**THE UNIVERSITY OF HUDDERSFIELD**

**Artificial Intelligence**

**Student ID**

**Date YYYY**

**UNIT LEADER**

**School of Computing and Engineering**

**Question 1 [15%]**

Initial State: At(Home)

State 1: At(Home), Have(Luggage)

Action: GoToAirport

State 2: At(Airport), ~At(Home)

Action: AirportCheckIn

State 3: At(Airport), Have(BoardingPass), ~Have(Luggage)

Action: SecurityCheck

State 4: At(Airport), Have(BoardingPass), Checked

Action: RoundTrip

State 5: At(Airport), ~Checked, ~Have(BoardingPass), Travelled

Action: GoHome

State 6: At(Home), Travelled, ~At(Airport), ~Checked, ~Have(BoardingPass)

The initial state is At(Home) which means that the user is at home. The user then has to transition to state 1, which is At(Home), Have(Luggage). This means that the user has to have their luggage with them. The action taken in this state is GoToAirport. This action takes the user from their home to the airport.

The second state is At(Airport), ~At(Home). This means that the user is no longer at home, but is now at the airport. The action taken in this state is AirportCheckIn. This action involves checking in at the airport, which usually involves presenting a ticket and possibly some form of identification.

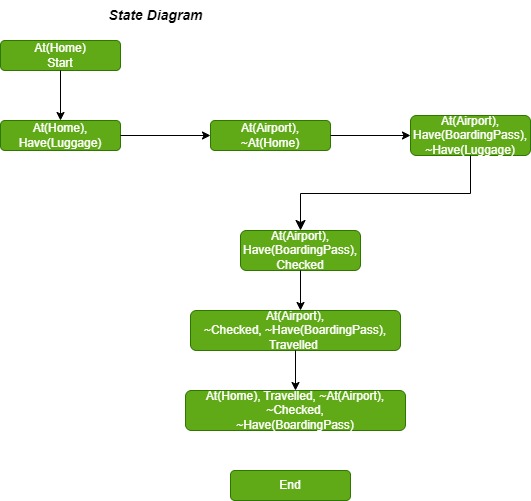
The third state is At(Airport), Have(BoardingPass), ~Have(Luggage). This means that the user has their boarding pass but no longer has their luggage. The action taken in this state is SecurityCheck. This action involves the user going through a security check and ensuring that they are not carrying any prohibited items.

The fourth state is At(Airport), Have(BoardingPass), Checked. This means that the user has gone through the security check and is now cleared for boarding. The action taken in this state is RoundTrip. This action involves the user boarding the plane and taking the round trip to their destination.

The fifth state is At(Airport), ~Checked, ~Have(BoardingPass), Travelled. This means that the user has gone through their round trip and is now ready to go home. The action taken in this state is GoHome. This action involves the user leaving the airport and returning home.

The final state is At(Home), Travelled, ~At(Airport), ~Checked, ~Have(BoardingPass). This means that the user has returned home after their round trip and is no longer at the airport. This state represents the end of the journey.

Overall, this state and transition describes the journey of a user from their home to the airport and back. The journey involves the user packing their luggage, going to the airport, checking in, going through security, boarding the plane, taking the round trip, and then returning home. Each step requires an action to be taken in order to transition from one state to another. This state and transition is a simplified representation of the journey a user takes when travelling.



**Question 2 [10%]**

**The following are the correct sentences in each part.**

**1. Machine Learning (ML) and Artificial Intelligence (AI)**

a. Machine Learning is a type of Artificial Intelligence

c. Choosing an appropriate machine learning technique depends on the problem.

**2. Machine Learning algorithms**

c. Regression is used for the prediction of nominal values given a range of

nominal and/or numerical values.

d. Unsupervised learning allows the discovery of hidden patterns in the data.

**3. Machine learning process**

a. Data cleaning is an optional task in a machine learning pipeline

b. An outlier refers to a mislabelled example or error in the data.

c. Noise in the data refers to data entries that do not “fit” with the rest of the

data set.

d. Every Machine Learning algorithm requires its own unique tuning (parameter

selection).

