated?

## Advantage and Disadvantage of each Clustering Algorithms (Partition and Hierarchal):

### Partition:

**Advantages:**

* Easy to implement
* With a large number of variables, K-Means may be computationally faster than hierarchical clustering (if K is small).
* k-Means may produce Higher clusters than hierarchical clustering
* An instance can change cluster (move to another cluster) when the

centroids are recomputed. **Disadvantages:**

* Difficult to predict the number of clusters (K-Value)
* Initial seeds have a strong impact on the final results
* The order of the data has an impact on the final results
* Sensitive to scale: rescaling your datasets (normalization or standardization) will completely change results. While this itself is not bad, not realizing that you have to spend extra a4en (on to scaling your data might be bad.

### Hierarchal:

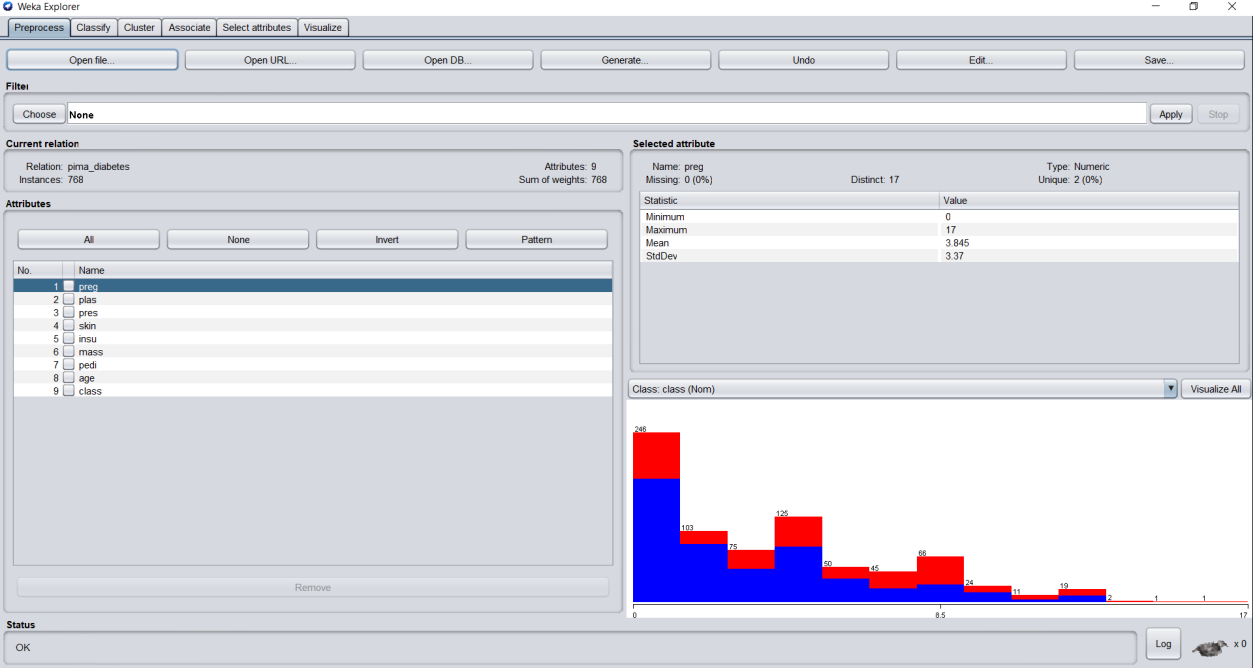
**Advantages**

1. No apriori information about the number of clusters required.
2. Easy to implement and gives best result in some cases.

