Maciej Medyk – COT6777 – Web Mining

Question 1 – [3.00pt]

What is Entropy? Please show the formula, and explain how to use Entropy to quantify the randomness of the system [0.50pt].

Entropy is a measure of randomness or impurity in a closed system set. Entropy is calculated by formula

$$\text{Entropy}(S) = \sum_{i=1}^{n} -p(i) \log_2 p(i)$$

Entropy can be used to evaluate the randomness of data in the set and if there is no randomness then entropy would be zero. If we take the randomness of the set with two labels + and – when both labels are evenly distributed meaning out of 8 labels 4 belong to + and other 4 belong to – then we will have entropy of 1.

What is Bayes Rule? Please explain how to use Bayes rules for classification (or decision making) [0.50pt].

Bayes Rule takes into consideration probability of an event based on the condition that is related to that event. The Bayes rule can be expressed by following formula.

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)}$$

Applying this rule to classification, we can estimate the probability of class based on the set of attributes. For the formula above we can find calculate posterior probability of how many classes we can find in given document. The $P(X_1 \mid C)$ is calculated using Bayes rule.

$$P(X_1,\ldots,X_5 \mid C) = P(X_1 \mid C) \bullet P(X_2 \mid C) \bullet \cdots \bullet P(X_5 \mid C)$$

What is the conditional independence assumption of the Naïve Bayes learning? Please explain the relationship between Naïve Bayes classifier and the Bayes Rule [0.50pt].

Conditional Independence Assumption assumes that the probability of observing the conjunction of attributes is equal to the product of individual probabilities $P(x_i \mid C_j)$ and it features detect term presence and those terms are independent of each other given the class. This means that not all attributes are necessary to lead to the outcome and outcome is not dependent on all attributes being present. Native Bayesian Classifier is based on Bayes theorem with independence assumption prediction. Bayes theorem provides a way of calculating posterior probability P(c | d). Naïve Bayes classifier assumes that effect of value of predictor d on given class c is independent of values of other predictors.

What is "Information Gain"? Please explain how to use "Information Gain" to construct a decision tree [0.50pt].

Information gain is the difference between original entropy of the system before division into smaller subsets based on division category. For example, we start with set that that has 5 positive labels and 4 negative labels. We calculate entropy for that set. Then we divided this set into subsets based one of the categories, let's say its wind. For wind being strong we have 3 positive and 3 negative labels and for wind weak we have 2 positive and 1 negative labels. We then calculate entropy of both subsets and deduct each from original entropy achieving information gain. Once we find the category we that has the biggest information gain we will use that category for next branch in the decision tree. The formula for information gain is as follows

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

What is the bias of the "Information Gain"? Please show one solution to reduce the bias of the information gain [0.50pt].

Information bias occurs when entropy gain is maximized due to category with unique, ungroupable values. For example if we chose the category like ID which could be an integer with and every one of those numbers will achieve perfect information gain where each subset will have entropy of zero; however, this shouldn't be used at the decision tree as each subset will have count of 1 value, this category is not a groupable category, and truly you won't be building a decision tree, but something more similar to a hash table where each key maps to one specific outcome. One way to avoid information bias is to check if all items in category are unique. If they are then this will lead to information bias. Before we should take category into consideration we should run a command that would check if category has unique values stored in the structured data by using `SELECT category FROM originalset GROUP BY category HAVING COUNT(category) > 1` or use Gain Ratio system instead. If we get a result then we know the items are not unique as at least one item in category appears more than once.

Please list the major steps of using binominal Naïve Bayes learning for text classification [0.50pt].

Binomial Naïve Bayes classifier is one where all features are individually binomial (binary variables) describing inputs and takes into account multiplicity of those binary features. To use it you have to extract vocabulary from documents and count number of documents. We should also remove all duplicate words from document. When classifier runs it generates indicator attached to each vocabulary resulting in 1 indicating presence in the document or 0 indicating absence. To effectively conduct learning process we have to put in a document that is short as Binomial Naïve Bayes classifier does not produce good results with long texts. The model estimates the fraction of documents containing vocabulary term rather than fraction of tokens containing the term as Multinomial Model does. Applying the model will result with conditional probabilities. Example training table is featured below

Boolean	Doc	Words	Class
Training	1	Chinese Beijing	c
	2	Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Tokyo Japan	?

Question 2 – [2.50pt]

Please manually construct a decision tree by using Information Gain Ratio as the attribute selection criteria (list the major steps of the tree constructions, and report the final decision tree) [2.00pt].

We want to calculate Information Gain using Information Gain Ration for each attribute criteria. We will consider Outlook, Temperature, Humidity, and Wind) considering original set. What we need to do is calculate Information Gain and Split Information to achieve Gain Ratio. In order to make the calculations we have to count how many times each yes or now falls into category and sub category in order to find entropy and weights for calculating information split. To take Wind as example on Strong label Yes gets 0 and No gets 1 while on Weak label Yes gets 3 and No gets 1. That will result with entropy of $1/5(-1/1*\log_2(1))=0$ for Yes outcome and entropy $4/5(-3/4*\log2(3/4)-1/4*\log(1/4))=0.649$ for No. Total Information Gain will be $0.971-0.649=0.322$. Then we need to calculate Information Split which would be $-1/5*\log_2(1/5)-4/5*\log_2(4/5)=0.722$ At this moment we simply divide Information Gain by Information Split in order to calculate Gain Ratio and we choose the highest Gain Ratio for next decision branch.

No Root										
Description	Original	Outlook			Temperature			Humidity		Wind
		Sunny	Overcast	Rain	Hot	Mild	Cold	High	Normal	Strong	Weak
Yes	3	1	1	1	1	1	1	2	1	0	3
No	2	2	0	0	1	0	1	2	0	1	1
Column Total	5	3	1	1	2	1	2	4	1	1	4
All Columns Total	5	5			5			5		5	
Ent(Yes)	0.442	0.528	0.000	0.000	0.500	0.000	0.500	0.500	0.000	0.000	0.311
Ent(No)	0.529	0.390	0.000	0.000	0.500	0.000	0.500	0.500	0.000	0.000	0.500
Ent(Yes) + Ent(No) Total	0.971	0.918	0.000	0.000	1.000	0.000	1.000	1.000	0.000	0.000	0.811
Weighted Entropy	0.971	0.551	0.000	0.000	0.400	0.000	0.400	0.800	0.000	0.000	0.649
Info Gain		0.420			0.171			0.171		0.322	
Split Info per Column		0.442	0.464	0.464	0.529	0.464	0.529	0.258	0.464	0.464	0.258
Total Split Information		1.371			1.522			0.722		0.722	
Gain Ratio (Info Gain / Split Info)		**0.306**			**0.112**			**0.237**		**0.446**	

