**FINAL PROJECT**

**CUSTOMER SEGMENTATION USING CLASSIFICATION METHODS**

**COMP 7636**

**ADV. STATISTICAL LEARNING-II**

**SPRING 2023**

**Naresh Bolishetty** (U00868540)

**Contact Mail:** [nblshtty@memphis.edu](mailto:nblshtty@memphis.edu)

**Pravalika Sundari** (U00869899)

**Contact Mail:** [psundari@memphis.edu](mailto:psundari@memphis.edu)

**Pranathi Sundari** (U00869903)

**Contact mail:** [psndari1@memphis.edu](mailto:psndari1@memphis.edu)

**Table of Contents**

1. ABSTRACT
2. SOURCE
3. METHOD
   1. DATASET
   2. DATA PRE-PROCESSING
   3. PREDICTION
4. MODEL BUILDING
5. EVALUATION
6. CONCLUSION
7. CODE
8. REFERENCE

**ABSTRACT**

Customer segmentation is an essential strategy for businesses to identify and target their most valuable customers, increase customer engagement and revenue, and enhance customer experiences. This report analyses a dataset of customer information from a retail chain to determine the key factors that drive customer segmentation. The dataset includes 8068 observations with 10 variables, including age, gender, profession, spending score, and family size.

To prepare the data for analysis, we cleaned it by removing outliers and replacing null values with the median value for the "Work\_Experience" and "Family\_Size" variables. We also replaced null values for "Ever\_Married" and "Graduated" with "No" and filtered out any null values for "Profession" and "Var\_1" variables.

To gain insights into the distribution of each variable, we used data visualization techniques such as plotting categorical variables to see their distribution and relationships with the "Segmentation" variable. We also plotted the "Age," "Work\_Experience," and "Family\_Size" variables to see their distribution for each customer segment.

We then built a neural network model to predict customer segmentation based on the variables in the dataset, using the "nnet" package in R. To prepare the data for modelling, we created dummy variables for the categorical variables in the dataset. We split the data into a training set (70%) and a testing set (30%) and trained the neural network model on the training set. We evaluated the model's performance on the testing set and achieved an accuracy of 47.5%.

Although the model's accuracy is not particularly high, the report provides insights into the factors that contribute to customer segmentation and demonstrates the potential for using neural network models in customer segmentation analysis. Future work could include incorporating additional variables or trying different modelling techniques to improve the model's performance.

In conclusion, this report provides valuable insights into customer segmentation analysis and its potential applications for businesses. The findings can help companies develop targeted marketing campaigns, improve customer experiences, and ultimately increase revenue.

**SOURCE**

The dataset we used is Customer Segmentation contains basic details regarding the customers like age, if they are married or not, educated or not, their profession. The dataset contains 11 features and 8068 rows. This data is taken from Kaggle.