### Disadvantages

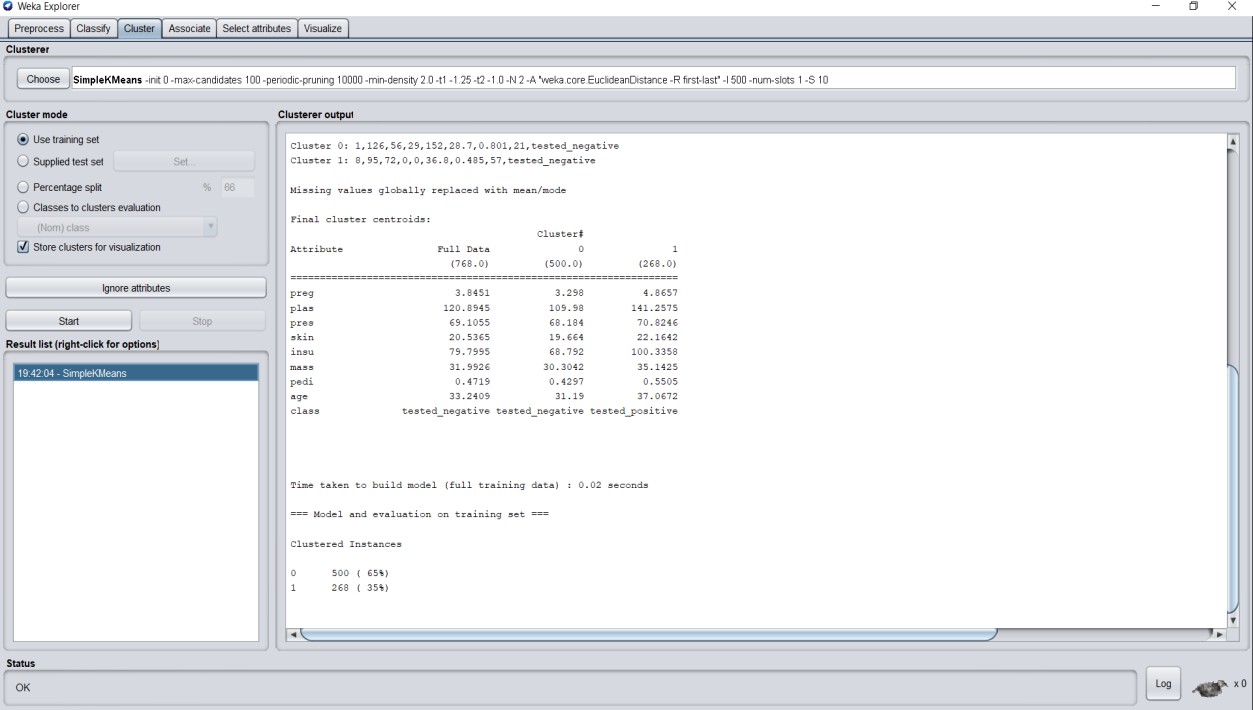
1. Algorithm can never undo what was done previously.
2. Time complexity of at least O (*n2 log n*) is required, where *‘n’* is the number of data points.
3. Based on the type of distance matrix chosen for merging different algorithms can suffer with one or more of the following:
4. Sensitivity to noise and outliers
5. Breaking large clusters

# Implementation: -



## **OUTPUT: -**

**i)K-Means:**



**Classifier Output:**

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 - periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A

"weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Relation: pima\_diabetes Instances: 768

Attributes: 9

preg plas pres skin insu mass pedi age class

Test mode: evaluate on training data

=== Clustering model (full training set) ===

kMeans

======

Number of iterations: 4

Within cluster sum of squared errors: 149.5177664581119

Initial starting points (random):

Cluster 0: 1,126,56,29,152,28.7,0.801,21,tested\_negative

Cluster 1: 8,95,72,0,0,36.8,0.485,57,tested\_negative

Missing values globally replaced with mean/mode

Final cluster centroids:

Cluster#

Attribute Full Data 0 1

(768.0) (500.0) (268.0)

============================================================

======

|  |  |  |  |
| --- | --- | --- | --- |
| preg | 3.8451 | 3.298 | 4.8657 |
| plas | 120.8945 | 109.98 | 141.2575 |
| pres | 69.1055 | 68.184 | 70.8246 |
| skin | 20.5365 | 19.664 | 22.1642 |
| insu | 79.7995 | 68.792 | 100.3358 |
| mass | 31.9926 | 30.3042 | 35.1425 |
| pedi | 0.4719 | 0.4297 | 0.5505 |
| age | 33.2409 | 31.19 | 37.0672 |

