Name: Shaunak Oke
Student no:18204307

Question 1)

a)

I used information gain filter and two variation of j48 wrappersubset eval on the dataset for feature selection

1)Using information gain filter on dataset we get the following ranking of the attributes.

Ranked attributes:
0.2679985 12 contact_duration
0.0622654 16 last_outcome
0.0612088 11 contact_month
0.0530756 14 num_days
0.0393945 9 contact
0.0267725 15 num_last
0.0220689 1 age
0.0203178 2 job
0.0157931 6 balance
0.0147436 8 personal_loan
0.0131754 13 num_contacts
0.0096835 7 housing_loan
0.0071353 3 status
0.0037221 4 education
0.000019 5 has_defaulted
0 10 contact_day

Selected attributes: 12,16,11,14,9,15,1,2,6,8,13,7,3,4,5,10 : 16

From this information one can say that we can ignore contact_day and has_defaulted straight away because of their extremely low values as compared to the rest of the ranked attributes. This may not be the case for the wrapper because we are using all the data from dataset as training set for the filtering operation.

Please note(In this dataset removing any number of lower ranked attributes lowers the total accuracy but to increase speed we can select the above mentioned attributes for removal)

Measuring attributes viability solely on their accuracy: The accuracy keeps on decreasing as we remove attributes. After additional removal of the education attribute the accuracy is highest for the filter.(this peak is still lower than the accuracy of the classifier on all the attributes)
For further filtering we can also filter out more attributes dependent on size or speed we want at the cost of accuracy and losing some information about relations between some attributes.

2)Using different wrappers

Using J48 trees with wrapper subset evaluation and greedy stepwise forward results in:
Selected attributes: 1,5,9,10,11,12,15,16 : 8

age
has_defaulted
contact
contact_day
contact_month
contact_duration
num_last
last_outcome

Using J48 with backwards elimination results in more attributes being selected as follows:
Selected attributes: 2,3,7,8,9,10,11,12,13,15,16 : 11

job
status
housing_loan
personal_loan
contact
contact_day
contact_month
contact_duration
num_contacts
num_last
last_outcome

J48 in this scenario uses decision trees and prunes the attributes that are not necessary. This results in the useful attributes selection by building the decision tree. This tree is built based on the use of entropy of the dataset before and after it was split upon a certain attribute. Basically it tries to maximize the information gain for splitting. After the tree is built it is pruned by the algorithm to remove useless attributes.

b) The results of wrapper and filter are different significantly. For example contact_day is the least important attribute in filter method while it is considered as necessary in J48 based wrapper. Wrappers based results are different because wrappers use cross validation to estimate accuracy of the subset of attributes. Wrappers based results are different because wrappers use some type of learning algorithm and they also use some specified type of classification algorithm .Filters on the other hand simply check the relation between the given attribute and the class to determine its importance. They work on the basis of a simple evaluation function that just ranks attributes based on their class values. Wrappers based results have been built using cross validation. Filters however are used on the entire training set.

Overall the information gain filters ranking is not adequate. Selection of any attributes depending on this feature will surely lower the accuracy of evaluation on the dataset. They however are faster than the wrapper method but for this given size of data , the size is small enough to not be relevant to the speed.

Wrappers on the other hands selected the attributes better in my opinion. They however tend to be more expensive than the filter method.

c)**Using cross fold validation with 10 folds and given classifiers on filters and wrappers.**
i)**Testing with j48 classifier**
Filters: After testing we can see the efficacy of wrappers. Based on the information gain filter removing lower ranked attributes slightly decreased the accuracy of operation.

Wrappers: on the other hand selected the right attributes. Selecting attributes based on backwards elimination j48 wrapper increased accuracy of the classifier by 5%(82-87)

ii)Testing with naïve bayes classifier

Filters: Information gain filter results in no change in accuracy of naive classifer.