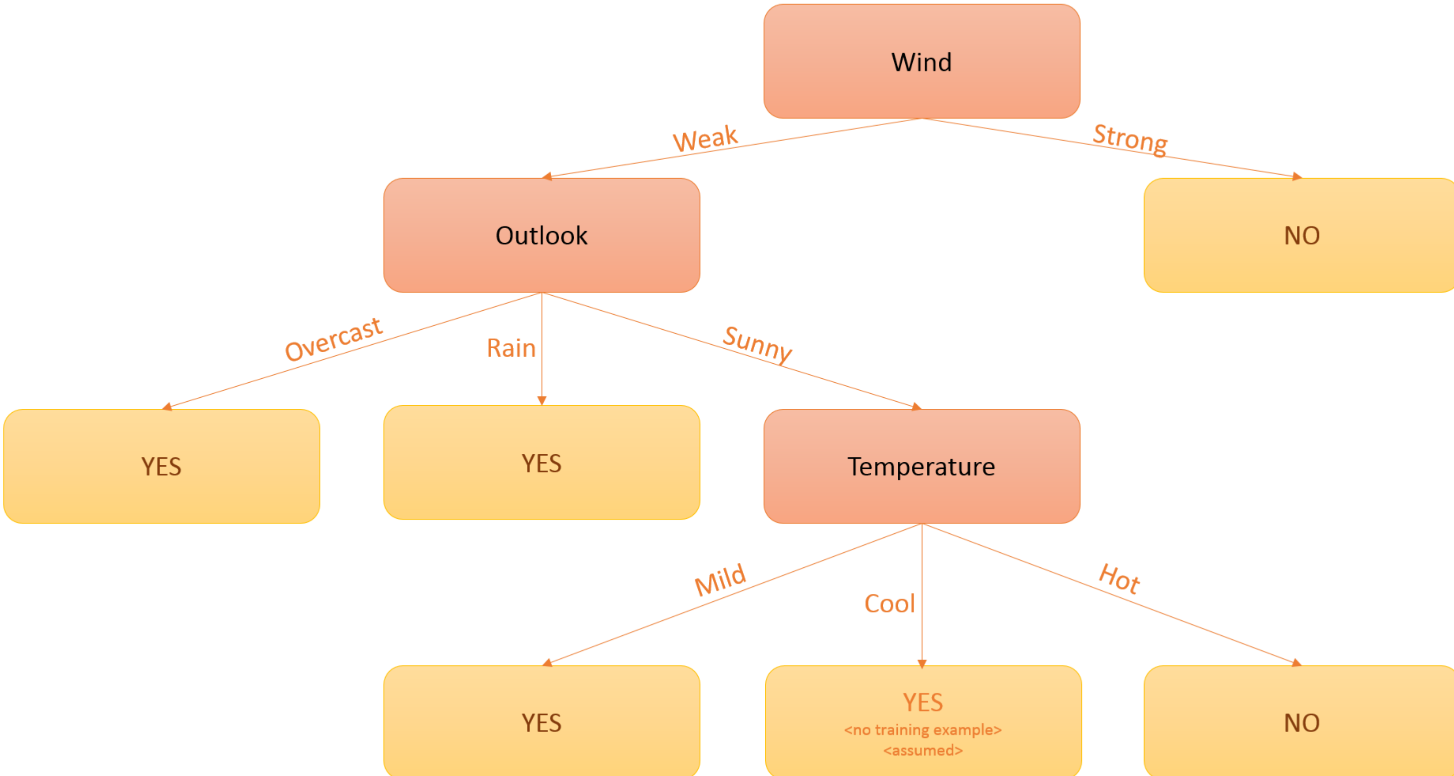
After calculating Information Gain we see that Outlook has the highest information gain from all 4 categories; however, the best Gain Ratio belongs to category Wind and Wind will be chosen as root node. We see that Strong label leads to NO while Weak leads to 3 Yes and 1 No so we need to calculate remaining categories (Outlook, Temperature, and Humidity) for Wind-Weak subset.

Wind-Weak									
Description	Wind Weak	Outlook			Temperature			Humidity	
		Sunny	Overcast	Rain	Hot	Mild	Cold	High	Normal
Yes	3	1	1	1	1	1	1	2	1
No	1	1	0	0	1	0	0	1	0
Column Total	4	2	1	1	2	1	1	3	1
All Columns Total	4	4			4			4	
Ent(Yes)	0.311	0.500	0.000	0.000	0.500	0.000	0.000	0.390	0.000
Ent(No)	0.500	0.500	0.000	0.000	0.500	0.000	0.000	0.528	0.000
Ent(Yes) + Ent(No) Total	0.811	1.000	0.000	0.000	1.000	0.000	0.000	0.918	0.000
Weighted Entropy	0.811	0.500	0.000	0.000	0.500	0.000	0.000	0.689	0.000
Info Gain		0.311			0.311			0.123	
Split Info per Column		0.500	0.500	0.500	0.500	0.500	0.500	0.311	0.500
Total Split Information		1.500			1.500			0.811	
Gain Ratio (Info Gain / Split Info)		**0.208**			**0.208**			**0.151**	

After calculating Information Gain we see that both Outlook and Temperature have highest Information Gain and Gain Ratio. We will chose Outlook as labels Overcast and Rain both lead to YES while Sunny leads to 1 Yes and 1 No; therefore, we need to do further calculation on Outlook Sunny.

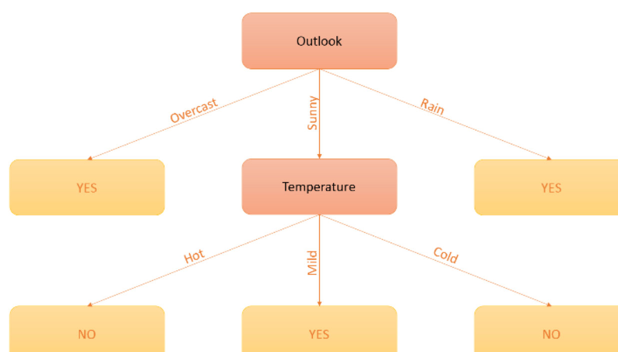
Outlook-Sunny						
Description	Outlook Sunny	Temperature			Humidity	
		Hot	Mild	Cold	High	Normal
Yes	1	0	1	0	1	0
No	1	1	0	0	1	0
Column Total	2	1	1	0	2	0
All Columns Total	2	2			2	
Ent(Yes)	0.500	0.000	0.000	0.000	0.500	0.000
Ent(No)	0.500	0.000	0.000	0.000	0.500	0.000
Ent(Yes) + Ent(No) Total	1.000	0.000	0.000	0.000	1.000	0.000
Weighted Entropy	1.000	0.000	0.000	0.000	1.000	0.000
Info Gain		1.000			0.000	
Split Info per Column		0.500	0.500	0.000	0.000	0.000
Total Split Information		1.000			0.000	
Gain Ratio (Info Gain / Split Info)		**1.000**			**0.000**	

After calculating Information Gain we see that Temperature have highest Information Gain and Gain Ratio. We will choose Temperature as Hot leads to NO and Mild leads to YES. Now we reached the time when we can build the decision tree.



It is known that C4.5 uses Information Gain Ratio measure for decision tree construction, does C4.5 search all hypothesis to find the best tree (explain why or why not) [0.25pt].

It seems like Gain Ratio looks at the best ratio of Information Gain vs how many decisions are made meaning how many sub labels the label contain. The bias here comes with the fact that the Gain Ratio tends to favor the labels with smaller split like wind having two sub categories of Weak and Strong rather than Outlook which has three subcategories of Overcast, Rain, and Sunny. As this is a good way to avoid the issue with attribute that might have all unique labels it does not guarantee that the tree found is the best tree. If we use Information Gain only we actually result at much smaller tree of height 3 rather than height 4. The C 4.5 does not check all possibilities to find best tree but addresses issues with regular tree generation as category with unique labels had largest information gains but also highest split.



Tree using only Information Gain

What is the inductive Bias of C4.5 [0.25pt].

It chooses first tree that is acceptable while not necessarily choosing the shortest, most optimal tree. This method is designed to avoid issues with entropy calculations and it chooses the attributes with highest Gain Ratio rather than Information Gain closest to the root.

Question 3 – [2.50pt]

Please manually construct a Naïve Bayes Classifier (list the major steps, including the values of the priori probability [1.00pt] and the conditional probabilities [1.00pt]. Please use m-estimate to calculate the conditional probabilities (m=1, and p equals to 1 divided by the number of attribute values for each attribute) [2.00pt].

Initially we have to calculate probability of class Yes and No where P(Yes) would be calculated as how many Yes occurs within N set. The result of P(Yes) = 9/15 and P(N)=6/15. Then we need to calculate probability for each category and label given class label. What that means is we need to calculate the probability of categories Outlook P(Sunny | Yes), P(Sunny | No), P(Overcast | Yes), P(Overcast | No), etc. This is done by calculating for example how many outlooks Sunny is associated with word Yes. We find that Sunny appears 2 out of total of 9 yes present.

Class			
	Outcome	P(Yes) = 0.594	P(Yes) = 7.033
Outlook			
	Sunny	P(Sunny | Yes) = 0.233	P(Sunny | Yes) = 0.476
	Overcast	P(Overcast | Yes) = 0.333	P(Overcast | Yes) = 0.190
	Rain	P(Rain | Yes) = 0.433	P(Rain | Yes) = 0.333
Temperature			
	Hot	P(Hot | Yes) = 0.233	P(Hot | Yes) = 0.333
	Mild	P(Mild | Yes) = 0.433	P(Mild | Yes) = 0.476
	Cool	P(Cool | Yes) = 0.333	P(Cool | Yes) = 0.190
Humidity			
	High	P(High | Yes) = 0.450	P(High | Yes) = 0.643
	Normal	P(Normal | Yes) = 0.550	P(Normal | Yes) = 0.357
Wind			
	Strong	P(Strong | Yes) = 0.450	P(Strong | Yes) = 0.500
	Weak	P(Weak | Yes) = 0.550	P(Weak | Yes) = 0.500

Please use your Naïve Bayes classifier to determine whether a person should play tennis or not, under conditions that "Outlook=Overcast & Temperature=Hot & Humidity =Normal& Wind=Weak" [0.50pt] .