class tested\_negative tested\_negative tested\_positive

Time taken to build model (full training data) : 0.02 seconds

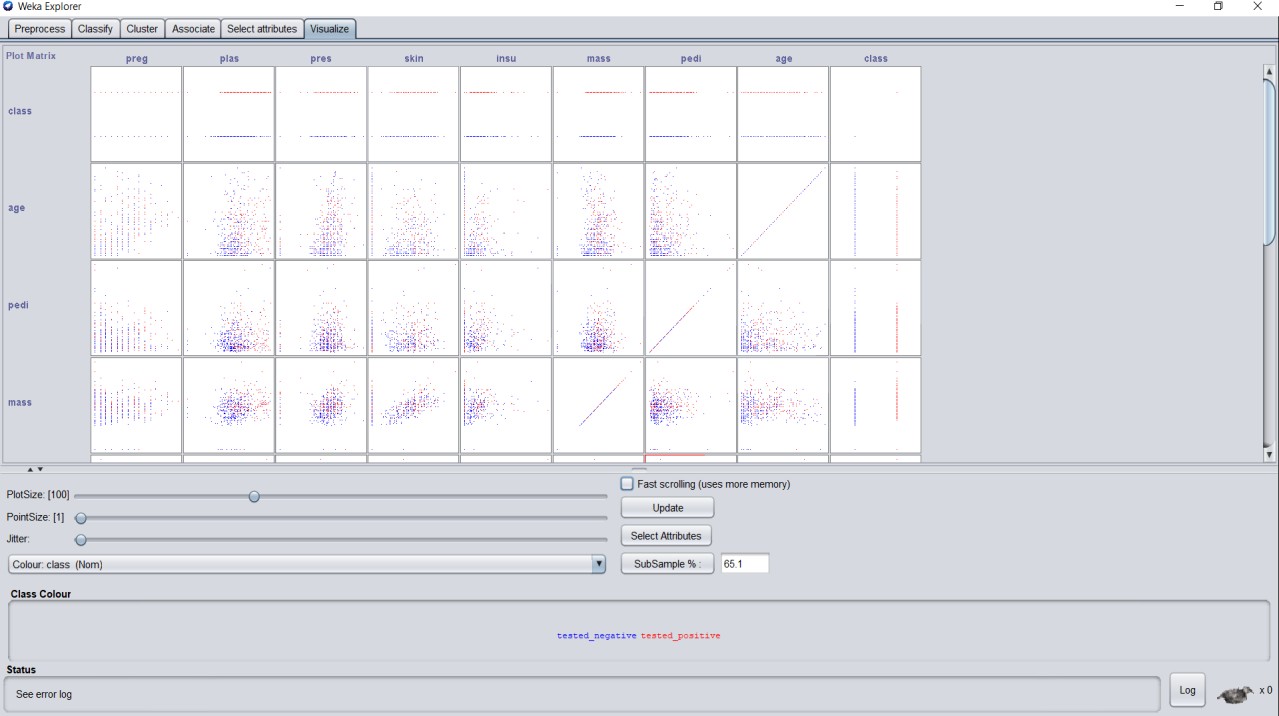
=== Model and evaluation on training set ===