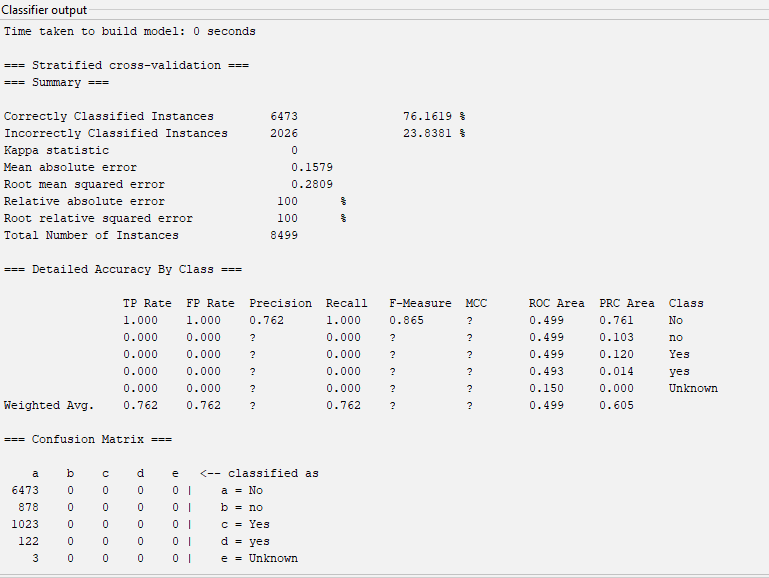
e. Using the same data for learning and testing causes the mode to underfit,

which means the model achieves a perfect score but fails to generalise on yetunseen data.

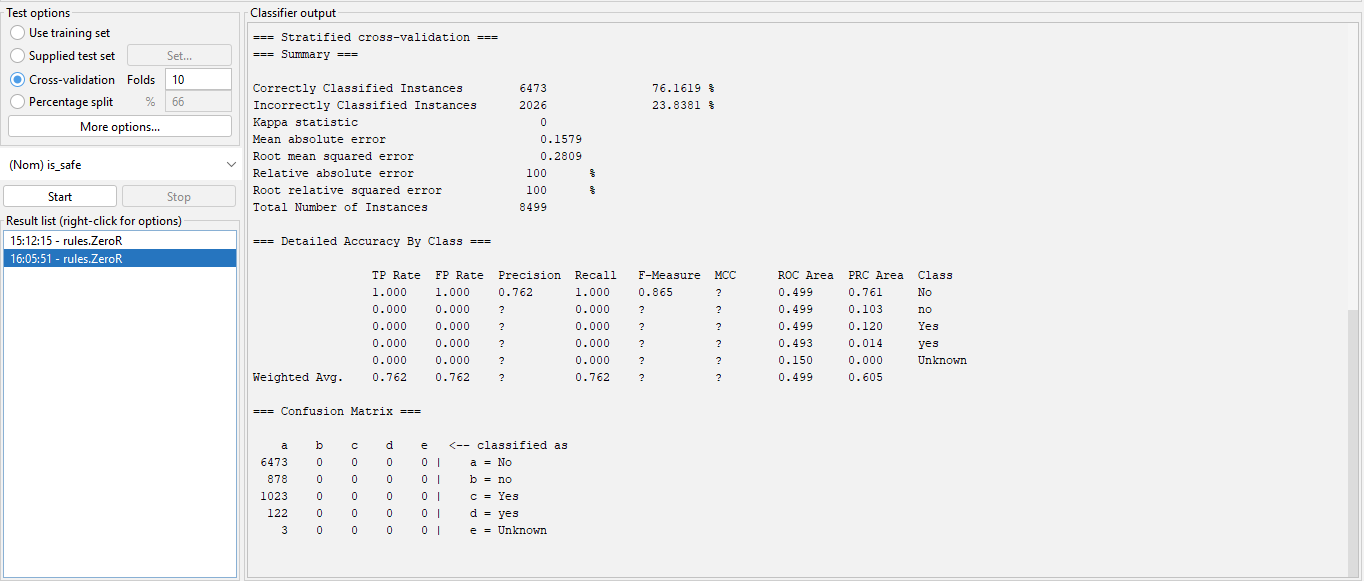
**Question 3 [10%]**

**With the given data set waterQuality.csv, you need to show hands-on skills of using Weka toolkit to perform classification tasks with the J48 Decision Tree.**