P(yes) = $0.333 \times 0.233 \times 0.550 \times 0.550 = 0.02353 = 2.353\%$

P(no) = $0.190 \times 0.333 \times 0.357 \times 0.500 = 0.01134 = 1.134\%$

Person should play tennis as P(yes) > P(no)

Question 4 – [2.00pt]

A patient takes a lab test and the result comes back positive. Assume the test returns a correct positive result in only 95% of the cases in which the disease is actually present, and a correct negative result in only 95% of the cases in which the disease is not present. Assume further that 0.001 of the entire population have this cancer. Please use Bayes Rule to derive the probability of the patient having the cancer given that his/her lab test is positive (list the major steps) [2.00pt].

Things we know from the question is listed below in the table.

P(cancer)	**0.001**
P(+|cancer)	**0.950**
P(+|¬ cancer)	**0.050**

P(¬ cancer)	**0.999**
P(−|cancer)	**0.050**
P(−|¬ cancer)	**0.950**

We can use the following formula to calculate probability of patient having cancer given his test is positive by

$$P(cancer \mid +) = \frac{P(+ \mid cancer)\, P(cancer)}{P(+)} = \frac{0.00095}{P(+)}$$

We still need to calculate P(+) which will equal

$$P(+) = (P(+ \mid cancer)\, P(cancer)) + (P(+ \mid \neg\,\text{cancer})\ \text{P}(\neg\text{cancer}))$$

$$P(cancer \mid +) = \frac{P(+ \mid cancer)\, P(cancer)}{P(+)} = \frac{P(+ \mid cancer)\, P(cancer)}{(P(+ \mid cancer)\, P(cancer)) + (P(+ \mid \neg\,\text{cancer})\ \text{P}(\neg\text{cancer}))}$$

$$P(cancer \mid +) = \frac{P(+ \mid cancer)\, P(cancer)}{(P(+ \mid cancer)\, P(cancer)) + (P(+ \mid \neg\,\text{cancer})\ \text{P}(\neg\text{cancer}))}$$

$$P(cancer \mid +) = \frac{0.950 \times 0.001}{(0.950 \times 0.001) + (0.050 \text{ x } 0.999)} = \frac{0.00095}{0.05090} = 0.018664 = 1.8664\%$$

Question 5 – [2.00pt]

Please calculate the expected numbers of Male-Republican, Male-Democrat, Male-Independent, Female-Republican, Female Democrat, and Female-Independent [1.00pt]

Probability	Probability	Expected (P*1000)
P(Republican) = 0.450	P(Republican, Male) = 0.180	**180**
P(Democrat) = 0.450	P(Republican, Female) = 0.270	**270**
P(Independent) = 0.100	P(Democrat, Male) = 0.180	**180**
P(Male) = 0.400	P(Democrat, Female) = 0.270	**270**
P(Female) = 0.600	P(Independent, Male) = 0.040	**40**
	P(Independent, Female) = 0.060	**60**

	Voting Preference			
Voters	Republican	Democrat	Independent	
Male Voter	200 (180)	150 (180)	50 (40)	400
Female Voter	250 (270)	300 (270)	50 (60)	600
	450	450	100	1000

Please calculate the Chi-Square value, and the corresponding p-value [0.50pt]

Voter Type	Observed	Expected	O-E	$(O-E)^2$	CHI $((O-E)^2)/E$
Republican, Male	200	180	20	400	2.2222
Republican, Female	250	270	-20	400	1.4815
Democrat, Male	150	180	-30	900	5.0000
Democrat, Female	300	270	30	900	3.3333
Independent, Male	50	40	10	100	2.5000
Independent, Female	50	60	-10	100	1.6667
Chi-Square statistic value =					**16.2037**
Probability Value using Formula (1-CHISQ.DIST(S15,2,TRUE))					**0.000303**

The Chi-Squared value is 16.2037
The Degree of Freedom value is (3-1)*(2-1) = 2
P-Value is 0.000303

Degrees of freedom (df)	χ^2 value[17]											
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83	
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82	16.20
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27	
P value (Probability)	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001	

P=0.000303

Results						
	Republican	Democrat	Independent			*Row Totals*
Male	200 (180.00) [2.22]	150 (180.00) [5.00]	50 (40.00) [2.50]			400
Female	250 (270.00) [1.48]	300 (270.00) [3.33]	50 (60.00) [1.67]			600
Column Totals	450	450	100			1000 (Grand Total)

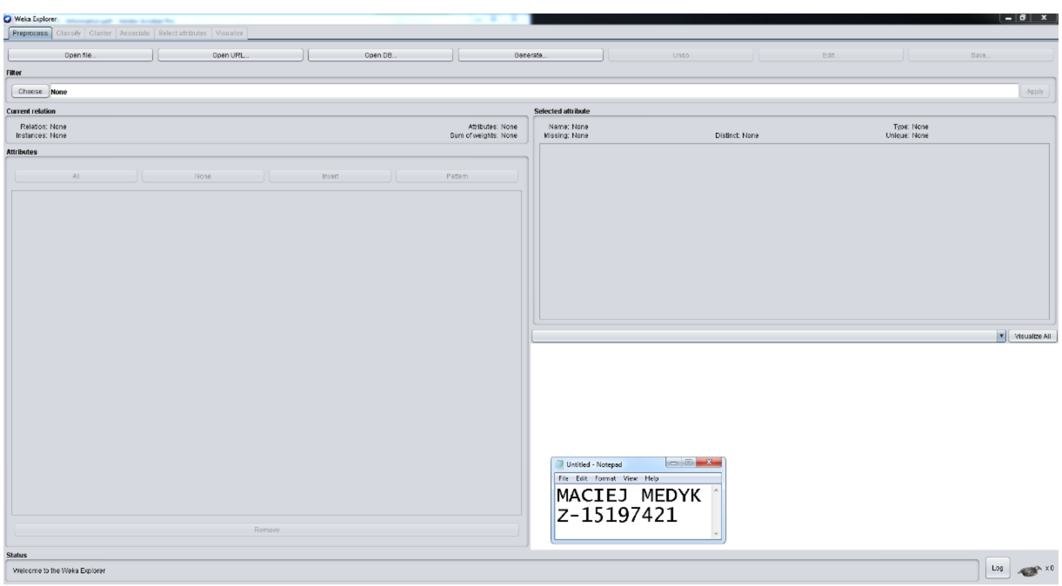
The chi-square statistic is 16.2037. The *p*-value is .000303.

Please explain whether there is a dependence between gender and the voting preference [0.50pt]

Null hypothesis states that variables (gender and voting) are independent. For analysis the Significance Level is 0.05 and our P-value is 0.000303. Since P-value < Significance Level we cannot accept the null hypothesis and have to derive at the conclusion that there exists a relationship between gender and voting preference.

