COMP1004 Assignment 1
Deadline: 1 Oct 2022 (SAT) 12 noon
Total Marks: 100

Submission instructions

- Submit a softcopy of your answer (.docx format) to blackboard before the deadline.
- Please include your student ID and name at the top of your submitted document.
- You should only provide your answers to the questions. There is no need to include the question text in your submitted document.
- Penalty for late submission: Submitted within 24 hours after the deadline (-30%); Submitted after 24 hours (0 mark).

Question 1 (33 marks)

a) [21 marks] Use the Teachable Machine (https://teachablemachine.withgoogle.com/) to train a model to predict whether you are drinking water or not drinking water. Note that you should appear in the images/videos used for training your model. Also, assume that "drinking water" is the positive class, "not drinking water" is the negative class and 50% is used as the threshold for prediction.

i) Discuss, with screenshots, how you train and test your model.

ii) Discuss, with screenshots, TWO different ways to confuse your model with false positive results.

ii) Discuss, with screenshots, ONE way to confuse your model with false negative result.

iv) Based on ii) and iii), illustrate how your model can be improved.

b) [12 marks] Answer the following questions with reference to the article "A visual introduction to machine learning" (available at http://www.r2d3.us/visual-intro-to-machine-learning-part-1/).

i) What is the target variable in the dataset?

ii) Can we improve the decision tree models by adding more layers? Explain your answer.

Question 2 (39 marks)

a) [22 marks] Watch the YouTube video "Stanford HAI 2019 Fall Conference - Keynote: The Coded Gaze with Joy Buolamwini" (available at https://www.youtube.com/watch?v=Mk5gLInf7So) to answer the following questions.

i) Describe the problems related to the performance of the facial recognition systems.
ii) What is/are the cause(s) of the problems in a(i)?
iii) How should the performance of facial recognition systems be evaluated?

b) [17 marks] Consider the following dataset for 10 customers for training a classifier using the boundary method (discussed in Lecture 2 slides) to predict whether a customer will purchase a product (1: purchase; 0: not purchase) based on the salary of the customer. Consider a model $\boldsymbol{salary + b = 0}$ (where **b** is an integer). Our goal is to maximize the classification accuracy of the training set, followed by minimizing the number of false negatives (you may assume that "purchase" is the positive class).

Customer ID	Salary (1000)	Purchase?
1	8	0
2	40	0
3	30	0
4	10	0
5	33	1
6	48	1
7	60	1
8	10	1
9	0	0
10	70	1

What should be the value of b for the best model? What is the classification accuracy, number of false negative(s) and false positive(s) for the best model?

	Your Answer
b	
Classification accuracy	
Number of false positive(s)	
Number of false negative(s)	

Question 3 (28 marks)

In this question, you should use the Orange data mining tool to predict the food type based on the Sweetness and Crunchiness of an ingredient.

a) [12 marks] Create the following dataset as training data.

	A	B	C	D
1	Ingredient	Sweetness	Crunchiness	Food Type
2	Apple	10	9	Fruit
3	Bacon	1	4	Protein
4	Banana	10	1	Fruit
5	Carot	7	10	Vegetable
6	Celery	3	10	Vegetable
7	Chesese	1	1	Protein
8	Grape	8	5	Fruit
9	Green Bean	3	7	Vegetable
10	Nuts	3	6	Protein
11	Orange	7	3	Fruit

Define a workflow to load the data and visualize the data in a scatterplot widget in Orange. Configure the scatterplot such that the x-axis shows the sweetness and the y-axis shows the Crunchiness of the ingredients. In the scatterplot, the different food types should be shown using different colours with a legend showing the type of food represented by different colours. Also, the ingredient of the food should be shown as labels beside each data point. You should adjust the zoom level such that all data points and labels are shown clearly in the scatterplot. Show the gridlines in the scatterplot.

Discuss your steps and show the screenshots of your Orange workflow and the output of your scatterplot. You should show both the scatterplot and the option panel on the left.

b) [16 marks] Refine the Orange workflow in (a) to predict the food type of an ingredient X with sweetness of 8 and crunchiness of 9 using kNN (k=1). For the kNN widget, select the options **Euclidean** for **Metric** and **Uniform** for **Weight**.

You should illustrate your steps, discuss the roles of the attributes of the dataset, capture a screenshot of your workflow and the model's prediction from the **Predictions** widget in Orange and state the predicted food type of X.