## Activity Sheet

### Learning outcomes

After solving these exercises, you will understand the following:

1. **Applying the Decision Trees using C5.0 and CART algorithms to solve classification and regression problems respectively**
2. **Understand and interpret the results generated from each algorithm in R**
3. **Comparison of the model performance in terms of Accuracy for Classification**
4. **Comparison of the model performance in terms of MAPE for regression**

**Note: R code has been provided wherever it is necessary. This code is for your reference only. Do not use the same code in R console.**

**Problem Statement:**

A large child education toy company which sells edutainment tablets and gaming systems both online and in retail stores wanted to extract and analyse the customer data to improve their business. They ideally want to treat customers based on their contribution rather than treating all alike. They have been operating for last few years and maintaining all kinds of data. The goal was to predict how much revenue the customer is likely to give.

### Steps to follow for Classification

1. Understand the problem statement cited above clearly
2. Consider the data provided ‘CustomerData1.csv’
3. Understand the data sources and data pre-processing steps
   a. Remove the attribute ‘CustomerID’
   b. Check for the missing values and fill the missing values using KNN Imputation.
   c. Convert all categorical attributes in to factor using as.factor()
   d. You may bin the numeric attributes if you think it is good for analysis and models
4. Target variable: ’Revenue’

5. **Regression Model**
   a. Split the dataset into train and test (70:30 ratio)
   b. Build model rpart using ‘Revenue’ attribute as it is library(rpart)
   ```
   rpart_model<-rpart(Revenue~.,data=train,method="anova")
   ```
   c. Plot the tree
   ```
    plot(rpart_model,main="Regression Tree for Revenue",margin=0.0001,uniform=TRUE)
   text(rpart_model,use.n=T,xpd=T,cex=0.8)
   ```
   d. Predict the Revenue for train and test datasets using the model generated
   ```
   predCartTrain=predict(rpart_model,newdata=train, type="vector")
   predCartTest=predict(rpart_model, newdata=test, type="vector")
   ```
   e. Check the evaluation metric : Error (MAPE) for regression model on train dataset and test datasets.
   ```
   regr.eval(train[,"Revenue"], predCartTrain)
   ```

```
regr.eval(test[,"Revenue"], predCartTest)
```

f. Check the rules obtained from the model

```
rpart_model
```

6. **Classification Models**

a. Bin the target attribute “Revenue” into 3 levels (High, Medium & Low).

```
 Rev_bin <- function(x){
if(x<180)
  bin <- 'low'
else if(x>=180 && x<350)
  bin <- 'medium'
else
  bin <- 'high'
return(bin)
}
```

b. Convert Target attribute as factor

c. Split the dataset into train and test (70:30 ratio)

d. C50 Model generation and understanding

i. Build model C50 on train dataset

```
C50_model=C5.0(Revenue~.,data=train,rules=T)
```

ii. Do the predictions on train and test datasets

```
 preds_C50_train = predict(C50_model, train, type="class")
preds_C50_test = predict(C50_model, test, type="class")
```

iii. Generate the confusion matrix for both train and test datasets

```
confmat_C50_train = table(train$Revenue, preds_C50_train)
confmat_C50_test  =  table(test$Revenue,  preds_C50_test)
```

iv. Compute the evaluation metric- Accuracy for train and test datasets iv. Importance of the attributes

e. Rpart Model generation and understanding

i. Build the model rpart on train dataset

```
rpart_model2<-rpart(Revenue~.,data=train,method="class")
plot(rpart_model2)
text(rpart_model2,xpd=T,cex=0.8)
```

ii. Do the predictions on train and test datasets

iii. Generate the confusion matrix for both train and test datasets

iv. Compute the evaluation metric- Accuracy for train and test datasets