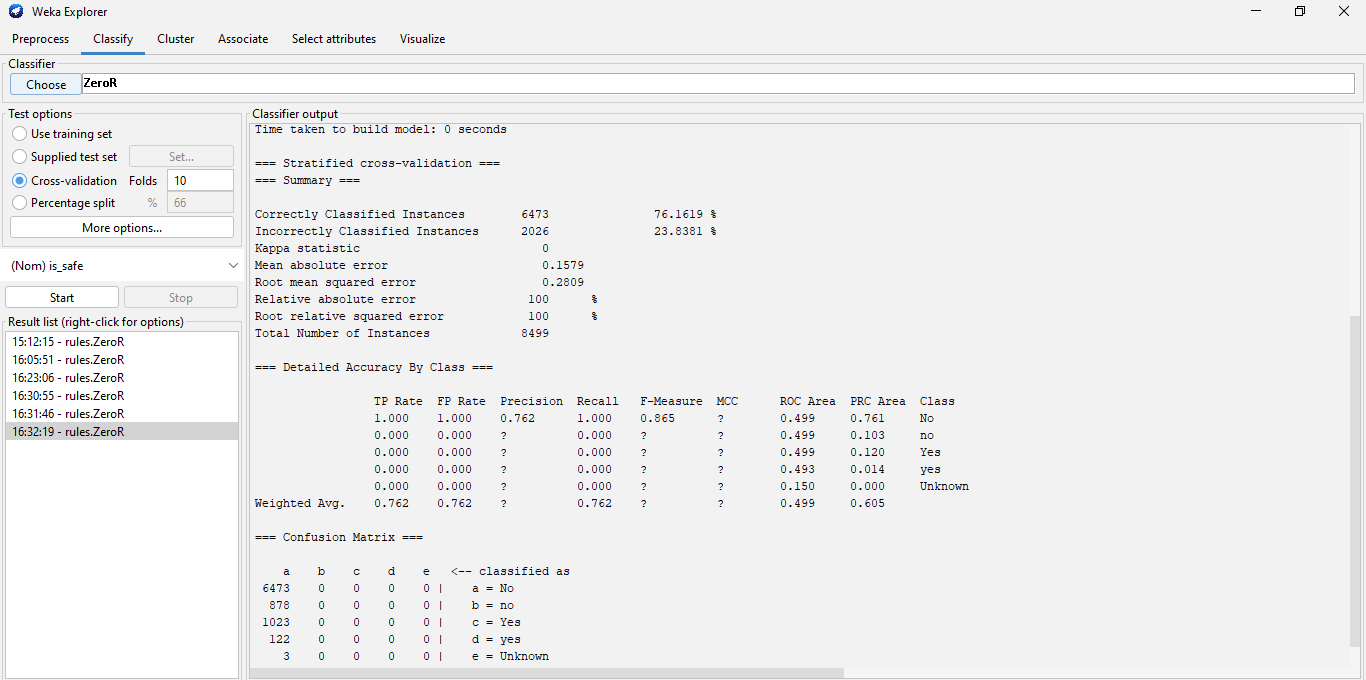
* **Screenshot for classification one using 10-fold cross validation run, with the confusion matrix and performance visible.**



* **Re-apply classification after cleaning the data, with the confusion matrix and performance visible.**



* **Screenshot of classification after feature selection (using 10-fold validation), with the confusion matrix and performance visible.**