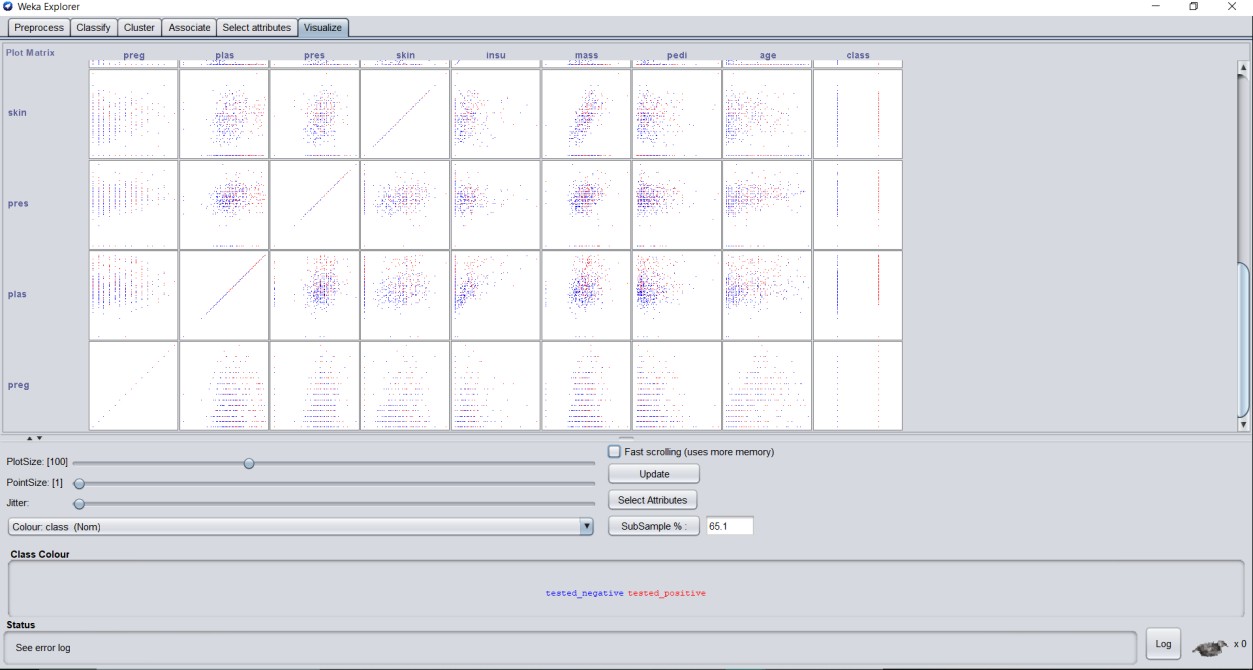
Clustered Instances

0 500 ( 65%)

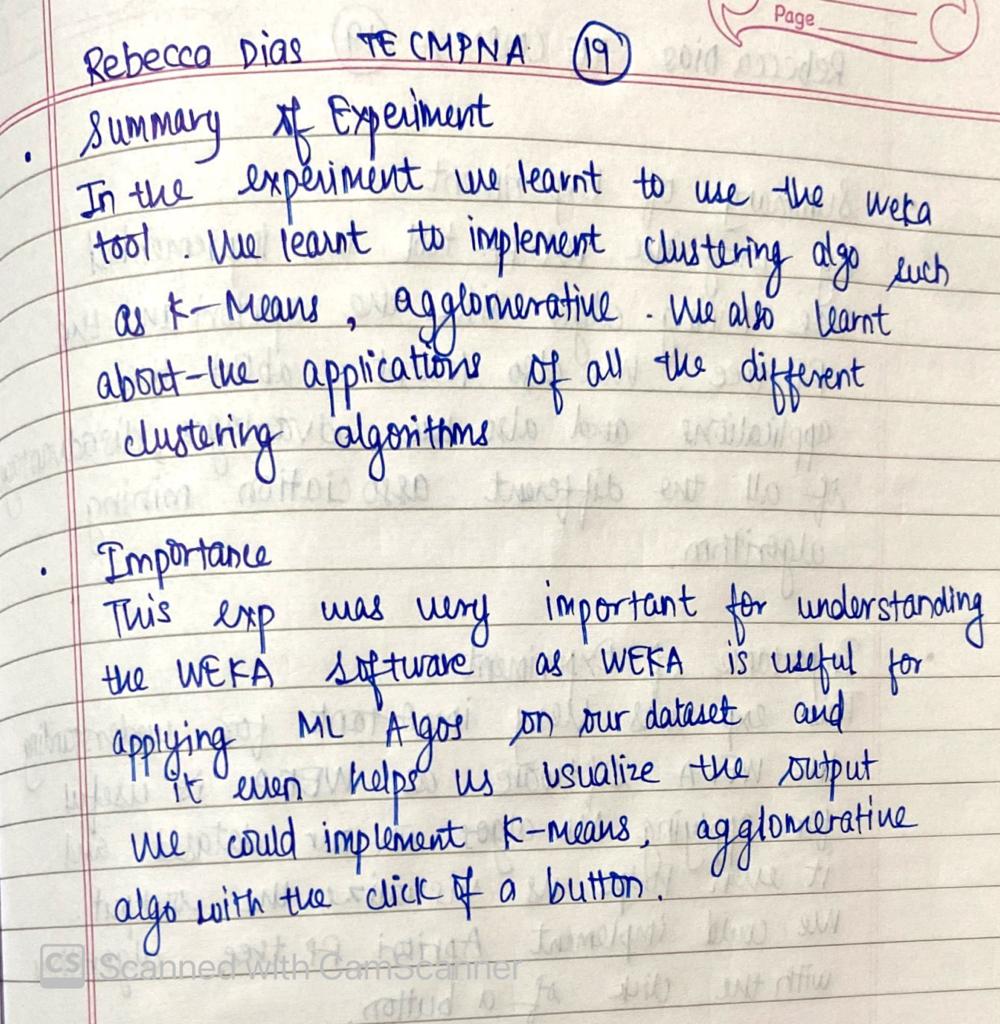
1 268 ( 35%)

