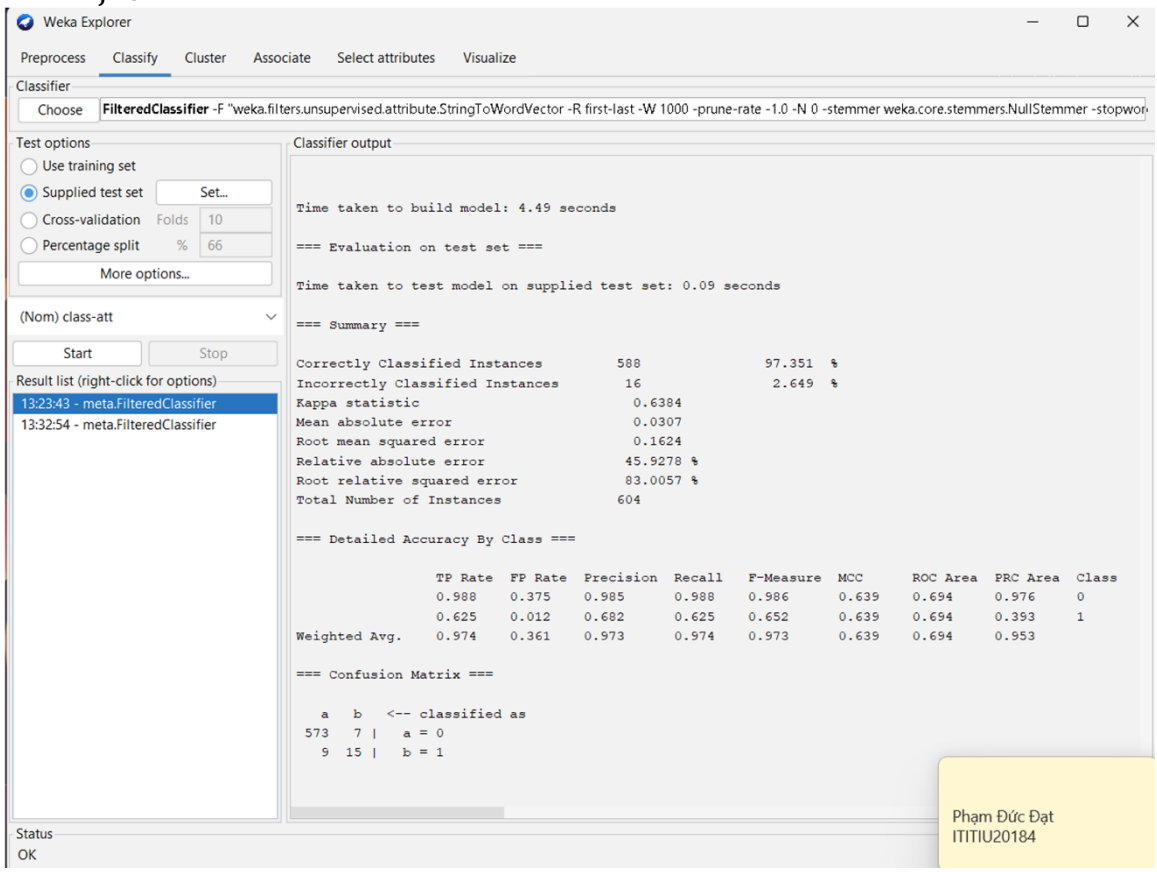
Name: Phạm Đức Đạt
ID Student: ITITIU20184

Question 1:

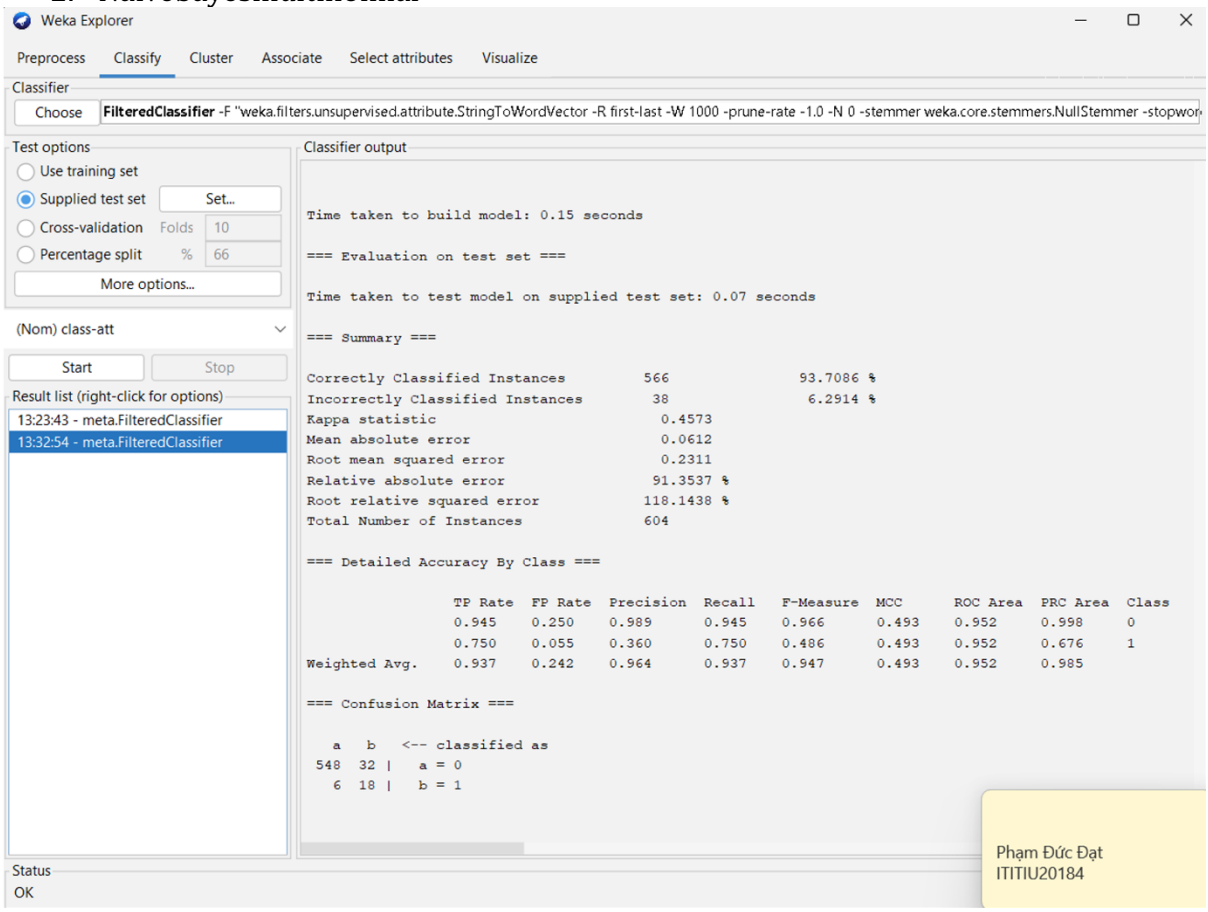
1. J48



- Correctly Classified Instances: 588 out of 604
- Percentage of Correct Classifications: 97.351%
- Incorrectly Classified Instances: 16 out of 604
- AUC (ROC Area): 0.694

⇨ Higher accuracy with 97.351% correctly classified instances.
⇨ Lower AUC (ROC Area) of 0.694.

2. Naivebayesmultinomial



- Correctly Classified Instances: 566 out of 604
- Percentage of Correct Classifications: 93.7086%
- Incorrectly Classified Instances: 38 out of 604
- AUC (ROC Area): 0.952

⇨ Lower accuracy with 93.7086% correctly classified instances.

⇨ Higher AUC (ROC Area) of 0.952.

Question 2:

a. Based on the formulas provided in Table 5.7, the best possible values for each output statistic are:

1. Lift Chart:
   - Best Possible Value: The highest lift value occurs when all true positives are ranked first, so the lift would be significantly greater than 1. Ideally, it can approach infinity if all positives are in a small subset of the data, making the denominator (subset size) very small.
   - Condition for Attainment: This occurs when the classifier perfectly ranks all true positives before any false positives or true negatives. This scenario is ideal and unlikely to occur in real-world data.
2. ROC Curve:
   - True Positive Rate (TP Rate or Sensitivity): The best possible value is 1 (100%).
   - Condition for Attainment: This occurs when the classifier correctly identifies all true positives, with no false negatives.
   - False Positive Rate (FP Rate): The best possible value is 0 (0%).

- Condition for Attainment: This is attained when the classifier does not incorrectly classify any negatives as positives.
- Overall Best Scenario: A TP rate of 1 and an FP rate of 0 represents perfect classification, where all positives are correctly identified and no negatives are mistakenly classified as positives.

3. Recall-Precision Curve:
   - Recall: The best possible value is 1 (100%).
   - Condition for Attainment: This occurs when the classifier correctly identifies all true positives, resulting in no false negatives.
   - Precision: The best possible value is 1 (100%).
   - Condition for Attainment: This is achieved when all instances classified as positive are true positives, meaning there are no false positives.
   - Overall Best Scenario: Both recall and precision are 1, indicating perfect performance where the classifier correctly identifies all positive instances without any errors.

b. Conditions for These Best Values:

- These ideal values are typically attained in scenarios where the classifier perfectly distinguishes between positive and negative instances, which is a rare occurrence in practical applications. These perfect values are most likely in highly controlled datasets or synthetic data designed to be easily separable. In real-world scenarios, there is usually some trade-off between these values, depending on the nature of the data and the classifier's ability to discriminate between classes.

Question 3:

The classifier “FilteredClassifier with StringToWordVector using J48” shows the best performance in both datasets

- for Corn dataset: classification accuracy is 97.35
- for Grain dataset: classification accuracy is 96.36

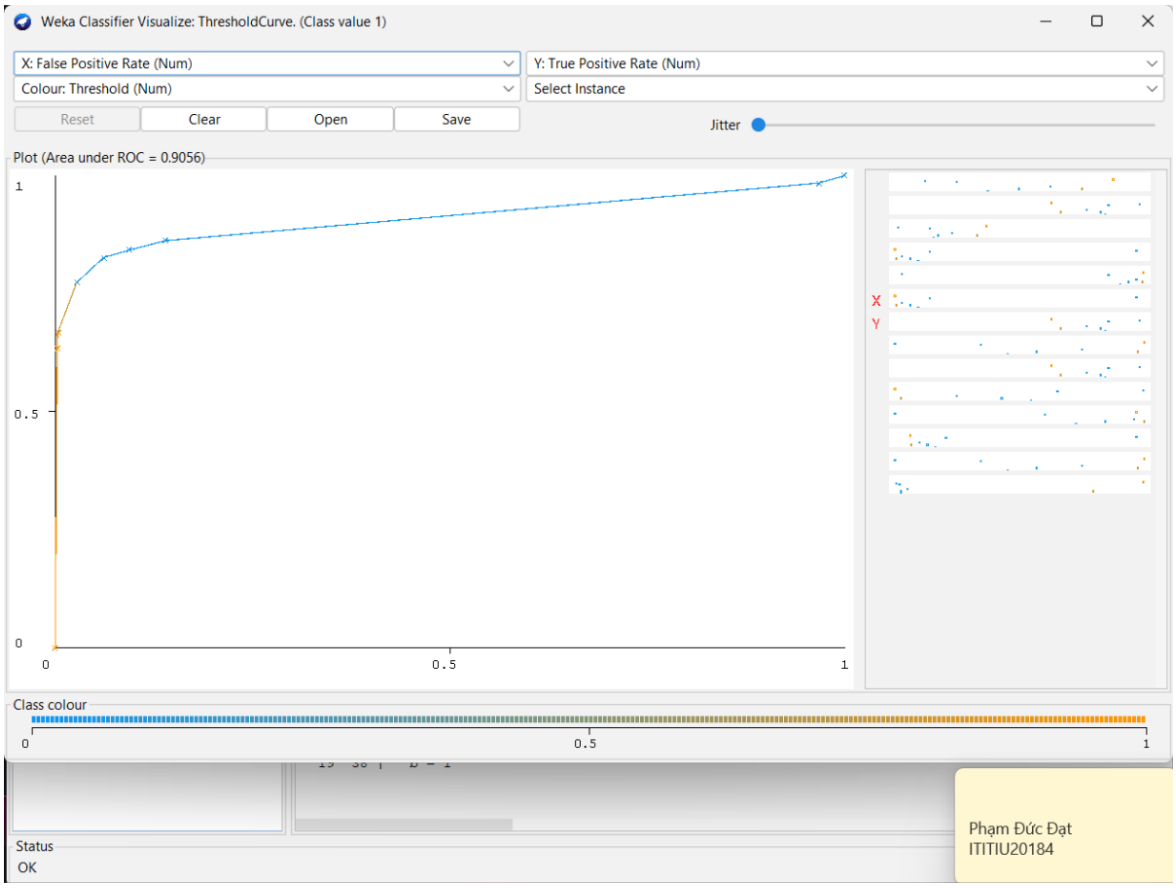
While the classifier “FilteredClassifier with StringToWordVector using NaiveBayesMultinomial” shows a little worse performance in both datasets:

