Clasificación k-NN con Iris

Introducción

El conjunto de datos de Iris contiene 150 observaciones y 5 variables. Tenemos 50 flores de cada especie.

Las variables longitud del sépalo, ancho del sépalo, longitud del pétalo y ancho del pétalo son variables cuantitativas que describen la longitud y el ancho de las partes de las flores en cm.
La variable Especie es categórica y consta de tres especies diferentes, a saber, setosa, versicolor y virginica.

Realizamos un análisis exploratorio del conjunto de datos y construimos un modelo de clasificación utilizando el método de los K vecinos más cercanos

Paquetes necesarios

Necesitamos los siguientes paquetes para nuestro análisis (si no los tienes instalados, hay que instalarlos primero):

```
> library(class)
> library(ggplot2)
> library(GGally)
```

Resumen de estadísticas

La siguiente tabla contiene el resumen de estadísticas del conjunto de datos. También podemos comprobar la varianza de cada variable.

```
> 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
```

```
> apply(iris[,1:4], 2, sd)
```

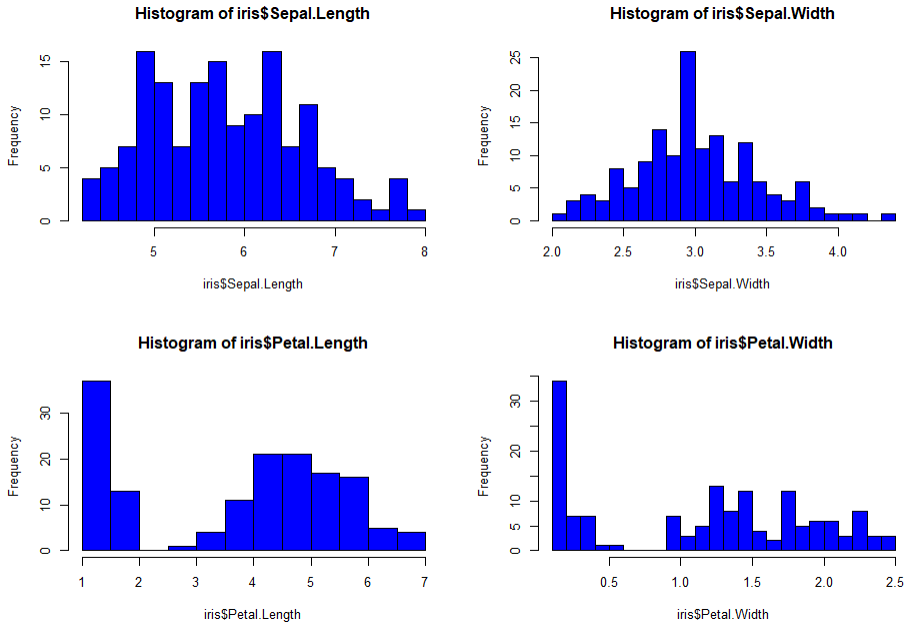
```
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
   0.8280661    0.4358663    1.7652982    0.7622377
```

Visualización de datos

Gráficos de histograma

A continuación, se muestra el histograma que muestra la distribución de las variables cuantitativas Longitud del sépalo, Ancho del sépalo, Longitud del pétalo y Ancho del pétalo.

