COMP47490 – Assignment 1

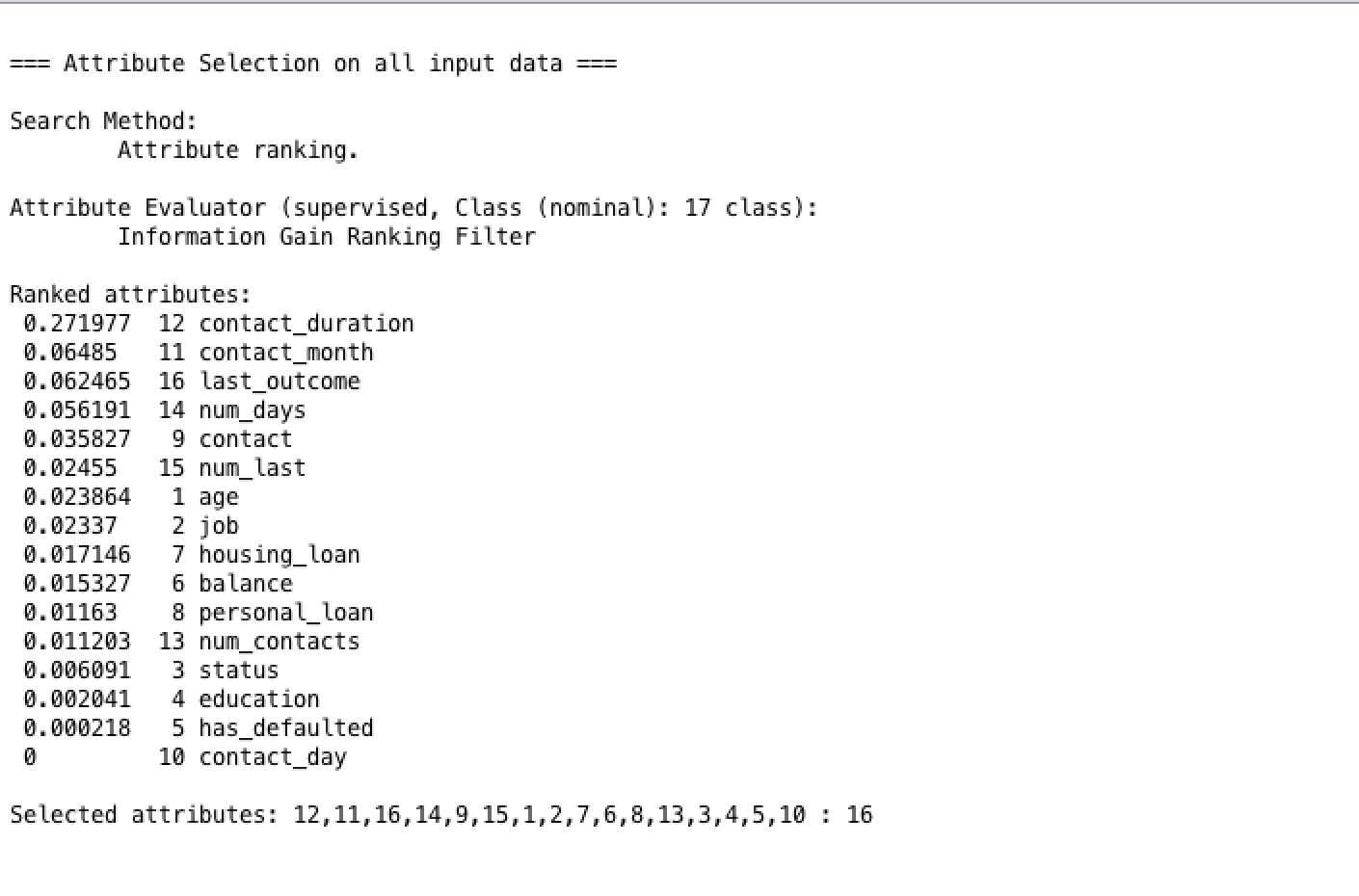
Viacheslav Borodin

15464068

Question 1)- Apply one filter and wrapper feature selection from those available in weka.

