COMP723 Data Mining and Knowledge Engineering
Assignment 2 – Data Mining (50%)

Due Date

This assignment may be completed individually or in groups of size 2.
Due Date: 30 October 2020, at 23:59 NZ time.
Submission: A soft copy needs to be submitted through Turnitin (a link for this purpose will be set up in Blackboard). When submitting the assessment **make the name(s) and student id(s) are indicated on the front page of the report**.

AIMS

The Aim of this assignment is two-fold. Firstly, in Part A, you are required to conduct a case study analysis of a data mining application from Industry, which will provide you with an insight into the ways in which data mining is used in practice.

Part A

You need to choose a **case study of data mining application** from one of the links at the end of this assignment and produce a report that describes the following:

DELIVERABLES

- Background information on the organisation that initiated the Data Mining application.
- A brief description of the target application (e.g. detecting credit card fraud, diagnosing heart disease, etc.) and the objectives of the data mining exercise undertaken.
- A description of the data used in the mining exercise (the level of detail published here will differ due to commercial sensitivity, hence flexibility will be used in the marking of this section).
- A description of the mining tools (data mining software) used, together with an identification (no details required) of the mining algorithms and how the mining algorithms were applied on the data. This description should include data pre processing steps and/or any tuning operations performed on the mining algorithms.
- Discussion of the outcomes and benefits **(be as specific as possible, talk about accuracy of results, potential or actual savings in dollar terms or time savings; do not talk in vague, general terms**) to the organisation that resulted from the mining exercise. This discussion should contain, in addition to the published material, your own reflection on the level of success achieved by the organisation in meeting their stated aims and objectives.

The total length should not exceed 2 pages. The criteria that will be used for assessment in Part A is as follows:

Criterion	Mark
Overall Quality of Presentation	**5**
Background of Organization, Application Objectives, Description of Data	**7**
Tools and Mining algorithms (Data Pre-Processing and/or Parameter Tuning)	**8**
Outcomes, Benefits and Reflection	**10**
Total	**30**

Part B

Question 1 *(Total of 30 marks)*

In this question you will investigate two different methods of combining two types of classifiers. The dataset that you will be experimenting with is the Diabetes dataset that has been used in the lab class:

https://raw.githubusercontent.com/jbrownlee/Datasets/master/pima-indians-diabetes.data.csv

You will produce two optimized classifiers and then combine them using a method called meta classification. This will be done in 5 steps. Use Python throughout and provide labelled and referenced code snippets for each of the 5 steps. Randomly split the dataset into 70% training and 30% testing and keep the sets consistent throughout the experiments.

Step 1: (4 marks)

Use the decision tree classifier and tune the max_depth parameter. Often the default setting of 2 produces poor results due to overfitting. Vary the max_depth parameter in the range [2, 20] in steps of 2 and record the classification accuracy on the test set for each value of this parameter.
Generate a two-column table (max_depth, test accuracy).

Step 2 (4 marks)

Now use the neural network (MLP) classifier and perform tuning on the learning rate parameter. Vary the learning rate in the range of [0.001, 0.01] in increments of 0.001 and record the classification accuracy on the test set for each value of this parameter.
Generate a two-column table (learning rate, test accuracy).

Step 3 (5 marks)

Use the best value of the max_depth parameter from step 1 and perform feature selection using Python's *SelectKBest()* method with the decision tree classifier. Vary the K parameter in the range [2..7] in increments of 1 and evaluate the accuracy on the test set for each value of K.
Generate a two column table (K, test accuracy).

Step 4 (5 marks)

Use the best value of the learning rate parameter from step 2 and perform feature selection using Python's SelectKBest() method with the MLP classifier. Vary the K parameter in the range [2..7] in increments of 1 and evaluate the accuracy on the test set for each value of the learning rate.
Generate a two-column table (K, test accuracy).

Step 5 (12 marks)

In this step we will combine the two classifiers into a single classifier. To do this first look up the *sklearn* documentation on the use of the *predict_proba()* method that returns a vector of probabilities for each class for a given test sample. For the diabetes dataset, this will return a vector of size 2 as there are two classes.

a) For each test sample apply *predict_proba()* for the decision tree model produced from step 3 and select the class that gives the highest probability value.
b) Repeat this process for the MLP model produced from step 4.

The two models are combined by classifying each test instance into the class that produced the highest probability from steps 5 a) and 5 b).
Generate the average classification accuracy on the test set that is produced by combining the two classifiers.

Question 2 *(Total of 40 marks)*

In this question you will explore different architectures for building a neural network. Once again you will use the Diabetes dataset.

1) Use the *sklearn.MLPClassifier* with default values for parameters and a single hidden layer with k=20 neurons. Use default values for all parameters other than the number of iterations which should be set to 150. Also, as is standard for an MLP classifier, we will assume a fully connected topology, that is, every neuron in a layer is connected to every other neuron in the next layer.

   Record the classification accuracy as it will be used as a baseline for comparison in later parts of this question. **(5 marks)**

2) We will now experiment with two hidden layers and experimentally determine the split of the number of neurons across each of the two layers that gives the highest classification accuracy. In part 1 of the question we had all k neurons in a single layer. In this part we will transfer neurons from the first hidden layer to the second

iteratively in step size of 1. Thus for example in the first iteration, the first hidden layer will have k-1 neurons whilst the second layer will have 1, in the second iteration k-2 neurons will be in the first layer with 2 in the second and so on. Summarise your classification accuracy results in a 20 by 2 table with the first column specifying the combination of neurons used (e.g. 17, 3) and the second column specifying the classification accuracy. **(7 marks)**

3) From the table created in part 2 of this question you will observe a variation in accuracy with the split of neurons across the two layers. Give explanations for some possible reasons for this variation. **(8 marks)**

4) By now you must be curious to see whether the trends that you noted in step 3 above hold true for other datasets as well. Use the car evaluation dataset from (https://archive.ics.uci.edu/ml/datasets/Car+Evaluation) to carry out the same experiment and generate the same table as in part 2 of this question. Comment on your results from this section. **(10 marks)**

5) Give reasons for any difference in trends between the tables you produced in parts 2 and 4 of this question. **(10 marks)**

**********End of Assignment Specification**********

Case Study Links

https://ieeexplore.ieee.org/document/806743
https://www.researchgate.net/journal/0167-9236_Decision_Support_Systems
https://www.researchgate.net/publication/221178905_Data_Mining_in_Healthcare_Information_Systems_Case_Study_of_a_Veterans_Administration_Spinal_Cord_Injury_Population/link/00b7d5304b166ea9ab000000/download
https://www.sciencedirect.com/science/article/abs/pii/S2211973616300149
https://ieeexplore.ieee.org/document/5738710
https://ieeexplore.ieee.org/document/902557
https://www.sciencedirect.com/science/article/pii/S0167923616302020
https://ieeexplore.ieee.org/document/5172596