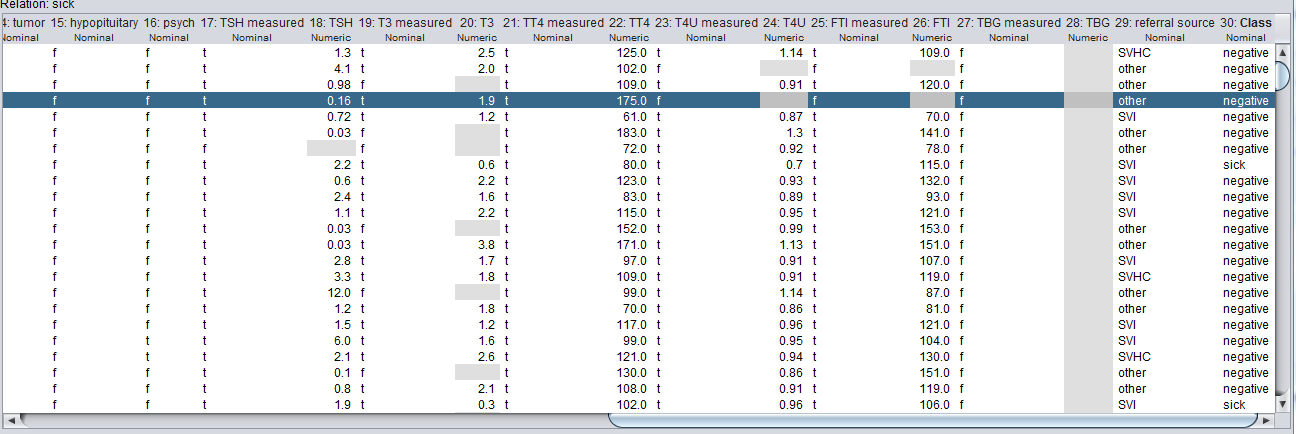
PRACTICAL 07

***Part A: Application of Discretization Filters***

1. **Perform the following tasks**

**1. Load the 'sick.arff' dataset**

****

1. **How many instances does this dataset have?**

**Ans :** There are 3772 Instances.

1. **How many attributes does it have?**

**Ans :** There are total 30 attributes.

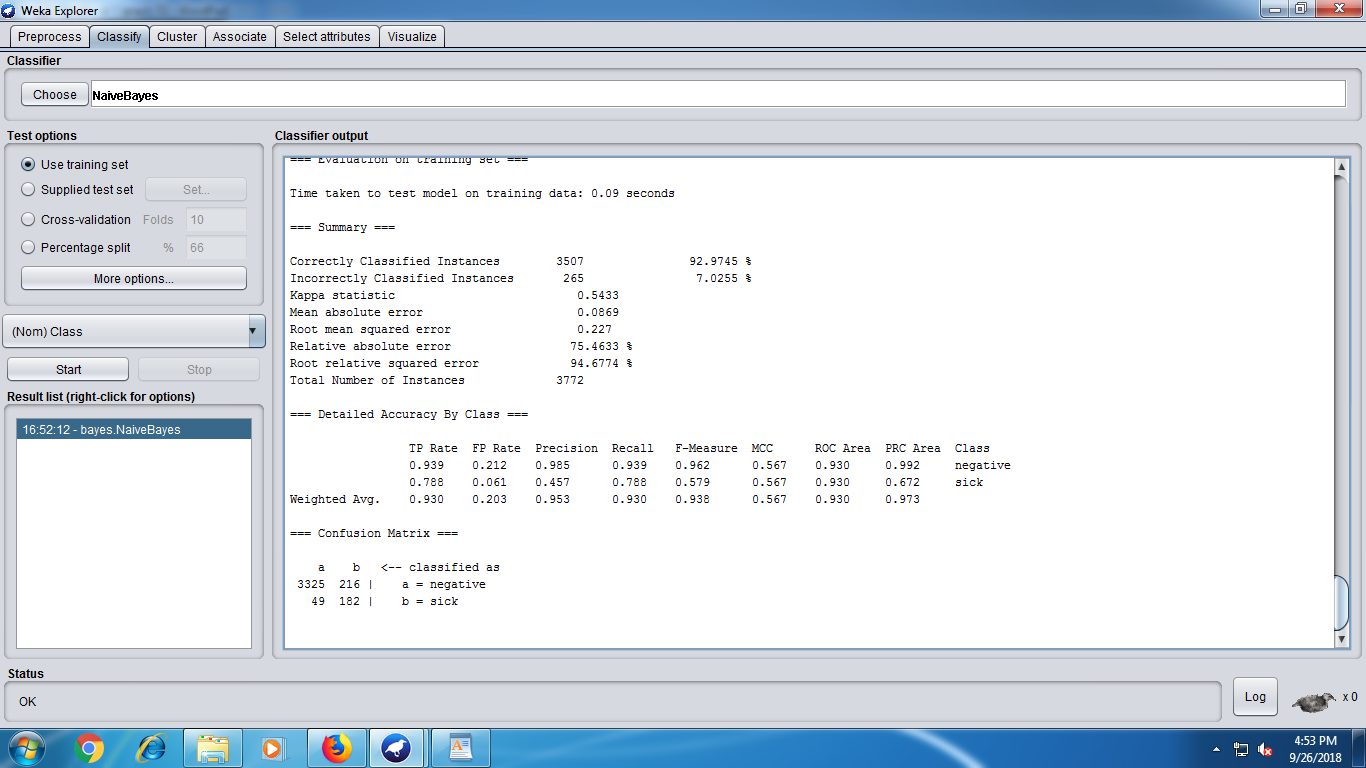
**4. Which is the class attribute and what are the characteristics of this attribute?**

**Ans :** The name of the class attribute is Class and the type of Classification is negative or sick.

**5. How many attributes are numerics? What are the attribute indexes of the numerica attributes?**

**Ans :** There are 7 numeric Attribute . The indexes of the attributes are 1, 18,20,22,24,26,28

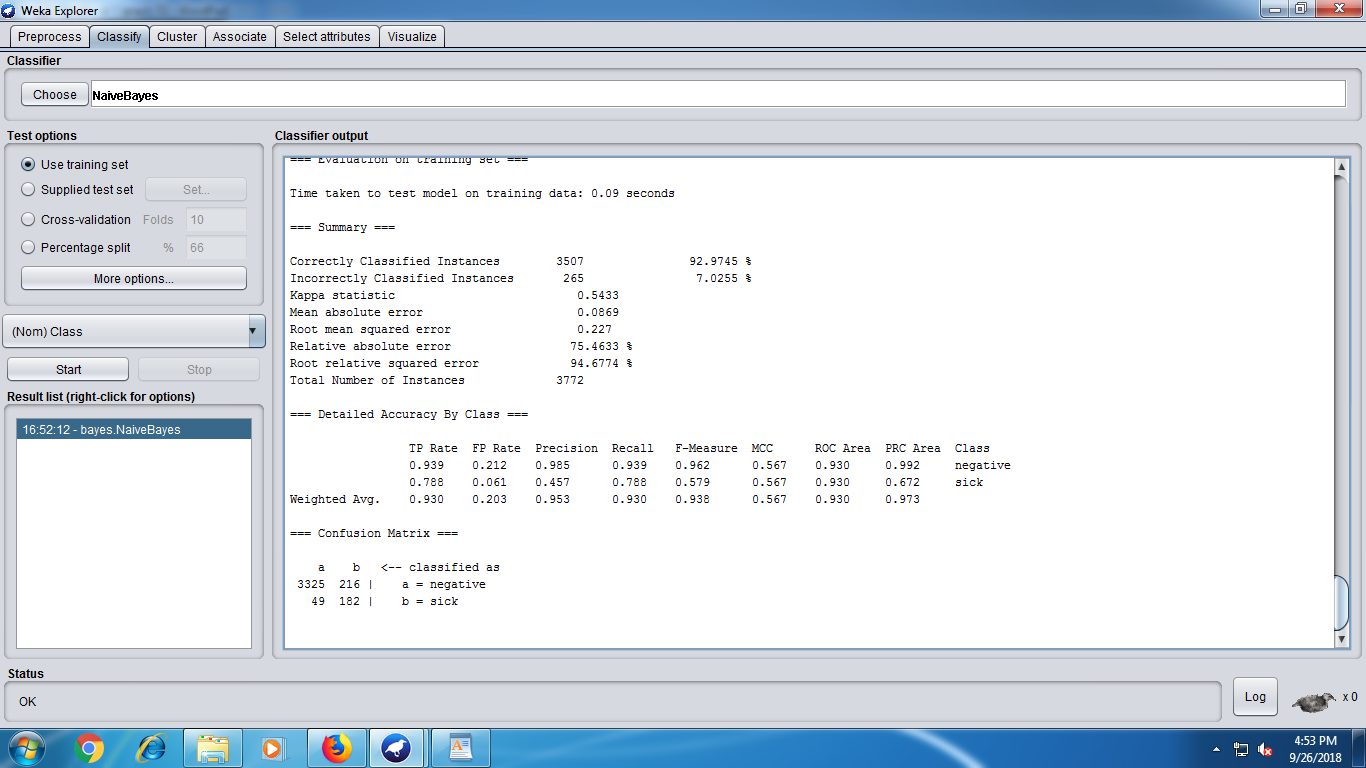
**6. Apply the Naive Bayes classifier. What is the accuracy of the classifier?**