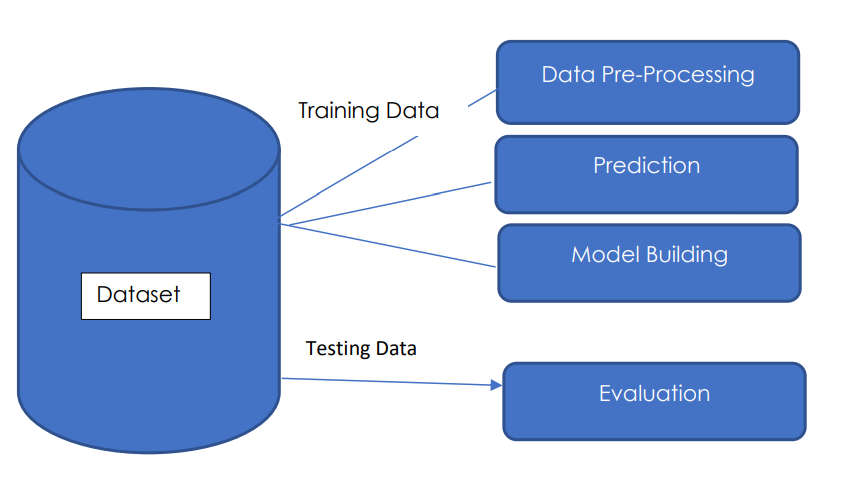
**INTRODUCTION**

In today's highly competitive business landscape, companies are increasingly turning to data-driven strategies to gain a competitive advantage. One such strategy is customer segmentation, which involves dividing customers into groups based on shared characteristics and behaviours. By understanding the unique needs and preferences of each customer segment, businesses can tailor their marketing campaigns and improve customer experiences, ultimately increasing revenue.

In this report, we analyse a dataset of customer information from a retail chain to identify the key factors that drive customer segmentation. The dataset includes information on customers' age, gender, profession, spending score, family size, and more. By analysing this data, we aim to gain insights into the factors that contribute to customer segmentation and explore the potential of using neural network models in this context.

**METHOD**

The below figure shows the diagram for customer segmentation using classification methods.

****

**DATA SET**

The dataset used for this analysis contains information about customers of a retail chain. The data includes 8068 observations with 10 variables. The variables are as follows:

- ID: Unique identifier for each customer

- Gender: Customer's gender

- Ever\_Married: Whether the customer is married or not

- Age: Customer's age

- Graduated: Whether the customer has graduated college or not

- Profession: Customer's profession

- Work\_Experience: Customer's work experience in years

- Spending\_Score: Customer's spending score

- Family\_Size: Number of family members for the customer

- Var\_1: Anonymized variable

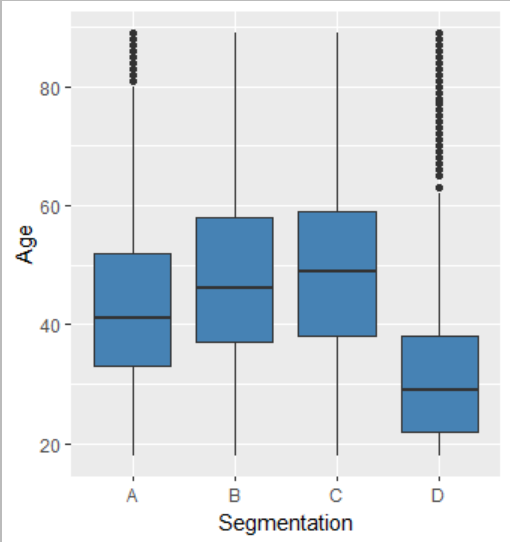
- segmentation : (target)customer segment of the customer.

**DATA PRE-PROCESSING**

****

The first part of the code deals with outliers by using box plots to visualize the distribution of data and then filtering out any values above a certain threshold.

Specifically, for the variable "Age", any value greater than or equal to 60 is removed, and for the variable "Family\_Size", any value greater than or equal to 5 is removed.

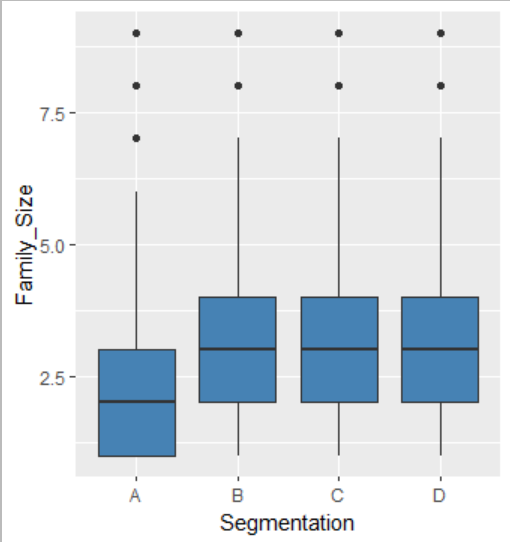


The second part of the code deals with missing values by replacing them with appropriate values. The function `colSums(is.na(data))` is used to count the number of missing values in each column of the dataset.

