Assignment 2 (Total: 30 marks) (2022S2)

1. Instructions

1. Answer all the questions in this assignment.
2. Total marks: 30 marks
3. **Submission date: 15 Jan 2023 23:59**

2. How to submit the answer

Your final submission should contain 5 files. All tasks should be completed using the Python Orange Tool.

1. 123456A_question1.ows(Python Orange)
2. 123456A_question2.ows(Python Orange)
3. 123456A_question3.ows(Python Orange)
4. 123456A_question4.ows(Python Orange)
5. 123456A_question5.ows(Python Orange)
6. 123456A_explanation.docx (word file)
    a. Insert and sign off the plagiarism declaration on the first page of the answer
    b. Ensure that you indicate which question and the part that you are explaining. E.g. Question (1b)
    c. Include all the screen capture of your python orange program, performance metrics or explanations required by the question.

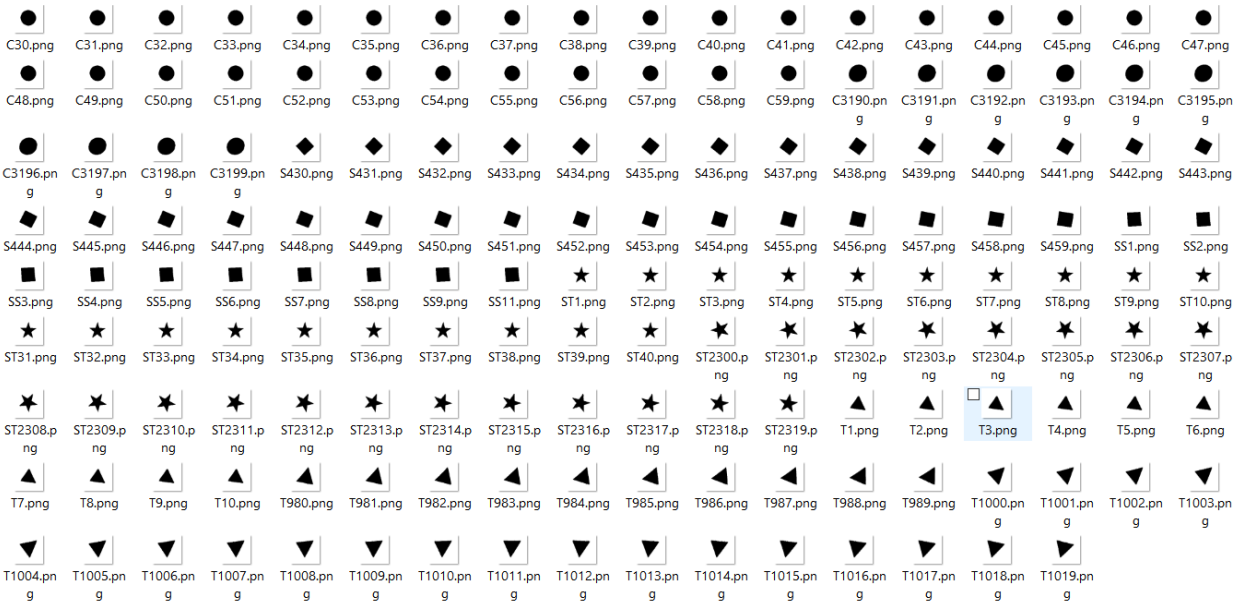
- Rename 123456A to your admin number
- Zip up the 6 files and rename the zip to your admin number.

3. Task

Question 1 Clustering of templates of different shapes (6 marks)

A template manufacturer has collected a pool of template images. Below show images of the different templates. These images are from four different types of templates produced by the production machine.

Below show images of different templates



a) Use python orange to cluster the templates into four groups. (1 mark)

b) Explain how the clustering model in section a) can group templates of similar features into the same group. (3 marks)

c)If the clustered result from a) is used to develop a machine learning classifier, what are the potential groups will be misclassified? Show and justify the problem described. (2 marks)

Question 2 Classification of Notes authenticity (5 marks)

Given a dataset (data_banknote_authentication.txt) with measurements of notes and their authenticity. Use it to train a classifier to detect real or fake notes.