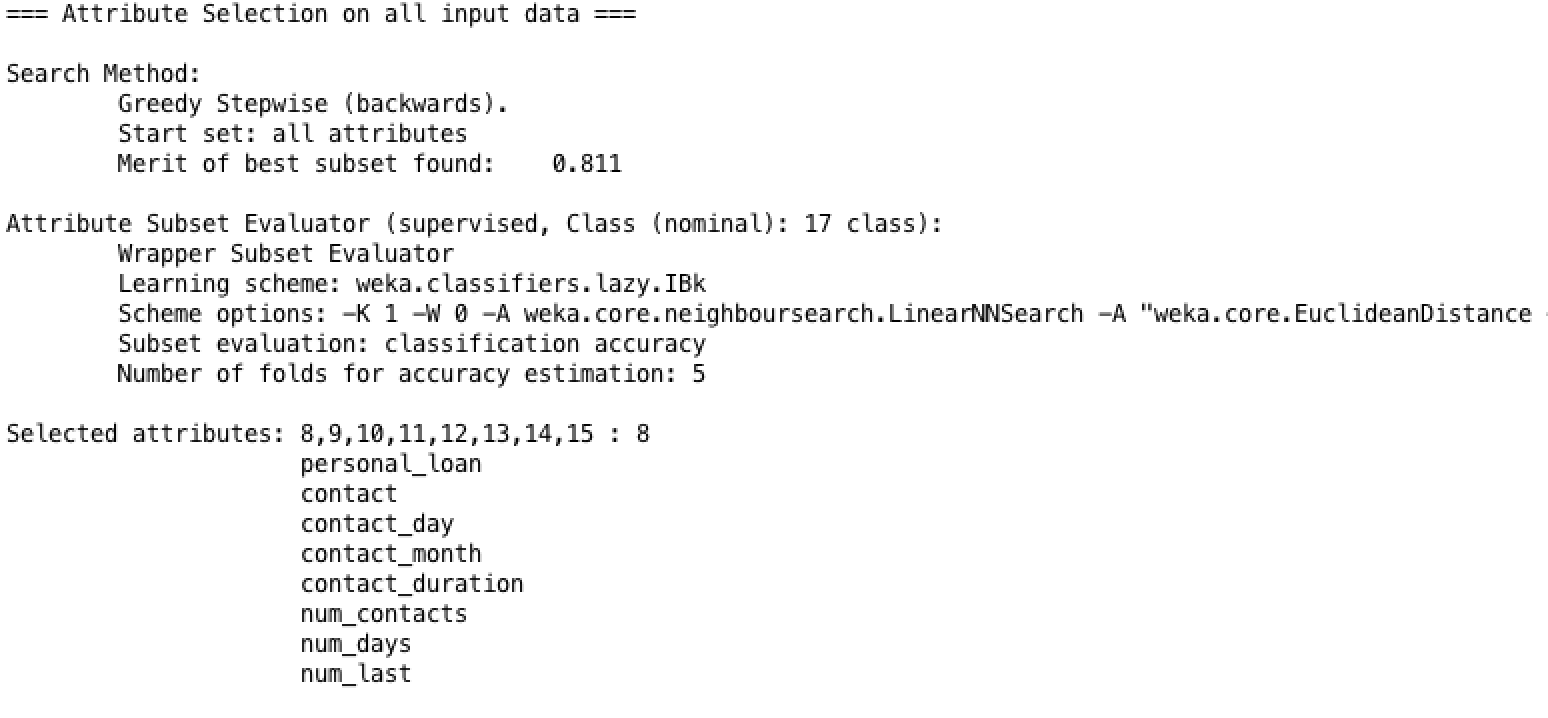
Filter feature selection:

Assign the attribute selection InfoGainAttributeEval as the Evaluator and the ranker as the search method. This will return the attribute ranking as shown below:



The feature technique selects all of the attributes going in descending order beginning with the highest ranked attribute to the lowest ranked attribute. In order to generate my result however I chose only the top 8 ranked attributes which were: 12,11,16,14,9,15,1,2. I used the ranker to select the attributes in order of their rank. However, seeing as the ranker returns the whole list of attributes, I only chose the top 8 attributes as there is a total of 16 attributes so I decided to choose only the top 8 of them. Also, another factor as to why I choose 8 attributes is because in the next section the wrapper returned me 8 attributes and it would be a fair comparison if the number of attributes is the same.