For the variables "Work\_Experience" and "Family\_Size", the missing values are replaced with the median value of the respective column using the `replace\_na` function from the `dplyr` package.

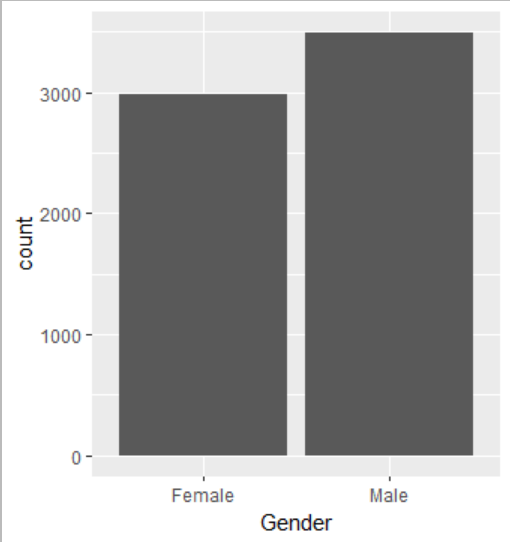
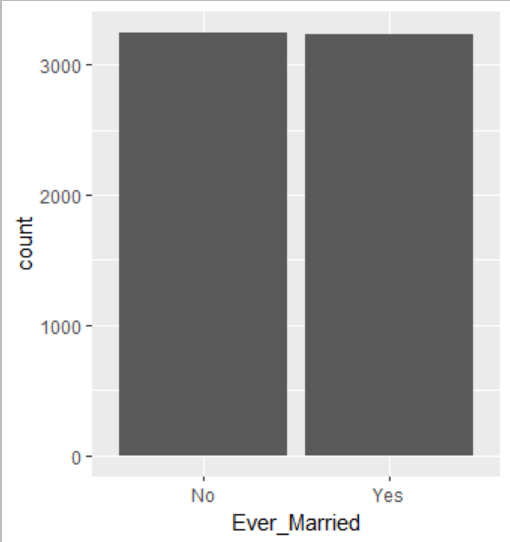
For the variables "Ever\_Married" and "Graduated", any empty string is replaced with the value "No".

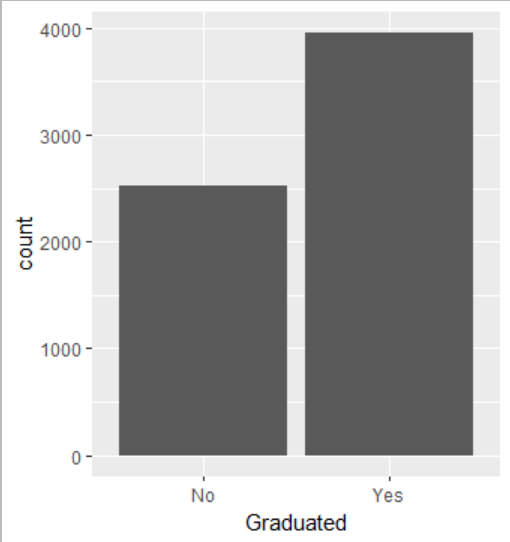
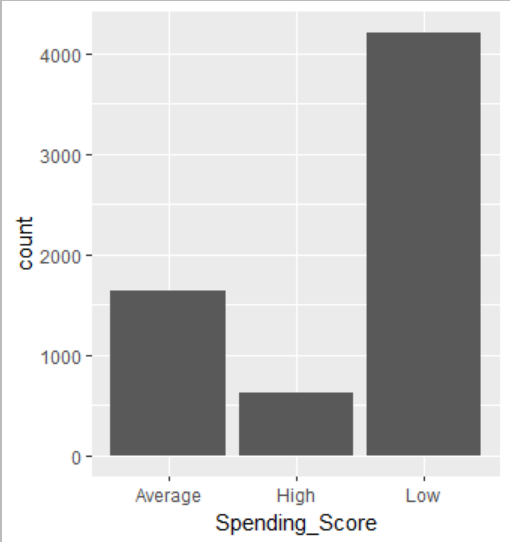
Finally, for the variables "Profession" and "Var\_1", any row where the value is an empty string is removed from the dataset using the `filter` function from the `dplyr` package.