- for Corn dataset: classification accuracy is 93.71
- for Grain dataset: classification accuracy is 90.73

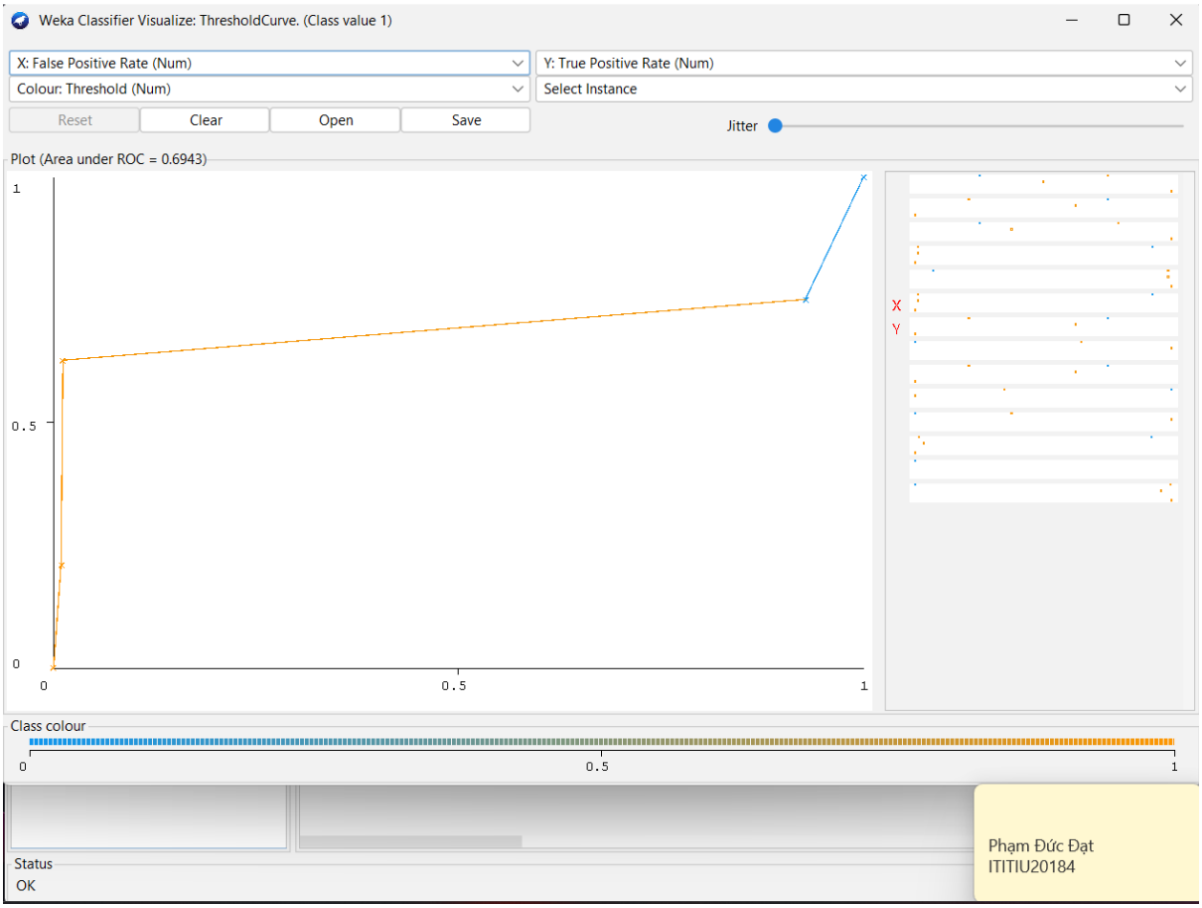
Meaning:

- Model Selection: The results suggest that J48 is a better model for these particular datasets, likely due to its ability to handle complex, non-linear relationships between words. This might make it more suitable for other similar text classification tasks where word interactions are significant.
- NaiveBayesMultinomial Considerations: Despite being outperformed by J48, NaiveBayesMultinomial is still a valuable model, especially for large text datasets or when interpretability and speed are crucial. Its lower accuracy might be due to its inability to model dependencies between words effectively in these datasets.

1. ROC curve of best classifier for class 1of Reuter Grain:



2. ROC curve of best classifier for class 1 of Reuter Corn:



Question 4:

Most extreme difference dataset:

1. For Corn Dataset:

- J48 Accuracy: 97.35%
- NaiveBayesMultinomial Accuracy: 93.71%

⇨ Difference: 97.35% - 93.71% = 3.64%

2. For Grain Dataset:

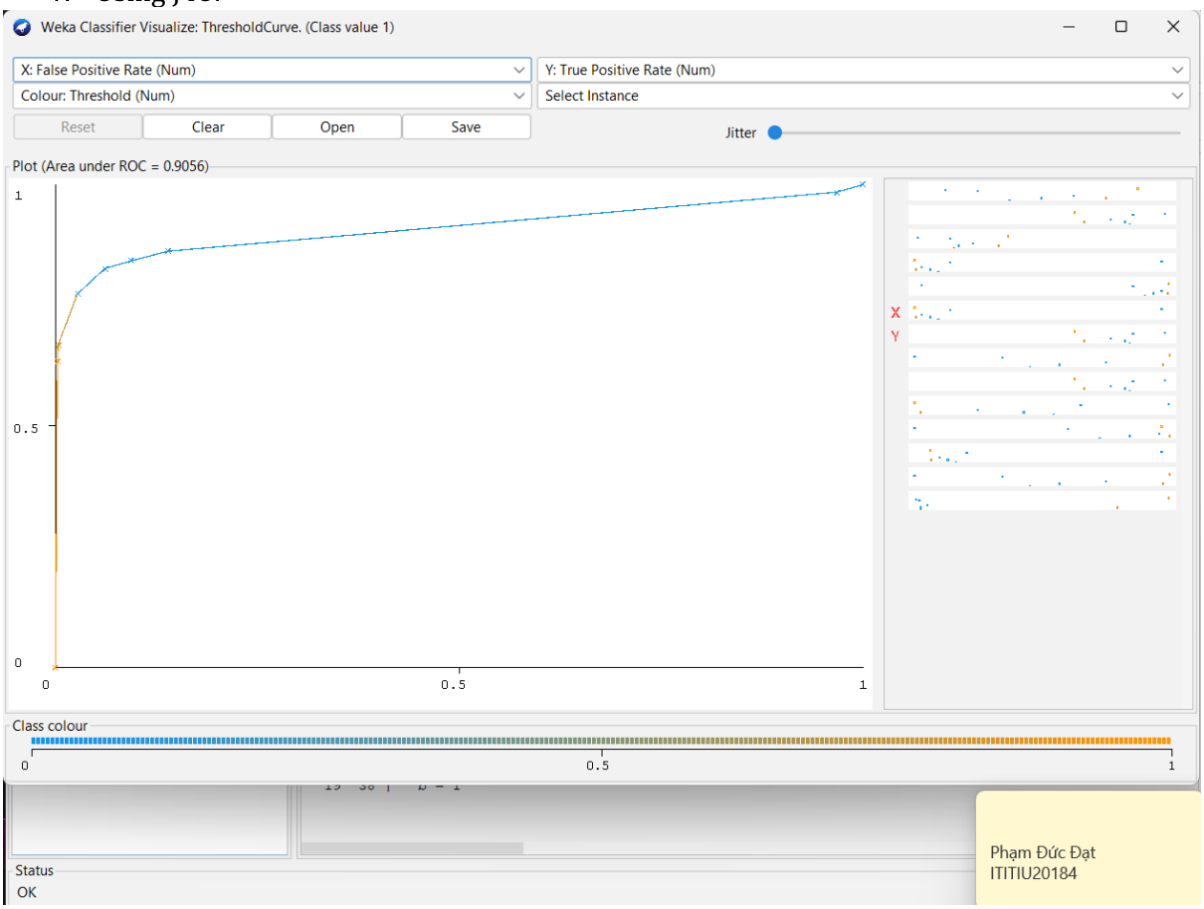
- J48 Accuracy: 96.36%
- NaiveBayesMultinomial Accuracy: 90.73%

⇨ Difference: 96.36% - 90.73% = 5.63%

The Reuters dataset that produced the most extreme difference in classification accuracy between the two classifiers is the Grain dataset.

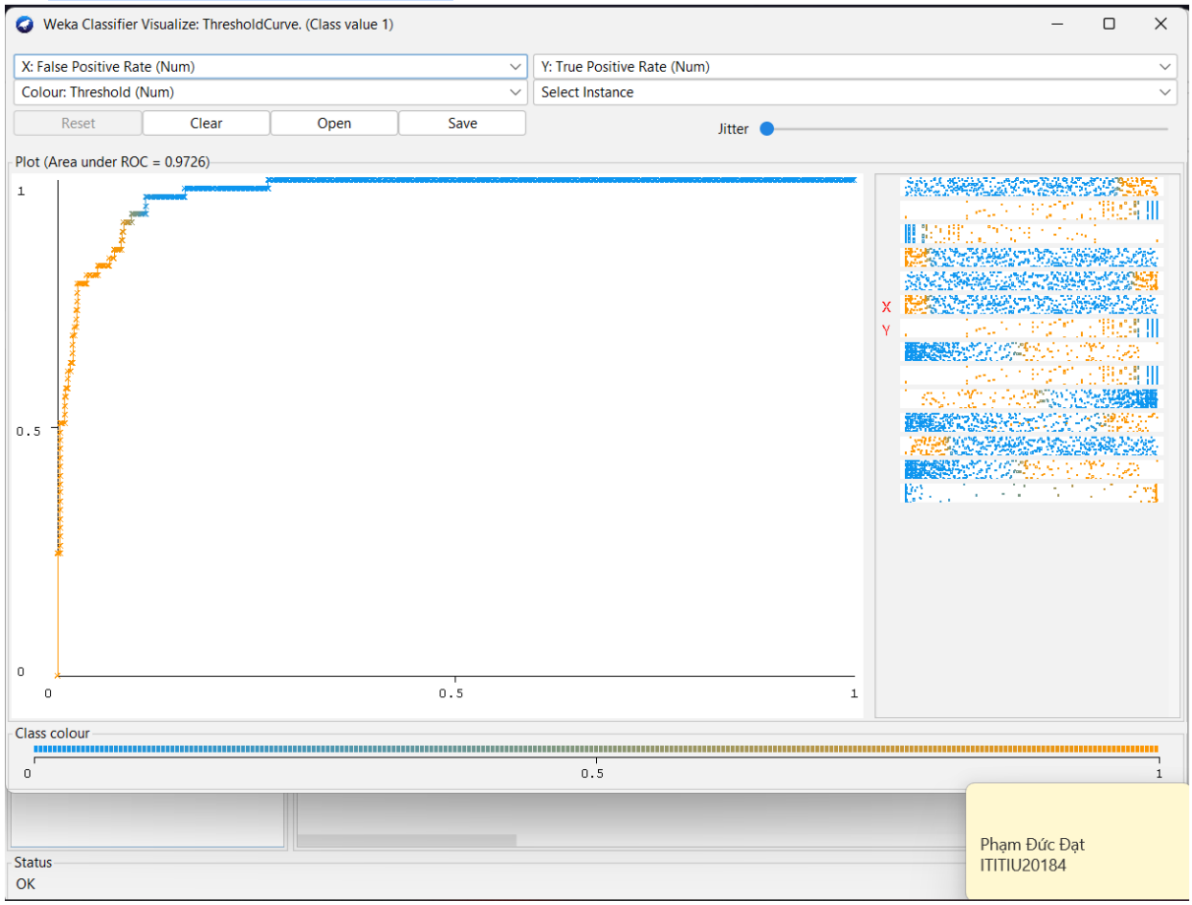
ROC curve for class 1:

1. Using J48:



- Area Under Curve: 0.9056
- Explanation: The ROC curve for J48 shows a reasonably high area under the curve (AUC), which is 0.9056. This indicates that the J48 classifier performs well in distinguishing between positive (class 1) and negative instances. However, while the curve shows good separation, there might be some cases where the classifier does not perfectly rank all instances correctly, as indicated by an AUC less than 1.