Wrapper feature selection: Assign the evaluator as the WrapperSubsetEval and the method to GreedyStepWise. In the search method of GreedyStepWise I chose the searchBackwards method to be true. The attributes that were selected are as follows:

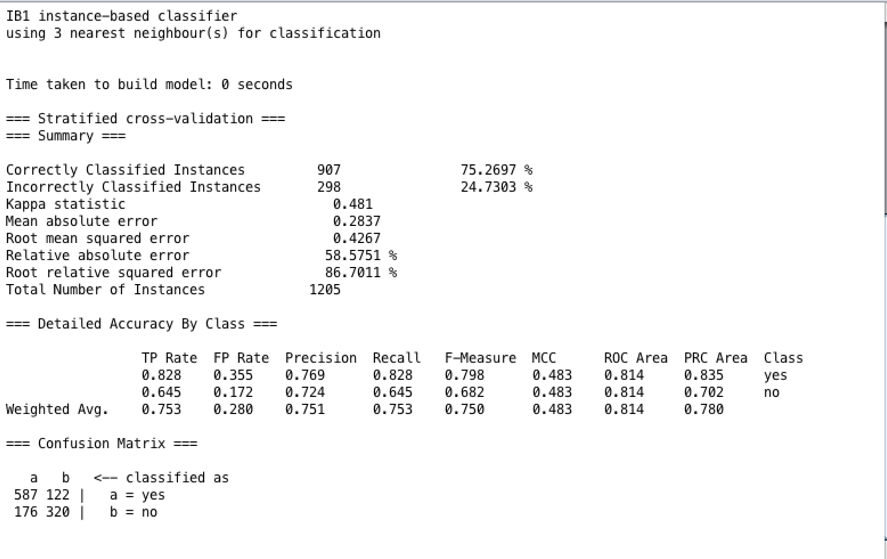


(b)- Report and discuss the differences between the feature subsets produced by the filter and wrapper techniques.

