Subject:	Machine Learning 3781
Assessment:	Project
Due date:	2024-03-11
Total:	100 Marks

Introduction

In this project, you are supposed to research and use some prior knowledge from other courses. You are expected to work in groups and collaborate.

Outline

1. For this exercise you do not need to do the Gain or Entropy calculations. There is a direct mapping between a Boolean function and its corresponding binary decision tree. The binary decision tree can usually by simplified as well to produce a simpler, more compact tree. Do not just write down the final, simplified tree. Show how you do the simplification.

   Give binary decision trees to represent the following Boolean functions:

   (a) $A \wedge \neg B$

   (b) $A \vee [B \wedge C]$

   (c) $[A \vee B] \wedge [C \vee D]$

2. Consider the training data for ENJOYSPORT as shown in Table 1:

Table1

Example	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes
2	Sunny	Warm	High	Strong	Warm	Same	Yes
3	Rainy	Cold	High	Strong	Warm	Change	No
4	Sunny	Warm	High	Strong	Cool	Change	Yes

Develop the decision tree that would be learned by ID3 algorithm, using the data in Table 1. Note that this decision tree will not be a binary tree (think about why this is the case). Show every step of your calculation, including all the steps of the Gain and Entropy calculations. Show the formulas that you use. Clearly explain the results and the reasons for the choices ID3 makes.

3. In the world we are experiencing now, people have become a lot more aware of medical test results. One question that most people would want to be answered: How accurate is a test for an infection?

There are terms often used in medical test results, and are vital in correctly interpreting a test result:

1. **Prevalence**: the ratio of the total population who is infected.

2. **True positive (TP)**: the ratio of the total tests that are accurately labelled as positive. This is dependent on the prevalence.

3. **True negative (TN)**: the ratio of the total tests that are accurately labelled as negative. This is dependent on the prevalence.

4. **Sensitivity:** (true positive rate TPR) the ratio of positive tests that are accurately labelled as positive. This is independent of the prevalence.

5. **Specificity:** (true negative rate TNR) the ratio of negative tests that are accurately labelled as negative. This is independent of the prevalence.

6. **Positive Predictive Value (PPV):** the probability of an infection given a positive test result.

7. **Negative Predictive Value (NPV):** the probability of no infection given a negative test result.

Consider the data on a medical test for SUPERBUG:
1. Out of every 10 000 people with a record of possible symptoms, more or less 100 people were diagnosed with SUPERBUG. These are confirmed cases, based on a combination of doctors' diagnoses, CT-scans, several different tests, and post mortem analyses.

2. It is known that for this test, 10 out of 100 positive test results are incorrect. This is the inverse sensitivity of the test.

3. It is known that for this test, 10 out of 50 negative test results are incorrect. This is the inverse specificity of the test.

You have just been tested for SUPERBUG, but are still waiting for your results. Obviously you will want to know what a positive or negative result will tell you about the likelihood that you have been infected with SUPERBUG, so that you can decide how to deal with the result.

Given the data above:

1. Define the variables you will use in your calculations.

2. Calculate the prevalence of SUPERBUG.

3. Calculate the sensitivity and specificity of the test.

4. Calculate the inverse sensitivity and inverse specificity of the test.

5. Calculate the false positive and false negative rate for the test.

6. Calculate the four prior probabilities, using your variable definitions.

Mark Allocation

Criteria	Weight
Qn 1.	**40**
1a)	5
1b)	15
1c)	20
Qn 2.	**20**
Qn 3.	**40**
Total	**100**

Additional Information

- All work must be done in groups of 3 to 4 people which are automatically assigned to you.
- One member can make a submission on behalf of the group.
- Belgium Campus have software that can **scan for plagiarism** and a student caught doing this will get 0 for this assignment.
- Late assignments will not be accepted; missing the deadline is an automatic 0.