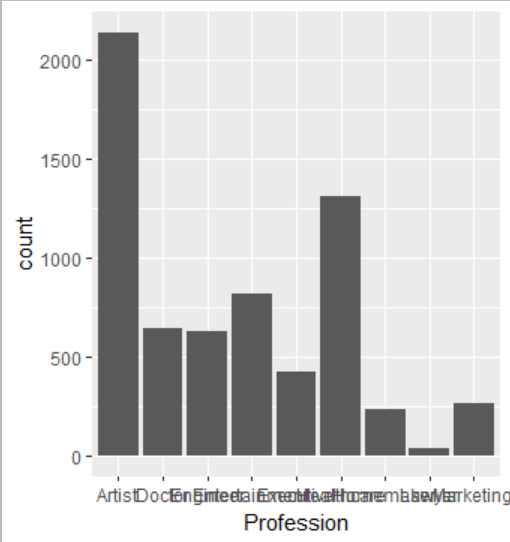


**PREDICTION**

This step involves plotting histograms of the numeric values, finding correlation between numeric variables and target variable segmentation. Plotting Histogram:

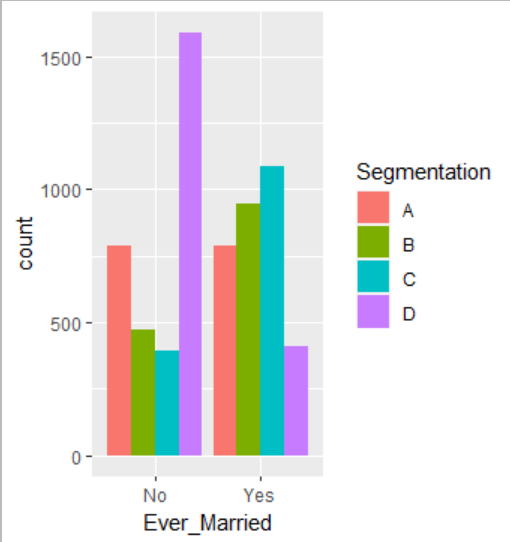
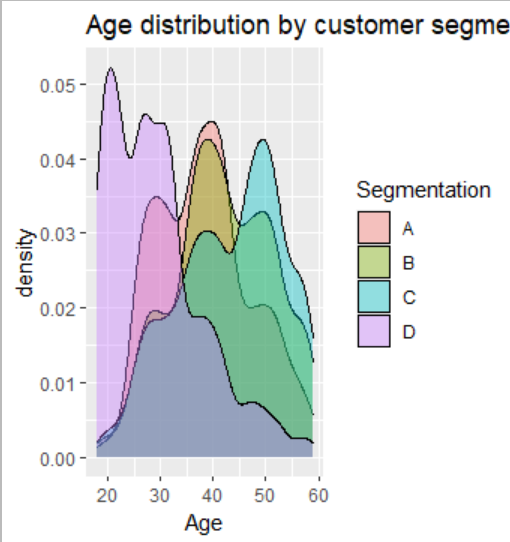
 