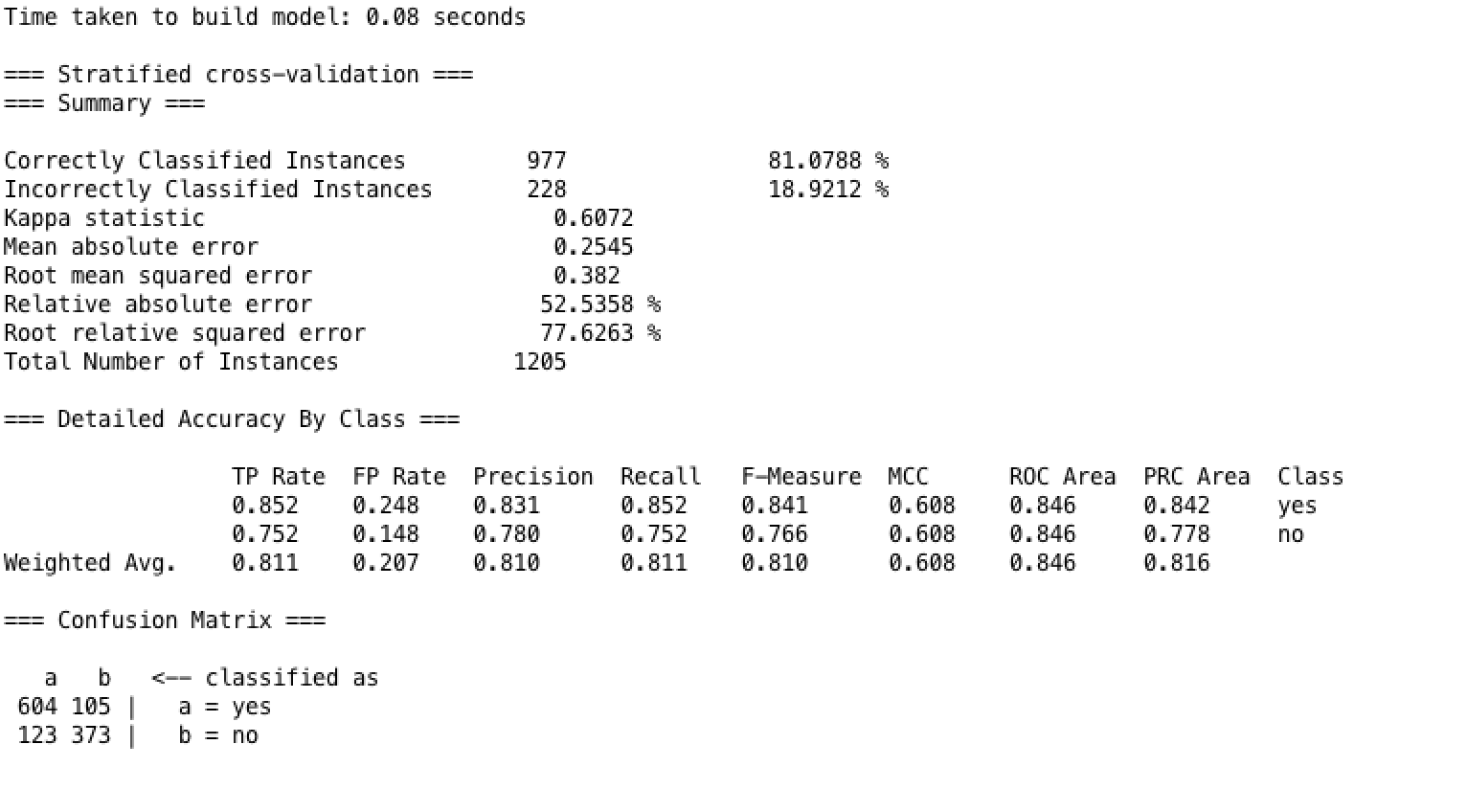
For my filter technique I ranked the attributes in their highest order and then only choose the top 8 of them as these were the top 50% of them. For the wrapper technique the method searchbackwards was set to true and again it returned me a list of 8 attributes. The attributes from both can be seen above and compares. For the filter the attributes that were chosen were job, age, contact and then every other attribute that was chosen was to do with the contact that the advertising campaign made with the customer and how long ago this was done. For the wrapper technique the attributes that were selected were all to do with the contact the campaign made with the customer rather than including personal details such as job and age. As we can see from the attribute values above that the attributes are pretty much identical to each other except for three attributes in the filter which are age, job and last\_outcome.

(c)- Evaluate and discuss performance of both of the above feature selection techniques when each is combined with two different classifiers of your choice.

