1. Dataset we are using is adult’s dataset. We are taking a subset of adults in order to perform k nearest neighbor. For k nearest neighbor preparation we had partitioned our subset Z in to training set and testing set.

The subset Z consist of 10 rows and 4 columns or fields or attributes of Z. They are age, sex, occupation, income. We partitioned our training set by taking a proportion less than or equal to 0.6 and testing set consists the values which are greater than 0.6.

The next step for k nearest neighbor is to remove the target variable from our training set and testing set. Here income in training set is detached and attached as our target variable and age, sex, occupation are our predictive variables. We need to remove target variable from testing and training because we are going to guess the new instance of the income variable through k nearest neighbor.

Training set and testing set are taken into sub training and sub testing variables. We are taking this because we are converting sub training and sub testing values from categorical to numerical to perform knn. We calculate a new variable instance obtained from knn for validation or evaluation we use testing set income variable again.

We installed compare package in order to compare our estimated or guessed values of k nearest neighbor to compare with our testing $ income values.

Here testing acts as validation or evaluation for k nearest neighbor.

By this we performed k nearest neighbor test for income levels <=50k. >50k

View(adults)

Z<-adults[1:10,c(1,7,10,15)]

Z$part<-runif(length(Z$income),min=0,max=1)

set.seed(1234)

training<-Z[Z$part<=0.4,]

testing<-Z[Z$part>0.4,]

View(training)

View(testing)

target<-training[,c("income")]

subtraining<-training[,c("age","sex","occupation")]

subtesting<-testing[,c("age","sex","occupation")]

subtraining$sex<-as.numeric(subtraining$sex)

subtraining$occupation<-as.numeric(subtraining$occupation)

subtesting$sex<-as.numeric(subtesting$sex)

subtesting$occupation<-as.numeric(subtesting$occupation)

install.packages("class")

library(class)

EstimatedORguessed<-knn(subtraining,subtesting,cl=target,k=3,prob=TRUE,use.all=FALSE)

install.packages("compare")

X1<- data.frame(testing[,4])

X2 <- data.frame(EstimatedORguessed)

comparison <-compare(X1,X2)

head(comparison)

testing...4.

1 <=50K.

2 <=50K.

3 <=50K.

4 >50K.

5 >50K.

$tC

EstimatedORguessed

1 <=50K.

2 <=50K.

3 <=50K.

4 <=50K.

5 <=50K.

So,from above analysis two values out five in estimated values are not equal with testing set. We can conclude that we need to improve our model as nearly 40 percent of data is not matching. (4&5 in testing not equal to 4&5 in estimatedORguessed values).

2. adults$part<-runif(length(adults$income),min=0,max=1)

Z<-adults[1:10,]

install.packages(c("rpart","rpart.plot","C50"))

library("rpart");library("rpart.plot");library("C50")

training<-adults[adults$part<=0.6,]

testing<-adults[adults$part>0.6,]

install.packages("tree")

library(tree)

training1<-training[,-14]

target1<-testing[,15]

cartfit<-rpart(income~.,training1)

print(cartfit)

rpart.plot(cartfit) 