Proyecto de Estadística Segunda Fase - Household

Loraine Monteagudo García L.MONTEAGUDO@ESTUDIANTES.MATCOM.UH.CU
Grupo C411

Amanda Marrero Santos A.MARRERO@ESTUDIANTES.MATCOM.UH.CU
Grupo C411

Manuel S. Fernández Arias M.FERNANDEZ2@ESTUDIANTES.MATCOM.UH.CU
Grupo C411

Tutor(es):
Msc. Dalia Díaz Sistach, *Universidad de la Habana, Facultad de Matemática y Computación*

1. Introducción

El estudio de la estadística es de vital importancia para el desarrollo de la sociedad moderna. Este proyecto usa los datos de UCI Machine Learning Repository, en particular el data set "Individual Household Electric Power Consumption". Se analiza un ejemplo real relacionado con el consumo eléctrico de una casa durante 4 años. Nuestro objetivo será realizar un estudio de estos datos usando técnicas de regresión, lo que nos permitirá determinar si existen relaciones lineales entre algunas de las variables dadas. Luego reduciremos la dimensión de los datos mediante el análisis de componentes principales (ACP), clusters y árboles de clasificación, haciendo además una interpretación de los mismos. Por último realizaremos un análisis de varianza o ANOVA para comparar las medias de una característica en varias poblaciones.

1.1 Datos

El conjunto de datos contiene 2075259 mediciones del consumo eléctrico recolectadas en una casa entre diciembre de 2006 y noviembre de 2010 (47 meses).

El data set contiene 9 variables que son:

- *date*: La fecha en el formato dd/mm/yy
- *time*: El tiempo en el formato hh:mm:ss
- *global_active_power*: la corriente global activa de una casa promediada por minuto (en kilowatt). La corriente activa global es la potencia consumida por dispositivos distintos de los dispositivos asignados a los submedidores. La potencia activa global es el consumo de energía real, es decir, la consumida por aparatos eléctricos distintos de los aparatos submedidos.
- *global_reactive_power*: la corriente global reactiva de una casa promediada por minuto (en kilowatt). La corriente reactiva global es la potencia que rebota y hace espuma sin ningún uso o fuga. Es el consumo de energía imaginario.
- *voltage*: el voltaje promediado por minuto (en volt)
- *global_intensity*: la intensidad promediada por minuto de toda la casa (en kilowatt). Es la magnitud de la energía consumida. También llamada como fuerza de la corriente.
- *sub_metering_1* (submedida1): Le corresponde a la cocina, conteniendo el lavavajillas, un horno y un microwave (en watt-hora de energía activa)
- *sub_metering_2* (submedida2): Le corresponde al cuarto de lavado, conteniendo la lavadora, una secadora, un refrigerador y una luz (en watt-hora de energía activa)
- *sub_metering_3* (submedida3: Le corresponde a un calentador de agua eléctrico y un aire acondicionado (en watt-hora de energía activa)

Como se puede apreciar, el data set tiene el inconveniente de que las variables están en medidas diferentes, lo que dificulta su comparación. Además, presenta valores perdidos, lo que ocasionó que se tuvieran que eliminar varias filas. Aún así, se tienen gran cantidad de datos con 2049280 observaciones reales, lo que dificultó su procesamiento y su análisis. Por ejemplo, varios diagramas de dispersión y boxplots mostraron muy poca información al tener tantos puntos para representar.

2. Técnicas de Regresión

La idea detrás de la regresión lineal simple es encontrar el modelo lineal que mejor se ajusta a los datos. Para decidir si nuestros datos se ajustan al modelo de Regresión Lineal Simple es necesario que se cumpla una serie de suposiciones a las que se les llama supuestos del modelo. Este modelo puede ser expresado como: $Y = X\beta + \epsilon$

Para elegir las variables con las que realizaremos la regresión primero veamos su relación mediante la matriz de correlación mostrada de manera simplificada en la Figura 1