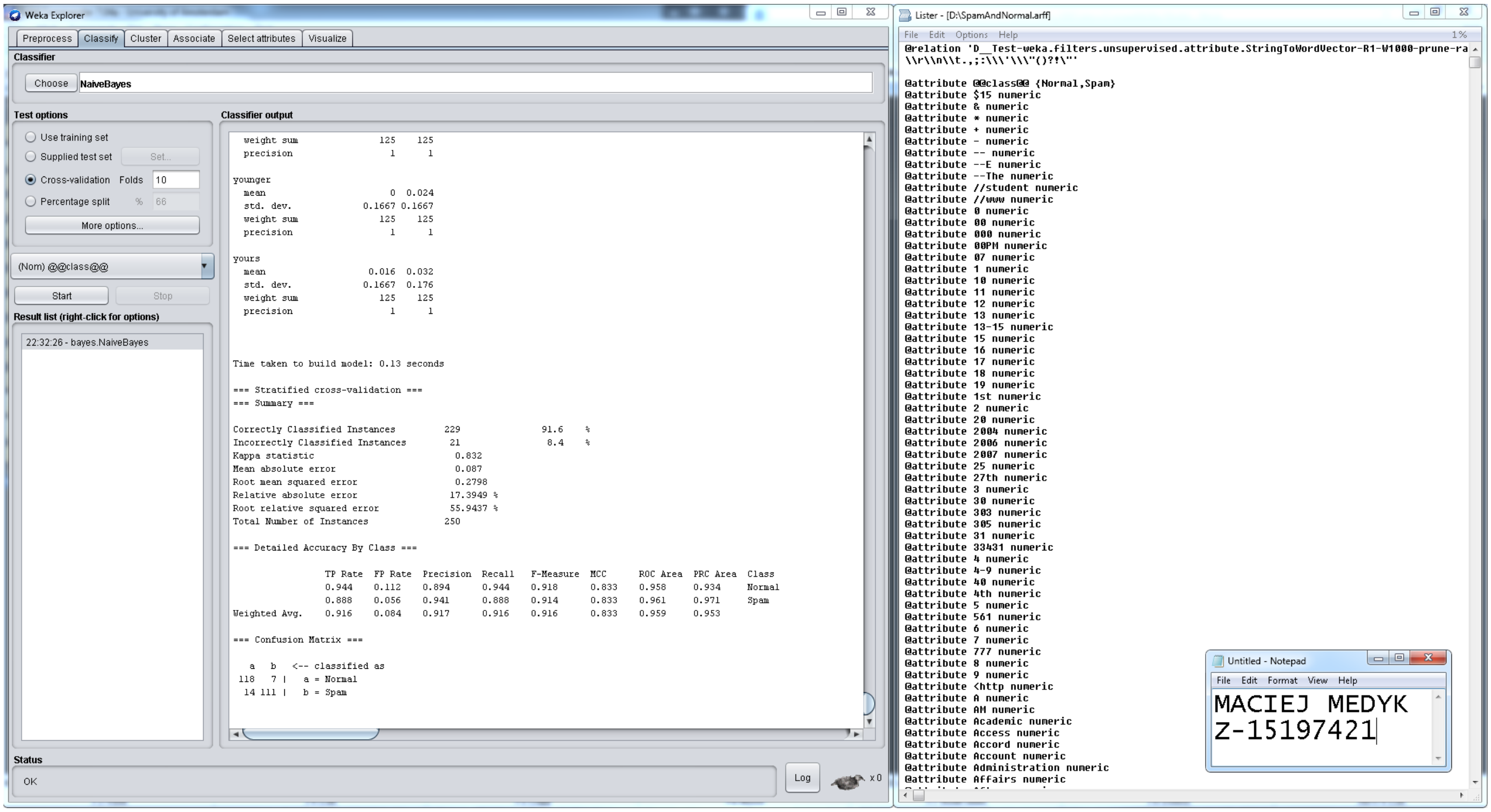
Question 6 – [3.00pt]

Please download and install WEKA (http://www.cs.waikato.ac.nz/ml/weka/), and show a screenshot that WEKA is running on your computer [1.00pt].



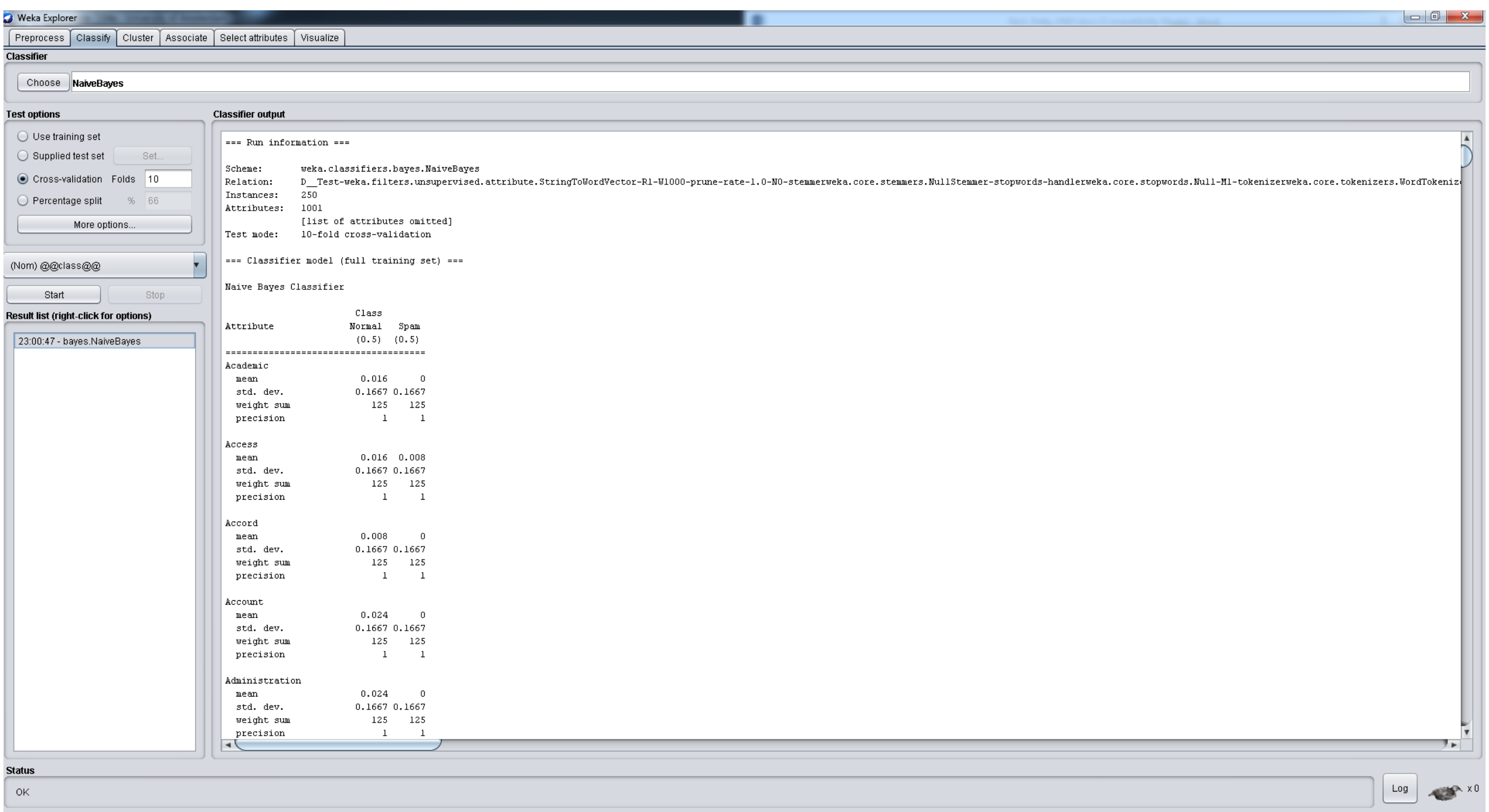
Please report the classification accuracy of your Naïve Bayes classifier, using 10-fold cross validation [1.00pt]

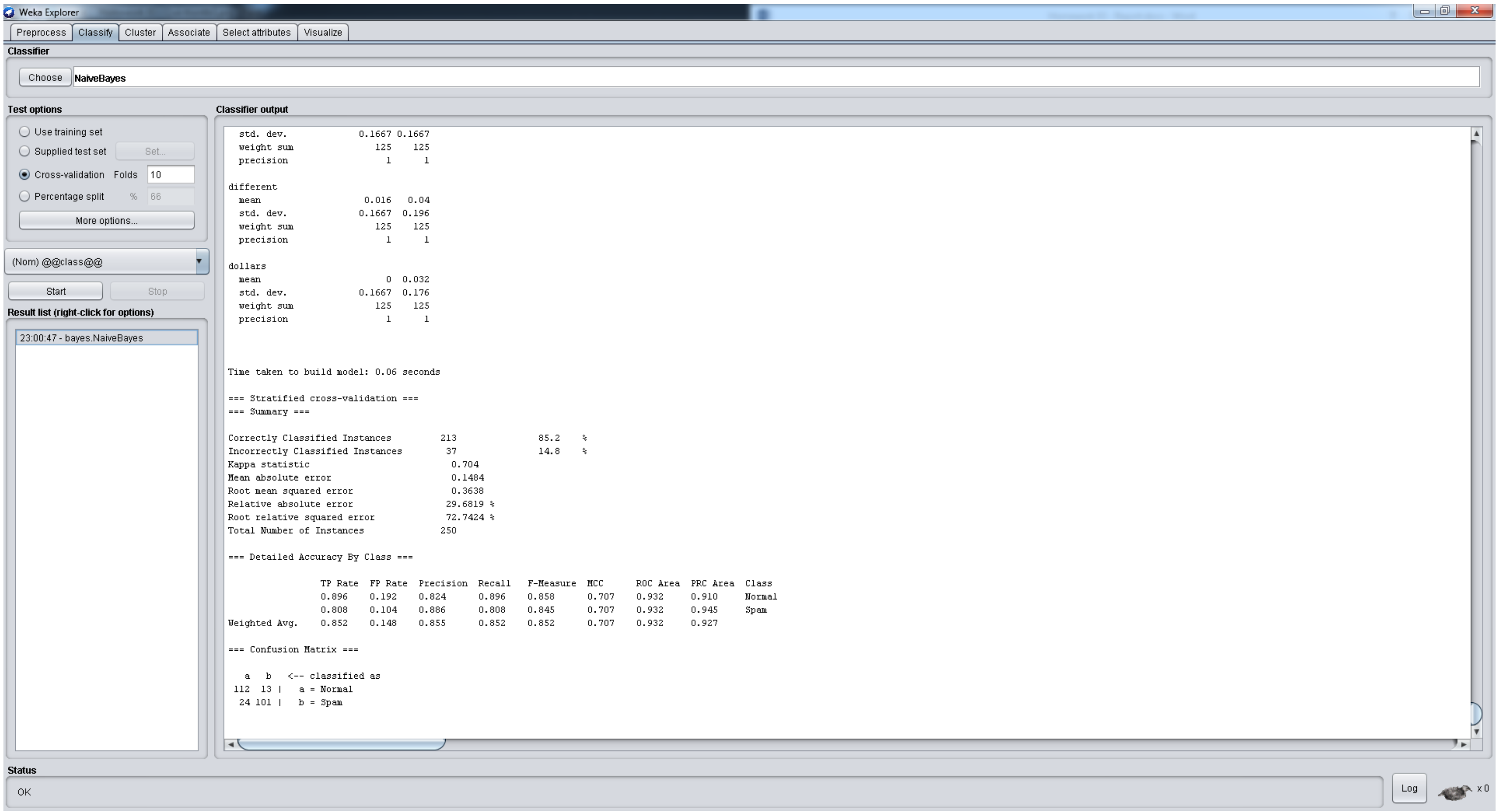
Report used 1789 attributes using 10-fold cross identification and classified 91.6% of instances correctly.