**2. Perform the following tasks:**

**1. Load the 'sick.arff' dataset.**

**2. Apply the supervised discretization filter.**



1. **What is the effect of this filter on the attributes?**

**Ans :** When supervised discretization is applied the numeric values are displayed in the form of ranges.

1. **How many distinct ranges have been created for each attribute?**

**Ans :** Age->3

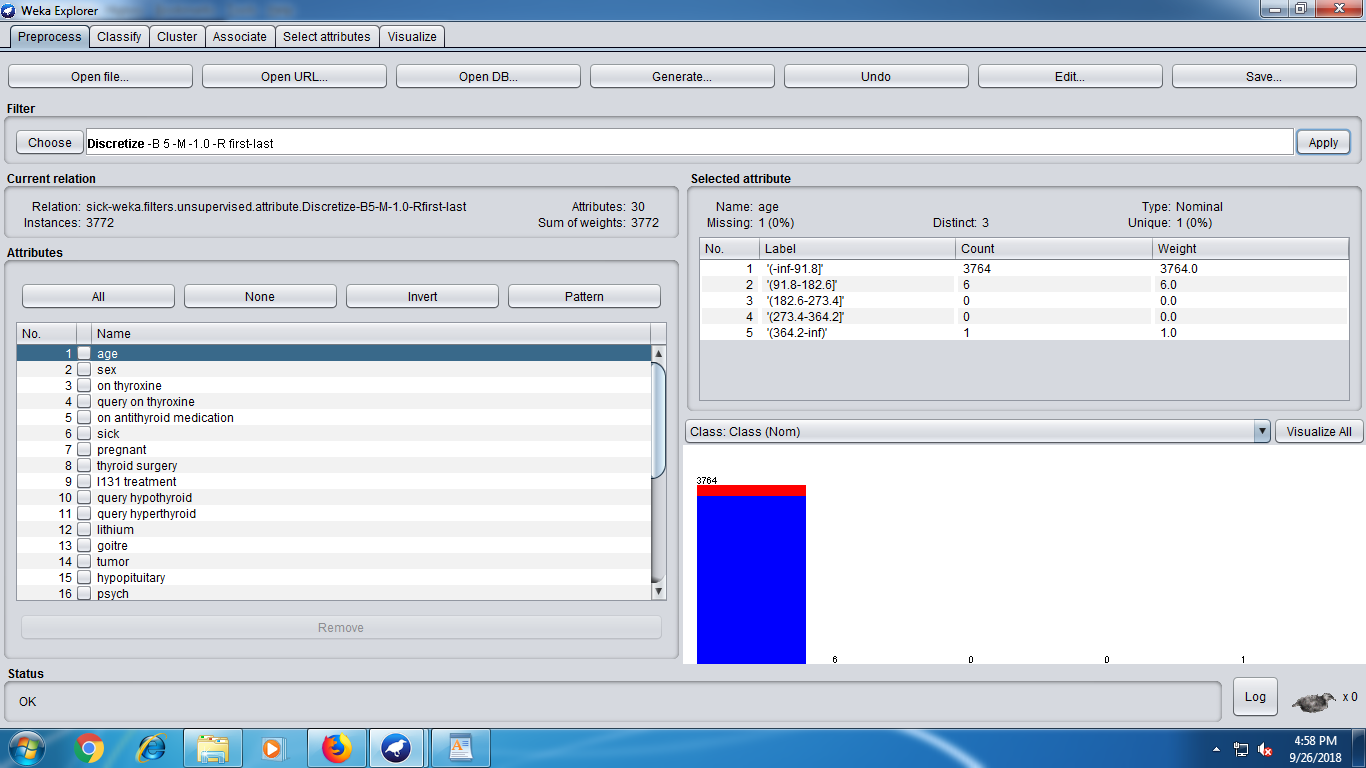
T3->2

TT4->2

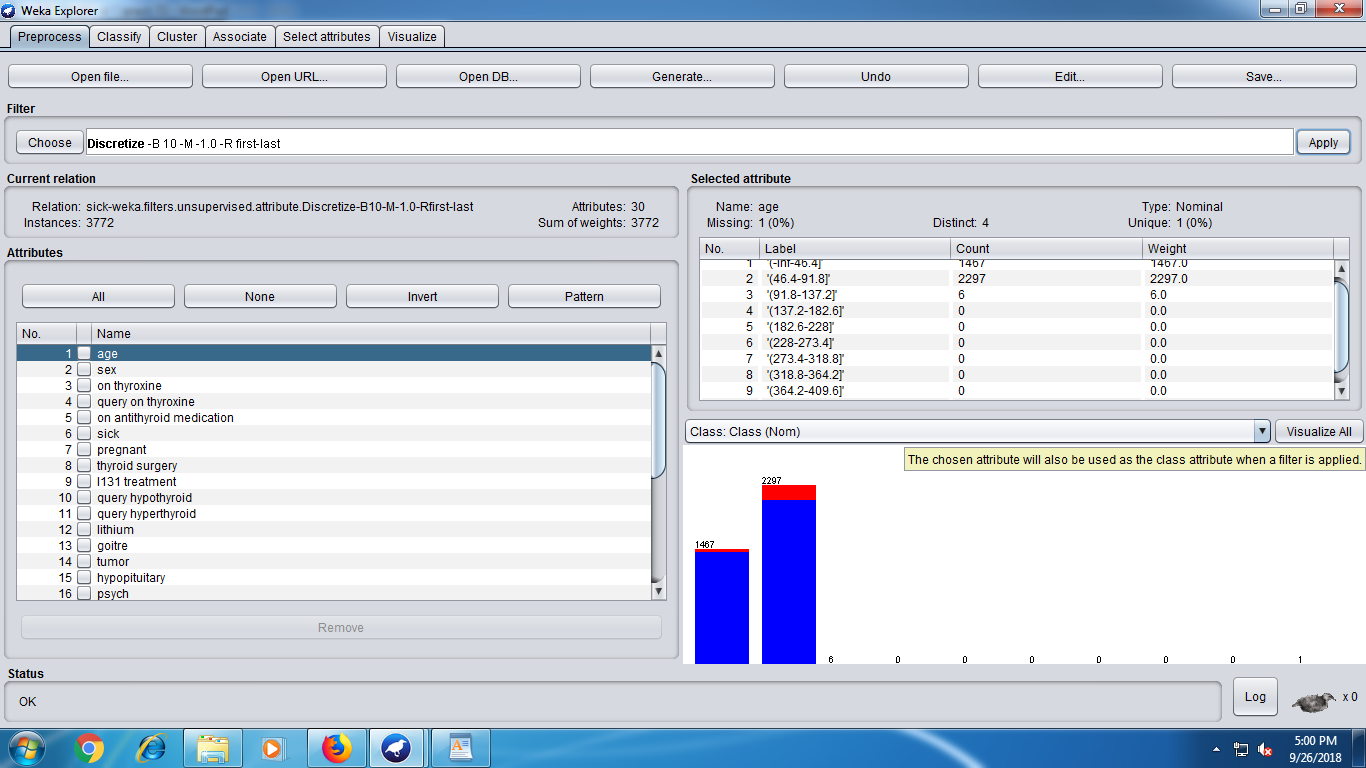
**5. Undo the filter applied in the previous step.**

**6. Apply the unsupervised discretization filter. Do this twice:**

**1. In this step, set 'bins'=5**



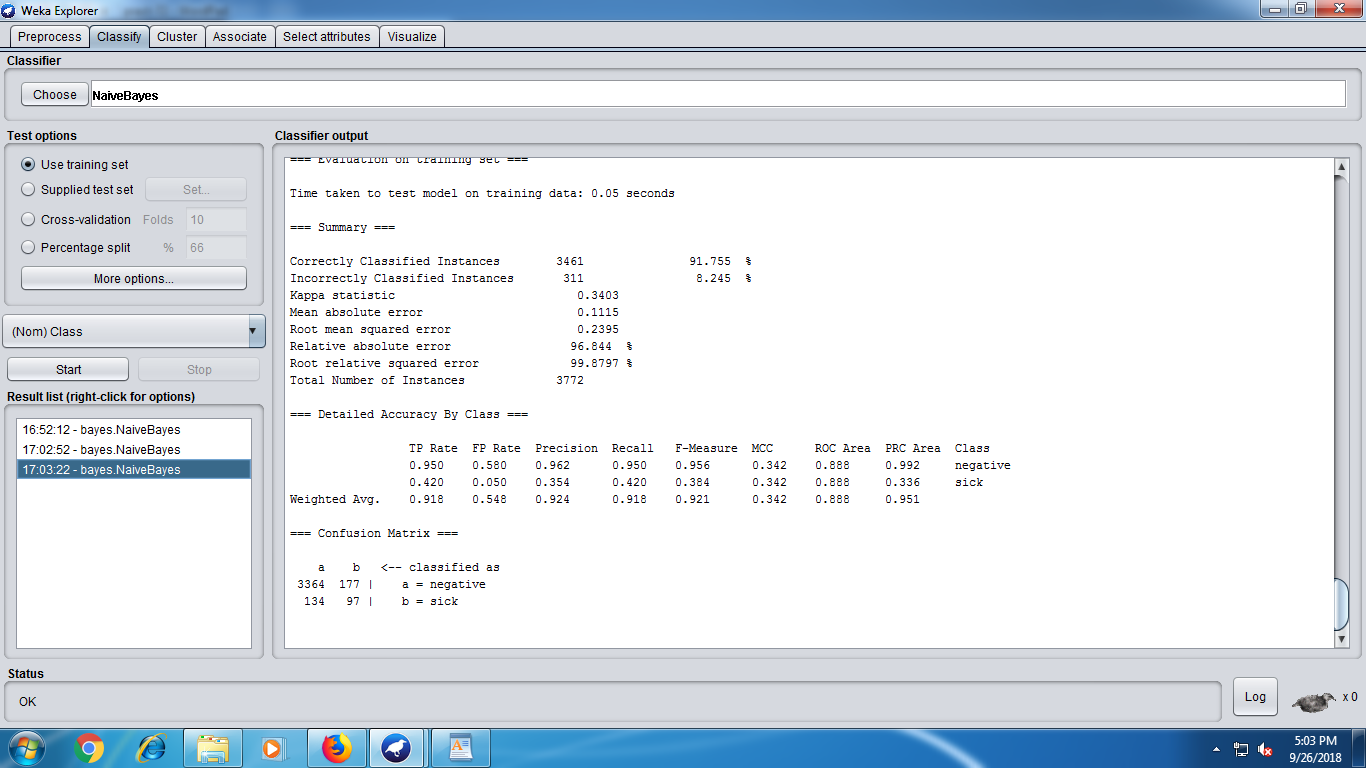
**2. In this step, set 'bins'=10**



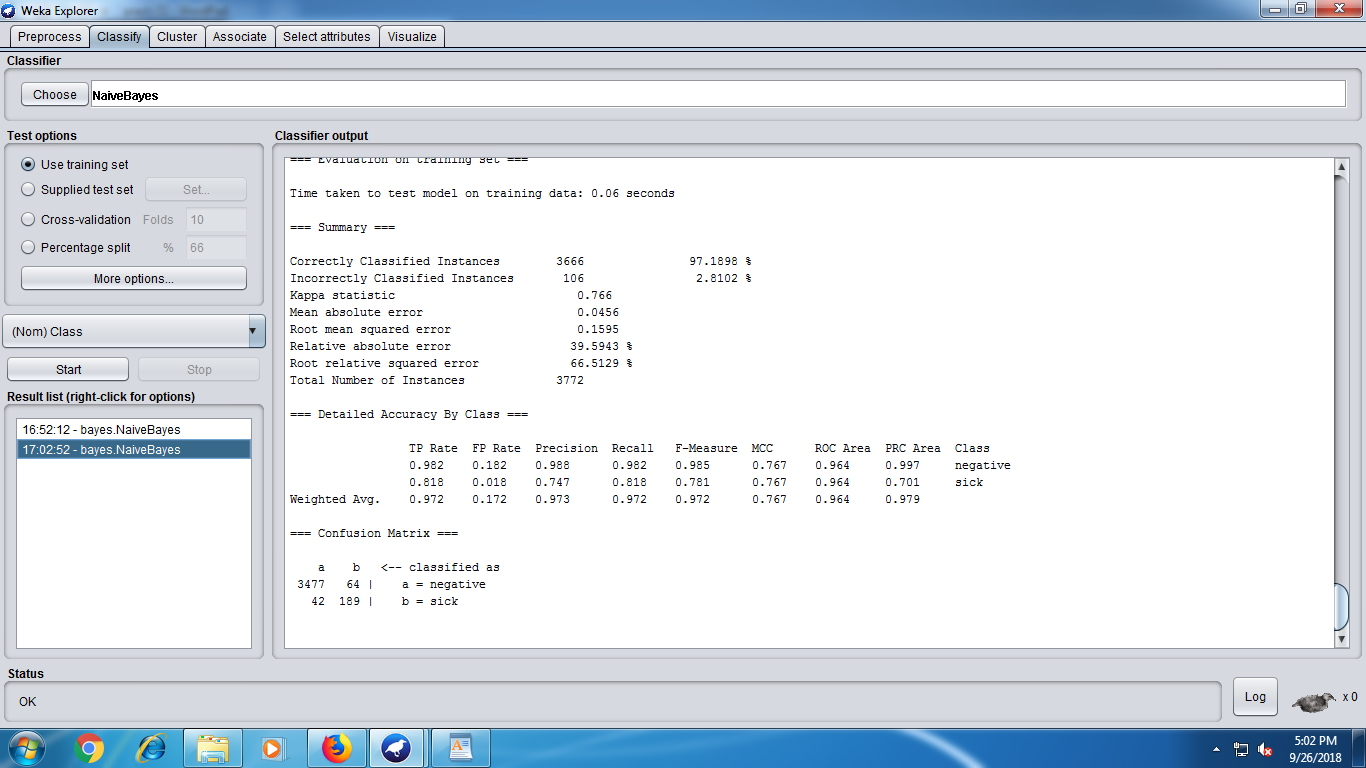
**3. What is the effect of the unsupervised filter filter on the datset?**

