**k-NN Algorithm**

# @authour Nikhil N Pandarge

# calling required packages

**Code:**

library(class)

library(ggplot2)

install.packages("GGally")

library(GGally)

install.packages("ggcorrplot")

library(ggcorrplot)

# data summary

data(iris)

summary(iris)

**Output:**

**> summary(iris)**

**Sepal.Length Sepal.Width Petal.Length Petal.Width Species**

**Min. :4.300 Min. :2.000 Min. :1.000 Min. :0.100 setosa :50**

**1st Qu.:5.100 1st Qu.:2.800 1st Qu.:1.600 1st Qu.:0.300 versicolor:50**

**Median :5.800 Median :3.000 Median :4.350 Median :1.300 virginica :50**

**Mean :5.843 Mean :3.057 Mean :3.758 Mean :1.199**

**3rd Qu.:6.400 3rd Qu.:3.300 3rd Qu.:5.100 3rd Qu.:1.800**

**Max. :7.900 Max. :4.400 Max. :6.900 Max. :2.500**

**Code:**

#------------------------

# Distribution of Variables

iris[,1:4] <- scale(iris[,1:4])

par(mfrow=c(2,2))

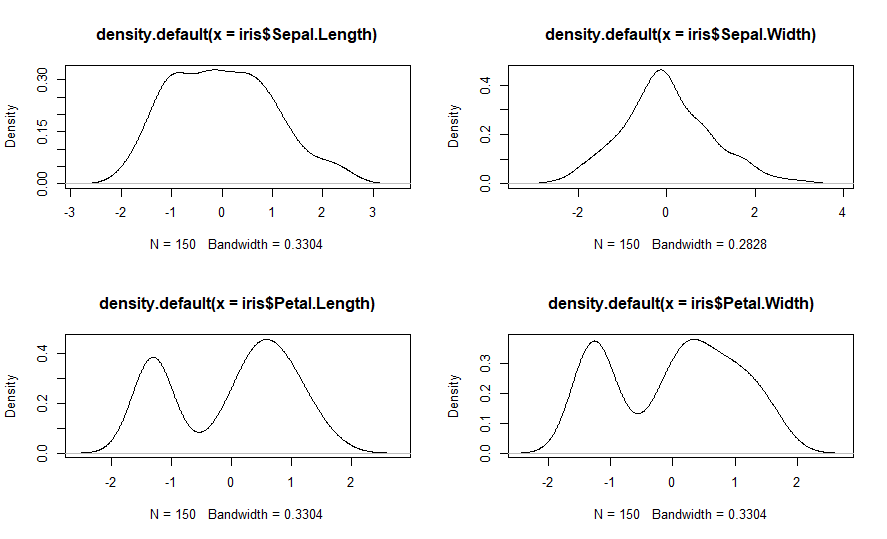
plot(density(iris$Sepal.Length), col=iris$Species)

plot(density(iris$Sepal.Width))

plot(density(iris$Petal.Length))

plot(density(iris$Petal.Width))

**Output:**



**Code:**

par(mfrow=c(2,2))

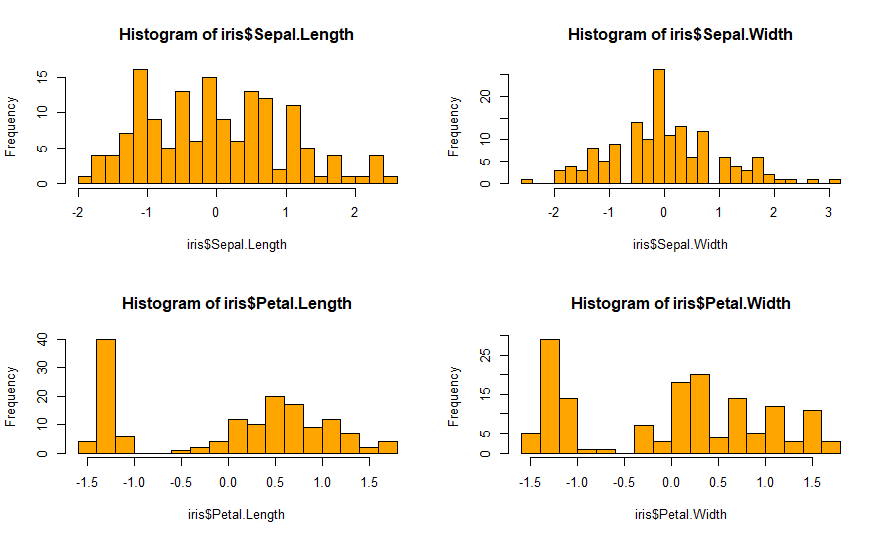
hist(iris$Sepal.Length, col="orange", breaks=20)

hist(iris$Sepal.Width, col="orange", breaks=20)

hist(iris$Petal.Length, col="orange", breaks=20)

hist(iris$Petal.Width, col="orange", breaks=20)