0 indicates a fake note and 1 indicates a real note.

```
m1,m2,m3,m4,label
3.6216,8.6661,-2.8073,-0.44699,0
4.5459,8.1674,-2.4586,-1.4621,0
3.866,-2.6383,1.9242,0.10645,0
3.4566,9.5228,-4.0112,-3.5944,0
0.32924,-4.4552,4.5718,-0.9888,0
4.3684,9.6718,-3.9606,-3.1625,0
3.5912,3.0129,0.72888,0.56421,0
2.0922,-6.81,8.4636,-0.60216,0
3.2032,5.7588,-0.75345,-0.61251,0
```

You need to provide.

a) A fake or real note classification program using SVM and KNN as the classifier model***. Both classifiers must achieve F1 accuracy of more than 0.9 with the test dataset***. (2 marks)

b)Explain what is the purpose of the training, validation and test data (1 mark)

c)If you are given only a choice to use one of the features(m1,m2,m3,m4), which feature will you choose to give the best classification? Justify your choice. (2 marks)

Question 3 Prediction of insurance price using Regression models (8 marks)

A health insurance company has the following data about their customer and the amount of insurance paid for each customer. They want to develop a machine learning application to help customers to predict the potential insurance amount. The Company Technical Director wanted to compare Machine Learning and Deep Learning training results.

You are tasked to do a prototype using Machine Learning linear regression and Deep learning regression model. Complete the following sections.

Here is a sample of some portion of the data(insurance.csv)

age	sex	bmi	children	smoker	region	charges
19	female	27.9	0	yes	southwest	16884.9
18	male	33.77	1	no	southeast	1725.55
28	male	33	3	no	southeast	4449.46
33	male	22.705	0	no	northwest	21984.4
32	male	28.88	0	no	northwest	3866.85
31	female	25.74	0	no	southeast	3756.62
46	female	33.44	1	no	southeast	8240.5
37	female	27.74	3	no	northwest	7281.50
37	male	29.83	2	no	northeast	6406.41
60	female	25.84	0	no	northwest	28923.1
25	male	26.22	0	no	northeast	2721.32
62	female	26.29	0	yes	southeast	27808.7
23	male	34.4	0	no	southwest	1826.84
56	female	39.82	0	no	southeast	11090.7

a) With the given dataset develop a **machine learning linear regression model** predictor. Split the dataset into a train and test set. Train the model using the training dataset, then test the trained model with the test set. You should achieve R2 of at least 0.7 using the test dataset. Screen capture the result and include it in the submission document (3 marks)

b) With the given dataset develop a **deep learning linear regression model** predictor. Split the dataset into a train and test set. Train the model using the training dataset, then test the trained model with the test set. You should achieve R2 of at least 0.8 using the test dataset. Screen capture the result and include it in the submission document (3 marks)

c)Use AI Ethics fairness principle to evaluate the insurance.csv dataset. Justify and comment on the potential fairness issue(s) that could happen when we use this dataset to develop a machine learning prediction application. (2 marks)

Questions 4 Topics Classification from the text (6 Marks)

Given a topics_dataset.tab dataset

The dataset contains sentences and the corresponding labels to indicate the topic of the sentence. The labels which indicate the topic are namely 1-World, 2-Sports, 3-Business, 4-Sci/Tech. Format the text dataset into the Bag of Word with python Orange. Then perform the following.

a) Use the data to train the Logistics Regression classifiers that can classify the 4 different topics from the sentence given. The classifier should have a F1>=0.8 and AUC>=0.8 based on the test data (2 marks)

b) Explain how the text encoding method Bag of Word extract features to be used in training the machine learning model(2 marks)

c)Given the scenario. (This hypothetical question does not refer to your previous result in a)

After performing training and testing, you notice that the accuracy of the training is 97% however your testing accuracy is only 65%. Describe what is the possible cause of the results. List and explain two possible ways to improve the test accuracy. (2 marks)

Questions 5 Image Classification Problems (5 Marks)

The folder **garbage** contains 6 subfolders of different garbage images namely cardboard, glass, metal, paper, plastic and trash.

a) Use the image data in the garbage directory to train a garbage classifier with python Orange models. (2 marks)

b) The model test set **F1 score must be at least 0.8**. Screen capture the result and include it in the submission document (1 mark)

c)From the F1 score you have achieved from b), can you conclude that all the 6 classes can perform well in the prediction for new unknown images? Show and explain your conclusion given.(2marks)

==========End of Assignment Question==========