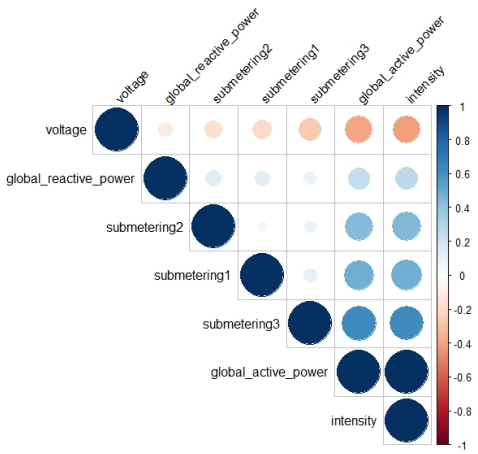

Figura 1: Forma gráfica de la matriz de correlación

Se observa una relación lineal entre la Corriente Global Activa y la Intensidad, siendo su coeficiente de correlación cercano a 1. Esta suposición se respalda por el diagrama de dispersión de la Intensidad y la Corriente Global Activa mostrada en la Figura 2.

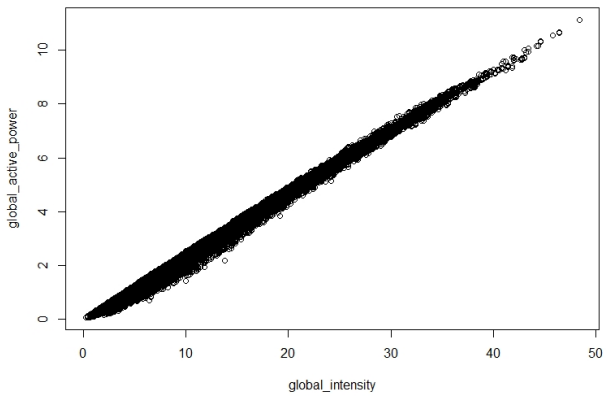

Figura 2: Diagrama de dispersión de la intensidad y la corriente global activa

Por lo tanto, se decide realizar la regresión lineal entre la Intensidad y la Corriente Global Activa. El resultado al ejecutar dicha regresión fue el siguiente:

```
Call:
lm(formula = global_active_power ~ global_intensity, data = household)

Residuals:
     Min       1Q   Median       3Q      Max
-1.07921 -0.01966  0.00645  0.02940  0.27324

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -8.076e-03  5.026e-05  -160.7   <2e-16 ***
global_intensity  2.376e-01  7.833e-06 30338.1   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04983 on 2049278 degrees of freedom
Multiple R-squared:  0.9978,    Adjusted R-squared:  0.9978
F-statistic: 9.204e+08 on 1 and 2049278 DF,  p-value: < 2.2e-16
```

Figura 3: Salida del Modelo de Regresión múltiple

Los valores de β_0 y β_1 se obtienen de la columna de estimados. El modelo con los coeficientes sustituidos sería:

$$gap = -8.07e^{-3} + 0.237 * global\_intensity$$

Análisis del *summary*:

- *Residuals*: Es el error entre la predicción del modelo y los resultados reales. Se observan valores relativamente pequeños, lo que contribuye a la robustez del modelo.

- *Coefficients*:

  - *Std Error*: tiene valores bajos, lo que parece indicar una buena precisión del estimador

  - *t-value* y $Pr(> t)$: Como $Pr(> t)$ es menor que 0.05 entonces la prueba es significativa

- *Performance Measures*

  - *Residual Standard Error*: Se obtienen pequeños valores

  - *Multiple / Adjusted R-Square*: Estos valores son bastantes cercanos a 1, por lo que el modelo explica bastante la variación.

  - *F-Statistic*: El valor del estadígrafo F es menor que 0.05 así que podemos afirmar que existe al menos una variable significativamente diferente a cero en el modelo.