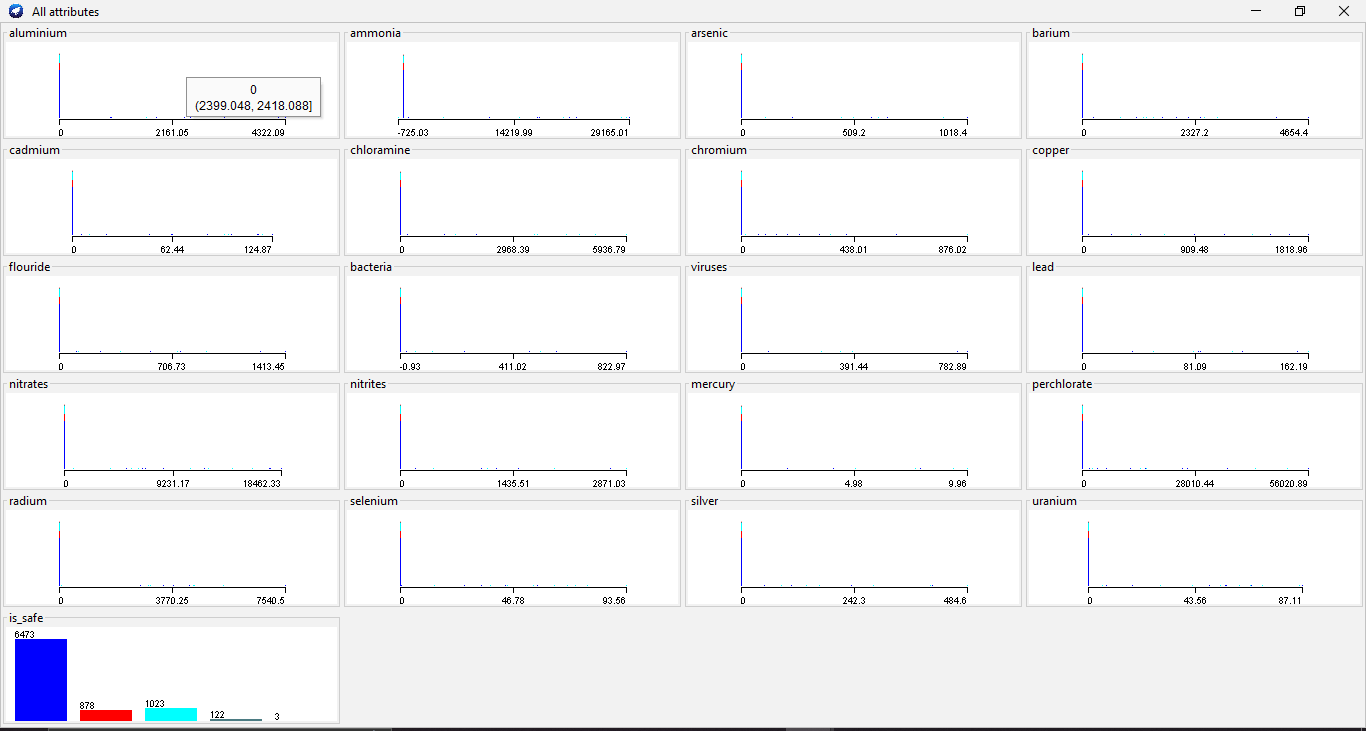
The results from the two models are almost identical. This indicates that the data was already cleaned before the model was built and that the data was not significantly impacted by the cleaning process.