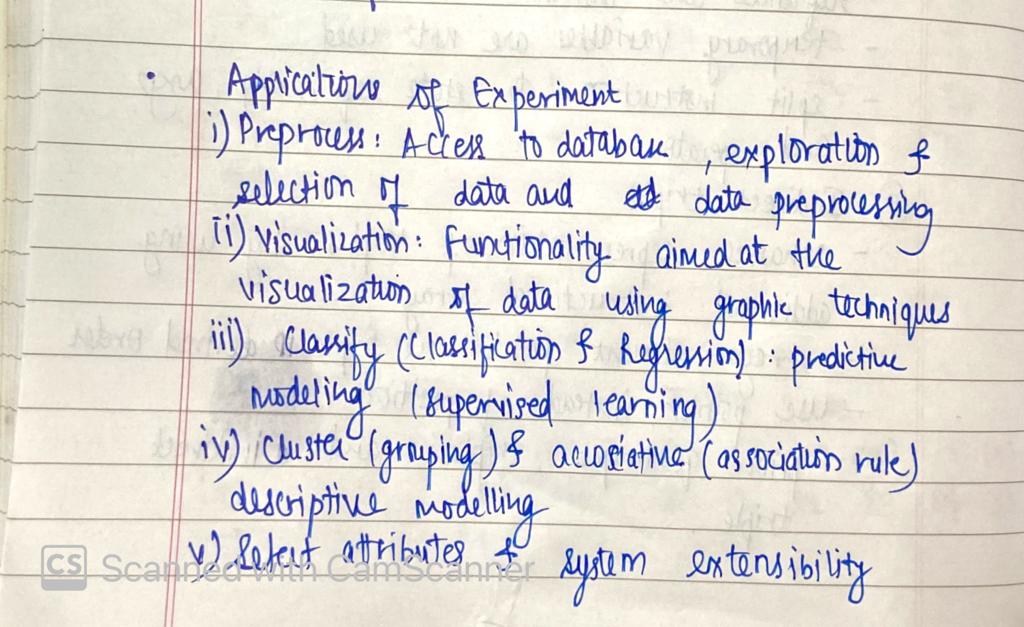
## **Visualization:**





**Conclusion: -**

****

****

# EXPERIMENT NO. 8C: Association Algorithm using WEKA

**Aim: -** Association Mining like Apriori, FP tree in WEKA.

**S/W Requirement: -** WEKA Tool

# Theory: -

## Application of Association Algorithms:

Association algorithm has various applications in machine learning and data mining. Below are some popular applications of association rule learning:

* + Market Basket Analysis: It is one of the popular examples and applications of association rule mining. This technique is commonly used by big retailers to determine the association between items.
  + Medical Diagnosis: With the help of association rules, patients can be cured easily, as it helps in identifying the probability of illness for a particular disease.
  + Protein Sequence: The association rules help in determining the synthesis of artificial Proteins.
  + It is also used for the Cata-log Design and Loss-leader Analysis and many more other applications.

