Assignment 4

**Title**: Predicting the class of the user by using the trip history dataset that is from a bike sharing service in the united states.

**Aim**: Gather the data set, use data analytic techniques on it and analyse and classify the data set by giving them class labels.

**Theory**: This experiment or assignment is carried out using decision tree algorithm to classify and make a classification model. Objective of this assignment should lead a student to analyse data, attribute, their relations and the outcome. As this assignment is carried out using decision tree, lets learn what is decision tree.

A decision tree is a decision support tool that uses a tree-like graph or model of decision and their possible consequences , including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements. A decision tree is a flowchart-like structure in which each internal node represents a “test” on an attribute. So in this assignment we are going to transform data in categorical form. After data transformation we need to create classification model using decision tree,(ID3,C4.5,Random Forest). So after classification model is created we can do comparative analysis of different classifiers with the base classifier (Decision Tree). After all the we need to have a graphical representation of the analysis that we did in order to have proper visual idea about our classification model.

**Code:**

import pandas as pd

importnumpy as np

importmatplotlib.pyplot as plt

fromsklearn.model\_selection import train\_test\_split

fromsklearn.metrics import accuracy\_score

fromsklearn import tree

df = pd.read\_csv(filepath\_or\_buffer = "capitalbikeshare.csv" )

df.head()

# breif info

df.info()

df.describe()

df["Member type"].unique()

# replace categorical data (Member type) with numbers

member\_type = {

"Member type": {

"Member":0,

"Casual":1,

"Unknown":2

}

}

df.replace(member\_type, inplace=True)

df.head()

df[["Duration","Start station number","End station number","Member type"]].boxplot()

len(df["Duration"].unique())

X = df[["Duration","Start station number","End station number"]]

y = df["Member type"]

X\_train, X\_test, y\_train,y\_test = train\_test\_split(X,y,test\_size = 0.30,random\_state = 42)

model = tree.DecisionTreeClassifier()

model.fit(X\_train,y\_train)

y\_pred = model.predict(X\_test)

accuracy = accuracy\_score(y\_test, y\_pred)

print("Accuracy of model : " , accuracy)

**Output**:

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 115597 entries, 0 to 115596

Data columns (total 9 columns):

Duration 115597 non-null int64

Start date 115597 non-null object

End date 115597 non-null object

Start station number 115597 non-null int64

Start station 115597 non-null object

End station number 115597 non-null int64

End station 115597 non-null object

Bike number 115597 non-null object

Member type 115597 non-null object

dtypes: int64(3), object(6)

memory usage: 7.9+ MB

Accuracy of model : 0.803316032295271