**Output:**

****

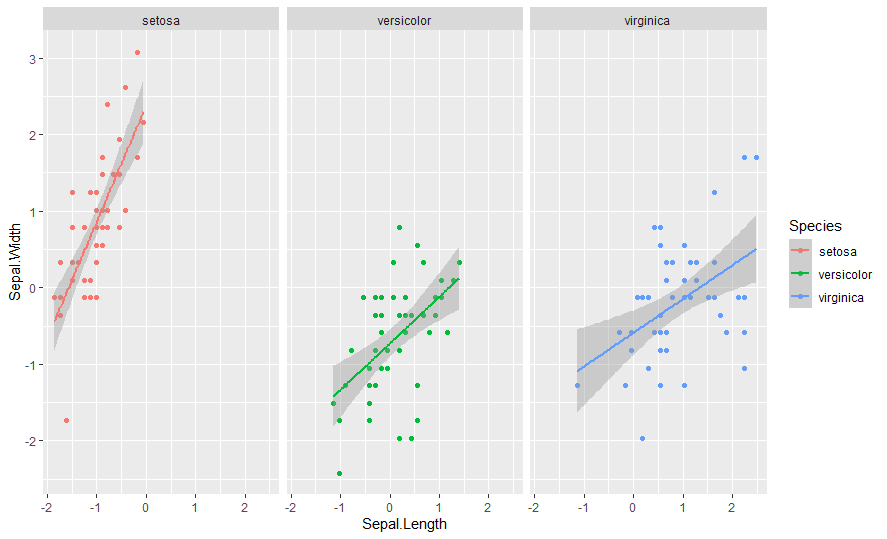
**Code:**

#-----------------------------

# Relationship between Variables:

ggplot(data = iris, aes(x = Sepal.Length, y = Sepal.Width, col = Species))+ geom\_point() + geom\_smooth(method="lm") + facet\_grid(.~iris$Species)

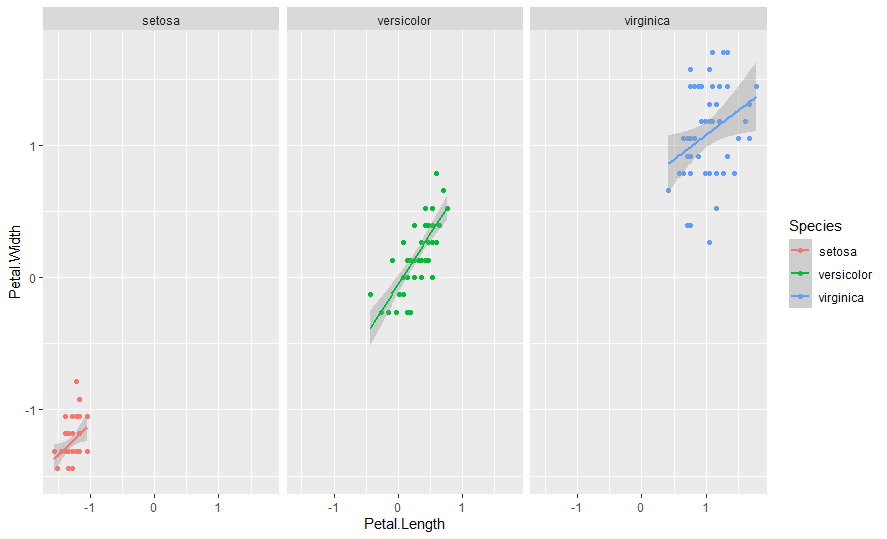
**Output:**



**Code:**

ggplot(data = iris, aes(x = Petal.Length, y = Petal.Width, col = Species)) + geom\_point() + geom\_smooth(method="lm") + facet\_grid(.~iris$Species)