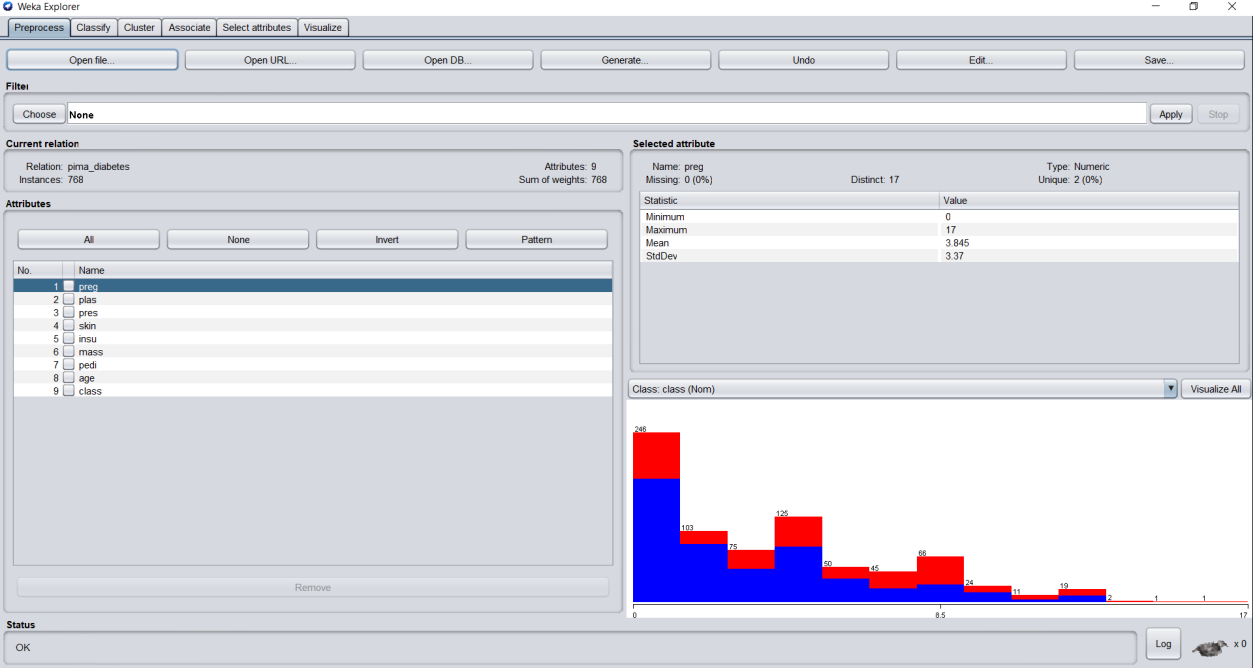
## Advantage and Disadvantage of Association Algorithms:

* **Advantages:**
* Uses large item set property
* Easily parallelized
* Easy to implement
* **Disadvantages:**
* Assumes transaction database is memory resident
* Requires many database scans