A la hora de analizar los residuos para comprobar que se cumplen los supuestos del modelo tenemos que analizar cuatro cuestiones con respecto a los residuos:

1. La media y la suma de los errores es cero.

   Como se puede observar en el siguiente código este supuesto se cumple:

   ```
   > mean(model$residuals)
   -8.040966e-16
   > sum(model$residuals)
   -1.643985e-09
   ```

2. Errores normalmente distribuidos.

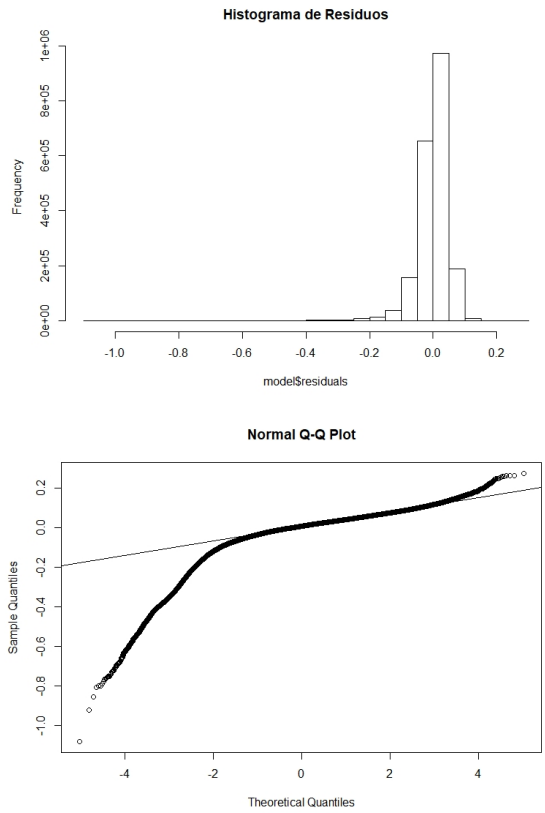

Figura 4: Salida de los residuos del modelo de Regresión Múltiple

El histograma de residuos y el QQ-plot mostrados en la Figura 2 son formas de evaluar visualmente si los residuos siguen una distribución normal. El histograma se asemeja a la distribución normal siguiendo el patrón de campana, sin embargo, el QQ-plot muestra bastante desviación con respecto a la línea normal, por lo tanto, para asumir la normalidad de los errores nos apoyamos en el test de Shapiro-Wilk:

```
        Shapiro-Wilk normality test

data:  sample(res, 5000)
W = 0.94654, p-value < 2.2e-16
```

Figura 5: Prueba de normalidad de Shapiro-Wilk

Se tomó una muestra de 5000 residuos ya que es la cantidad máxima de valores que puede recibir este test. Como $p - value = 2.2e^{16} < 0.05$ entonces no se puede rechazar la hipótesis nula, por lo que los errores no siguen una distribución normal.

Solo por un propósito ilustrativo continuaremos analizando los supuestos pero ya al incumplirse el de normalidad podemos concluir que el modelo no es válido.

3. Independencia de los residuos

La prueba Durbin-Watson se utiliza para verificar si los residuos son independientes. La hipótesis nula de esta prueba es que los errores son independientes.

```
        Durbin-Watson test

data:  data.anova
DW = 0.14871, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

Figura 6: Prueba para la independencia de los residuos de Dubin-Watson

Como se observa en la Figura 6 el $p-value$ de esta prueba es menor que $2.2e-16 < 0.05$ entonces podemos rechazar la hipótesis nula, por lo que los errores no son independientes, incumpliéndose el supuesto de independencia.

4. La varianza de los errores es constante (Homocedasticidad)

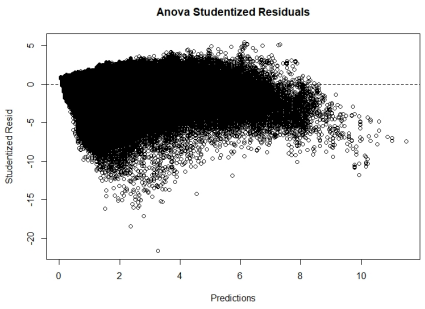

Figura 7: Gráfico estandarizado de los residuos

Por el gráfico de los residuos estandarizados mostrado en la Figura 7 no parece que el supuesto de Homocedasticidad se cumpla.

Lo comprobamos con la prueba de Breusch-Pagan:

