Installing Packages

install.packages(“e1071”) install.packages(“caTools”) install.packages(“caret”)

Loading package

library(e1071) library(caTools) library(caret)

#LDA analysis library (MASS) df40 <- LDA df40$Race <$ $-as.numeric(as.factor(df$40Race)) set.seed(123) training.individuals <- df40$Race$class==test.transform$Race)

model <- lda(Race~., data = train.transform) model attributes (model) p1 <- predict(model, train.transform)$classtab < -table(Predicted = p1, Actual = train.transform$Race) tab sum(diag(tab))/sum(tab) p2 <- predict(model, test.transform)$classtab1 < -table(Predicted = p2, Actual = test.transform$Race) tab1 sum(diag(tab1))/sum(tab1)

#Naive Bayes Test by Job df3 <-NBtest3 Job3 <- as.factor(df3$Job) df4=as.data.frame(df3) # Splitting data into train and test data split <- sample.split(df4, SplitRatio = 0.7) train_cl <- subset(df4, split == “TRUE”) test_cl <- subset(df4, split == “FALSE”)

Feature Scaling

train_scale <- scale(train_cl) test_scale <- scale(test_cl)

Fitting Naive Bayes Model to training dataset

set.seed(1) # Setting Seed classifier_cl <- naiveBayes(Job ~ ., data = train_cl) classifier_cl

Predicting on test data

y_pred <- predict(classifier_cl, newdata = test_cl)

Confusion Matrix

cm <- table(test_cl$Job, y_pred) cm

Model Evaluation

confusionMatrix(cm)

#Naive Bayes Test by Race (did not work) df10 <-NBtest df10$Race < -as.numeric(as.factor(df$10Race)) df10 df11=as.data.frame(df10) # Splitting data into train and test data split <- sample.split(df11, SplitRatio = 0.7) train_cl2 <- subset(df11, split == “TRUE”) test_cl2 <- subset(df11, split == “FALSE”)

Feature Scaling

train_scale2 <- scale(train_cl2) test_scale2 <- scale(test_cl2)

Fitting Naive Bayes Model

to training dataset

set.seed(1) # Setting Seed classifier_cl2 <- naiveBayes(Race ~ ., data = train_cl2) classifier_cl2

Predicting on test data’

y_pred <- predict(classifier_cl2, newdata = test_cl2)

Confusion Matrix

cm <- table(test_cl2$Race, y_pred) cm

u <- union(y_pred, test_cl2) t <- table(factor(y_pred, u), factor(test_cl2, u)) confusionMatrix(t)

Model Evaluation

confusionMatrix(cm)

#KMeans cluster analysis library (tidyverse) library (cluster) library (factoextra) df30<-LDA df30<- na.omit(df30) df30$Race < -as.numeric(as.factor(df$30Race)) df30 <- scale(df30) head(df30) distance <- get_dist(df30) fviz_dist(distance, gradient = list(low = “#00AFBB”, mid = “white”, high = “#FC4E07”)) k2 <- kmeans(df30, centers = 2, nstart = 25) str(k2) k2 fviz_cluster(k2, data = df30) df30 %>% as_tibble() %>% mutate(cluster = k2$cluster, Race = row.names(LDA)$)cluster, Race = row.names(LDA)) %>% ggplot(aes(Race, Hairstylist, color = factor(cluster), label = Race)) + geom_text() df30 %>% as_tibble() %>% mutate(cluster = k2$cluster, Race = row.names(LDA)$)cluster, Race = row.names(LDA)) %>% ggplot(aes(Race, Teacher, color = factor(cluster), label = Race)) + geom_text()

k3 <- kmeans(df30, centers = 3, nstart = 25) k4 <- kmeans(df30, centers = 4, nstart = 25) k5 <- kmeans(df30, centers = 5, nstart = 25)

plots to compare

p1 <- fviz_cluster(k2, geom = “point”, data = df30) + ggtitle(“k = 2”) p2 <- fviz_cluster(k3, geom = “point”, data = df30) + ggtitle(“k = 3”) p3 <- fviz_cluster(k4, geom = “point”, data = df30) + ggtitle(“k = 4”) p4 <- fviz_cluster(k5, geom = “point”, data = df30) + ggtitle(“k = 5”)

library(gridExtra) grid.arrange(p1, p2, p3, p4, nrow = 2)

df30 %>% as_tibble() %>% mutate(cluster = k5$cluster, Race = row.names(LDA)) %>% ggplot(aes(Race, Engineer, color = factor(cluster), label = Race)) + geom_text()

set.seed(123)

fviz_nbclust(df30, kmeans, method = “wss”)

set.seed(123) final <- kmeans(df30, 2, nstart = 25) print(final) fviz_cluster(final, data = df30)

LDA %>% mutate(Cluster = final*cluster*)cluster) %>% group_by(Cluster) %>% summarise_all(“mean”)

library (MASS) #Original exploratory sections library(ggplot2) library(dplyr) filter (Twogroups, Job == “Engineer”) ggplot(Twogroups)+ aes (x=Race, y = Total) + geom_boxplot (fill = “orchid”) + theme_minimal() hist(subset(Twogroups, Race ==“Was”)$Total, main = "Total females, White/Asian", xlab = "In Engineering") hist(subset(Twogroups, Race =="WOC")$Total, main = “Total females, Women of Color”, xlab = “In Engineering”) shapiro.test(subset(Twogroups, Race == “Was”)$Total) shapiro.test(subset(Twogroups, Race == "WOC")$Total)

#Group means test test <- wilcox.test(Twogroups*Total Twogroups*Race) test

filter (Twogroups2, Job == “Engineer”) ggplot(Twogroups2)+ aes (x=Race, y = Total) + geom_boxplot (fill = “orchid”) + theme_minimal() hist(subset(Twogroups2, Race ==“White”)$Total, main = "Total females, White", xlab = "In Engineering") hist(subset(Twogroups2, Race =="WOC")$Total, main = “Total females, Women of Color”, xlab = “In Engineering”) shapiro.test(subset(Twogroups2, Race == “White”)$Total) shapiro.test(subset(Twogroups2, Race == "WOC")$Total) test <- wilcox.test(Twogroups2*Total Twogroups*2Race) test

hist(subset(Twogroups2, Job ==“White”)$Total, main = “Total females, White”, xlab = “In Engineering”) res <- prop.test

#Exploratory charts library (tidyverse) ggplot(Twogroups, aes (x=Job, y = Total))+ geom_boxplot() ggplot(Twogroups, aes (x=Job, y = Total, color = Race))+ geom_boxplot() install.packages(“ggplot2”) library(ggplot2) subset (Twogroups2, Job ==“Engineer”) ggplot(Twogroups2, aes(x = Race, y = Total)) + geom_bar(stat=“identity”, fill =“blue”) +

```
xlab(‘WOC + Asian’) + ylab(‘Total Female Engineers’) + geom_text(aes(label =
signif(sum(Total)), nudge_y = 2000000)
```