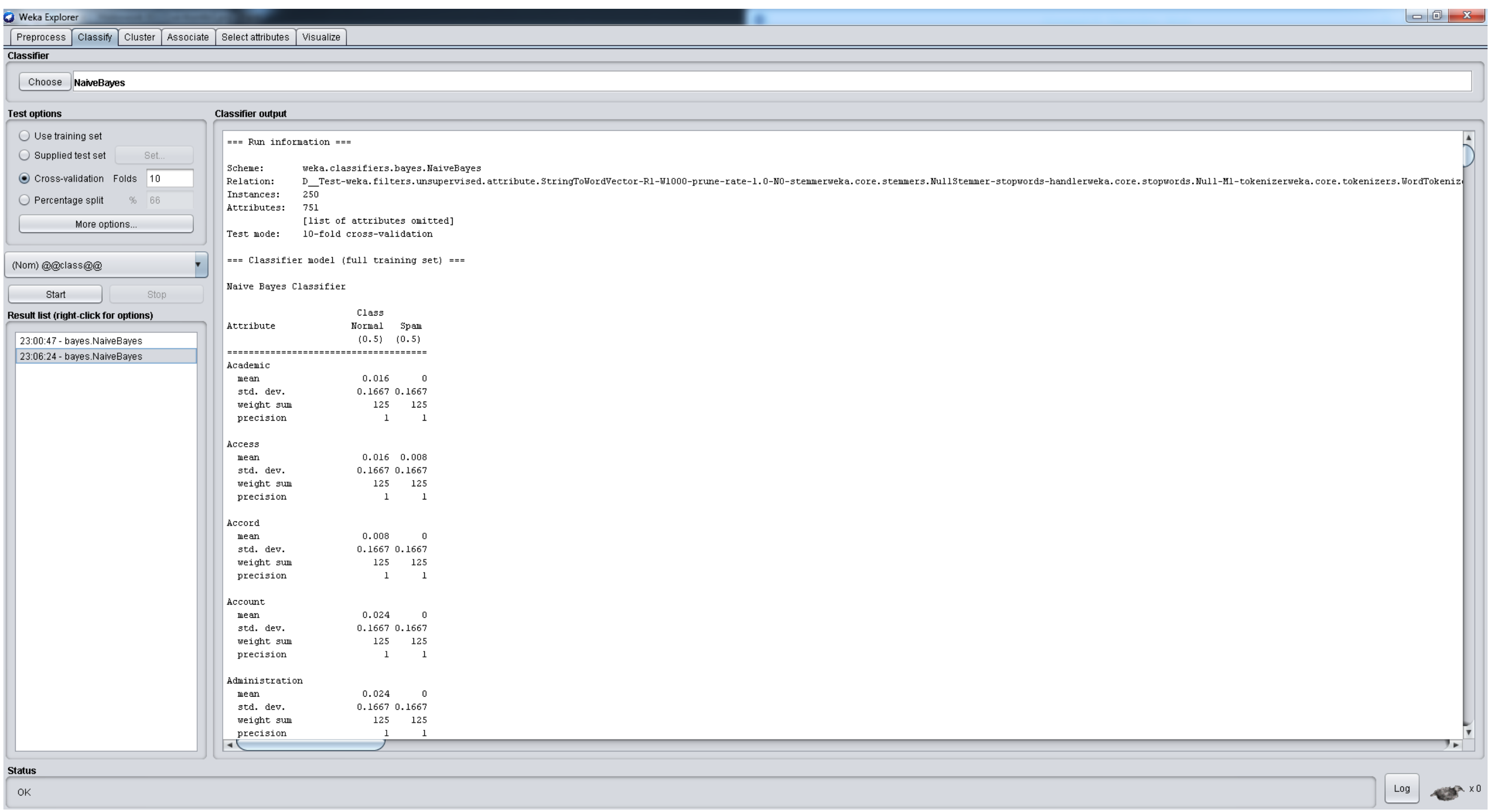
Please use Information Gain to select 1000, 750, 500, 250, 100 features/keywords, respectively, and report the Naïve Bayes classification results for each classifiers. [1.00pt]

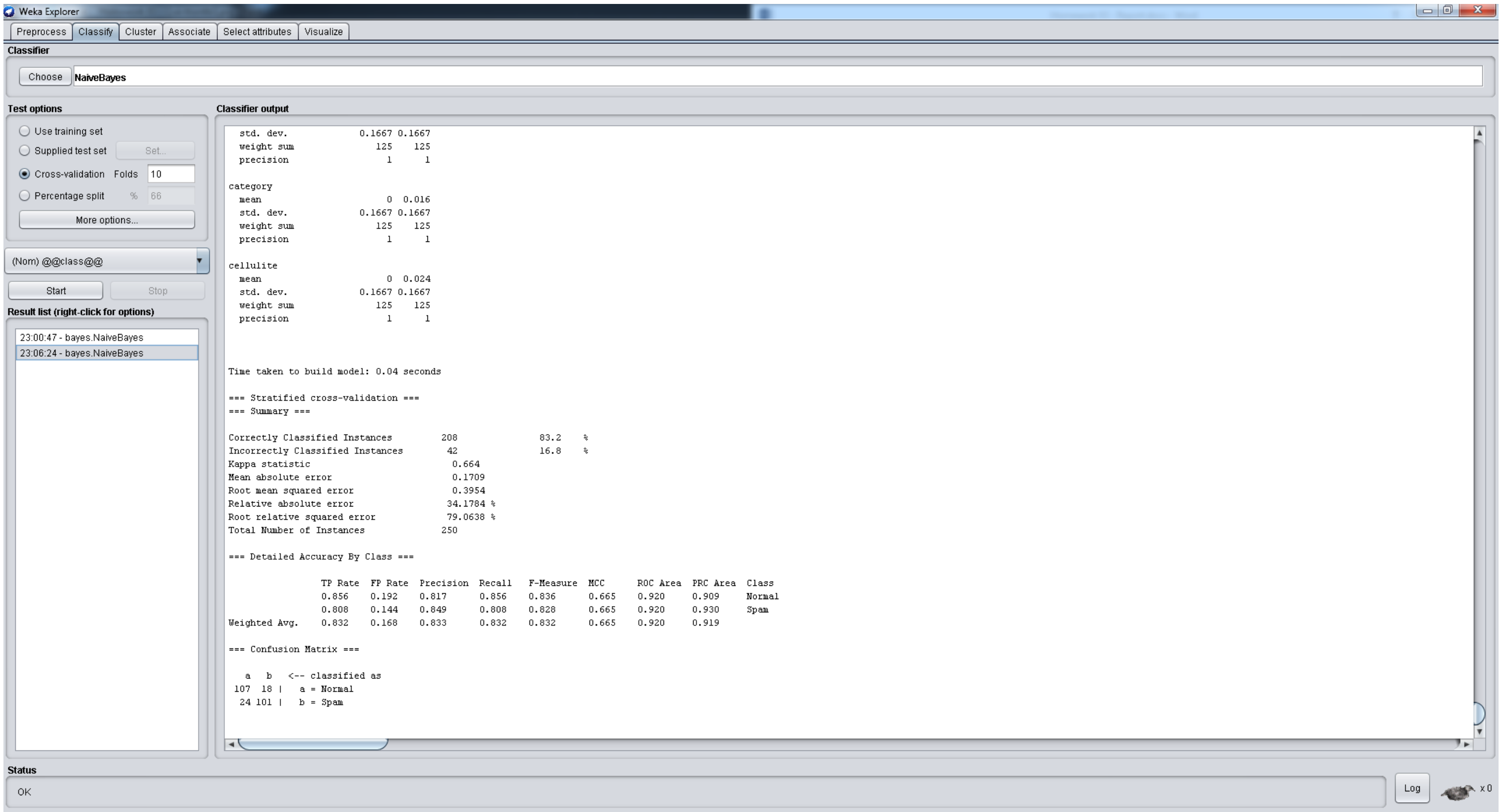
Report used 1000 attributes using 10-fold cross identification and classified 85.2% of instances correctly.