**7. Run the the Naive Bayes classifier after apply the following filters**

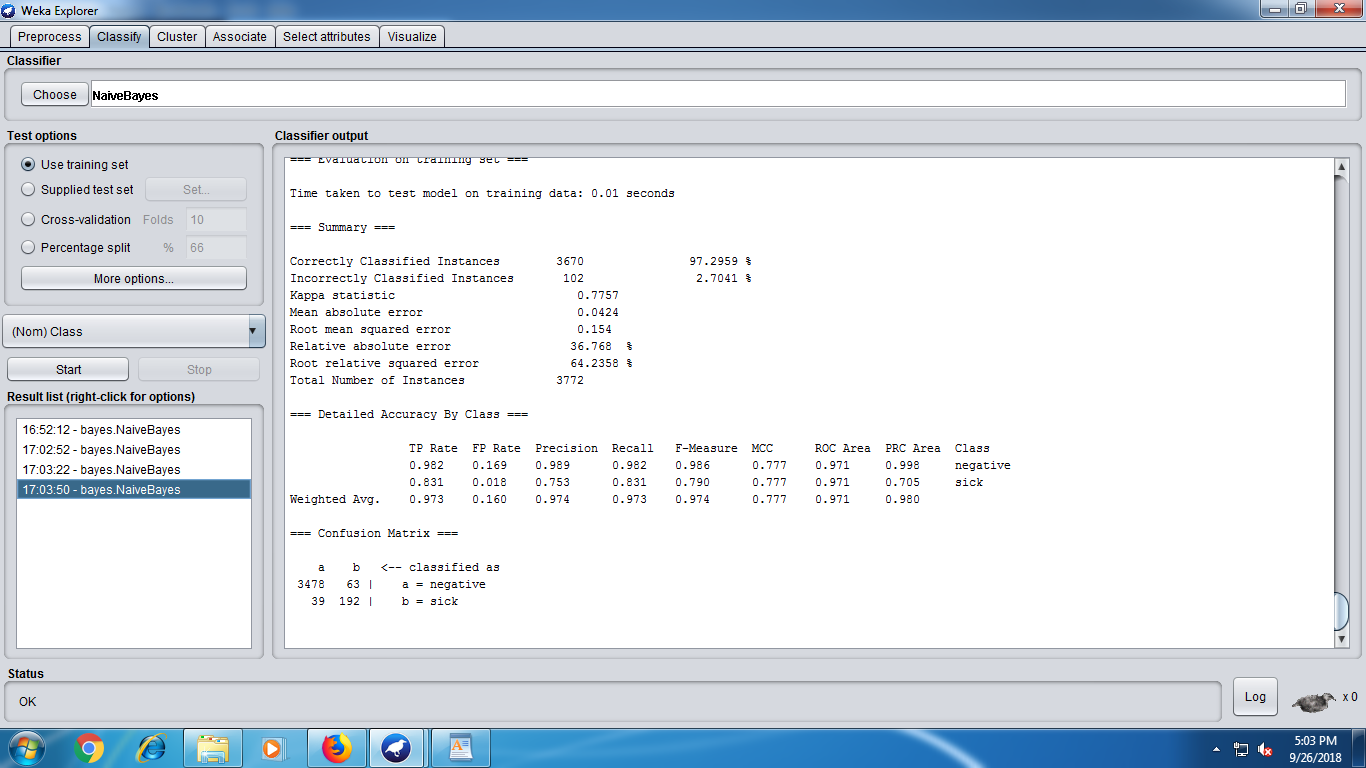
**1. Unsupervised discretized with 'bins'=5**



**2. Unsupervised discretized with 'bins'=10**



**3. Unsupervised discretized with 'bins''=20.**



**8. Compare the accuracy of the following cases**

**1. Naive Bayes without discretization filters**

**2. Naive Bayes with a supervised discretization filter**

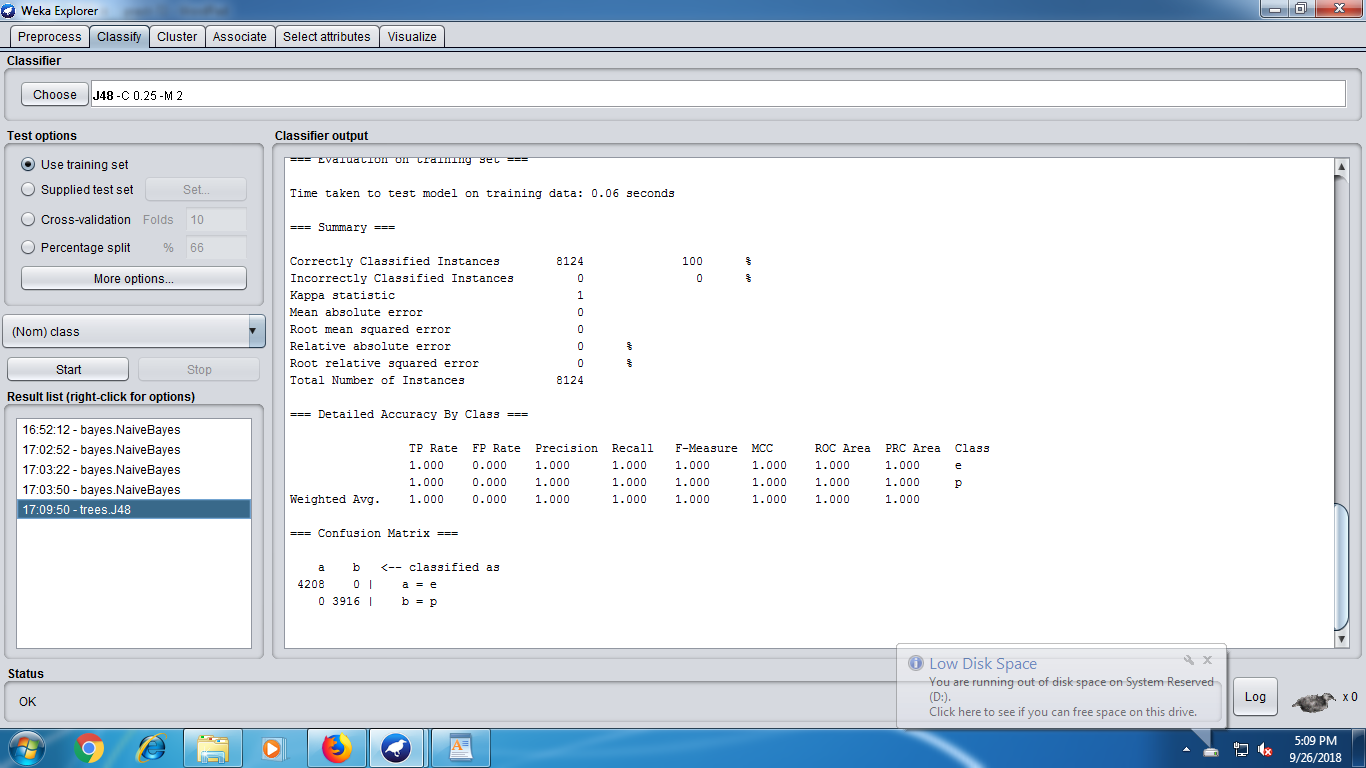
**3. Naive Bayes with an unsupervised discretization filter with different values for the 'bins attributes.**