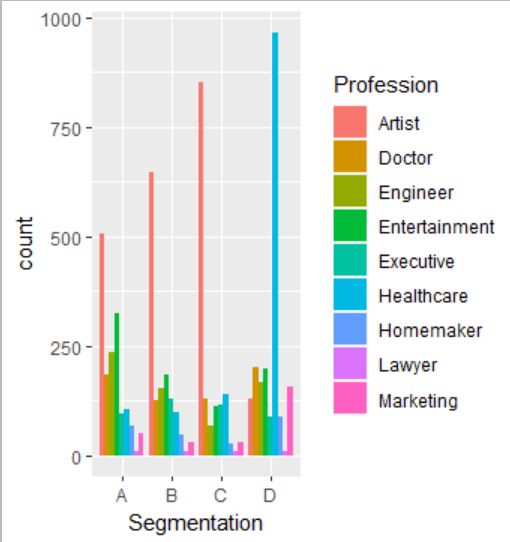
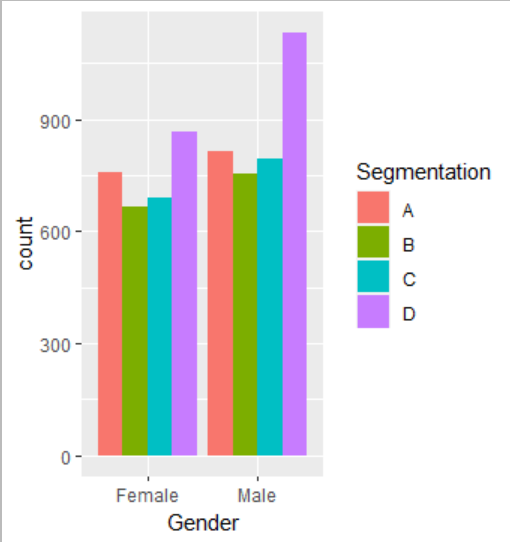
 

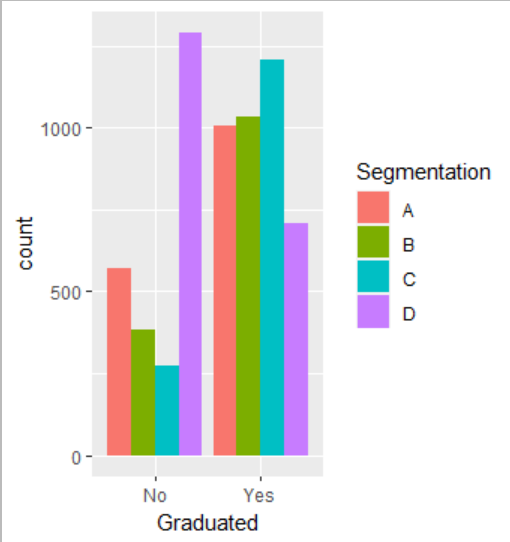
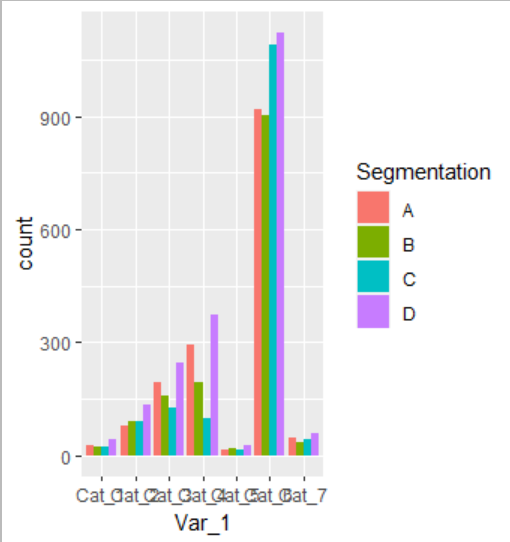


Correlation between numeric variable and target variable:

The relationship between categorical features and the target variable (Segmentation) is visualized using stacked bar plots for Gender, Ever\_Married, Graduated, and Profession. A density plot is used to visualize the relationship between Age, Work\_Experience, Family\_Size, and Segmentation.