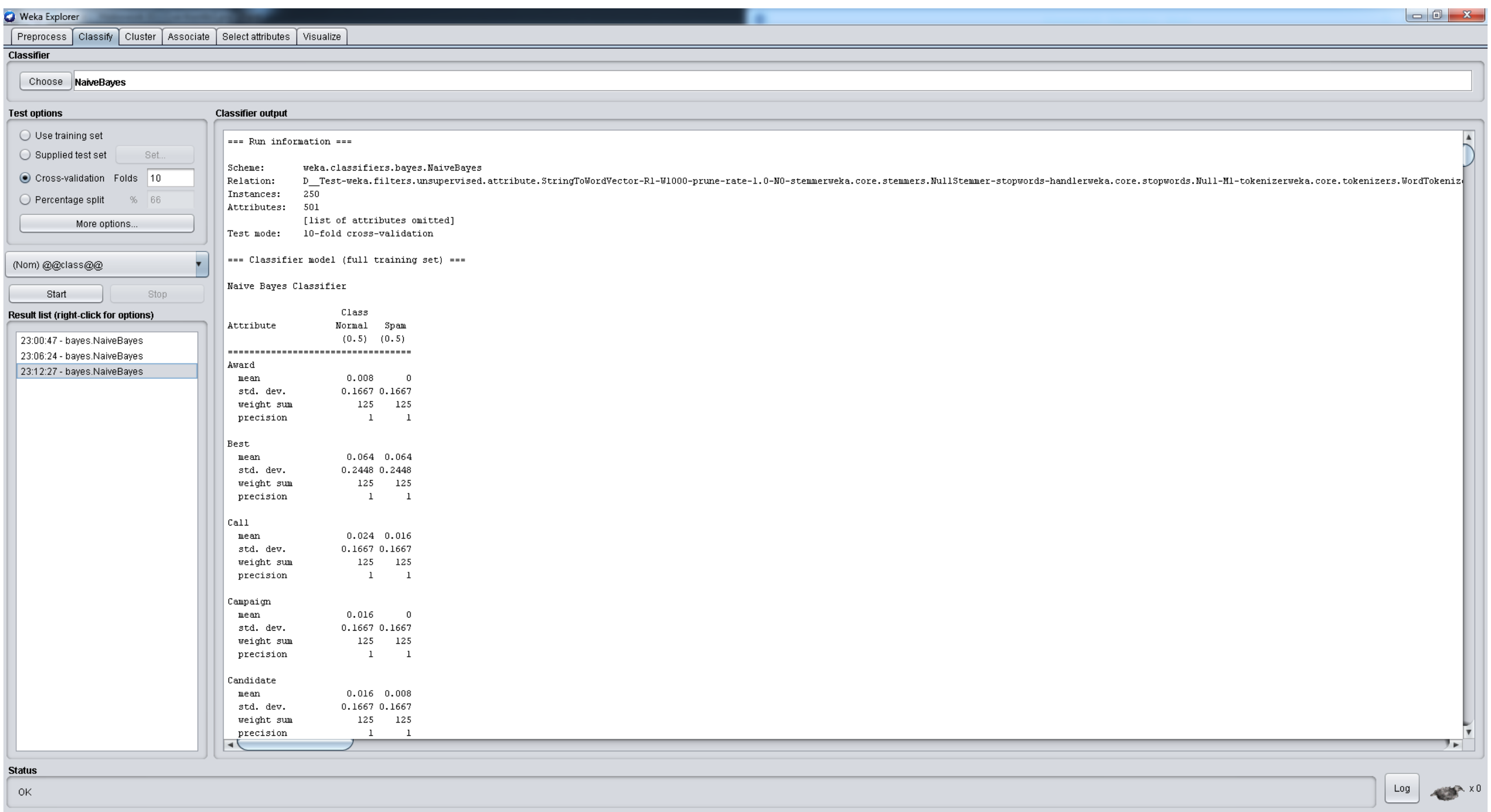


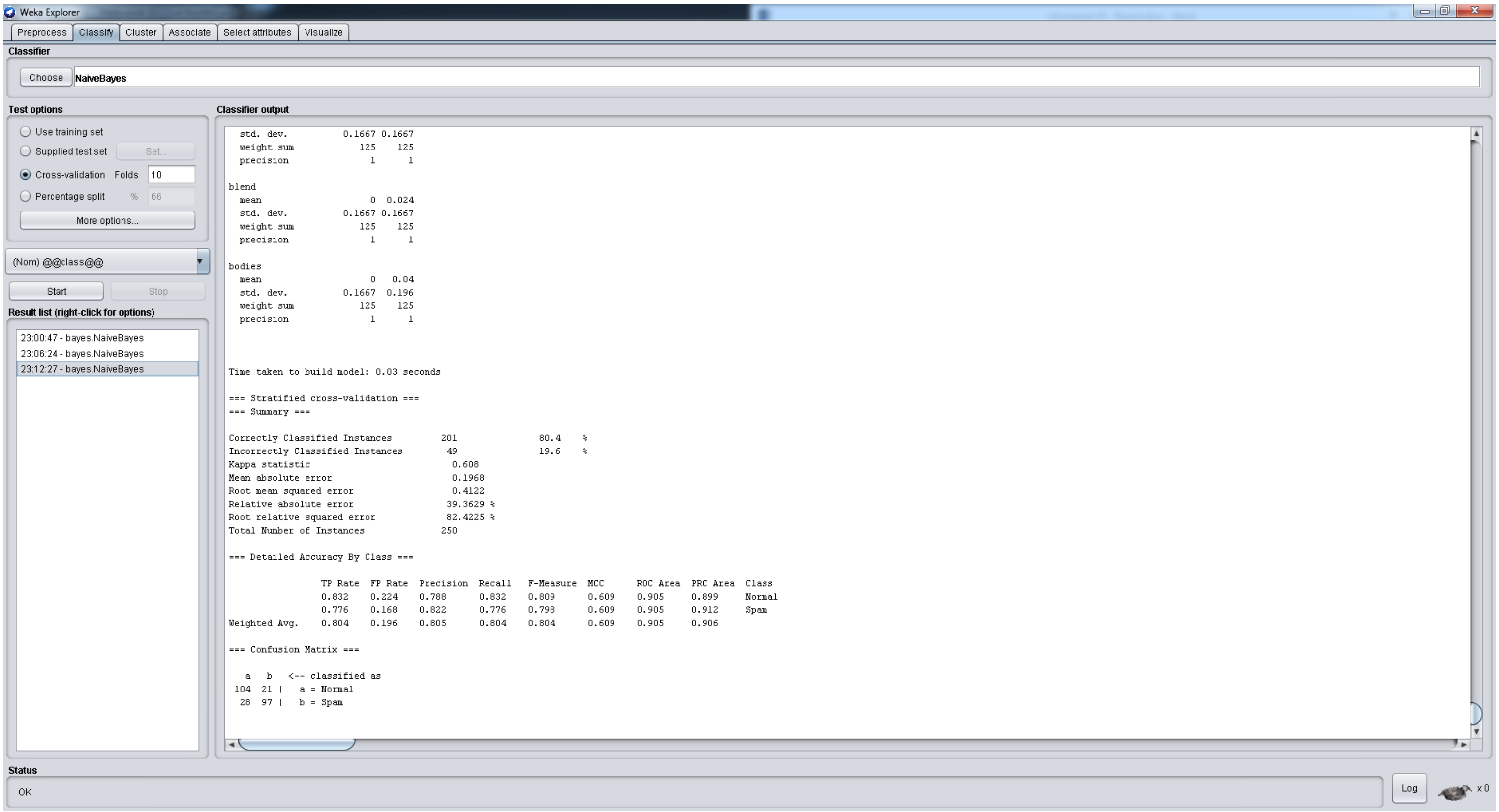
Report used 750 attributes using 10-fold cross identification and classified 83.2% of instances correctly.