## How Association helps in Boosting the Business Profit?

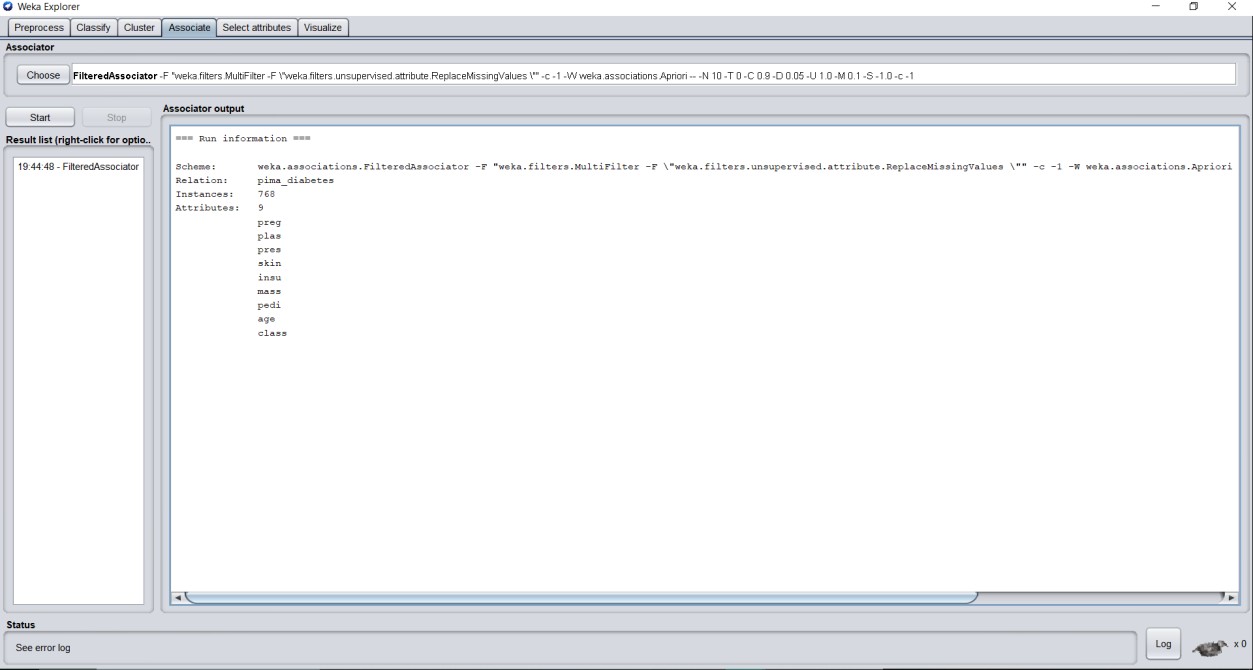
In data mining, **association** rules are useful for analysing and predicting customer behaviour. They play an important part in customer analytics, market basket analysis, product clustering, cata-log design and store layout. **Association Algorithm** usually contains or deals with a large number of transactions. For example, customers buying a lot of goods from a grocery store, by applying this method of the **algorithm** the grocery stores can enhance their sales performance and could work effectively.

# Implementation: -



## **OUTPUT: -**

i)Apriori:



**Classifier Output:**

=== Run information ===

Scheme: weka.associations.FilteredAssociator -F "weka.filters.MultiFilter -F

\"weka.filters.unsupervised.attribute.ReplaceMissingValues \"" -c -1 -W weka.associations.Apriori -- -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -