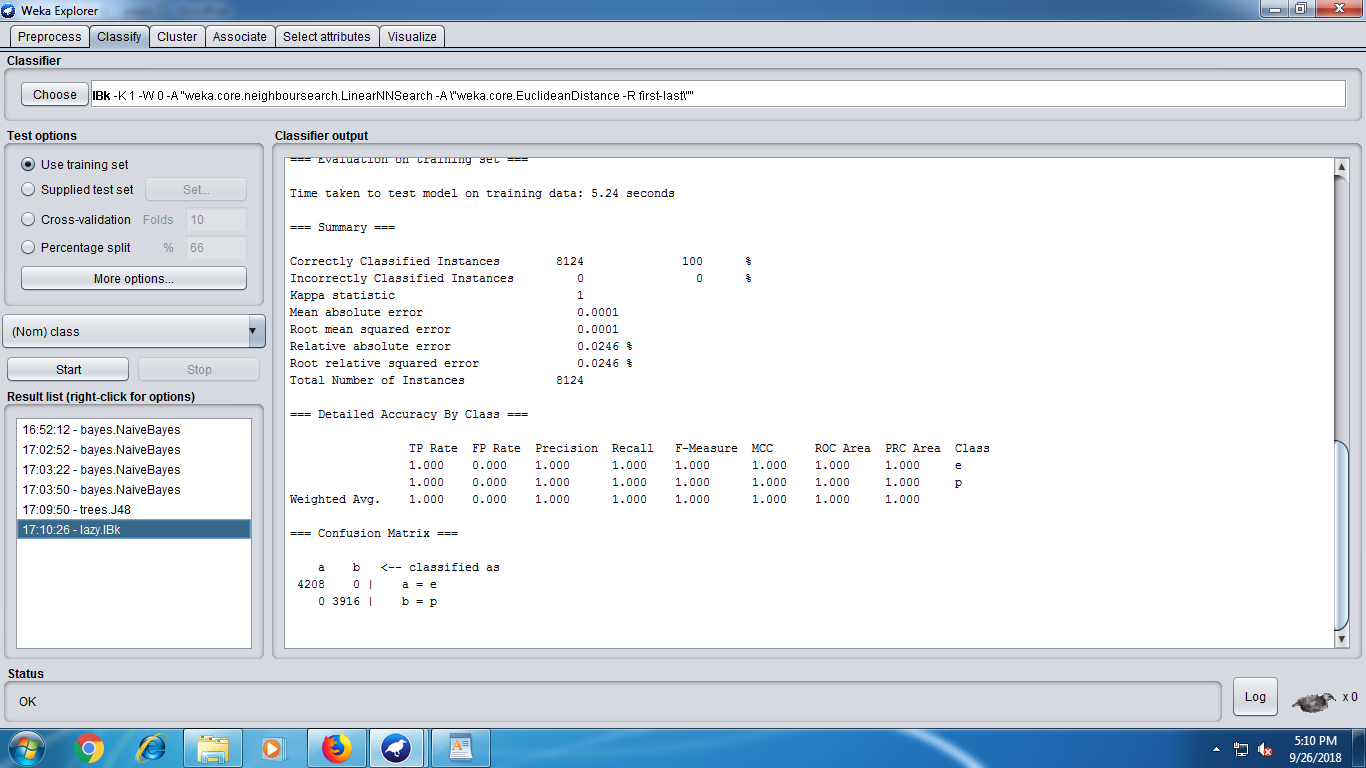
***Part B: Attribute Selection***

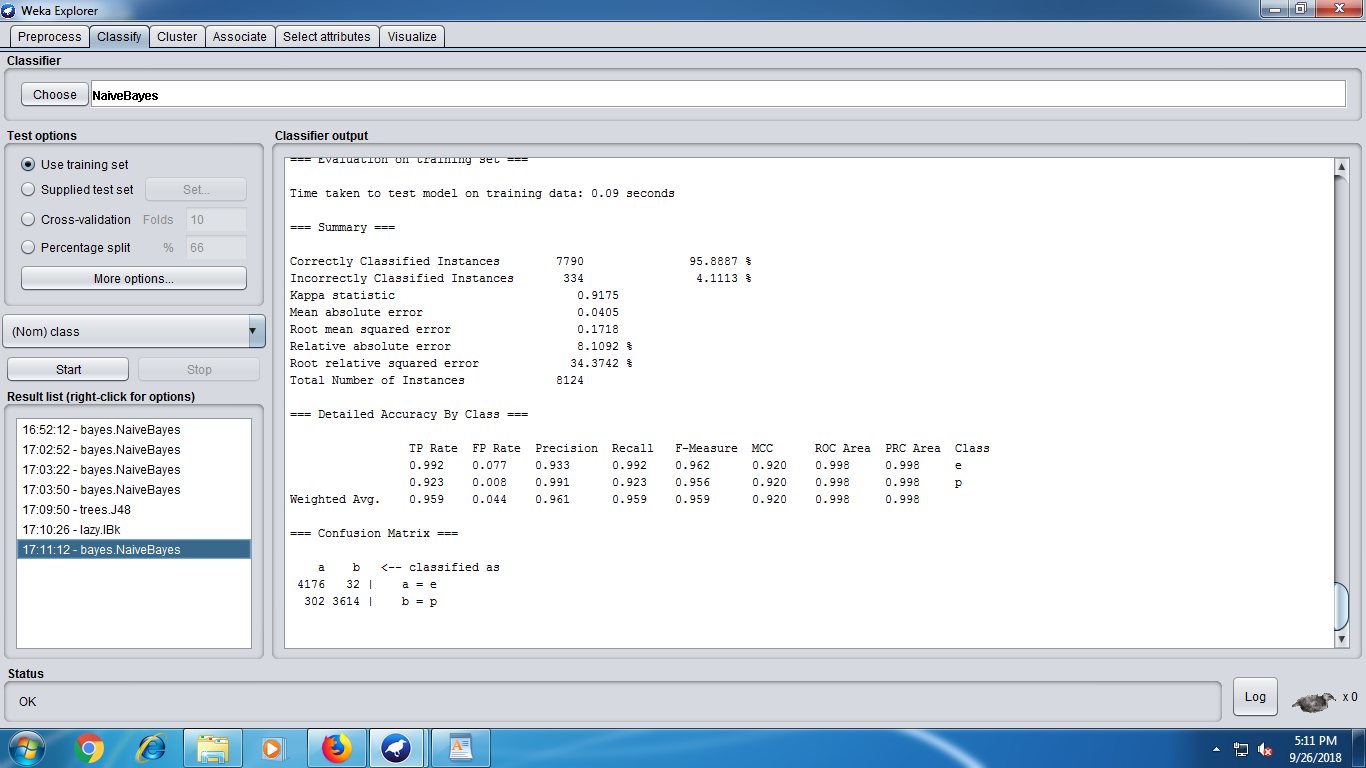
**1. Perform the following tasks:**

**1. Load the 'mushroom.arff' dataset**

**2. Run the J48, 1Bk, and the Naive Bayes classifiers.**







**3. What is the accuracy of each of these classifiers?**

**Ans :** 100 – J48

100 - IBK

95.88 – Naïve Bayes

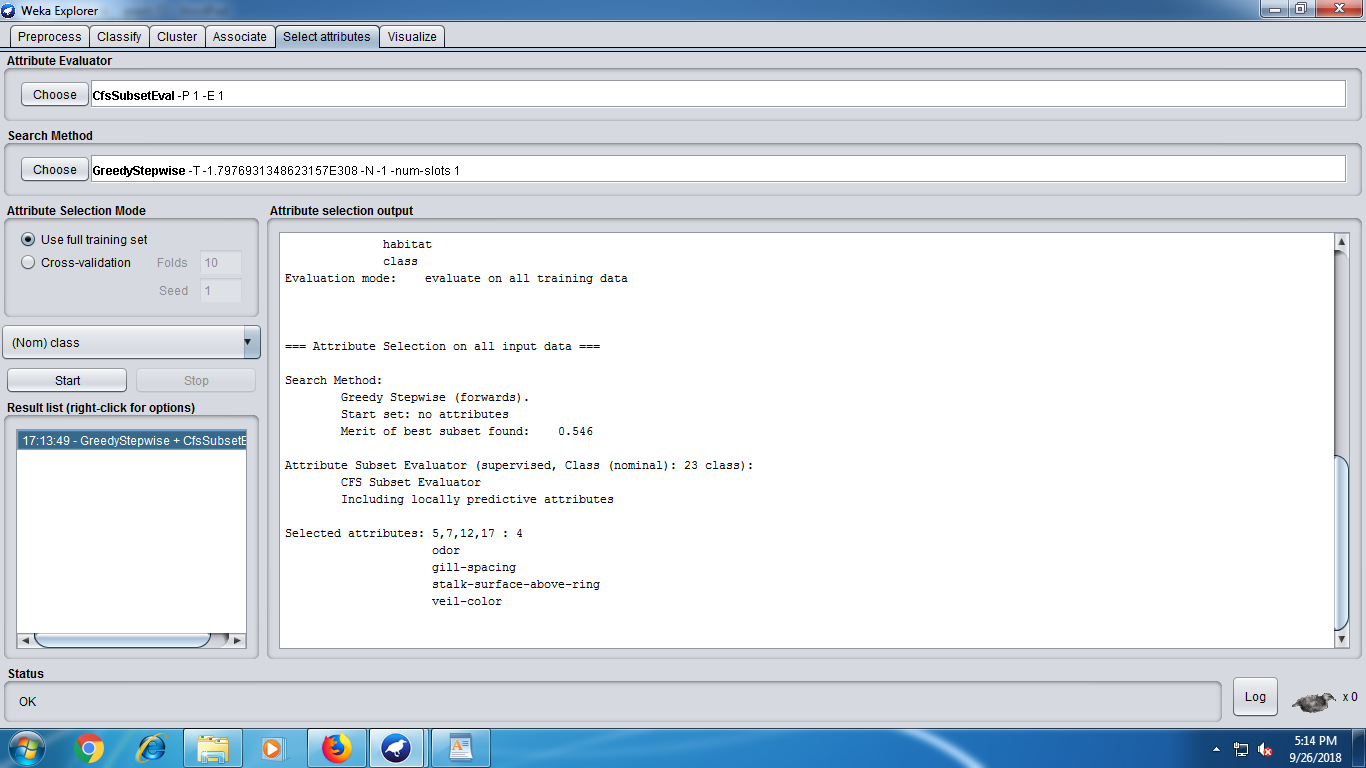
**2. Perform the following tasks:**

**1. Go to the 'Select Attributes' panel**

**2. Set attribute evaluator to CFSSubsetEval**

**3. Set the search method to 'Greedy Stepwise'**

**4. Analyze the results window**



1. **Record the attribute numbers of the most important attributes**

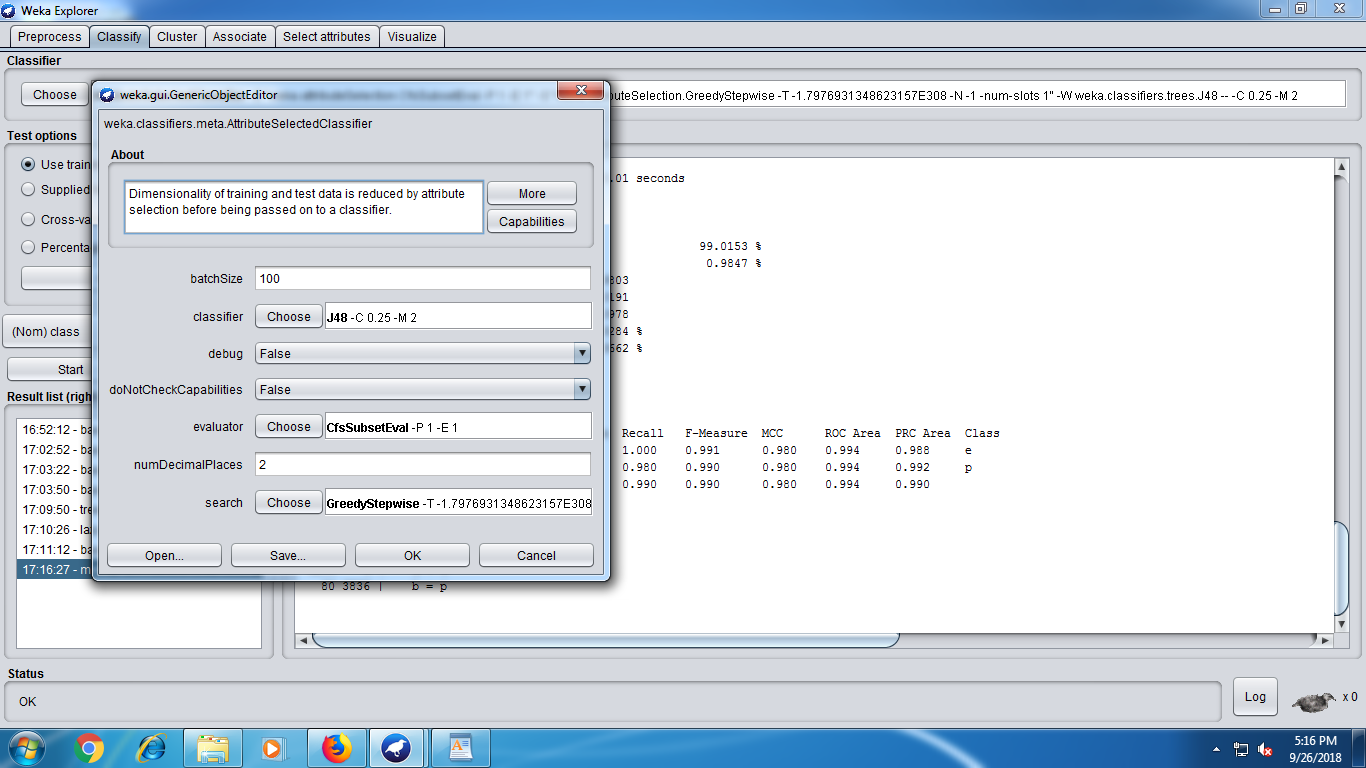
**Ans:** Attribute Number : 5 ,7 ,12 , 17