**Output:**

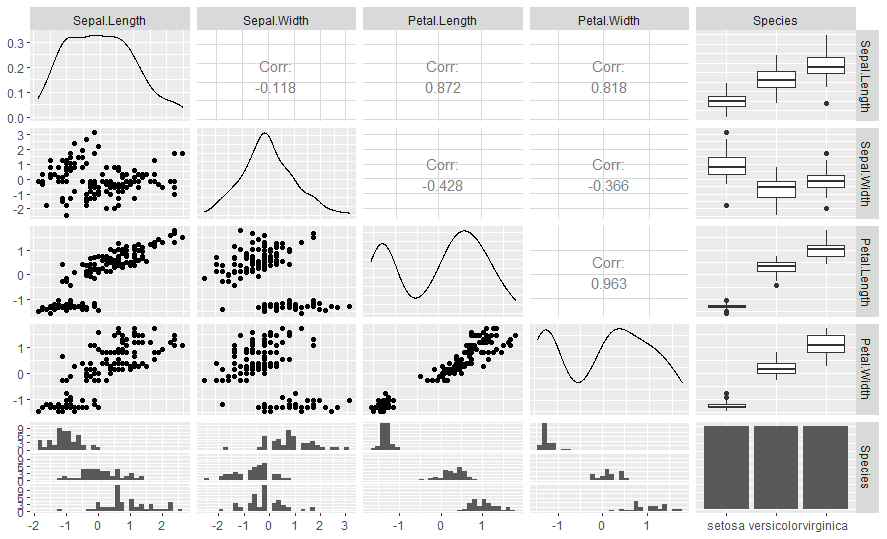


**Code**:

# use the correlation matrix to check the strength of correlation.

ggpairs(iris)

**Output:**

****

**Code:**

# Classification into Different Species using the K-Nearest Neighbor Method

set.seed(12366894)

setosa<- rbind(iris[iris$Species=="setosa",])

versicolor<- rbind(iris[iris$Species=="versicolor",])

virginica<- rbind(iris[iris$Species=="virginica",])

ind <- sample(1:nrow(setosa), nrow(setosa)\*0.6)

iris.train<- rbind(setosa[ind,], versicolor[ind,], virginica[ind,])

iris.test<- rbind(setosa[-ind,], versicolor[-ind,], virginica[-ind,])

# Choosing the Value of K

error <- c()

for (i in 1:15)

{

knn.fit <- knn(train = iris.train[,1:4], test = iris.test[,1:4], cl = iris.train$Species, k = i)

error[i] = 1- mean(knn.fit == iris.test$Species)

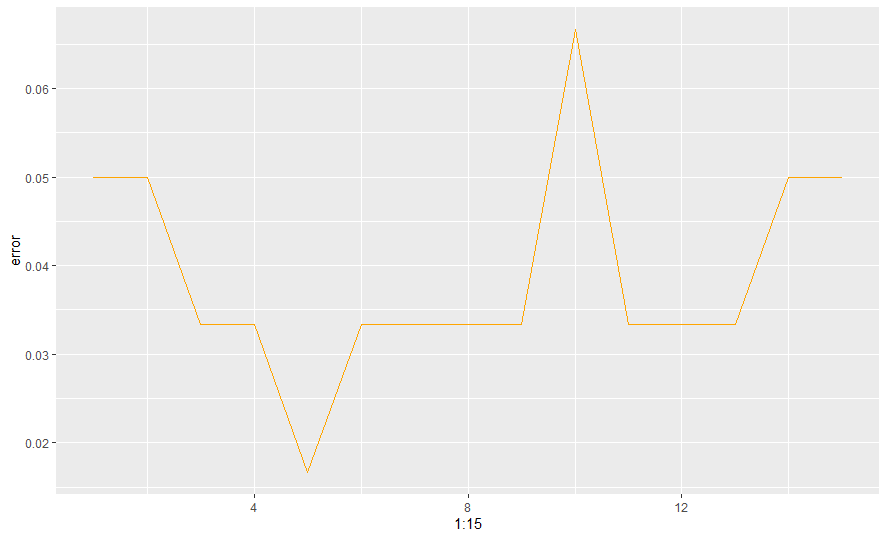
}

ggplot(data = data.frame(error), aes(x = 1:15, y = error)) +

geom\_line(color = "orange")

# Prediction Accuracy using the chosen K value

**Output:**

****

**Code:**

#------------------------

set.seed(12366894)

iris\_pred <- knn(train = iris.train[,1:4], test = iris.test[,1:4], cl = iris.train$Species, k=5)

table(iris.test$Species,iris\_pred)

**Output:**

**> table(iris.test$Species,iris\_pred)**

**iris\_pred**

**setosa versicolor virginica**

**setosa 20 0 0**

**versicolor 0 20 0**

**virginica 0 1 19**