

Big Data.

Anexo 1

José Angel Lozano Pineda

Contenido

[1) Introducción 3](#_Toc59730017)

[2) Ejercicio 1 3](#_Toc59730018)

[3)Ejercicio 2 3](#_Toc59730019)

[a) Apartado a 3](#_Toc59730020)

[b) Apartado b 4](#_Toc59730021)

[c) Apartado c 4](#_Toc59730022)

[3) Ejercicio 3 4](#_Toc59730023)

# 1) Introducción

En este documento se encontrará brevemente mostrada la organización del proyecto y los ficheros generados. Para ello, vamos a seguir el orden del documento de requisitos del proyecto.

El proyecto consta de tres carpetas para la realización de los ejercicios, y una carpeta ficheros, en la cual se encuentran los ficheros necesarios para realizar los ejercicios, y se crean los ficheros dependientes para pasos próximos, de esta manera se evita tener que estar manualmente relocalizando ficheros para poder utilizar en los diferentes pasos, automatizando este proceso.

No entraremos en aspectos técnicos dado que el código esta comentado en detalle.

# 2) Ejercicio 1

Este ejercicio esta desarrollado dentro del proyecto de scala, en el fichero “Parte 2/1a-2b-2c/Parte2/src/Ejercicio1.scala”

Este ejercicio necesita el fichero consumo.csv para realizar el procesamiento de los datos, generando como resultado el fichero “predicción.csv”

# 3)Ejercicio 2

## a) Apartado a

Este ejercicio esta desarrollado en R, en el fichero “Parte 2/ejercicio2a/ejercicio2a.R”.

En este ejericcio hacemos uso del fichero “prediccion.csv” generado anteriormente, desde el cual hacemos una partición para los datos de entrenamiento y prueba, creando los ficheros “2a-training\_set.csv” y “2a-test\_set.csv”.

Finalmente, tras el proceso de entrenamiento del modelo, se genera una predicción para el conjunto de datos de entrenamiento, con el cual calculamos errores, y el conjunto de test, almacenados en los ficheros “2a-prediction\_training\_set.csv” y “2a-prediction\_test\_set.csv” respectivamente, utilizados para posteriormente poder analizarlos y crear graficas.

## b) Apartado b

Este ejercicio esta junto al del ejercicio 1, dentro del proyecto de scala, en el fichero “Parte 2/1a-2b-2c/Parte2/src/Ejercicio2b.scala”.

Este ejercicio usa el mismo set de datos que el apartado a, por lo que se usa los ficheros generados “2a-training\_set.csv” y “2a-test\_set.csv” y adicionalmente, ya que posteriormente se guarda los datos de la tabla de comparación, necesita las predicciones realizadas con nnet del fichero “2a-prediction\_test\_set.csv”

Tras el entrenamiento del modelo, exportamos la definición de este, con todos los arboles que se crean, en el fichero “2b-random\_forest.txt”, de la misma manera que la predicción para los datos de pruebas, dentro de “2b-test\_prediction.csv”, las predichas para el set de entrenamiento en “2b-training\_prediction.csv” y la comparativa de los resultados de nnet y random forest, dentro del fichero “2b-comparativa\_RandomTree\_NNET.csv”

## c) Apartado c

Este ejercicio esta junto al del ejercicio 1 y 2b, dentro del proyecto de scala, en el fichero “Parte 2/1a-2b-2c/Parte2/src/Ejercicio2c.scala”.

En este ejercicio ase usa el fichero “consumo.csv”, con el que se crearan los 5 clusters, y genera los ficheros “consumo.csv”, donde se almacena para cada entrada del fichero “consumo.csv” el cluster al que pertenece con las columnas Sensor, fecha y label que se piden en el ejercicio, el fichero “2c-clusters-medias.csv” que es similar al anterior, pero que incluye la media de las medidas, para la generación de las gráficas, y finalmente, el fichero “2c-clusters-centros” que tiene la media de las dimensiones de los centroides de los clusters.

# 3) Ejercicio 3

En este apartado generamos todos los grafos necesarios para el informe y análisis de los resultados, que esta dividido en dos partes, un proyecto Python y un script en R.

El proyecto de Python tiene un script simple y el fichero de requisitos, ya que lo que hace el script es formatear el fichero “2b-random\_forest.txt” y ejecutar el comando de la librería eurekatrees con el fichero formateado, y esta genera en html todos los arboles del modelo creado en el apartado pasado, y los almacena en la carpeta “Parte 2/1a-2b-2c/Parte2/ejercicio3/Graficas/random\_forest”.

Por otro lado, el script de R genera el resto de gráficas, haciendo uso de la librería ggplot2 y los ficheros de predicciones que hemos ido almacenando en todos los apartados pasados.