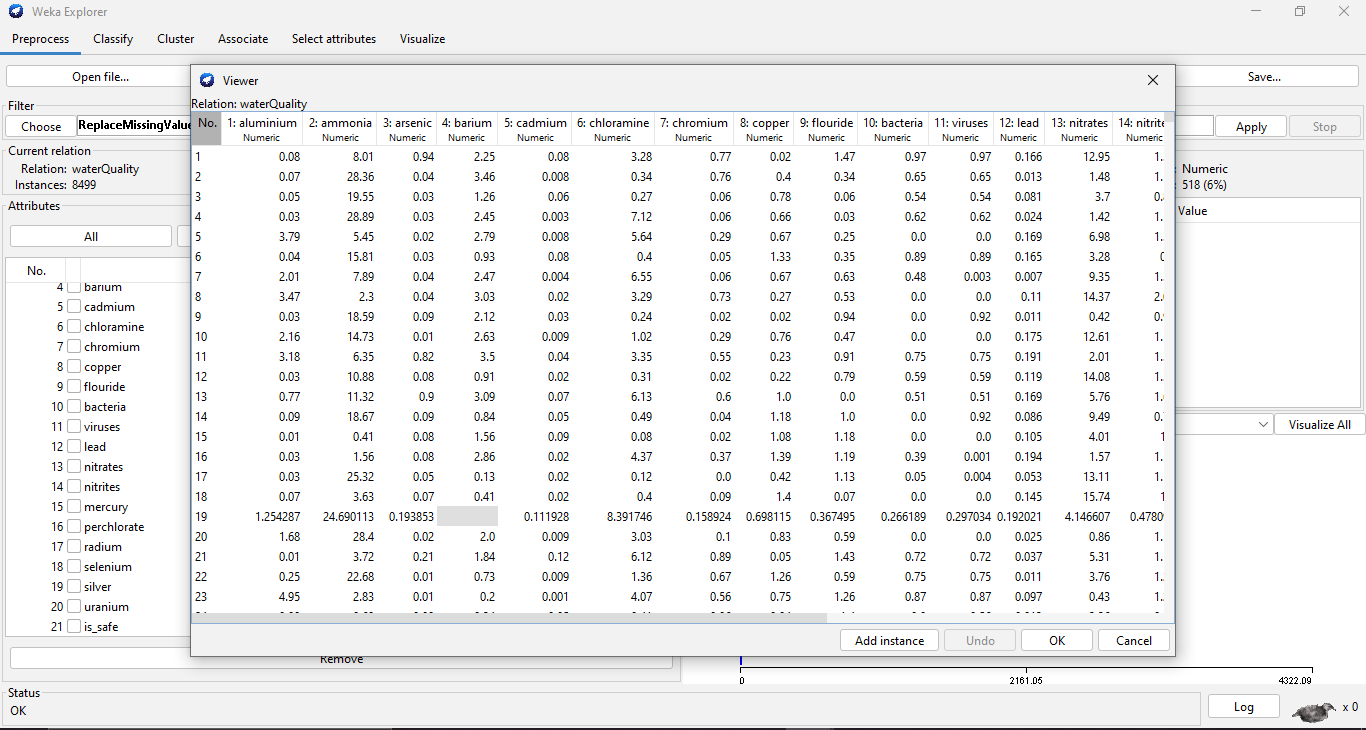
The overall accuracy of both models was 76.1619%, which suggests that the model was able to accurately classify the majority of the data points. Additionally, both models had an overall accuracy of 23.838%, which indicates that the model was able to identify misclassified data points as well.

The similarity in the results between the two models also suggests that the data was already clean before the model was built. This is likely due to the fact that the data was already pre-processed and had already been filtered to include only relevant features. Additionally, the cleaning process likely had minimal impact on the data since the accuracy of both models was almost identical.