```
        studentized Breusch-Pagan test

data:  model.fit
BP = 74708, df = 1, p-value < 2.2e-16
```

Figura 8: Gráfico estandarizado de los residuos

Como el $p - value = 2.2e - 16 < 0.05$ se puede rechazar la hipótesis nula, por lo tanto no se cumple la Homocedasticidad tampoco.

Con el objetivo de encontrar un modelo de regresión lineal que cumpliese los supuestos se probaron distintas combinaciones de variables adicionales aplicando regresión múltiple. Además, se intentó excluir registros con el objetivo de disminuir el tamaño de la muestra. Sin embargo, no se encontró una combinación que cumpliese con dichos supuestos.

3. Reducción de dimensión

3.1 Análisis de Componentes Principales (ACP)

Una de las formas de realizar la reducción de la dimensión es a través del análisis de componentes principales (ACP). Se busca obtener de las n variables de un sistema un subconjunto de tamaño k, las cuales provean tanta información como las n variables originales. Este análisis tiene el objetivo de revelar relaciones entre variables de las que no se tenía sospecha, permitiendo interpretaciones que realizando un análisis ordinario quizá no encontraríamos. Por ejemplo, podemos obtener mediante este estudio variables incorrelacionadas entre sí.

En nuestro data set hay un total de 9 variables, el análisis de los componentes principales se realiza con las variables que tienen un valor numérico, por lo que excluimos la fecha y la hora. Por lo tanto analizamos: la corriente global activa, la corriente global reactiva, el voltaje, la intensidad global y otras mediciones secundarias como el consumo activo de energía en watt-hora de la cocina (submedida1), el cuarto de lavado (submedida2) y el correspondiente a un calentador de agua y un aire acondicionado (submedida3).

Para determinar la relación entre las variables tenemos que calcular la matriz de correlación que se muestra en la Figura 9.

Por desgracia esta matriz es muy grande por lo que resulta fácil perdernos, para un mejor análisis se usa la función `symnum` obteniendo:

	GAP	GRP	I	V	S1	S2	S3
Corriente Activa	1						
Corriente Reactiva		1					
Intensidad	B		1				
Voltaje	.		.	1			
Submedida1	.		.		1		
Submedida2	.		.			1	
Submedida3	,		,				1

[1] 0 ‘ ’ 0.3 ‘.’ 0.6 ‘,’ 0.8 ‘+’ 0.9 ‘*’ 0.95 ‘B’ 1

Cuadro 1: Matriz de correlación usando `symnum`

Como se puede observar en el Cuadro 1 y en la Figura 1 no es altamente correlacionada la matriz, exceptuando por algunas variables como la intensidad y la corriente global activa que están altamente relacionadas.

Importancia de los componentes:

Se puede observar en la Figura 10 que con las 4 primeras componentes podemos explicar el 85 % de la muestra, sin embargo, es suficiente con una proporción acumulativa superior al 70 %, por lo que solo nos quedaríamos con las tres primeras. Según el criterio Kaiser elegiríamos las componentes q tengan valor propio λ_m mayor que 1, sin embargo, esto solo lo cumple la primera componente, pero se puede tomar un punto de corte $\lambda^* = 0.7$ en cuyo caso se cumpliría para las 5 primeras. De acuerdo al análisis realizado y al gráfico de la Figura 11 nos quedamos con las 3 primeras ya que no se aprecia que haya mucha variabilidad en las componentes restantes.

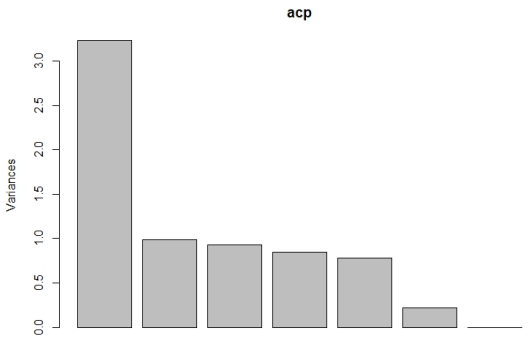

Figura 11: Gráfico de las Componentes Principales

Ahora analizamos la matriz de valores propios y así sabremos qué variables son importantes para cada componente y en qué medida.

	PC1	PC2	PC3
Corriente Activa	-0.5381	0.0511	0.0298
Corriente Reactiva	-0.1868	-0.6704	-0.3403
Intensidad	-0.5398	0.0317	0.0191
Voltaje	0.2961	-0.1424	-0.0595