**6. Run the meta classifier AttributeSelectedClassifier using the following:**

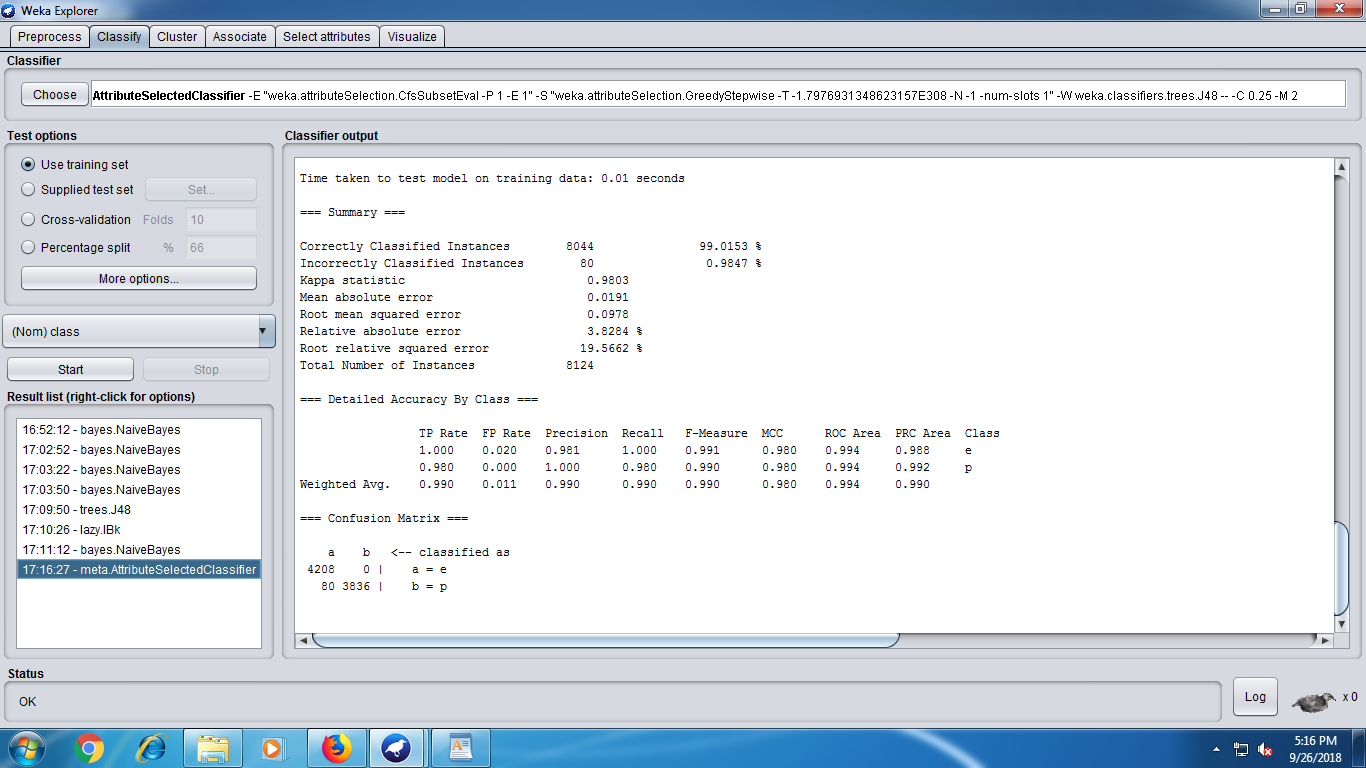
**1. CFSSubsetEval**

**2. GreedStepwise**

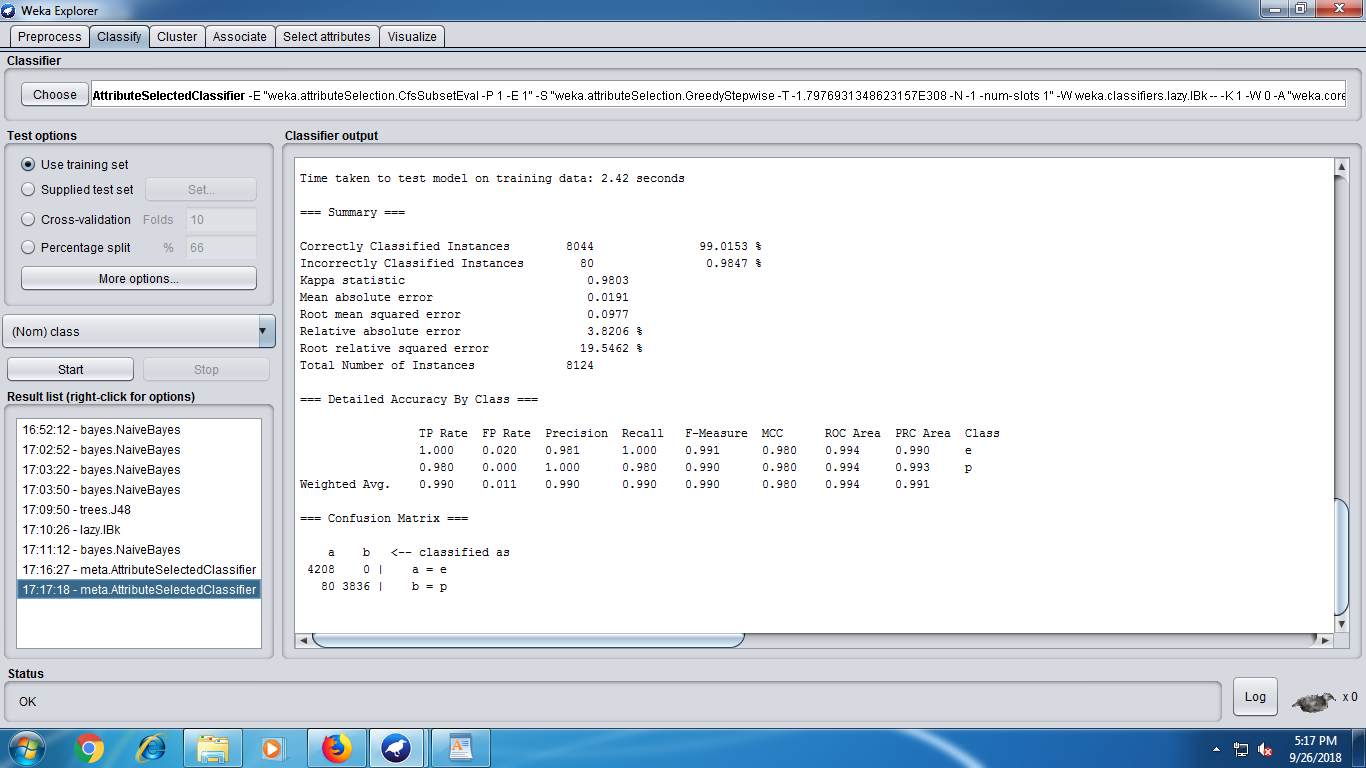
**3. J48, 1Bk, and NaiveBayes**



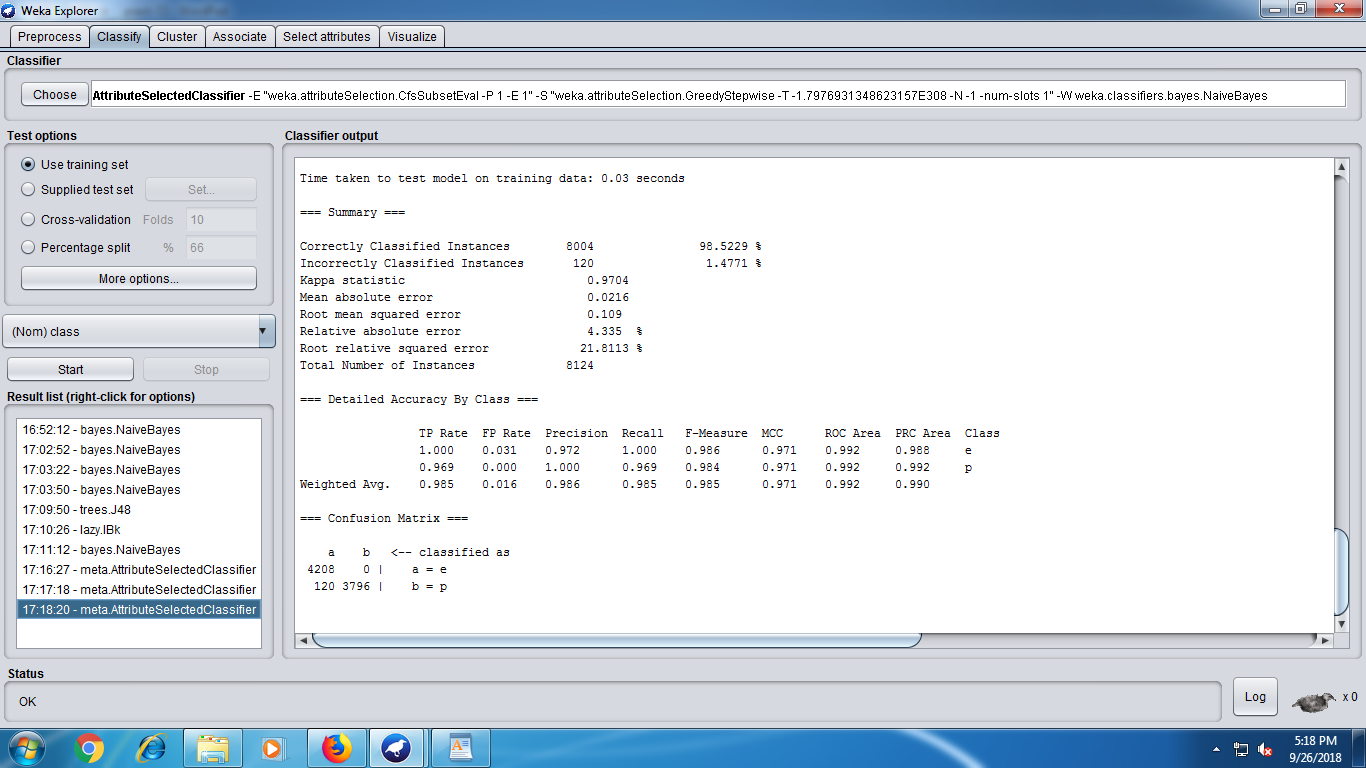
J48 :



IBK :



Naïve Bayes :