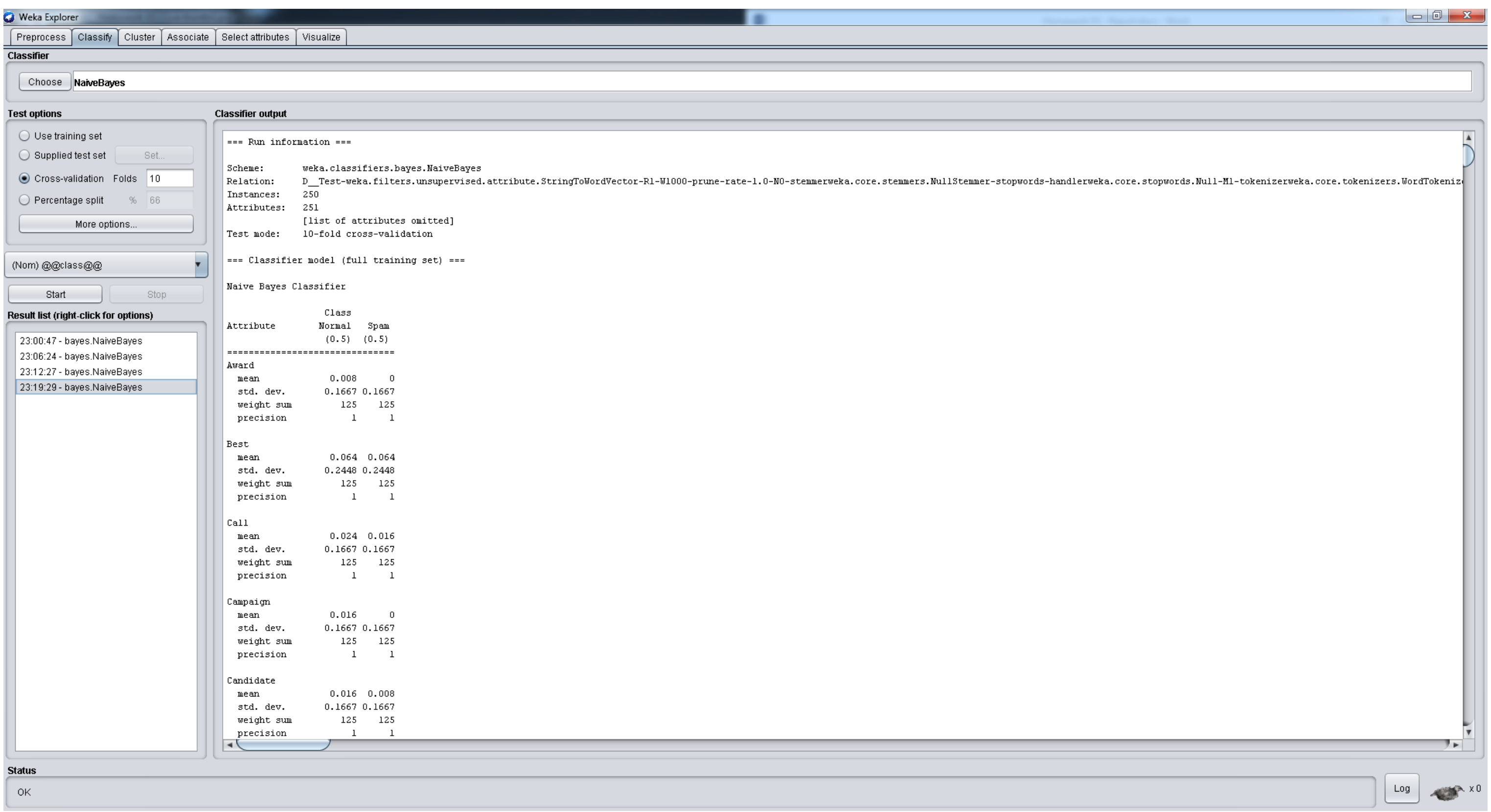


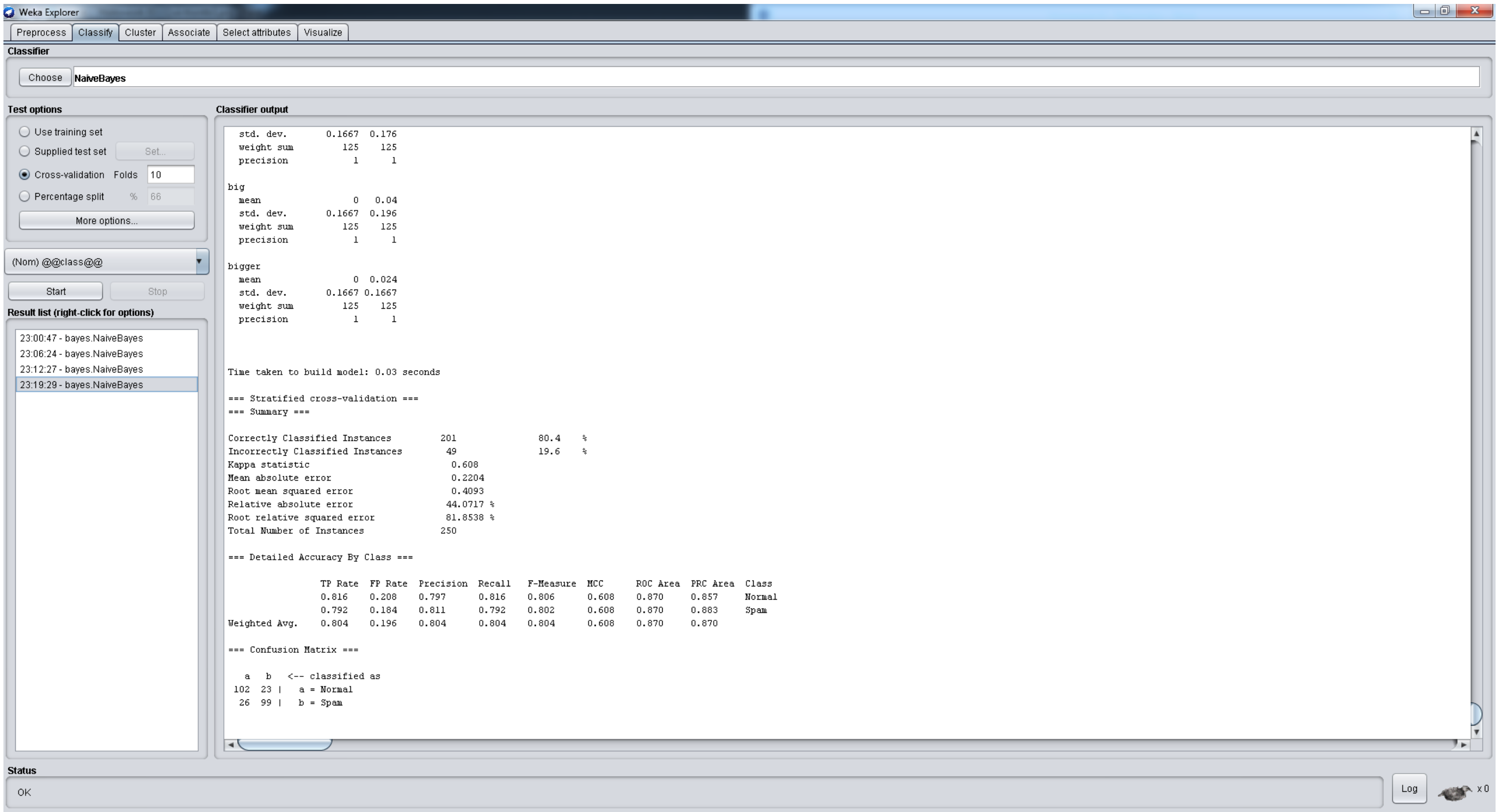
Report used 500 attributes using 10-fold cross identification and classified 80.4% of instances correctly.