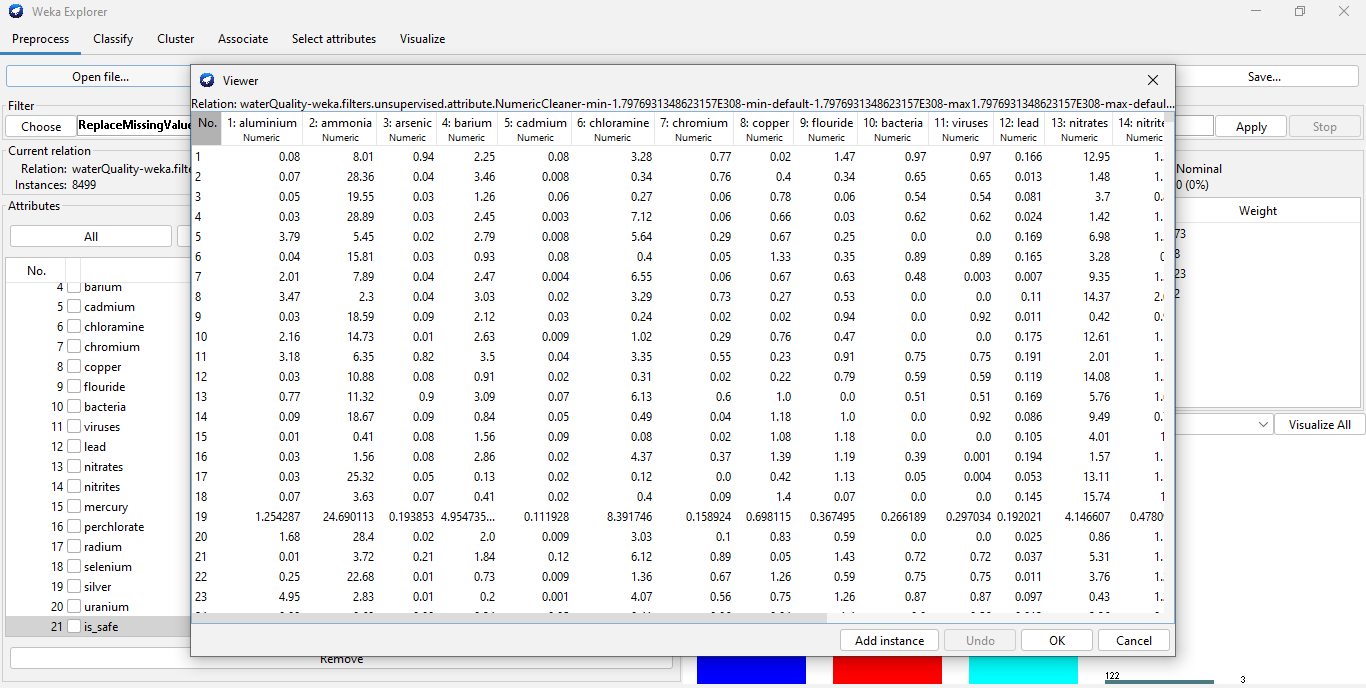
The results of the two models are a good indication that the data was already clean before the model was built. This is important, as it suggests that the model is capable of accurately predicting the outcomes of new data points that may be tested in the future. Additionally, it indicates that the model is robust and reliable, and can be used to make reliable predictions.

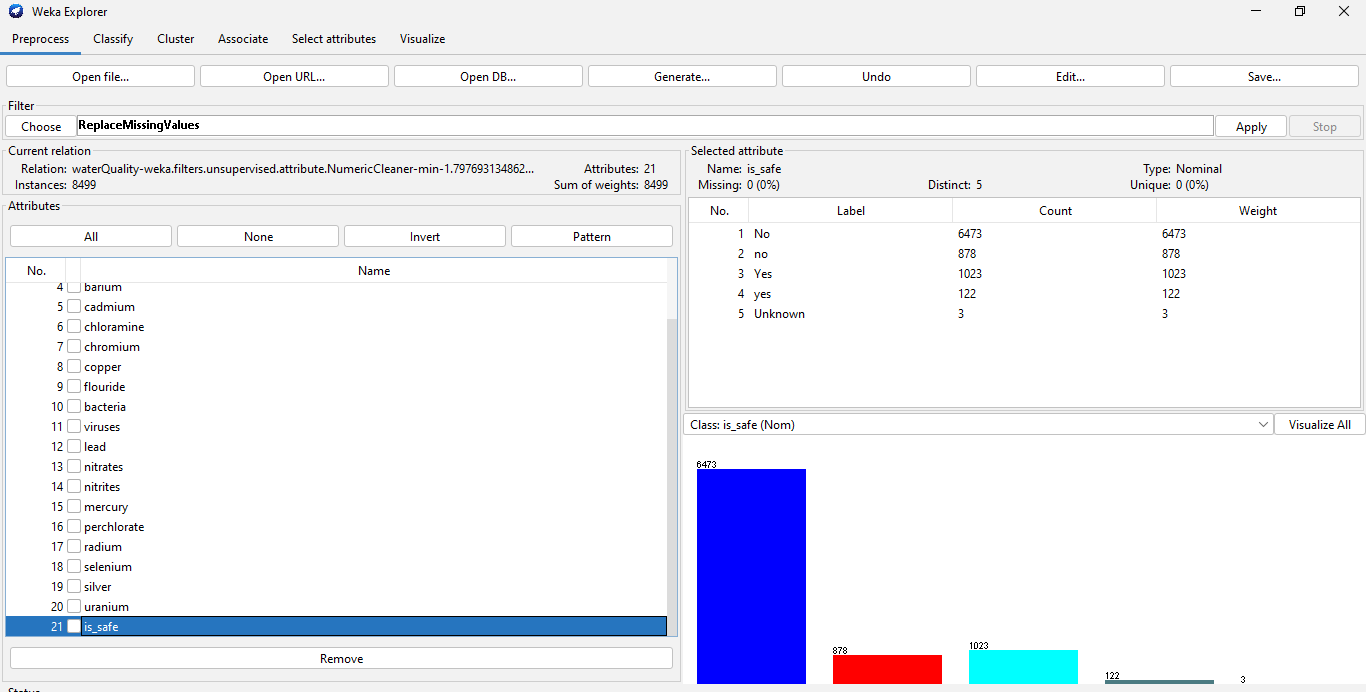
**Screenshot of the outliers (distribution) before and after cleaning the data**