**7. Record the accuracy of the classifiers**

**8. What are the benefits of attribute selection?**

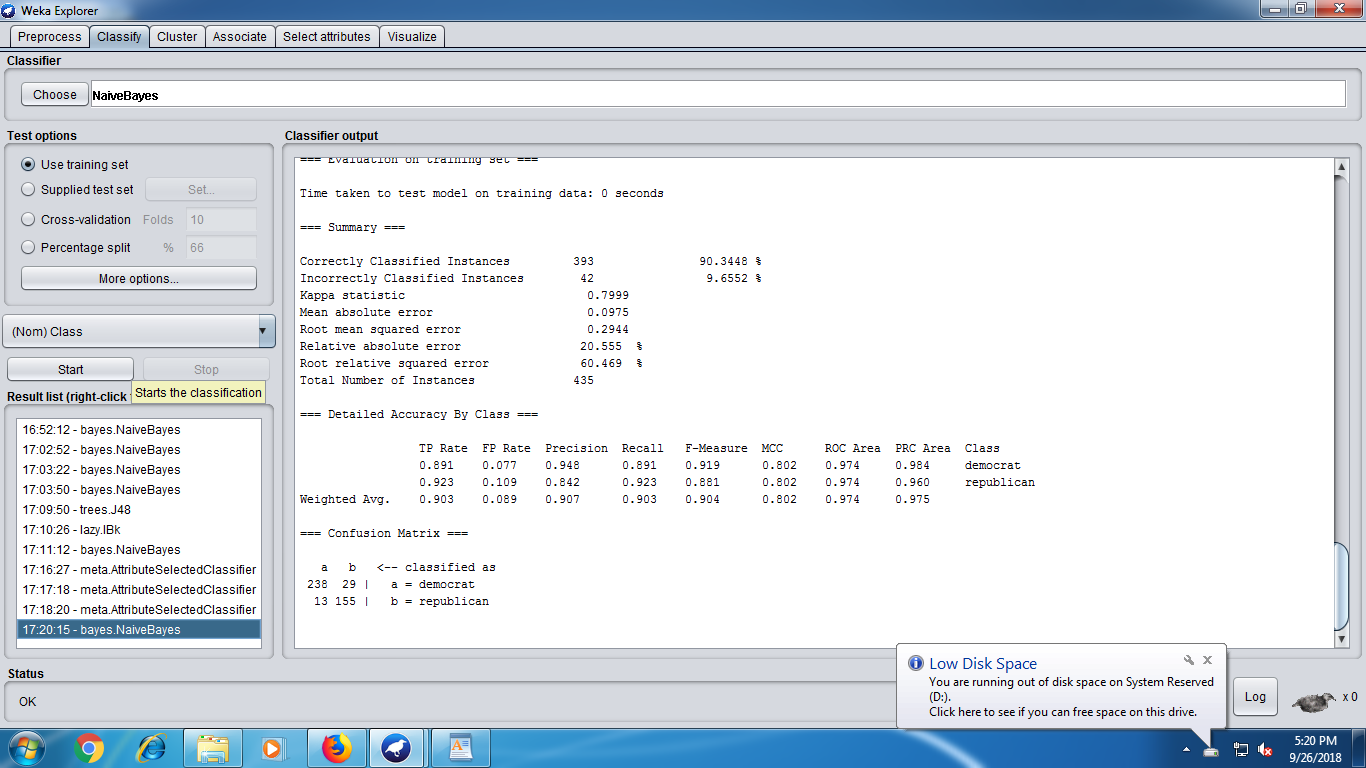
**Ans :** The benefits of attribute selection is increasing the efficiency of training and reducing the training time.