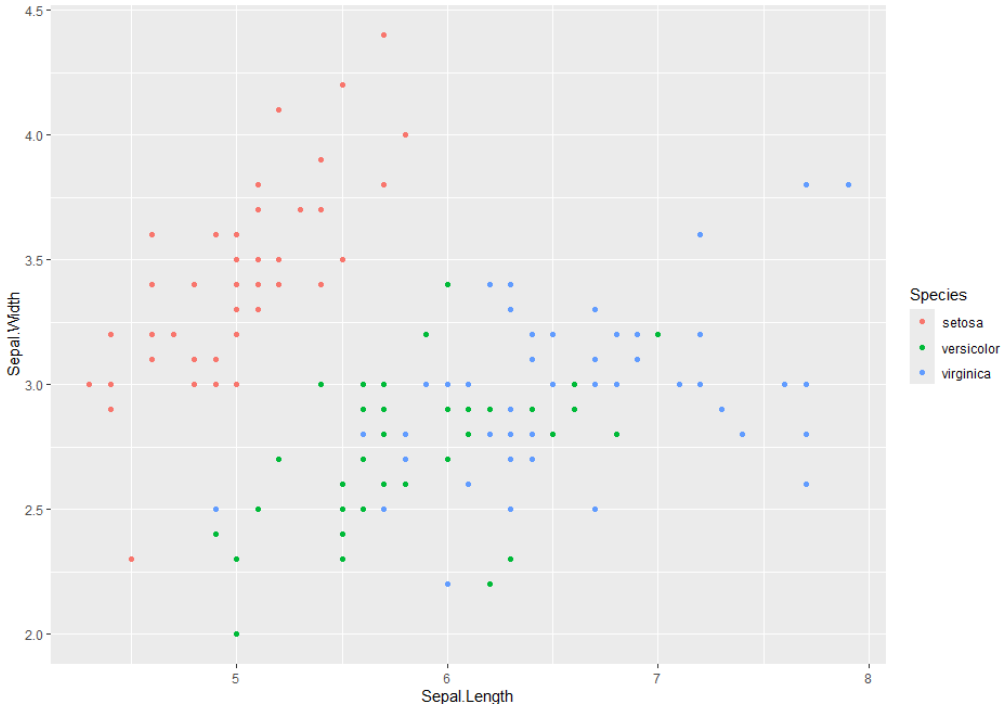
```
> par(mfrow=c(2,2))
> hist(iris$Sepal.Length, col="blue", breaks=20)
> hist(iris$Sepal.Width, col="blue", breaks=20)
> hist(iris$Petal.Length, col="blue", breaks=20)
> hist(iris$Petal.Width, col="blue", breaks=20)
```



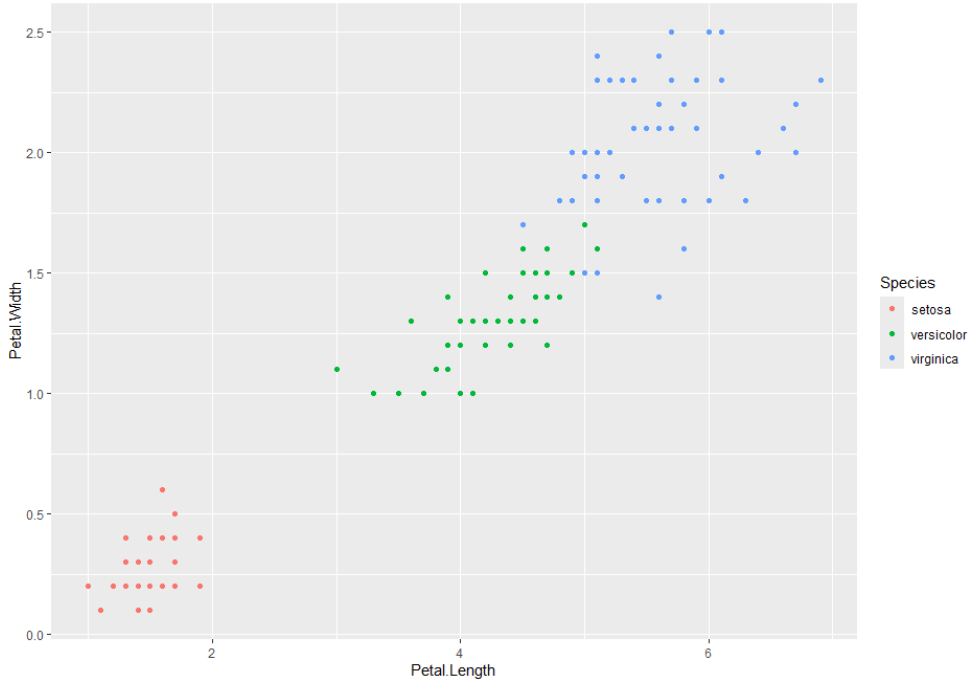
Gráficos de dispersión

Comprobación de la distribución de la anchura del sépalo frente a la longitud del sépalo y de la anchura del pétalo frente a la longitud del pétalo.
Virginica tiene el valor máximo de anchura y longitud del pétalo

```
> ggplot(data = iris, aes(x = Sepal.Length, y = Sepal.Width, col = Species)) +
+   geom_point()
```