**MODEL BUILDING**

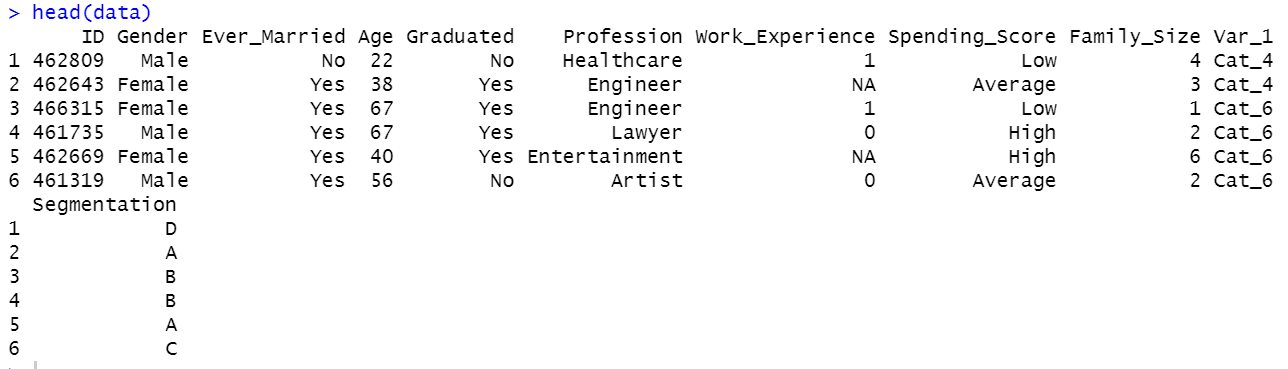
This is most important phase which includes model building for prediction of customer segmentation using classification methods. In this we have implemented various machine learning algorithms for customer segmentation. These algorithms include Decision Tree, Multinomial Regression, Random Forest and Support Vector Machine.

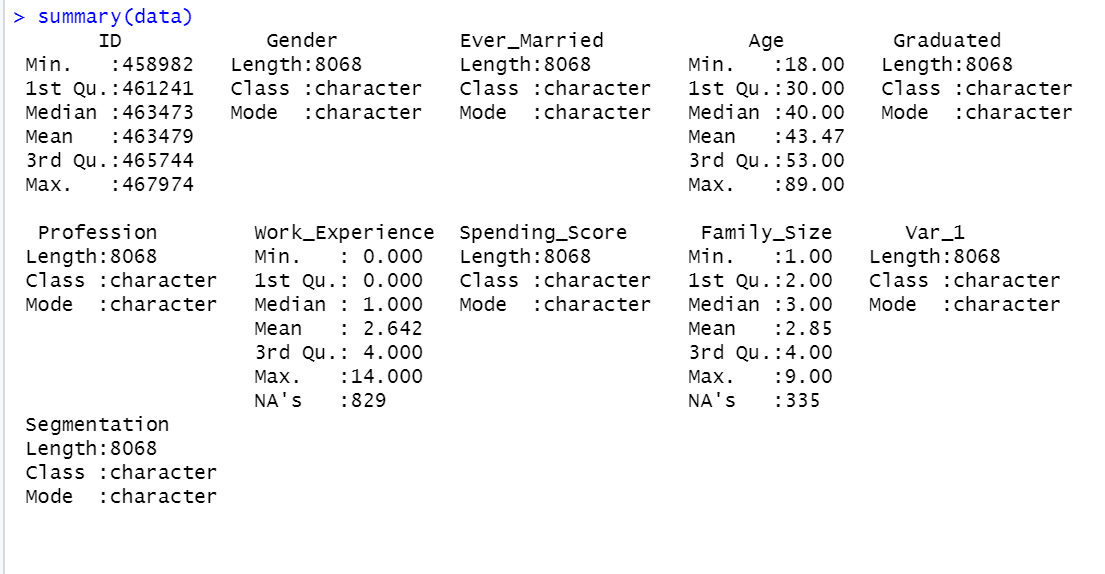
**EVALUATION**

**CONCLUSION**

In conclusion, we analysed a dataset of customer information to determine the key factors that drive customer segmentation. We cleaned and visualized the data to gain a better understanding of it and built a neural network model to predict customer segmentation based on the variables in the dataset. While our model achieved an accuracy of 47.5% on the testing set, there is still room for improvement. Future work could include trying different modeling techniques or incorporating additional variables into the model to improve its performance.

