**Orange Data Mining Software Report**

**Group 8 – Submitted by Akarsh Pokkunuru – 801043808**

The complete project workflow is shown below:

**A close up of a map

Description automatically generated**

The orange project workflow contains the following modules:

1. Association rule mining
2. Decision tree classifier
3. Hierarchical clustering

In all the above three tasks, Iris open dataset is used. The iris dataset feature distribution is shown below. There are three classes in the dataset and these classes are divided based on features such as petal length, petal width, sepal length and sepal width of various flowers.

A screenshot of a social media post

Description automatically generated

**Association Rule mining:**

The first task handled is association rule mining. The project workflow is as shown below:

A close up of a map

Description automatically generated

The iris dataset object contains the actual data in the form of rows and columns. We begin with

discretizing the dataset with equal frequency and number of intervals is set to 3.A screenshot of a cell phone

Description automatically generated

After discretizing the dataset, we find the frequent itemsets with a minimum support of 3. The result of dataset and its corresponding support is shown below.A screenshot of a social media post

Description automatically generated

Then we mine the rules with a minimum support of 3 and minimum confidence of 80 %. The result for the Iris dataset is as shown below. The complete set of results can be found in the project file.

A screenshot of a computer screen

Description automatically generated

**Decision Tree Classifier:**

The next task is to classify the features of the Iris dataset using a decision tree. The project workflow is shown below:

A close up of a map

Description automatically generated

We begin with a random data sampler and divide the dataset into test and train sets. 70% of the data goes into training and remaining 30% of data was used for testing.

A screenshot of a cell phone

Description automatically generated

After train-test split, we setup decision tree as follows:

A screenshot of a cell phone

Description automatically generated

The test and score module gives various metrics such as classification accuracy, Area under Curve, precision and recall scores, which are commonly used as metrics for performance of any data mining algorithms. These scores are shown below.

The overall classification accuracy is 92.4%.

A screenshot of a social media post

Description automatically generated

The corresponding confusion matrix is shown below:

A screenshot of a social media post

Description automatically generated

The results of the decision tree are plotted on a scatter plot for better clarity and is show below:

We can clearly see the classification boundaries for all the three classes.

A screenshot of a cell phone

Description automatically generated

Finally, the decision tree itself is visualized below:

A screenshot of a cell phone

Description automatically generated

**Hierarchical Clustering:**

The clustering workflow is shown as below:

A picture containing clock

Description automatically generated

Clustering is performed initially by using the cosine distance metric and computing the distance between all its features.

A screenshot of a cell phone

Description automatically generated

The distance is then used to compute the clusters as follows:

A screenshot of a computer

Description automatically generated

The red and blue are manually selected to form two clusters for the sake of simplicity of visualization. The corresponding dendrograms are also visualized using average hierarchical clustering technique.

Thus, three tasks were performed on the Iris dataset. The project file named “*group\_8*\_orange” is also attached for a full inspection of results.