2. Using NaiveBayesMultinomia



- Area Under Curve: 0.9056
- Explanation: The ROC curve for NaiveBayesMultinomial is even closer to the ideal case with an AUC of 0.9726. This higher AUC indicates that NaiveBayesMultinomial is more effective at ranking positive instances higher than negative ones, meaning it has a better overall capability to distinguish between the two classes. The curve almost reaches the top-left corner, suggesting near-perfect performance in this ranking task. **The higher AUC indicates better performance in distinguishing between classes.**

Question 5:

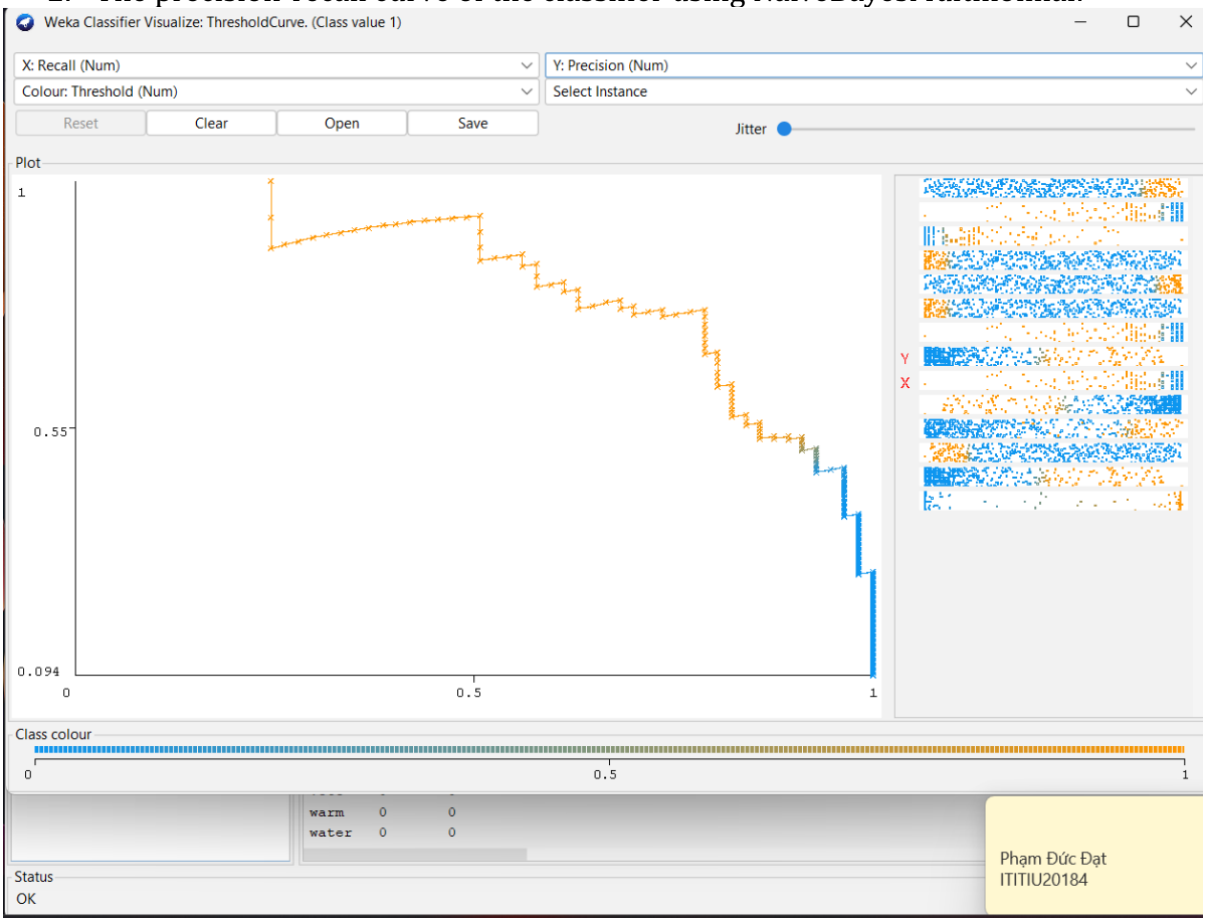
An ideal ROC curve for perfect performance would rise vertically from the origin (0,0) to the top-left corner (0,1) and then move horizontally to the top-right corner (1,1). This would indicate a true positive rate of 1 and a false positive rate of 0, resulting in an area under the curve (AUC) of 1.0. When two curves don't overlap at all means model has an ideal measure of separability. It is perfectly able to distinguish between positive class and negative class.

Question 6:

1. The ideal precision-recall curve for perfect performance:

The ideal precision-recall curve for perfect performance would be a horizontal line at a precision of 1 across all recall values until recall reaches 1. This indicates that the model perfectly identifies all positive instances without any false positives.

2. The precision-recall curve of the classifier using NaiveBayesMultinomial:



This precision-recall curve indicates good performance, with high precision at various levels of recall. This complements the high AUC of the ROC curve, suggesting the model performs well overall, especially in scenarios where class imbalance might be a concern.