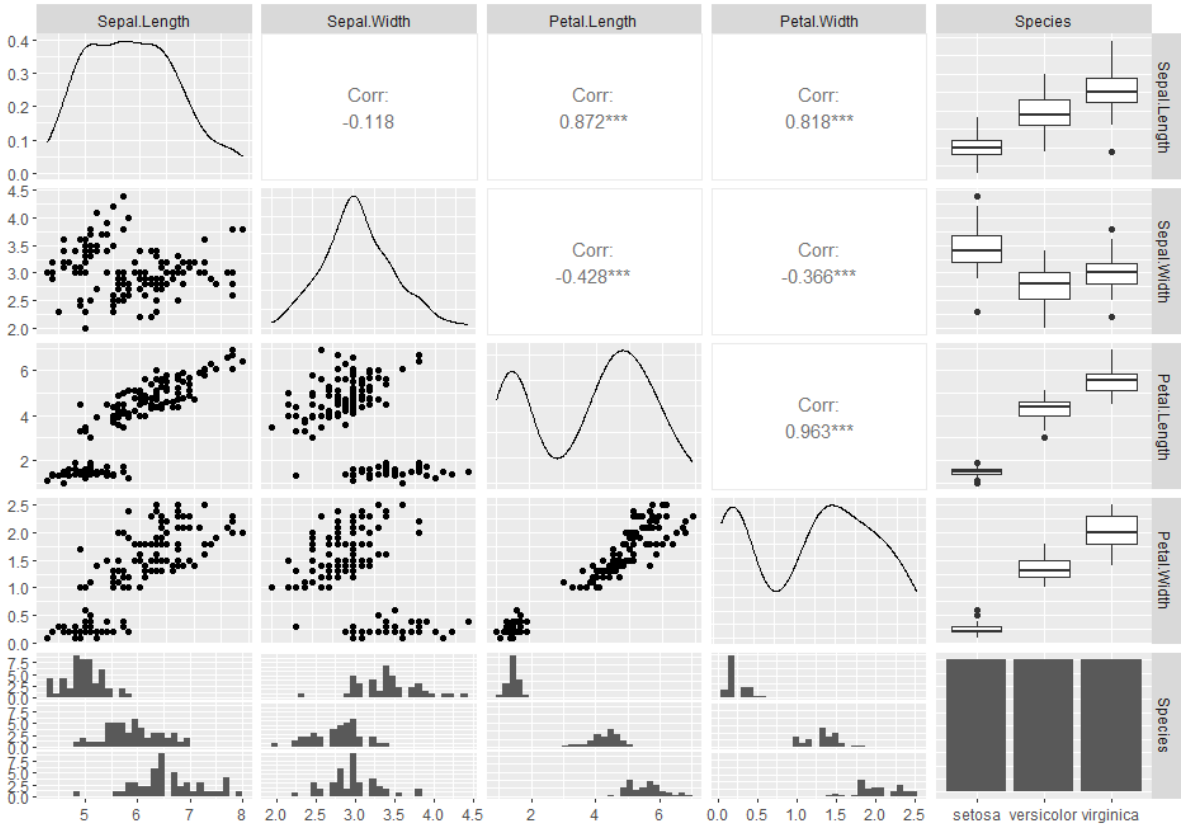

```
> ggplot(data = iris, aes(x = Petal.Length, y = Petal.Width, col = Species)) +
+   geom_point()
```



Matriz de correlación

```
> ggpairs(iris)
```

A partir del gráfico de correlación, observamos que:



Existe una fuerte correlación positiva entre la longitud del pétalo y el ancho del pétalo con un coeficiente de correlación de 0.963.
Existe una fuerte correlación positiva entre el ancho del pétalo y la longitud del sépalo con un coeficiente de correlación de 0.818.
Existe una fuerte correlación positiva entre la longitud del pétalo y la longitud del sépalo con un coeficiente de correlación de 0.872.