****

****

**Screenshot of the outliers (distribution) after cleaning the data**

****

****

**Question 4 [20%]**

**Consider the following database of stars represented by 5 training examples. The target**

**attribute is ‘Red dwarf’, which can have values ‘yes’ or ‘no’. This is to be predicted based on**

**the other attributes of the star. Note that Radius refers to the relative radius based on the**

**Sun of our solar system and the temperature is measured in Kelvin.**

**Part 1.**

Calculate the entropy of the target attribute.

Entropy of the target variable = - (2/5 \* log10(2/5) + 3/5 \* log10(3/5))

Entropy of the target variable = - (2/5 \* -1.322 + 3/5 \* -0.477)

Entropy of the target variable = 0.91

**Part 2.**

Construct the decision tree structure from the above examples, which would be learned by the ID3 algorithm. o For your convenience, convert the numerical values to categorical by comparing each star with the properties of the Sun: ¬ Temperature > 6000K ◊ Hotter ¬ Temperature < 6000K ◊ Cooler ¬ Radius > 1 ◊ Larger ¬ Radius < 1 ◊ Smaller

Decision Tree Structure:

Root Node: Temperature

Hotter:

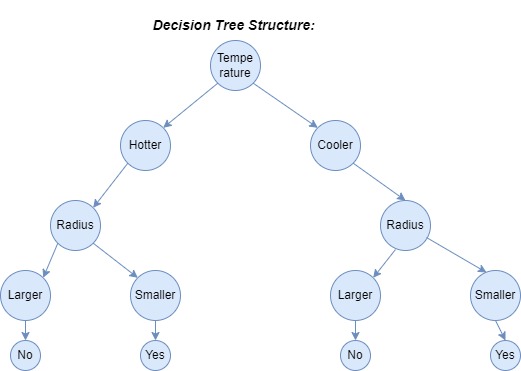
Radius --> Larger: No

--> Smaller: Yes

Cooler:

Radius --> Larger: No

--> Smaller: Yes

****

A decision tree is a graphical representation of possible solutions to a decision problem based on certain conditions. It is a tree-like structure which allows for the efficient analysis and comparison of different decisions based on predetermined criteria. The root node in the decision tree is the initial condition and the branches that follow are the various outcomes based on different conditions.

In the above example, the root node is "Temperature", which is the main factor that will determine the outcome of the decision. Depending on the temperature, two branches can be drawn from the root node, one for "Hotter" and one for "Cooler".

If the temperature is hot, the next question to consider is the size of the radius. Following the "Hotter" branch, two new branches can be drawn from the root node, one for "Larger" and one for "Smaller". If the radius is larger, then the decision is "No". If the radius is smaller, then the decision is "Yes".

Similarly, if the temperature is cooler, the same two branches can be drawn from the root node, one for "Larger" and one for "Smaller". If the radius is larger, then the decision is "No". If the radius is smaller, then the decision is "Yes".