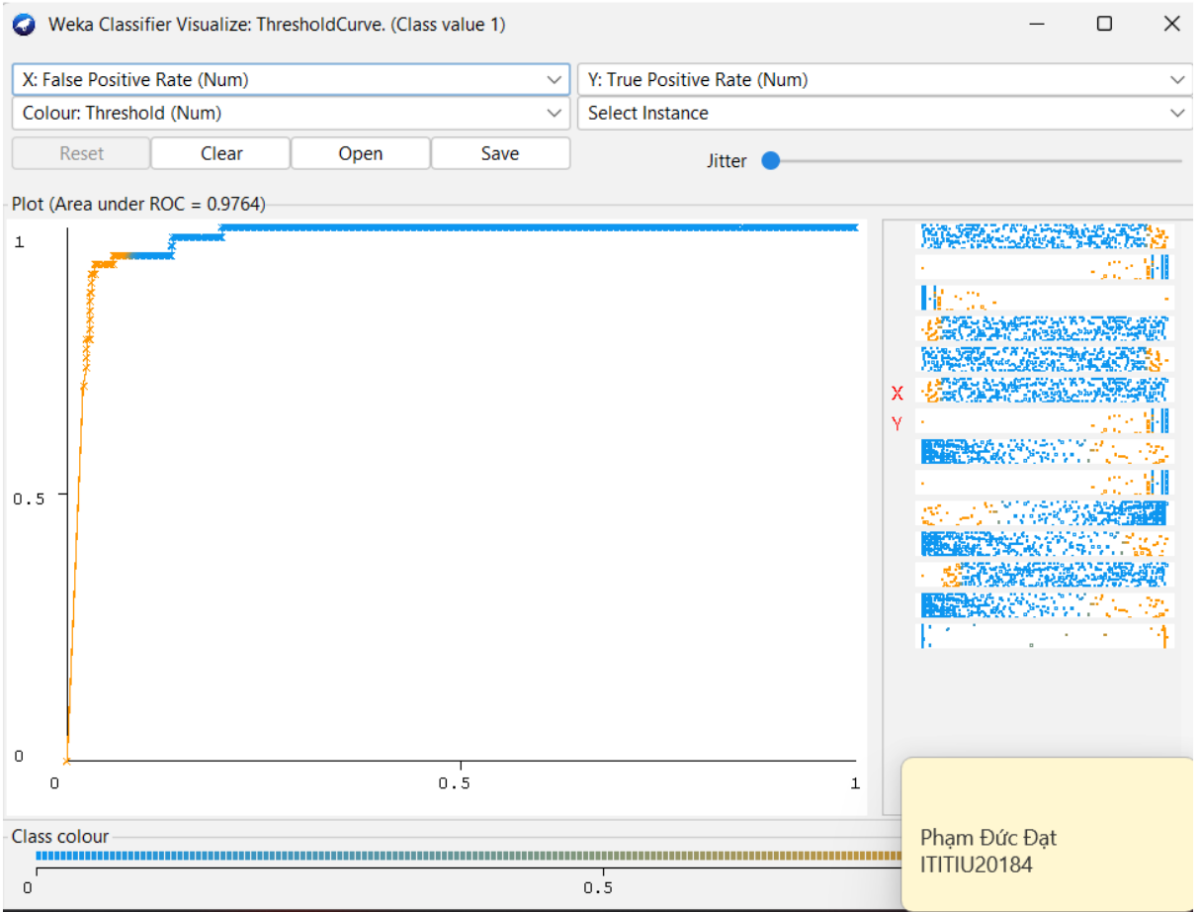
Question 7:

1. **Improved ROC curve for class 1 of Grain dataset:**

- Set IDFTransform and TFTransform attributes to be True in StringToWordVector filter
- Classifier: Weka's AttributeSelectedClassifier, using ranking with ClassifierAttributeEval and the Ranker search, the classifier is FilteredClassifier with StringToWordVector using NaiveBayesMultinomial.

Comment:

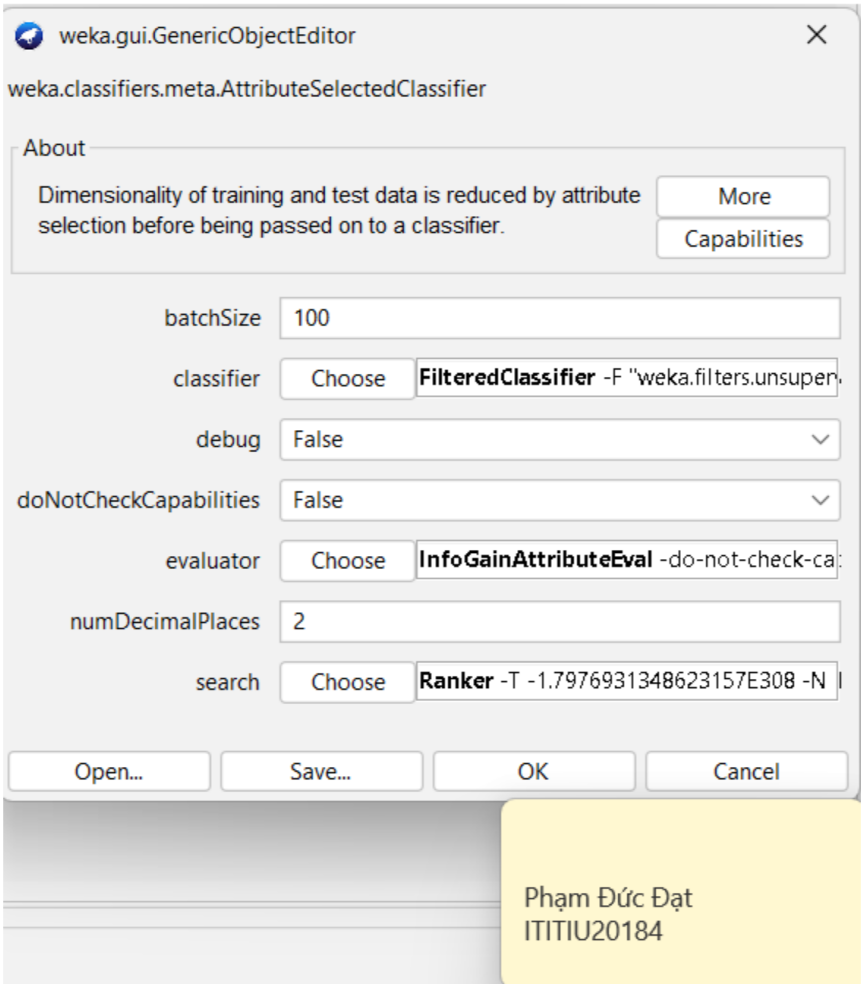
- The Area Under the Curve (AUC) improved to **0.9764**.
- This is an improvement compared to the previous AUC of **0.9726** obtained without applying these transformations.