**Using 3 nearest neighbour for classification:**

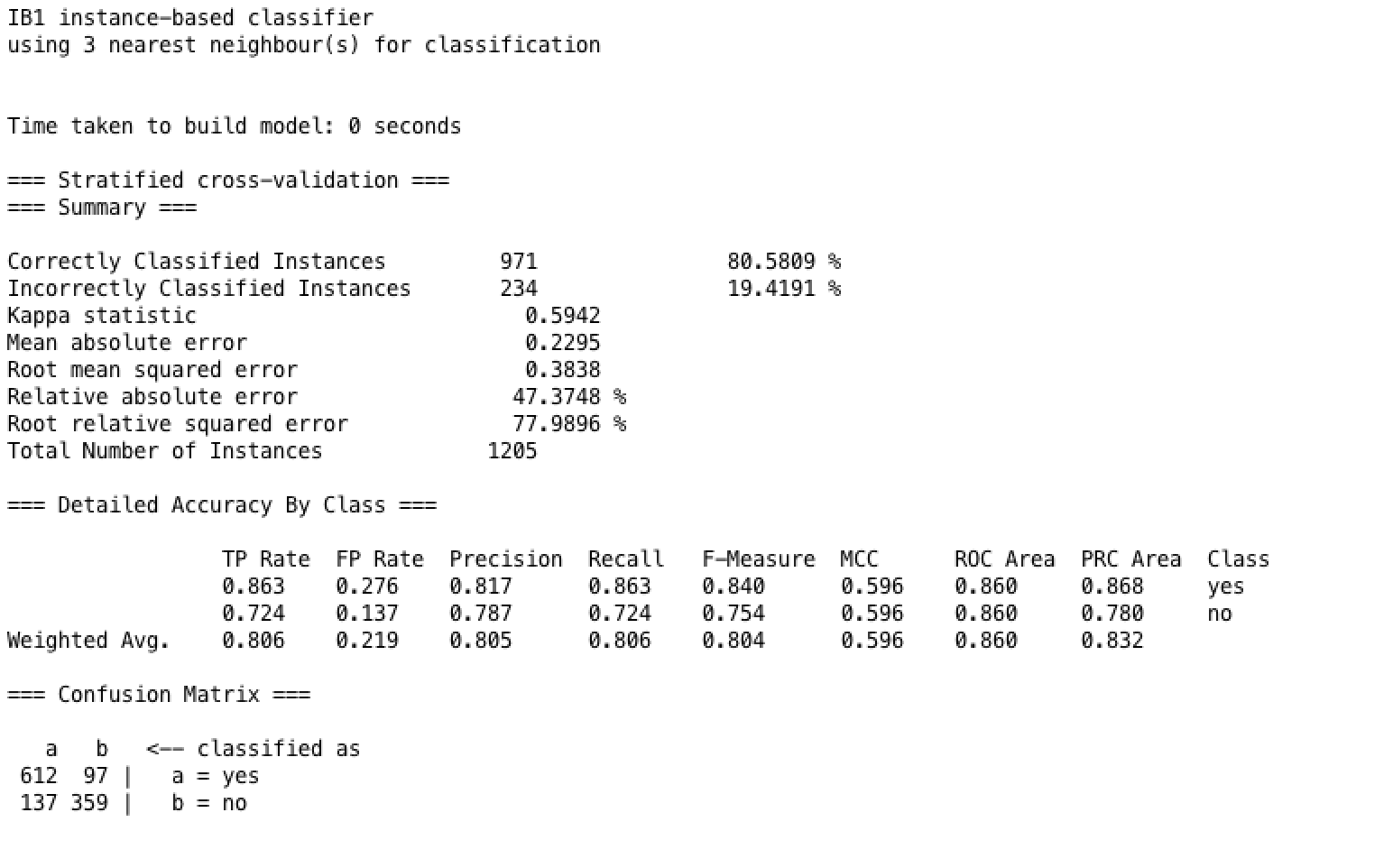


**Using Decision tree J48 for classification:**

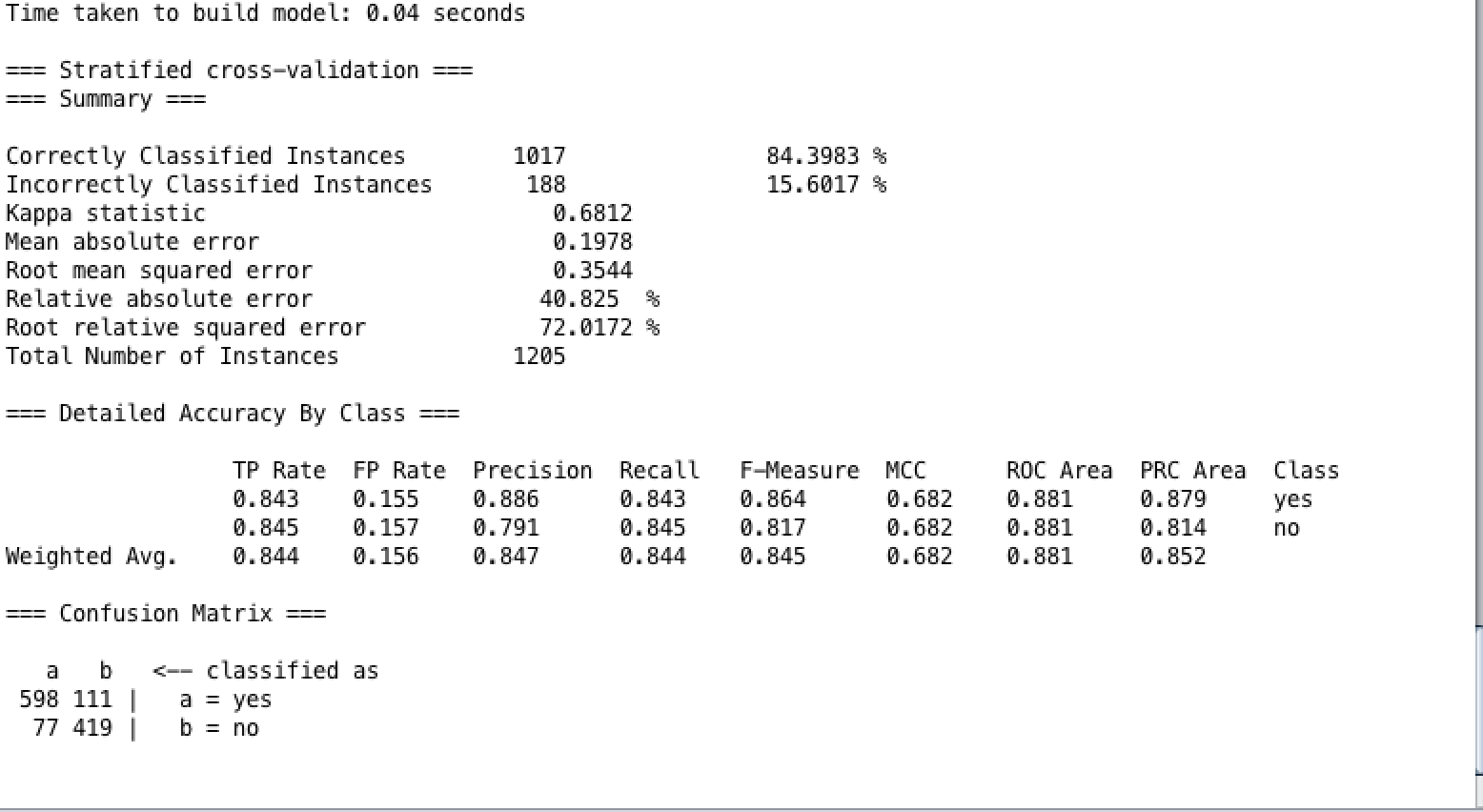
****

Wrapper:

**Using 3 nearest neighbour:**

****

**Using Decision tree J48 for classification:**



**Comparison of KNN classifier:** For this question I used the KNN to set it to 3 nearest neighbours. As we can see from the results that were produced from the KNN classifier the wrapper seems to have a higher accuracy as the Correctly classified instances is 80.5809% in comparison to 75.2697% using the filter. This indicates that the wrapper has a 81% accuracy rate in comparison to a 75% accuracy rate of the filter subset. The result for the filter technique tells us that there are 122 false positives and 176 false negatives as seen in the matrix at the very bottom of the result. The result for the wrapper indicates to us that there are 97 false positives and 137 false negatives. A false positive is the prediction that a customer would subscribe to the new service that is being offered by the bank but didn’t end up subscribing. A false negative would tell us the customer wouldn’t subscribe to the new service offered by the bank but ended up subscribing. So, after analysing these results we can see if can see that the filter technique had a higher accuracy of people who would subscribe to the new service even though not anticipated to do so initially. However, the accuracy rating is higher for the wrapper technique.

**Comparison of J48 Classifier:** For this question I used the decision tree J48 from the tree classifier in order to compare the results of the filter against the results of the wrapper. First of all, I will have a look at the accuracy of each. The accuracy of the filter decision tree is 81.0788 whereas the accuracy of the wrapper decision is 84.3983. So, the accuracy is 81 vs 84 percent. Indicating that the wrapper technique is more accurate. The matrix again can show us which combination produces more accurate results. The wrapper has 111 false positives and 77 false negatives, and the filter has 105 false positives and 123 false negatives. In this analyzation the filter has a higher accuracy of people subscribing to the service but initially not anticipated to do so. If we have a look at the average mean error the error is higher in the filter technique with it being 0.2545 in comparison to 0.1978 in the wrapper technique. The absolute error for the filter technique is also higher meaning that the filter seems to have higher errors in comparison to the wrapper techniques.