Clasificación mediante K_NN

División del conjunto de datos

Ahora dividimos el conjunto de datos de Iris en un conjunto de datos de entrenamiento y uno de prueba para aplicar la clasificación K_NN. El 80 % de los datos se utiliza para el entrenamiento, mientras que la clasificación K_NN se prueba en el 20 % restante de los datos.

```
> set.seed(12420352)
> iris[,1:4] <- scale(iris[,1:4])
> 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)*0.8)
> iris.train<- rbind(setosa[ind,], versicolor[ind,], virginica[ind,])
> iris.test<- rbind(setosa[-ind,], versicolor[-ind,], virginica[-ind,])
> iris[,1:4] <- scale(iris[,1:4])
```

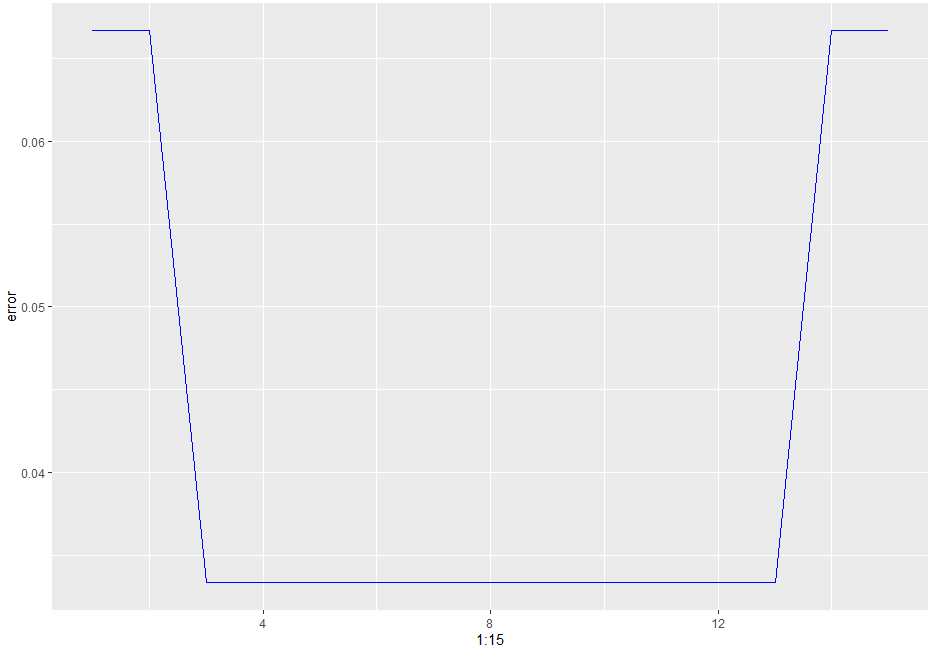
Búsqueda del valor óptimo de K

El siguiente gráfico muestra el error de clasificación para diferentes valores de k. Vemos que el error disminuye inicialmente, pero luego comienza a permanecer constante y aumenta.

```
> 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)
+ }
```

En el gráfico siguiente, vemos que el error mínimo se produce cuando el valor de k es igual a 3 y permanece constante hasta 13. Elegimos el modelo menos complejo y optamos por k = 3.

```
> ggplot(data = data.frame(error), aes(x = 1:15, y = error)) +
+   geom_line(color = "Blue")
```



Matriz de confusión

El error mínimo se observa en k=3. Obtenemos la matriz de confusión y la precisión del modelo:

```
> iris_pred <- knn(train = iris.train[,1:4], test = iris.test[,1:4], cl = iris.train$Species, k=3)
> table(iris.test$Species,iris_pred)
```

	iris_pred		
	setosa	versicolor	virginica
setosa	10	0	0
versicolor	0	9	1
virginica	0	0	10

Observamos que la precisión de predicción cuando k=3 es superior al 90%.

Por lo tanto podemos crear un clasificador k-NN que ofrece una predicción superior al 90% en el conjunto de los datos de prueba.