Wrapper: Backwards elimination J48 wrapper increased accuracy of the naïve classifier dramatically by almost 10%(78 to 87)

i**ii)Selection of Combination:**Before deciding which combination to use we need to determine the parameters of the experiment. If speed is essential or the dataset is large and we can say there are less correlations between attributes then we can use filters.

For our case however the dataset size is small so we can focus on accuracy and select the combination with the best accuracy. **Naïve bayes classifier on J48 attributes subset wrapper** results in the most accuracy (**87.4689%**) so we can select this combination for our analysis.

(question 2 next page)

Question 2

a)

Overfitting occurs when the classifier model is fitted too closely on the given training data. This leads to surprisingly accurate results on testing data but such a model underperforms in the case of any data outside the experimental data. In more general terms the model has learned the bad attributes or noise from the training set.

This leads to the model unable to predict classes accurately incase of new data. Whether model is overfit or not may not be that apparent unless tested on new data. This leads to confusion. There is also the question that till what extent is the model overfitted to the training data.

Cross validation is subjectively the best method to avoid overfitting. In this technique we simply divide the data into k(say) subsets. Each subset is then divided into k-1 subsets that are used for training and one subset for testing. At the end the accuracy of all the folds is averaged to find the total accuracy of the model. This method also has the benefit of freedom to adjust k to properly route out any overfitting of the model to the training data.

b)

Test set and validation set.
Test set: It is used for assessing the accuracy of the classifier after the model has been fully built. This set is used to assess how the model performs on different data other than training data.

Validation set: It is used to tune the parameters of the fully built model to fine tune the model to show the accurate result and remove the minor noise/distortions. It performs the final tuning the model. After this set we can say that the model is completely built. After this step we can start testing the data on actual test/real world data.

c)

F-measure is an alternative method to accuracy to determine tests accuracy. Normal accuracy give same weightage to both positive and negative labels. This might not be desirable for an information retrieval system. It is usually used to adjust precision and recall together and assess how they together produce a balanced measure for a good classifier. Generally ROC curve is plotted to assess precision vs recall.

F measure is calculated by taking the weighted harmonic mean of precision and recall.
For balanced f score
F_1=(2.precision.recall)/(precision+recall)
General formula for f measure is
$F=(1+\beta^2).Precision.recall/(\beta^2.Precison + recall)$

d)

Feature selection vs Feature transformation:
Feature selection: It is used to select a subset of useful attributes from the given features.

Usually when we start with a dataset, not all the features are useful, not all of the features contribute to the building of a proper module. These features must be found and removed for accurate and fast models. At the same time this technique tries to keep maximum amount of information while keeping the new dataset more compact and useful.
Example: Filters and wrappers.

Feature transformation: This technique on the other hand tries to convert/transform the features into more compact features. Usually old features are converted to new features based on various techniques. These features store less data than before but at the same time usually have more discriminatory power than the earlier dataset. These can be achieved by combining multiple features or transformation of one feature using normalization or similar techniques.
Example: Principal Component Analysis(PCA) ,range transformation, z score normalization.

e)
Entropy refers to the amount of disorder/uncertainty for a given source of information. In context of decision trees at the root node initially all the class labels are present together. When we split the node using some decisions and features we generate child nodes each with their own class labels. When the leaf nodes have more pure data as compared to the purity of the root node we say the entropy of the data has decreased. If however the dataset still has uncertainty the entropy stays the same or even increases.

Information gain refers to the impact of a certain feature selection .It is used to calculate how much information we get by splitting a certain feature.
It is calculated by the difference of original entropy and entropy after split(weighted on proportion of size)

More information gain implies that the feature splits the data better than the feature with lower information gain.

f)
Usually odd value of k is used for binary classification. Overall choice of k usually depends on the dataset.
We can usually test accuracy of some given k with cross fold validation.

Another technique is to test the accuracy for some given k.
For example start with k=1 then k=5 then k=49 then k=99.
If we find that say k=49 gets highest accuracy then we can test k near k=49 slowly increasing k till accuracy is maximum.
This technique is however a bit intensive if we need to test for a huge test dataset.