After comparing the results of both the filter and the wrapper techniques I believe the combination of the wrapper and the decision tree is most suitable for this dataset as this has the highest accuracy rate in comparison to other combinations. The second-best combination is of the wrapper and the KNN classifier which also shows a higher accuracy rate than the filter techniques. Therefore, the wrapper with decision tree is my choice of combination for this dataset along with wrapper and KNN.

Question 2)-

(a)  Explain what is meant by *overfitting* in the context of classification. Why is overfitting considered to be a problem? Briefly explain some techniques you might use to avoid or mitigate overfitting.

Overfitting is a problem when the model is fitted too closely to the training data. Overfitting happens when the model learns the detail and the noise of the training data to the point that it negatively impacts the performance of the new data. The model isn’t able to differentiate between the noise that is picked up in the training data and these situations don’t apply to the new data meaning that the model isn’t able to generalise the situations that aren’t presented in the training data.

Some techniques that may help to avoid overfitting are as follows: To help avoid overfitting one way is to gather a larger training data set. Sometimes in order to differentiate from the training data it might be useful to have quite a large sample in order to have accurate results, however this may not work all the time. Another way to stop overfitting would be to have a cross validation technique to stop overfitting. The idea behind this is to use the initial training data then divide this data into k-folds. After dividing the data into k-folds then the holdout method is repeated k times. Each time the k-subset is used as the validation set and the k-1 is used to form another training set. Another way is to terminate the data early on. This happens when the model starts to learn the distribution of data and then starts to overfit. By recognising when the data starts to deviate and shift it is possible to stop the learning process before the overfitting happens.

(b)-Explain the difference between a *test set* and a *validation set*.

Test set : A test set is used in order to test a program after it has been trained on in the initial training set.

Validation set: The validation set is used in order to avoid overfitting. This dataset is used after the training dataset and is used to test the data and to determine if the algorithm can identify new examples and values.

(c)- Describe the F-measure used in the context of evaluating classifier performance.

The F measure is a measure of the harmonic mean between performance and a recall. It is an explanation of how precise the classifier is. If the precision is high and the recall is low this will mean that the classifier is very accurate but will miss instances that are difficult to classify. For the best results the performance and the recall should be more or less equal.

(d)- Explain the differences between feature selection and feature transformation approaches for dimension reduction.

Feature transformation: This is the transformation of the data to improve results. This is when the original data stored in the dataset get transformed to a new smaller set that will hold as much information as possible. An example of feature transformation would be Principal Component Analysis. This is an unsupervised algorithm which creates linear combinations of the original features.

Feature selection : This is a selection of features which must be a subset of the original set of features. Feature selection may be used for shorter training times, simplify models so that they are easier to understand, to reduce overfitting and also to improve accuracy. An example of feature selection is Pearson’s correlation to measure the linear dependence between two variables.

(e)- Explain the use of entropy and information gain in the decision tree model.

Entropy : The entropy is a measure of the uncertainty of the probability distribution around types of information. The entropy is a measure of randomness. The entropy shows us the uncertainty about the decisions from a tree. If the tree nodes have a high level of uncertainty this would also return a higher entropy in comparison to a node that has a low level of uncertainty have a lower level of uncertainty. If all the examples have the same label this will provide an entropy value of 0. The more uncertain we are the higher the entropy will be.

Information gain : gives a measure of how much information a feature gives and about a class. It also tells us how important an attribute is.

(f)- Suggest how we might choose an appropriate value for the parameter k when building a KNN classifier.

In order to choose a parameter k for a kNN classifier it is worth noting the size of the k value. If the value k is small this will mean that there will be a higher chance of noise issues in the final result. If the value is larger this will be more expensive however it will help solve the noise problem within the classifier. In order to choose the optimal K value for the KNN classifier it is best to use a cross validation technique.