c -1

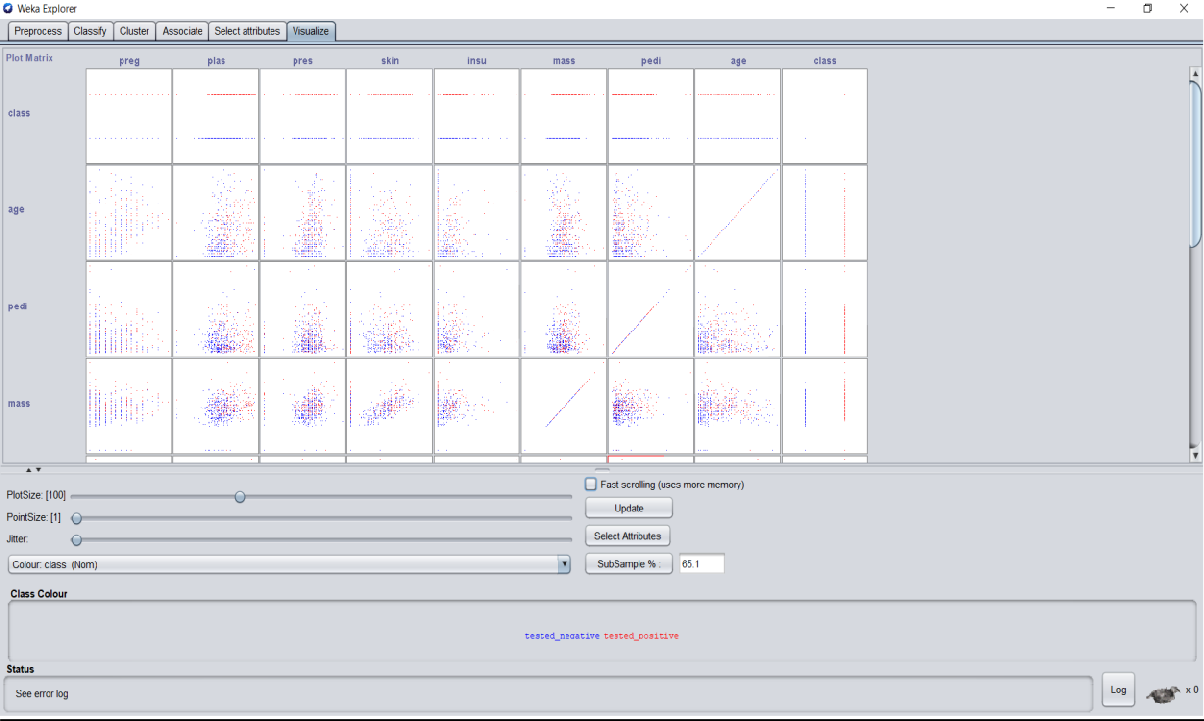
Relation: pima\_diabetes Instances: 768

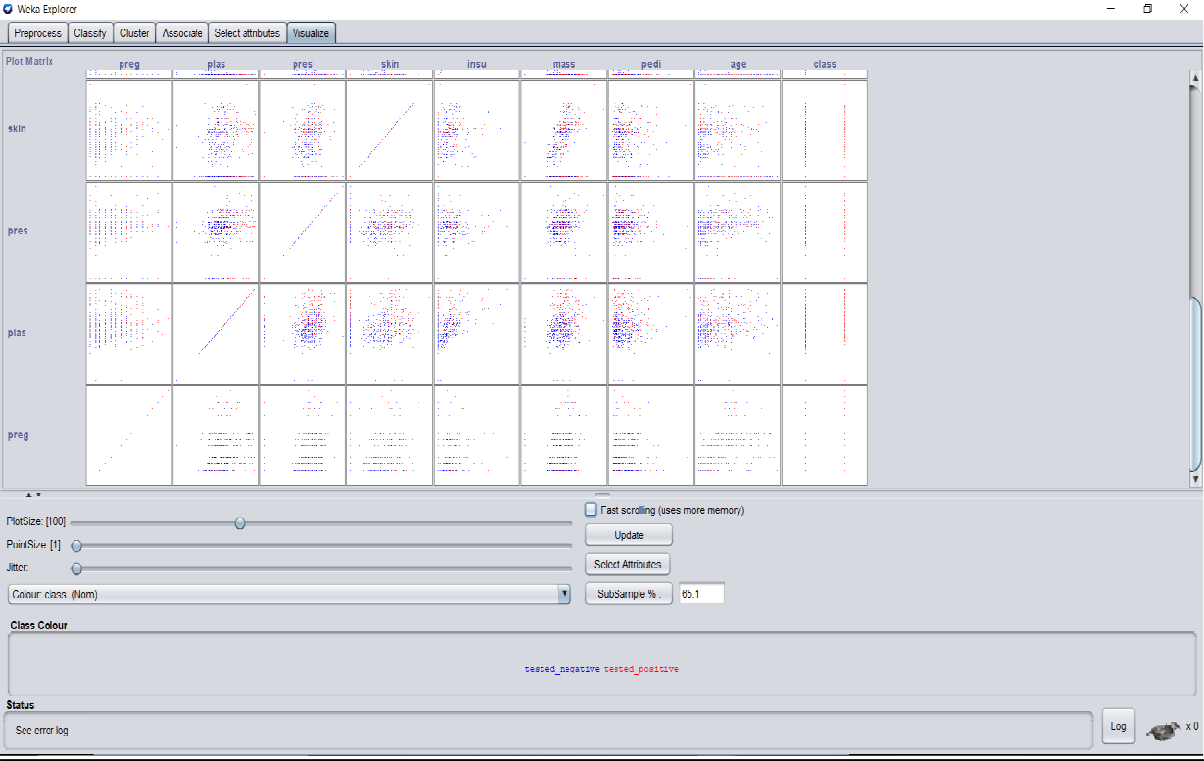
Attributes: 9

preg plas pres skin insu mass pedi

age class

## **Visualization:**





**Conclusion: -**