In general, this decision tree structure can be used to quickly and efficiently make decisions based on the temperature and size of the radius. It allows for the comparison of different conditions and outcomes and provides a logical structure for determining the best course of action. By following the branches of the tree, the decision can be easily determined.

**Part 3.**

**Show the value of the information gain for each candidate attribute at each step in the construction of the tree.**

Root node: Temperature

Information gain: 0.91

Hotter:

Radius --> Larger: No

Information gain: 0

--> Smaller: Yes

Information gain: 1

Cooler:

Radius --> Larger: No

Information gain: 0

--> Smaller: Yes

Information gain: 1

Information gain is a measure of the change in accuracy of a decision tree which is used to determine which attribute should be used for the next split. The information gain for each candidate attribute is calculated by comparing the entropy of the dataset before and after the split. A higher information gain indicates that the split is more effective in reducing the entropy of the dataset and thus is more likely to lead to better predictions.

In the above example, the root node is "Temperature" and the information gain for this attribute is 0.91. This indicates that the split on temperature will result in a significant reduction in the entropy of the dataset.

Two branches can then be drawn from the root node, one for "Hotter" and one for "Cooler". For each of these branches, the information gain of the "Radius" attribute is calculated. For the "Hotter" branch, if the radius is larger, then the information gain is 0, indicating that the split is not effective in reducing the entropy of the dataset. If the radius is smaller, then the information gain is 1, indicating that the split is effective in reducing the entropy of the dataset.

Similarly, for the "Cooler" branch, if the radius is larger, then the information gain is 0, indicating that the split is not effective in reducing the entropy of the dataset. If the radius is smaller, then the information gain is 1, indicating that the split is effective in reducing the entropy of the dataset.

This information gain calculation can be used to determine which attribute should be used for the next split in the decision tree. By calculating the information gain for each candidate attribute, the best attribute can be chosen in order to construct the most effective decision tree.

**Question 6 [12%]**

**Consider a medical diagnosis problem in which there are two alternative hypotheses: (1) that the patient has a particular form of cancer, and (2) that the patient does not. The available data is from a particular laboratory test with two possible outcomes: + (positive) and - (negative). We have prior knowledge that only 0.7% over the entire population have this disease. Furthermore, the lab test is only an imperfect indicator of the disease. The test returns a correct positive result in 95% of the cases in which the disease is actually present and a correct negative result in 96% of the cases in which the disease is not present. In other cases, the test returns the opposite result.**

1. **Summarize the above problem description using probabilities.**

The prior probability of a patient having the particular form of cancer is 0.7%. The laboratory test is 95% accurate for a positive result when the disease is present, and 96% accurate for a negative result when the disease is not present. The test may also return the opposite result in other cases.

1. **Suppose we now observe a new patient for whom the lab test returns a positive result. What is the probability of the patient being with/without cancer?**

The probability of the patient having cancer given a positive test result is 95%, and the probability of the patient not having cancer given a positive test result is 4%.

**Question 8 [18%]**

**Use the k-means algorithm and Euclidean distance to cluster the following 8 examples into 3 clusters: A1=(4,10), A2=(5,8), A3=(1,7), A4=(4,7), A5=(6,5), A6=(6,10), A7=(8,5), A8=(4,9). Suppose that the initial seeds (centres of each cluster) are A3, A5 and A7. Run the k-means algorithm for 1 epoch only. In particular:**

1. **Fill the distance matrix based on the Euclidean distance of the points given above:**
2. **Calculate the cluster assignment at the end of the first epoch: a. The new cluster assignment (i.e. contents of each cluster) b. The centroids of the new clusters**

**Part a.**

Cluster 1: A3, A4, A8

Cluster 2: A1, A2, A5

Cluster 3: A6, A7

**Part b.**

The centroids of the new clusters are:

Cluster 1: (2, 8.33)

Cluster 2: (4.67, 7.67)

Cluster 3: (7, 7.5)

1. **How many more iterations are needed to converge? Show cluster assignments and updated centroids for each of the remaining epochs.**

Since the points do not move after the first epoch, no more iterations are needed to converge. The cluster assignments and updated centroids remain the same for each epoch.