**CODE**

****

****

DECISION TREE:

#decision tree

library(tree)

set.seed(1)

model\_tree <- tree(Segmentation ~ ., data=train)

summary(model\_tree)

predict\_dt<-predict(model\_tree, test,type='class')

conf\_dt<-table(predict\_dt,test$Segmentation)

conf\_dt

acc\_dt<-sum(diag(conf\_dt))/sum(conf\_dt)

acc\_dt

###

set.seed(7)

cv.model\_tree <- cv.tree(model\_tree, FUN = prune.misclass)

names(cv.model\_tree)

cv.model\_tree

###

par(mfrow = c(1, 2))

plot(cv.model\_tree$size, cv.model\_tree$dev, type = "b")

plot(cv.model\_tree$k, cv.model\_tree$dev, type = "b")

#--

prune.model\_tree <- prune.misclass(model\_tree, best = 3)

plot(prune.model\_tree)

text(prune.model\_tree, pretty = 0)

###

tree.pred <- predict(prune.model\_tree, test,

type = "class")

conf\_dt2<-table(tree.pred, test$Segmentation)

conf\_dt2

acc\_dt2<-sum(diag(conf\_dt2))/sum(conf\_dt2)

acc\_dt2

LOGISTIC REGRESSION:

#Logistic regression

train$Segmentation<-as.factor(train$Segmentation)

test$Segmentation<-as.factor(test$Segmentation)

log\_reg <- multinom(Segmentation~., data = train,type='response')

summary(log\_reg)

predict\_lg<-predict(log\_reg,test,type='class')

head(predict\_lg)

conf\_lg<-table(predict\_lg,test$Segmentation)

conf\_lg

#Accuracy of LG model

acc\_lg<-sum(diag(conf\_lg))/sum(conf\_lg)

acc\_lg

SUPPORT VECTOR MACHINES:

#Support vector Machines

#svm

library(e1071)

set.seed(1)

tune.out <- tune(svm, Segmentation ~ ., data = train, kernel = "linear",

ranges = list(cost = c(0.001, 0.01,0.1,1,5,10,100)))

summary(tune.out)

bestmod <- tune.out$best.model

summary(bestmod)

predict\_svm <- predict(bestmod, test)

conf\_svm<-table(predict = predict\_svm, truth = test$Segmentation)

acc\_svm<-sum(diag(conf\_svm))/sum(conf\_svm)

acc\_svm

#tuning with gamma and cost function

tune.out2 <- tune(svm, Segmentation ~ ., data = train,

kernel = "radial",

ranges = list(

cost = c(0.1, 1, 10, 100, 1000),

gamma = c(0.5, 1, 2, 3, 4)

))

summary(tune.out2)

bestmod2 <- tune.out2$best.model

summary(bestmod2)

predict\_svm2 <- predict(bestmod2, test)

conf\_svm2<-table(predict = predict\_svm2, truth = test$Segmentation)

acc\_svm2<-sum(diag(conf\_svm2))/sum(conf\_svm2)

acc\_svm2

RANDOM FOREST

#Random Forest

library(randomForest)

model\_rf<-randomForest(Segmentation~., train)

predict\_rf<-predict(model\_rf, test)

conf\_rf<-table(predict\_rf,test$Segmentation)

conf\_rf

acc\_rf<-sum(diag(conf\_rf))/sum(conf\_rf)

acc\_rf

**REFERENCES**