EXPERIMENT NO. 5 (Part-I)

Aim: To perform ETL process for building a data warehouse using ORANGE tool.

1. Consider bridges.mt1.tab and bridges.mt2.tab files as input and do as directed:

- Concatenate the files
- Apply preprocessing for missing values
- Save the new file
- Open the new file and view it.

2. Use the above file to

- Discretize the variables and find the type of river that is maximum in number, which type of material is majorly used?
- Consider only erected, type, river, purpose and length columns and save it as file- F1.tab

3. Do as directed:

- Import auto-mpg.tab file. Make 2 parts of the file columns.
  1. cylinders, weight, acceleration, model_year and car_name.
  2. mpg, origin, horsepower, displacement and car_name.
- Consider these 2 parts and merge to get the original file.

Answer the following questions:

1. In which year maximum cars were built?
2. Write a python script to display first 3 rows of auto-mpg.
3. Write a python script to find the maximum displacement, its index and the corresponding row.
4. Conclude about the number of cylinders. Add grouping of model_year in box plot to make conclusions.
5. Show analysis using box plot.

Python Script:

```
import Orange
data= Orange.data.Table("auto-mpg")
print(data.domain)
for d in data[:3]:
print (d)
p=[1]
for i in range(0,398):
p.append(data[i]['displacement'])

for i in range(0,398):
print(p)

print('max value is:')
print(max(p))
print('index of max value is:')
i=p.index(max(p))
print(i)

data= Orange.data.Table("auto-mpg")
print(data.domain)
for q in data[i-1]:
print (q)
```

EXPERIMENT NO. 5 (Part-II)

Aim

Introduction to the Weka machine learning toolkit.

Questions:

1. Press the Explorer button on the main panel and load the weather dataset and answer the following questions
    1. How many instances are there in the dataset?
    2. State the names of the attributes along with their types and values.
    3. What is the class attribute?
    4. In the histogram on the bottom-right, which attributes are plotted on the X,Y-axes? How do you change the attributes plotted on the X,Y-axes?
    5. How will you determine how many instances of each class are present in the data
    6. What happens with the Visualize All button is pressed?
    7. How will you view the instances in the dataset? How will you save the changes?

2. Load the weather dataset and perform the following tasks:
    1. Use the unsupervised filter RemoveWithValues to remove all instances where the attribute ‘humidity’ has the value ‘high’?
    2. Undo the effect of the filter.
    3. Answer the following questions:
        1. What is meant by filtering in Weka?
        2. Which panel is used for filtering a dataset?
        3. What are the two main types of filters in Weka?
        4. What is the difference between the two types of filters? What is the difference between and attribute filter and an instance filter?

Part A: Application of Discretization Filters

1. Perform the following tasks
    1. Load the 'sick.arff' dataset
    2. How many instances does this dataset have?
    3. How many attributes does it have?
    4. Which is the class attribute and what are the characteristics of this attribute?
    5. How many attributes are numerics? What are the attribute indexes of the numerica attributes?
    6. Apply the Naive Bayes classifier. What is the accuracy of the classifier?
2. Perform the following tasks:
    1. Load the 'sick.arff' dataset.
    2. Apply the supervised discretization filter.
    3. What is the effect of this filter on the attributes?
    4. How many distinct ranges have been created for each attribute?
    5. Undo the filter applied in the previous step.
    6. Apply the unsupervised discretization filter. Do this twice:
        1. In this step, set 'bins'=5
        2. In this step, set 'bins'=10
        3. What is the effect of the unsupervised filter filter on the datset?
    7. Run the the Naive Bayes classifier after apply the following filters
        1. Unsupervised discretized with 'bins'=5
        2. Unsupervised discretized with 'bins'=10
        3. Unsupervised discretized with 'bins"=20.
    8. Compare the accuracy of the following cases

1. Naive Bayes without discretization filters
2. Naive Bayes with a supervised discretization filter
3. Naive Bayes with an unsupervised discretization filter with different values for the 'bins attributes.

Part B: Attribute Selection

1. Perform the following tasks:
    1. Load the 'mushroom.arff' dataset
    2. Run the J48, 1Bk, and the Naive Bayes classifiers.
    3. What is the accuracy of each of these classifiers?
2. Perform the following tasks:
    1. Go to the 'Select Attributes' panel
    2. Set attribute evaluator to CFSSubsetEval
    3. Set the search method to 'Greedy Stepwise'
    4. Analyze the results window
    5. Record the attribute numbers of the most important attributes
    6. Run the meta classifier AttributeSelectedClassifier using the following:
        1. CFSSubsetEval
        2. GreedStepwise
        3. J48, 1Bk, and NaiveBayes
    7. Record the accuracy of the classifiers
    8. What are the benefits of attribute selection?

Part C

1. Perform the following tasks:
    1. Load the 'vote.arff' dataset.
    2. Run the J48, 1Bk, and Naive Bayes classifiers.
    3. Record the accuracies.
2. Perform the following tasks:
    1. Go to the 'Select Attributes' panel
    2. Set attribute evaluator to 'WrapperSubsetEval'
3. Set search method to "RankSearch'
4. Set attribute evaluator to 'InfoGainAttributeEval'
5. Analyze the results
6. Run the metaclassifier AttributeSelectedClassifier using the following:
    1. WrapperSubsetEval
    2. RankSearch
    3. InfoGainAttributeEval
7. Sampling
    1. Load the 'letter.arff' dataset
    2. Take any attribute and record the min, max, mean, and standard deviation of the attribute
    3. Apply the Resample filter with 'sampleSizePercent' set to 50 percent
    4. What is the size of the filtered dataset. Observe the min, max, mean, and standard deviation Of the attribute that was selected in step 2. What is the percentage change in the values?
    5. Give the benefit of sampling a large dataset.