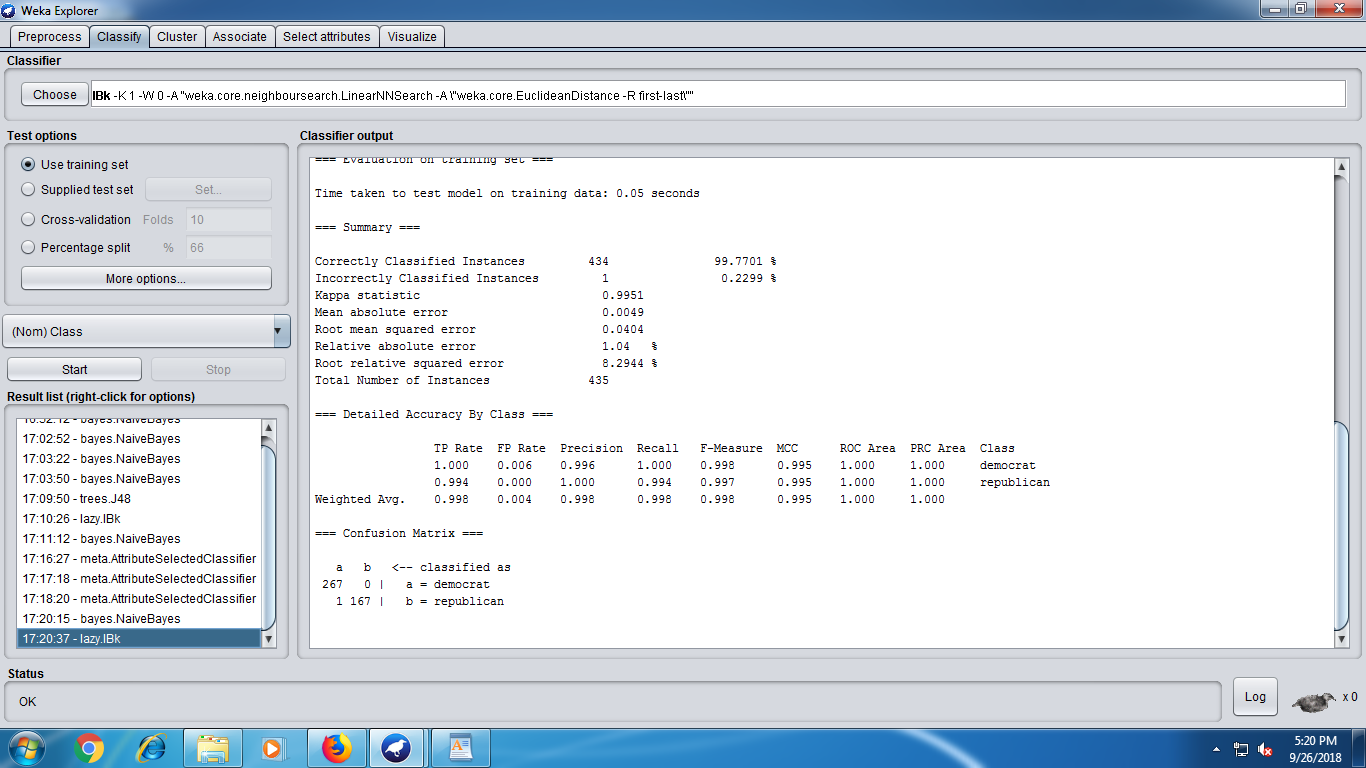
***Part C***

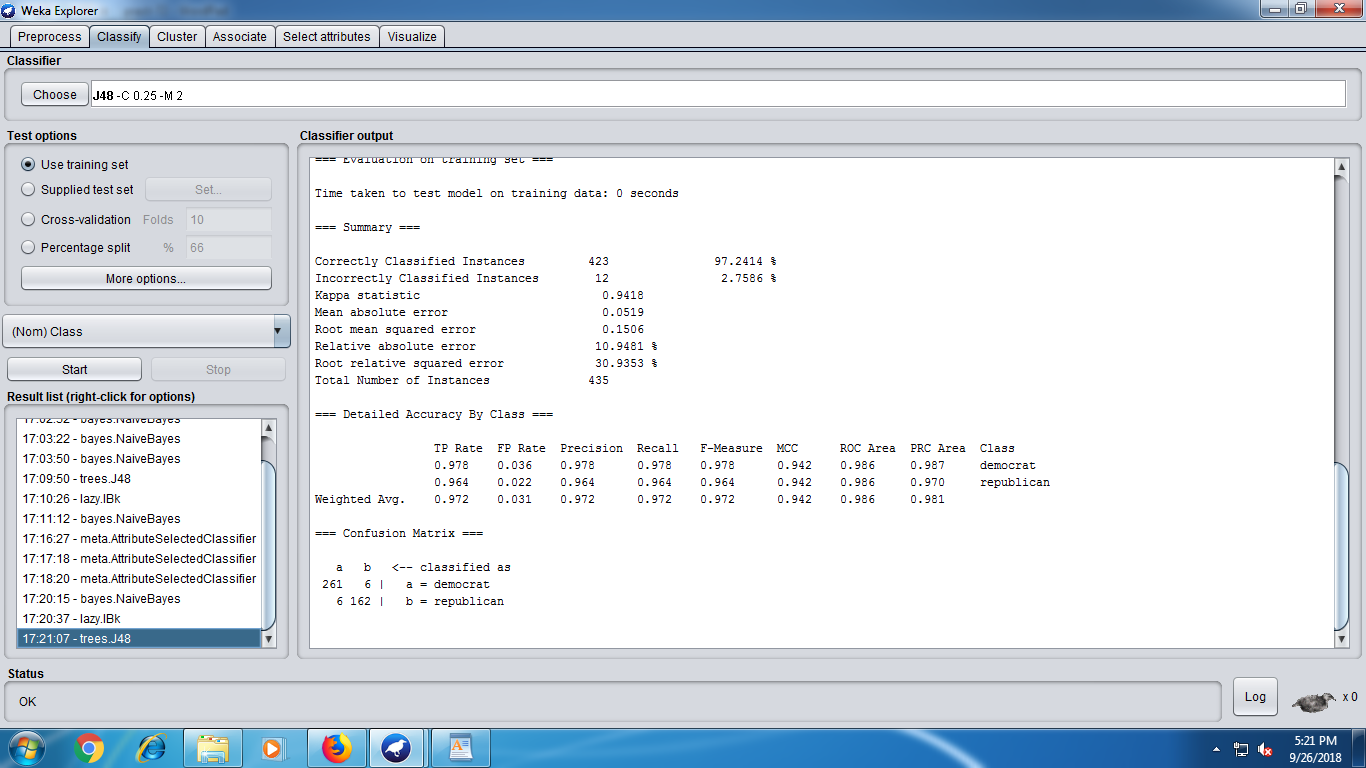
**1. Perform the following tasks:**

**1. Load the 'vote.arff' dataset.**

**2. Run the J48, 1Bk, and Naive Bayes classifiers.**







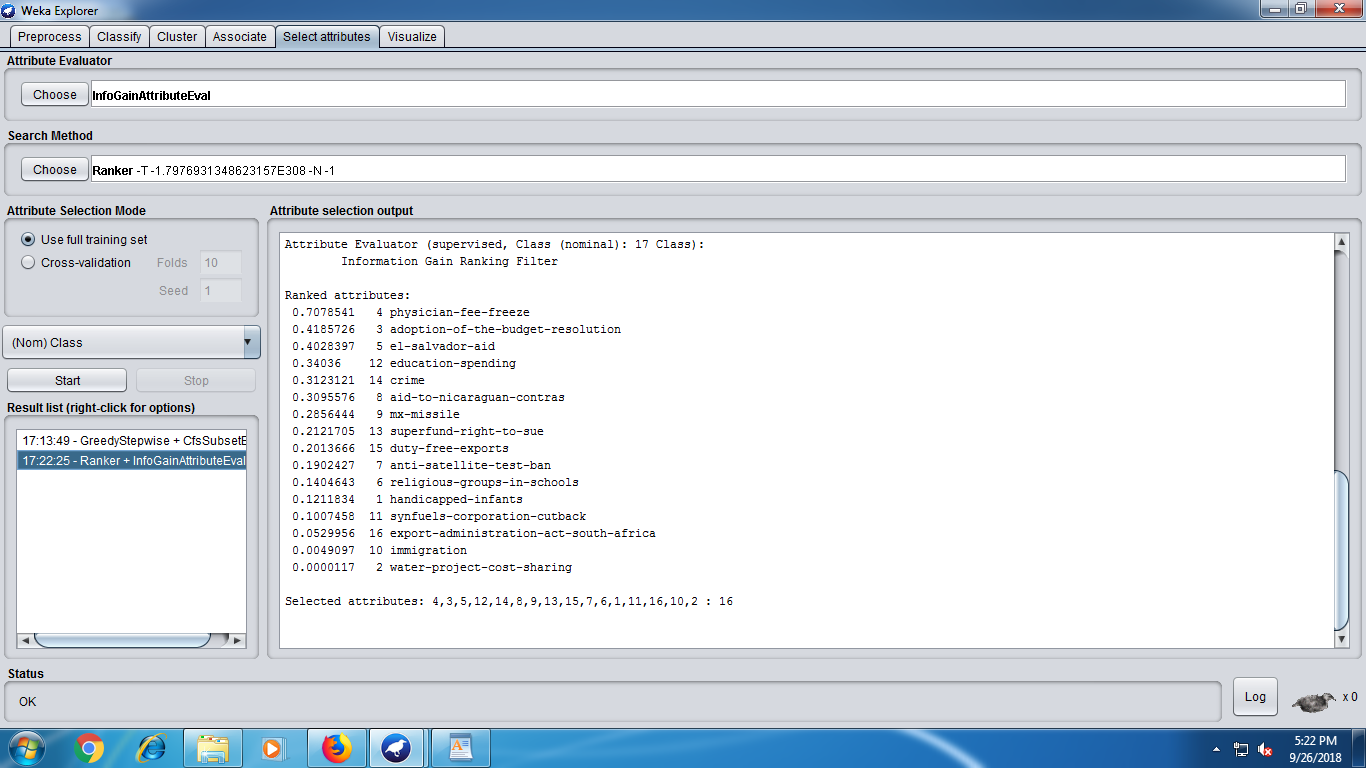
**3. Record the accuracies.**

**2. Perform the following tasks:**

**1. Go to the 'Select Attributes' panel**

**2. Set attribute evaluator to 'WrapperSubsetEval'**

**3. Set search method to ''RankSearch'**



**4. Set attribute evaluator to 'InfoGainAttributeEval'**

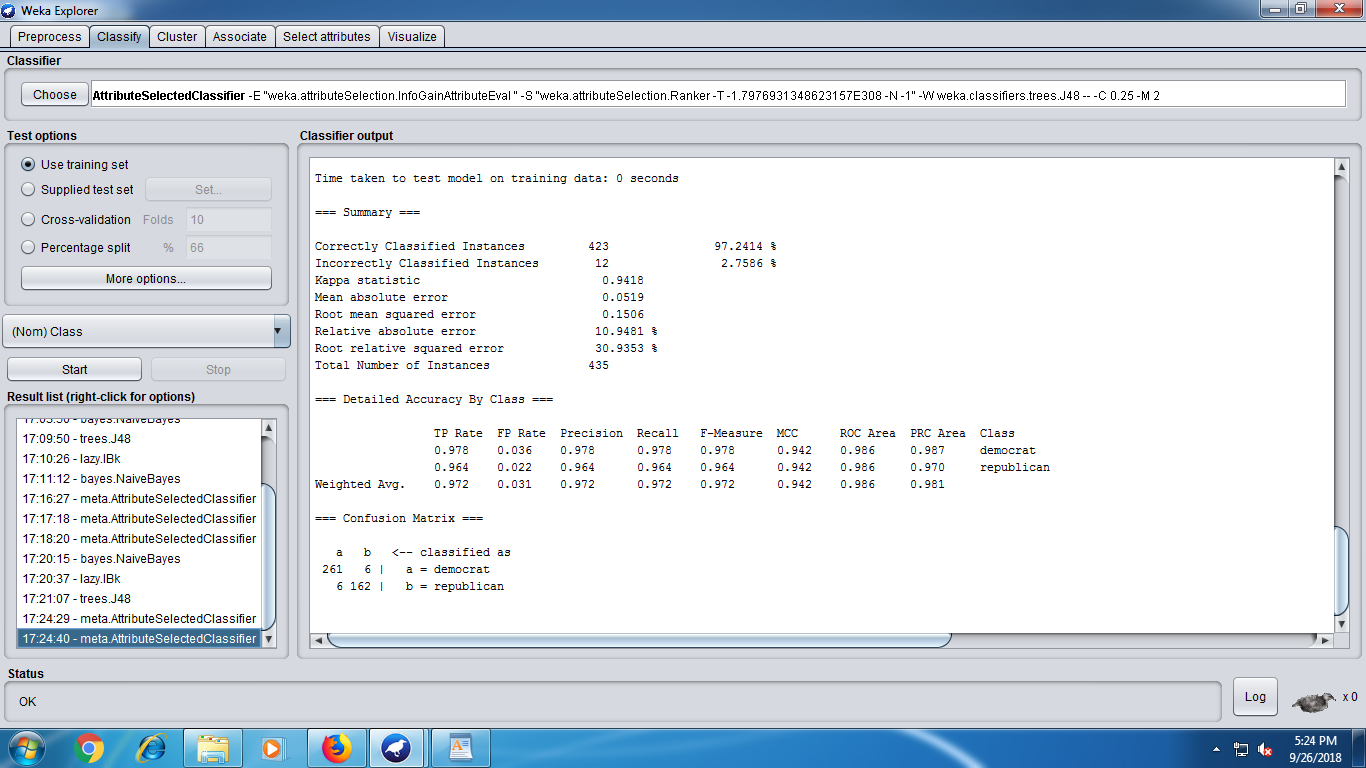
**5. Analyze the results**

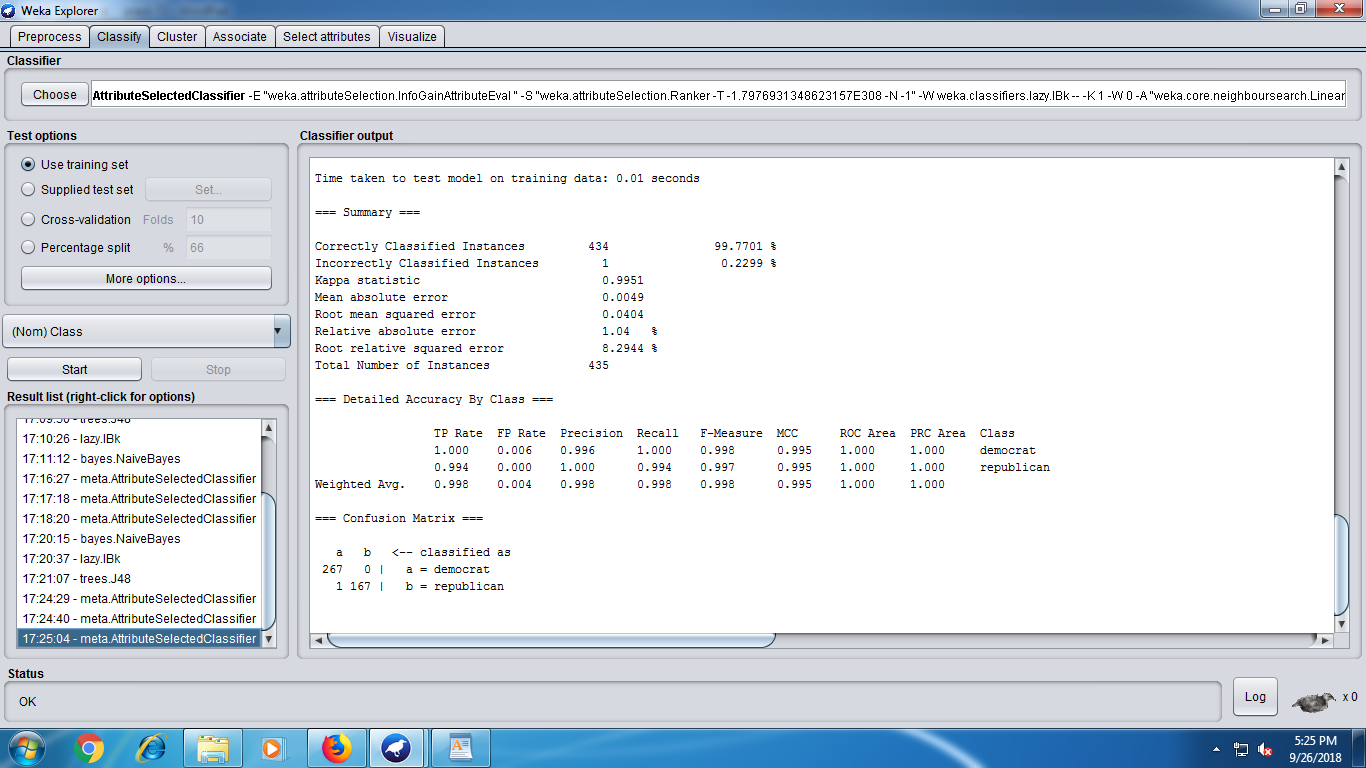
**6. Run the metaclassifier AttributeSelectedClassifier using the following:**