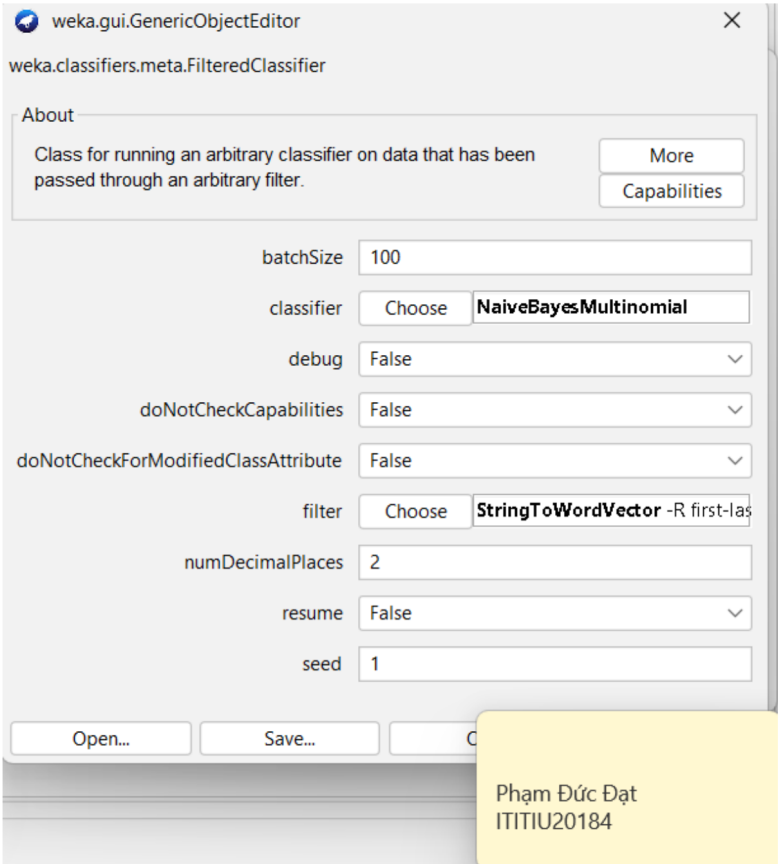


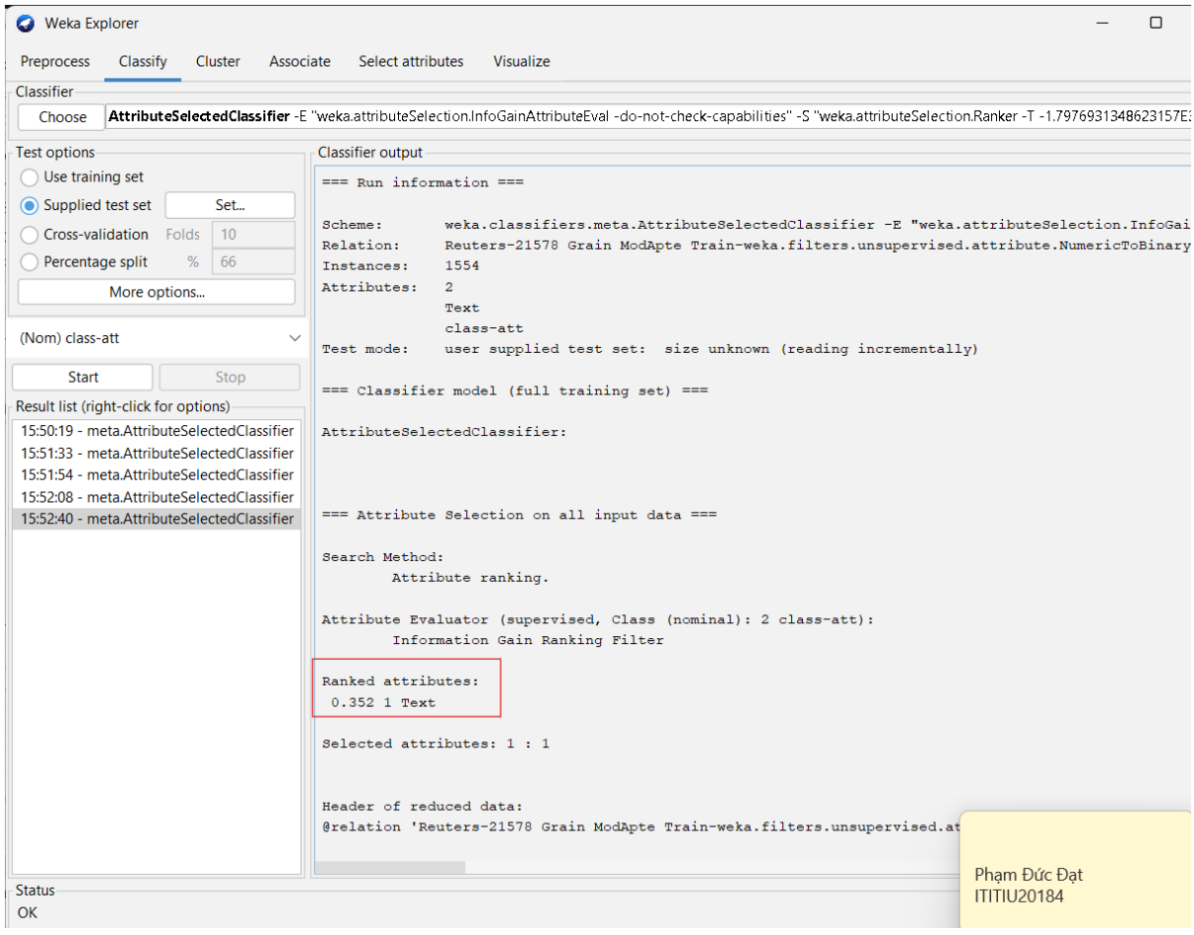
2. **Improved ROC curve for class 1 of Corn dataset:**

