COLEGIO UNIVERSITARIO DE ESTUDIOS FINANCIEROS CUNEF
Máster en Data Science para Finanzas
Técnicas de Agrupación y de Reducción de la Dimensión

Nombre: Paúl Córdova

Informe Ejecutivo

El presente informe tiene la finalidad de analizar el set de datos iris; logrando así, una comprensión de la información e identificando los aspectos más relevantes. En este proceso se determinó que las variables largo y ancho del pétalo actúan como diferenciadores en relación a las especies de las flores; lo que permitió gestionar técnicas de clasificación con alta precisión.

Informe Nº 1 – iris Dataset

La base de datos “iris” es una de las utilizadas tanto por su practicidad y la calidad de la información que posee. Se observó que el conjunto de datos (Anderson,1935) cuenta con 150 observaciones y 5 variables, en este sentido, se alberga información de flores según: largo del sépalo, ancho del sépalo, largo del pétalo, ancho del pétalo y especie (iris setosa, iris virginica e iris versicolor). Al realizar un análisis descriptivo, se destacó que la media en el largo y ancho del sépalo fue superior entre 1 y 2 centímetros con respecto al largo y ancho del pétalo respectivamente.

En un análisis por variable se identificó que el largo y ancho del sépalo aparentan una distribución normal en sus datos (simetría en la información); mientras que, el largo y ancho del pétalo concentran su información en sus valores iniciales y finales (distinguiéndose dos distribuciones). Simultáneamente, se realizó una comparación gráfica en base a la variable cualitativa (especie), lo cual permitió identificar diferencias según el tamaño (largo y ancho) del sépalo; siendo más evidentes al analizar por el tamaño del pétalo.

A partir de las observaciones preliminares, se determinó a la variable especie como un diferenciador en el conjunto de datos. Por consiguiente, se construyó un gráfico de densidad conjunto de cada variable aplicando una distinción según la especie, como resultado se obtuvo que las variables del largo y ancho del pétalo, apartan la distribución de la especie iris setosa con respecto a las dos restantes; y se corroboró mediante un análisis de dispersión entre las variables continuas. Para reforzar este aspecto, se realizó el Análisis Discriminante Lineal (ADL) con el objetivo de destacar la clasificación de las flores según su especie, evidenciando que los datos permiten una clasificación acertada con apenas un error del 2% (3 errores en 150).

Para concluir, se observó que las variables largo y ancho del pétalo actúan como diferenciadores en relación a las especies de las flores; aspecto que permitió gestionar técnicas de clasificación en los datos (ADL) con una alta precisión.

Referencias bibliográficas o digitales:

- The data were collected by Anderson, Edgar (1935). The irises of the Gaspe Peninsula, Bulletin of the American Iris Society, 59, 2–5.

ANEXOS

A. Script Análisis Iris Dataset

```
### Práctica IRIS DATASET ###
data("iris")

#Identificamos las variables y dimensiones
str(iris)
dim(iris)

#Análisis descriptivo variables
for (i in 1:4) {
  est<-summary(iris[, i])
  print(names(iris[i]))
  print(est)
}

#Histograma según variable
for (i in 1:4) {
  h <-hist(iris[,i], main=names(iris[i]),
           xlab="Tamaño", ylab="Frecuencia")
}

#Densidad según variable
for (i in 1:4) {
  density= density(iris[,i])
  plot(density, main = names(iris[i]))
}

#Diagrama de caja según especies
for (i in 1:4) {
  boxplot(iris[, i]~Species, data=iris, main=names(iris[i]),
          xlab="Especie",ylab="Tamaño")
}

#Instalación y carga de paquetes
install.packages("sm")
library(sm)
attach(iris)

#Densidad según especie (en conjunto) por variable
especie.f <- factor(iris$Species, levels= c(1,2,3),
                labels = c("setosa", "versicolor", "virginica"))
for (i in 1:4) {
  sm.density.compare(iris[,i], iris$Species, xlab="Tamaño")
  title(main=names(iris[i]))
}

#Gráfic dispersión según variables (distinguiendo especie)
pairs(iris[1:4], main = "Dispersión según especie",
      pch = 21, bg = c("red", "black", "brown")[unclass(iris$Species)])

#Análisis Discriminante Lineal
install.packages("MASS")
library(MASS)

ADL<- lda(Species~Sepal.Length+Sepal.Width+Petal.Length+Petal.Width,
          data=iris)
ADL

prediccion <- predict(object = ADL, newdata = iris[, 1:4])
str(prediccion)
table(iris$Species, prediccion$class, dnn = c("Clase True", "Clase Prediction"))

p.error<-(3/150)*100
p.error #El porcentaje de error es de 2%
```