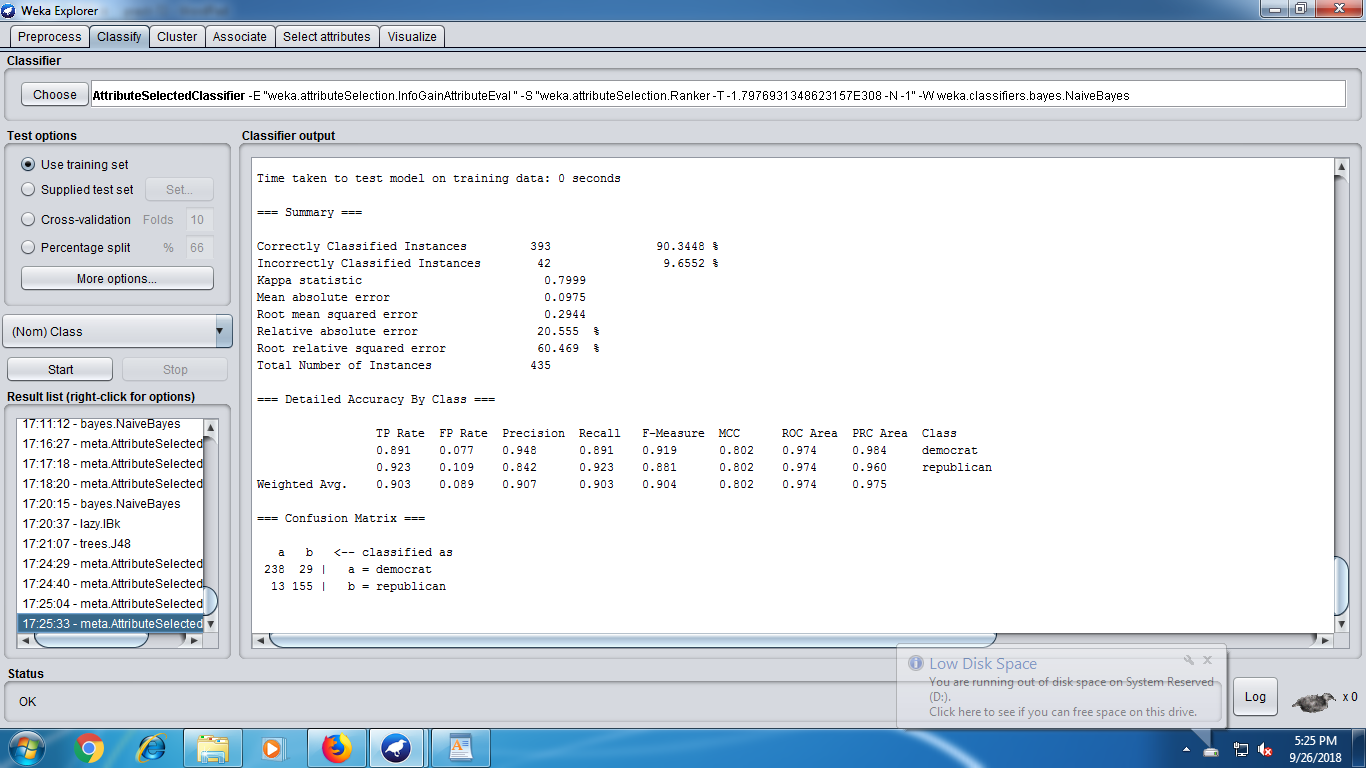
**1. WrapperSubsetEval**

**2. RankSearch**

**3. InfoGainAttributeEval**



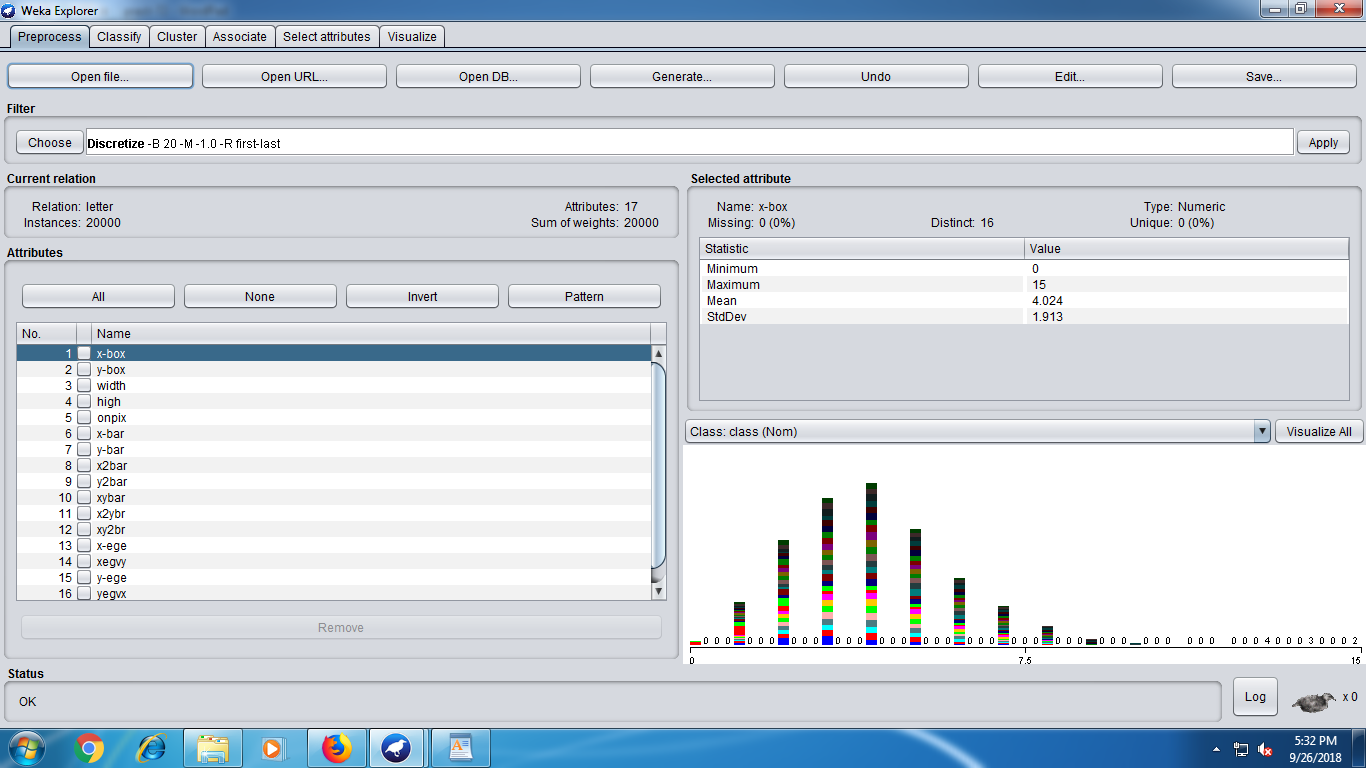




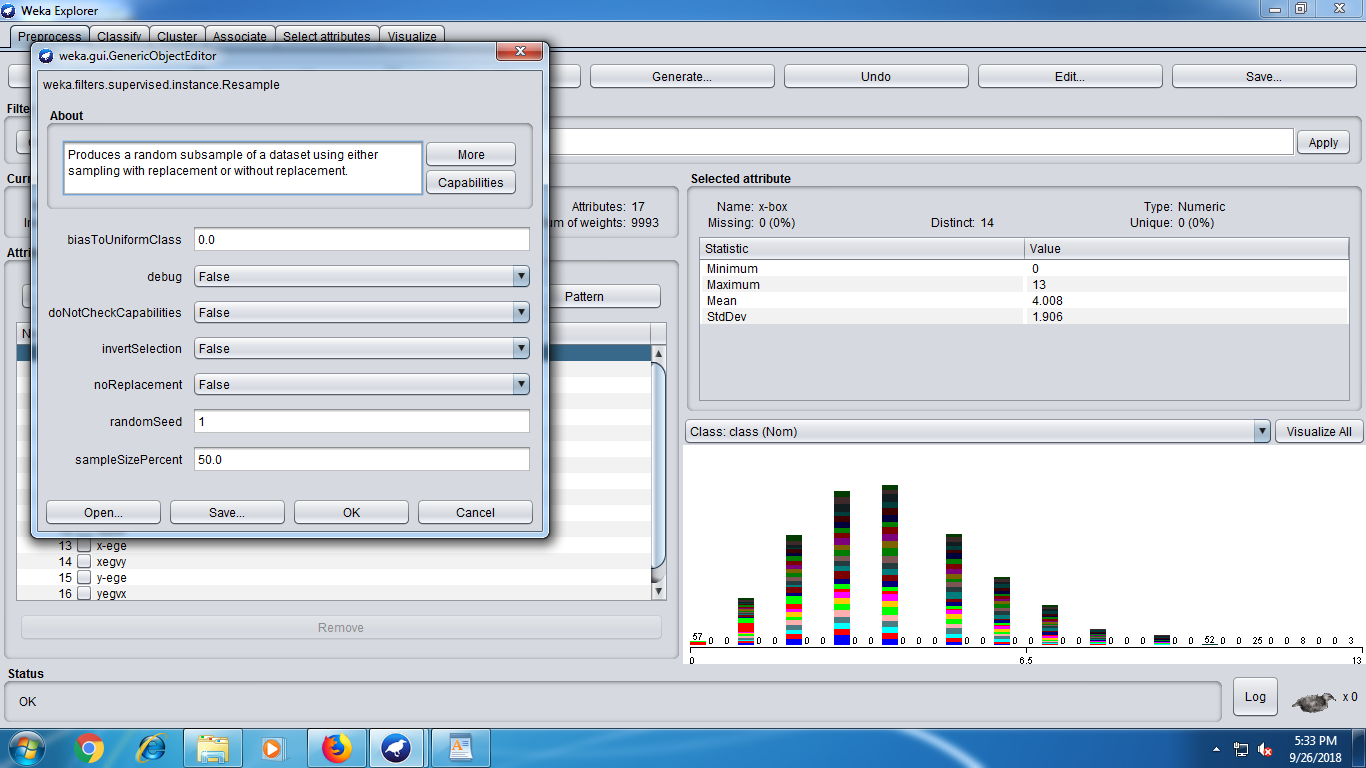
**7. Sampling**

**1. Load the 'letter.arff' dataset**

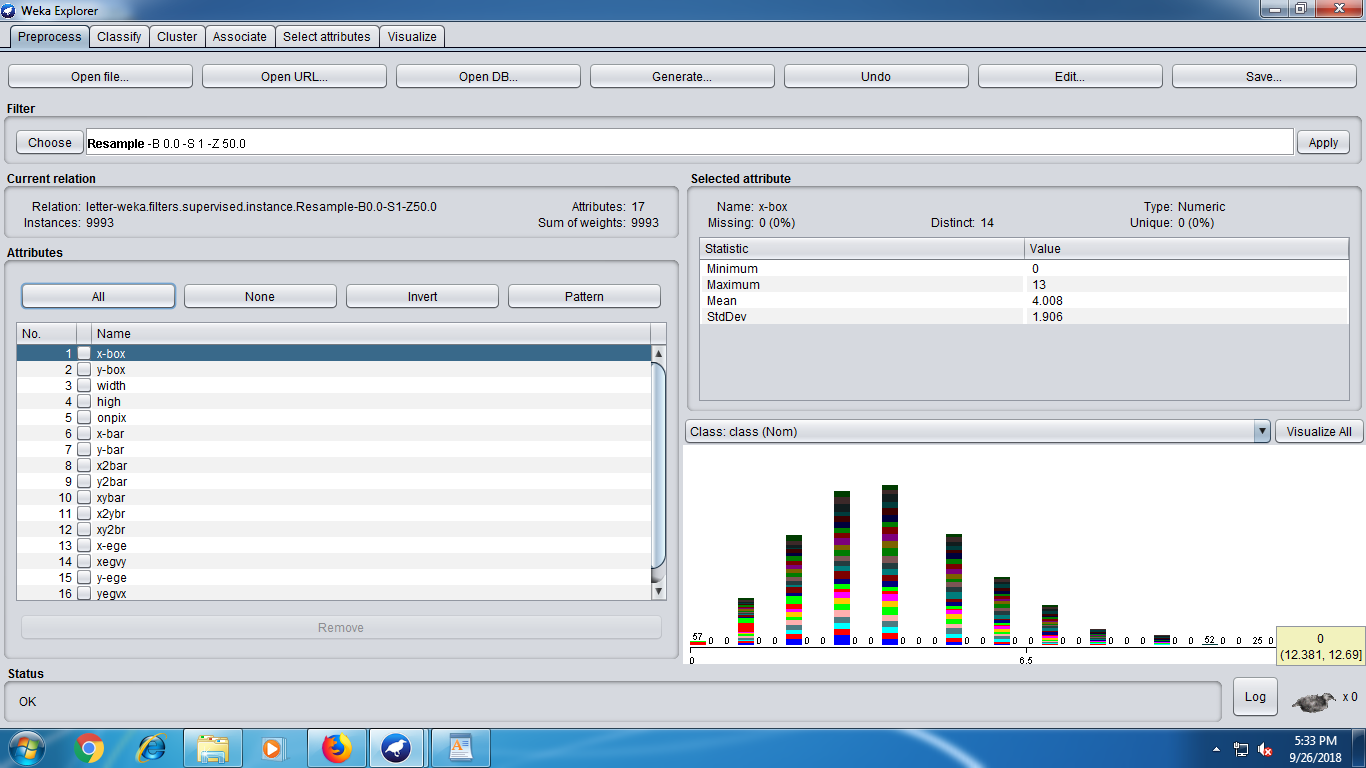
**2. Take any attribute and record the min, max, mean, and standard deviation of the attribute**



**3. Apply the Resample filter with 'sampleSizePercent' set to 50 percent**



1. **What is the size of the filtered dataset. Observe the min, max, mean, and standard deviation Of the attribute that was selected in step 2. What is the percentage change in the values?**



The Change is zero. This suggest that the data sample is best to analyse the results and to further test the new instances.

**5. Give the benefit of sampling a large dataset.**

**Ans :**

1.Sampling saves time to a great extent by reducing the volume of data. Youdo not go through each of the individual items.

2. Sampling Avoids monotony in works. You do not have to repeat the query again and again to all the individual data.

3. When you have limited time, survey without using sampling becomes impossible. It allows us to get near-accurate results in much lesser time

4. When you use proper methods, you are likely to achieve higher level of accuracy by using sampling than without using sampling in some cases due to reduction in monotony, data handling issues etc.

1. By using sampling, you can get detailed information on the data even by employing small amount of resources.