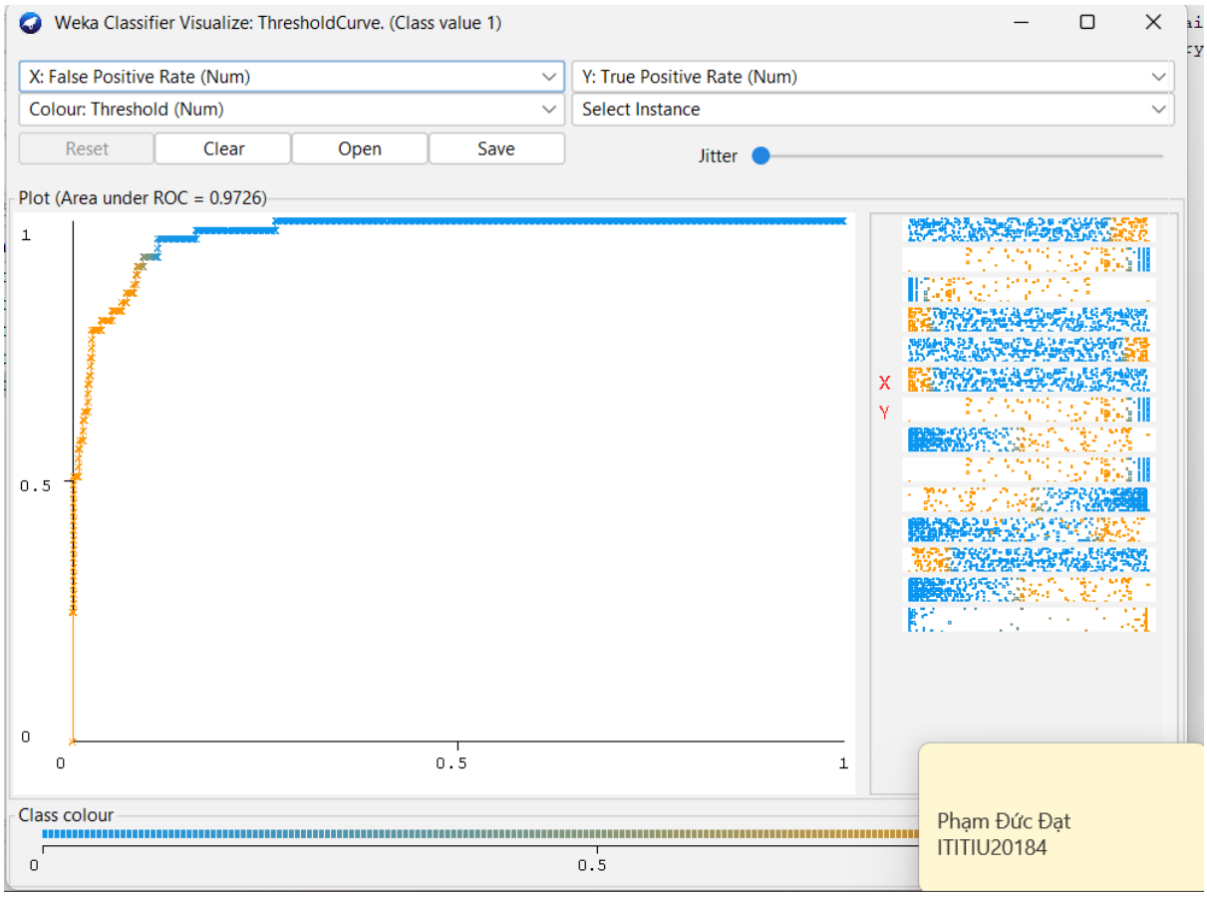
- Using default options for StringToWordVector and NaiveBayesMultinomial as the classifier returns the best performance.

Question 8:


weka.gui.GenericObjectEditor
weka.classifiers.meta.AttributeSelectedClassifier
About
Dimensionality of training and test data is reduced by attribute selection before being passed on to a classifier.
More
Capabilities
batchSize
100
classifier
Choose
FilteredClassifier -F "weka.filters.unsuper
debug
False
doNotCheckCapabilities
False
evaluator
Choose
InfoGainAttributeEval -do-not-check-ca
numDecimalPlaces
2
search
Choose
Ranker -T -1.7976931348623157E308 -N
Open...
Save...
OK
Cancel
Phạm Đức Đạt
ITITIU20184


weka.gui.GenericObjectEditor
weka.classifiers.meta.FilteredClassifier
About
Class for running an arbitrary classifier on data that has been passed through an arbitrary filter.
More
Capabilities
batchSize
100
classifier
Choose
NaiveBayesMultinomial
debug
False
doNotCheckCapabilities
False
doNotCheckForModifiedClassAttribute
False
filter
Choose
StringToWordVector -R first-las
numDecimalPlaces
2
resume
False
seed
1
Open...
Save...
Phạm Đức Đạt
ITITIU20184


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
AttributeSelectedClassifier -E "weka.attributeSelection.InfoGainAttributeEval -do-not-check-capabilities" -S "weka.attributeSelection.Ranker -T -1.7976931348623157E
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class-att
Start
Stop
Result list (right-click for options)
15:50:19 - meta.AttributeSelectedClassifier
15:51:33 - meta.AttributeSelectedClassifier
15:51:54 - meta.AttributeSelectedClassifier
15:52:08 - meta.AttributeSelectedClassifier
15:52:40 - meta.AttributeSelectedClassifier
Classifier output
=== Run information ===
Scheme: weka.classifiers.meta.AttributeSelectedClassifier -E "weka.attributeSelection.InfoGai
Relation: Reuters-21578 Grain ModApte Train-weka.filters.unsupervised.attribute.NumericToBinary
Instances: 1554
Attributes: 2
Text
class-att
Test mode: user supplied test set: size unknown (reading incrementally)
=== Classifier model (full training set) ===
AttributeSelectedClassifier:
=== Attribute Selection on all input data ===
Search Method:
Attribute ranking.
Attribute Evaluator (supervised, Class (nominal): 2 class-att):
Information Gain Ranking Filter
Ranked attributes:
0.352 1 Text
Selected attributes: 1 : 1
Header of reduced data:
@relation 'Reuters-21578 Grain ModApte Train-weka.filters.unsupervised.at
Status
OK
Phạm Đức Đạt
ITITIU20184



- The AttributeSelectedClassifier using InfoGainAttributeEval to rank attributes and varied the numToSelect field in the Ranker.
- The highest AUC value obtained was 0.9726, indicating that the optimal number of selected attributes provided the best classification performance with the NaiveBayesMultinomial classifier and StringToWordVector.
- The exact optimal number of attributes would be the value of numToSelect that resulted in this AUC.