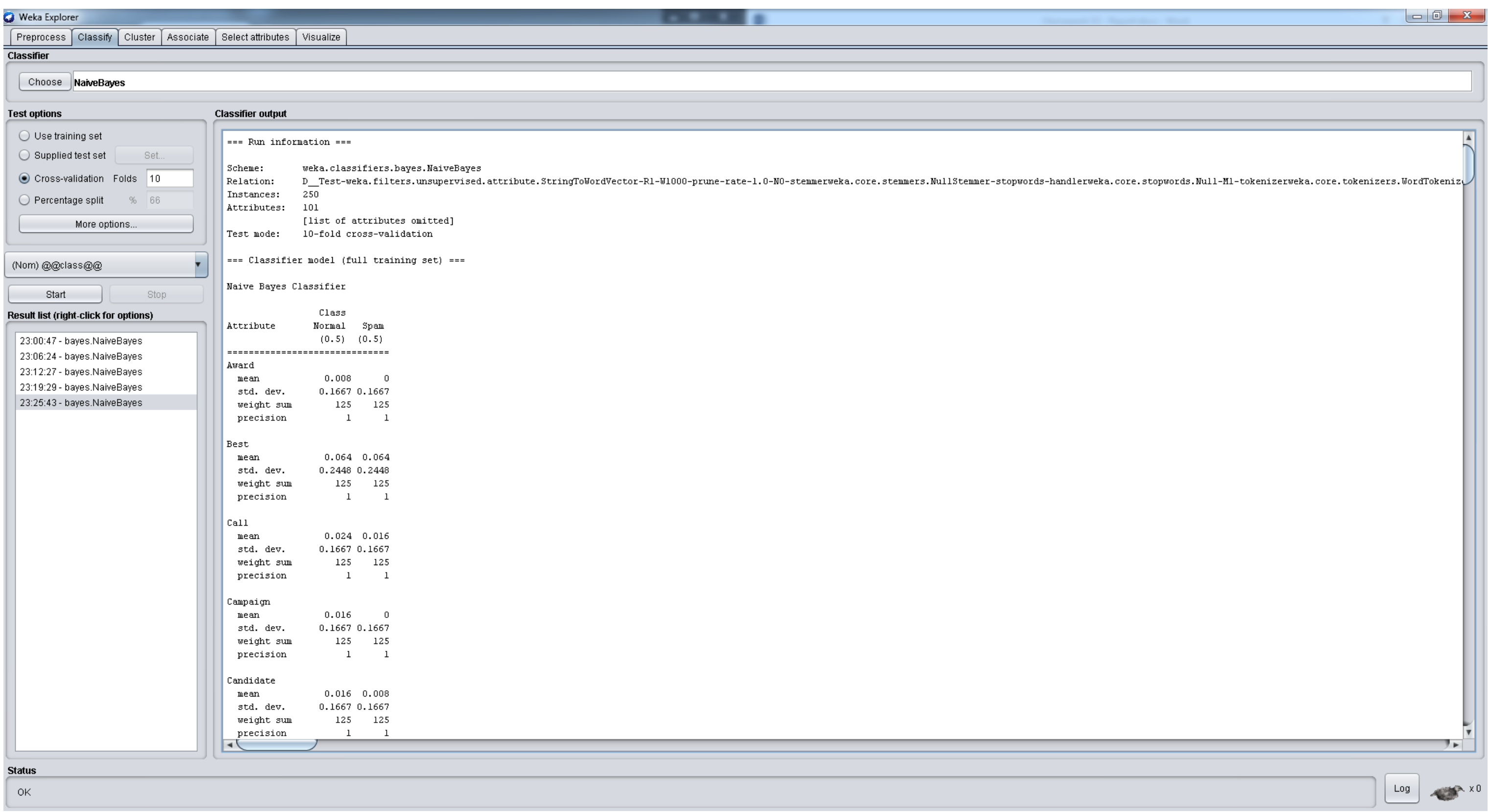


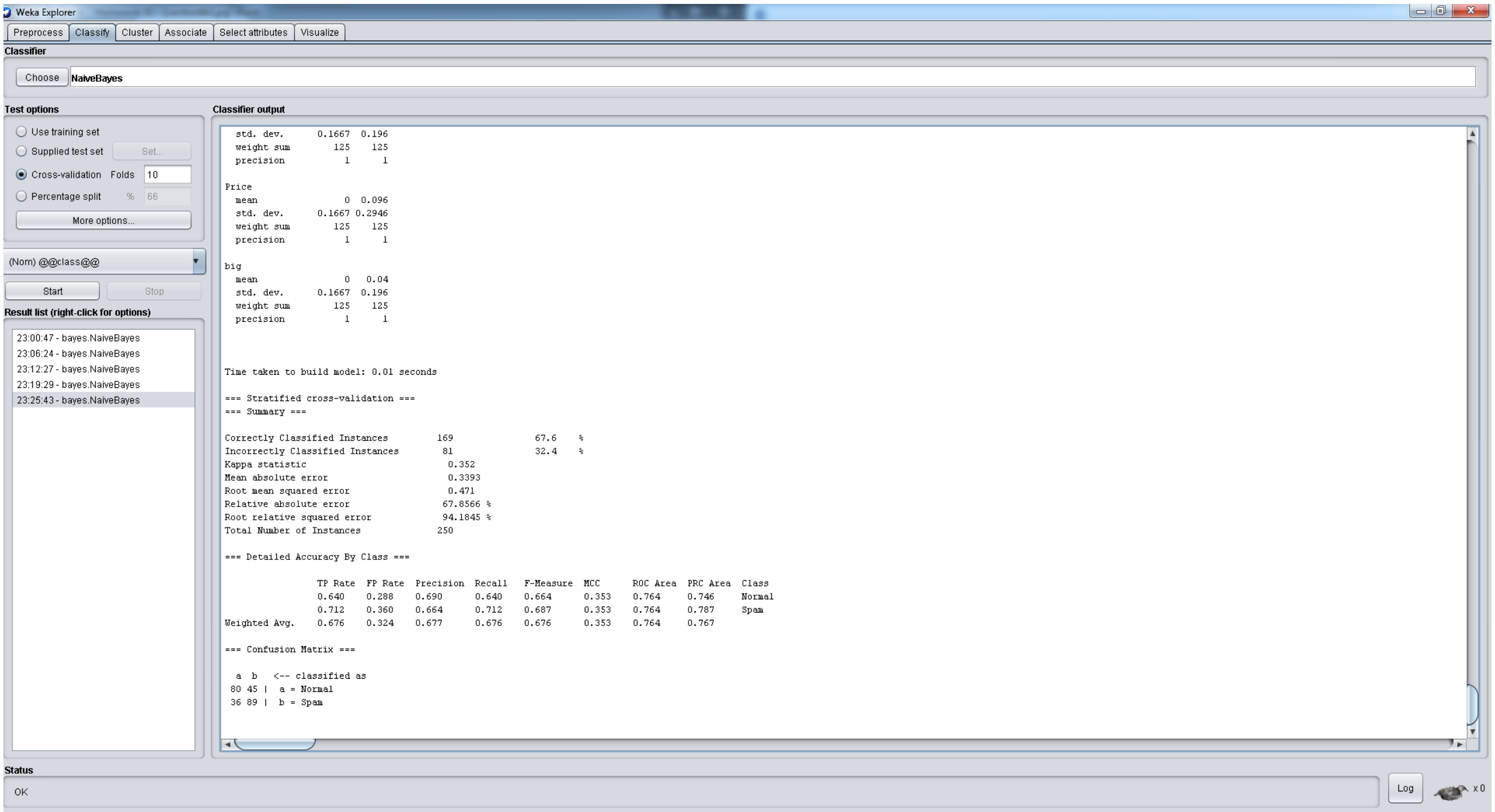
Report used 250 attributes using 10-fold cross identification and classified 80.4% of instances correctly.





Report used 100 attributes using 10-fold cross identification and classified 67.6% of instances correctly.





All files with extension ARFF used for conducting experiment are available under this Dropbox Link below
Link : https://www.dropbox.com/sh/sqqy13fu6ut6zmk/AABAbHZ0L4dI6nJgFJAPhdM8a?dl=0