Submedida1	-0.2953	0.1005	-0.7409
Submedida2	-0.2661	-0.5733	0.5005
Submedida3	-0.3721	0.4336	0.2828

Cuadro 2: Matriz de valores propios

Comencemos por la primera componente, tomamos el mayor valor propio modular 0.2953 y lo dividimos entre 2, esto da 0.148 todo valor propio cuyo modulo esté por encima de 0.148 en la columna de PC1 nos dará las variables que conforman esta componente. Por tanto, la interpretación sería que la PC1 está caracterizada por un impacto negativo de todas las variables involucradas. Lo que quiere decir que es una muestra con baja corriente global activa, baja corriente global reactiva, baja intensidad y la corriente consumida en las submediciones 1, 2 y 3 son bajas también.

Siguiendo el mismo análisis con la segunda componente el máximo valor propio es 0.4336, por lo que existe una baja corriente global reactiva y un bajo consumo de energía en la submedición 2 correspondiente al cuarto de lavado.

La tercera componente nos muestra un alto consumo en la submedida 3 correspondiente al calentador de agua eléctrico y el aire acondicionado y bajos valores de corriente global reactiva y en la submedida 1 correspondiente a la cocina.

3.2 Clusters

Otra técnica popular de reducción de dimensiones es el clúster. La idea básica del clúster es clasificar objetos

	global_active_power	global_reactive_power	intensity	voltage	submetering1	submetering2	submetering3
global_active_power	1.0000	0.2470	0.9989	-0.3998	0.4844	0.4346	0.6386
global_reactive_power	0.2470	1.0000	0.2661	-0.1122	0.1231	0.1392	0.0896
intensity	0.9989	0.2661	1.0000	-0.4114	0.4893	0.4403	0.6265
voltage	-0.3998	-0.1122	-0.4114	1.0000	-0.1960	-0.1674	-0.2682
submetering1	0.4844	0.1231	0.4893	-0.1960	1.0000	0.0547	0.1026
submetering2	0.4346	0.1392	0.4403	-0.1674	0.0547	1.0000	0.0809
submetering3	0.6386	0.0896	0.6265	-0.2682	0.1026	0.0809	1.0000

Figura 9: Matriz de correlación

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	1.7972	0.9937	0.9657	0.9215	0.8845	0.46642	0.02713
Proportion of Variance	0.4614	0.1411	0.1332	0.1213	0.1118	0.03108	0.00011
Cumulative Proportion	0.4614	0.6025	0.7357	0.8570	0.9688	0.99989	1.00000

Figura 10: Importancia de los componentes

formando grupos que sean lo más homogéneos posibles dentro de sí mismos y heterogéneos entre sí. Por tanto, el número de clusters dependen de lo que consideremos como similar. El análisis de clúster es una tarea de clasificación que puede ser vista desde dos puntos de vistas diferentes:

1. Clúster por particiones: se usan cuando conocemos cuántos grupos hay. Producen una partición de los objetos en un número especificado de grupos siguiendo un criterio de optimización. El algoritmo usado para generar este tipo de clusters es el de *K-Means*

2. Clúster Jerárquico: se usan cuando no conocemos cuántos grupos hay. Producen una secuencia de particiones, juntando o separando clusters. En cada paso se juntan o separan dos clusters siguiendo algún criterio especificado.

Para realizar el análisis de los datos por clúster al igual que con ACP, se eligieron las variables que tienen un valor numérico, excluyendo la fecha y la hora.

Antes de elegir el tipo de clúster vamos a estandarizar los datos para evitar errores en la clasificación por cuestiones de variabilidad en las unidades de medida. Para esto usaremos la fórmula:

$$Z_i = \frac{X_i - \mu_i}{\sigma_i}$$

Como se puede observar en la Figura 12, los datos estandarizados tienen el mismo comportamiento que los que no fueron estandarizados, la diferencia está en la escala de las mediciones que es similar. Trabajaremos con las mediciones escaladas.

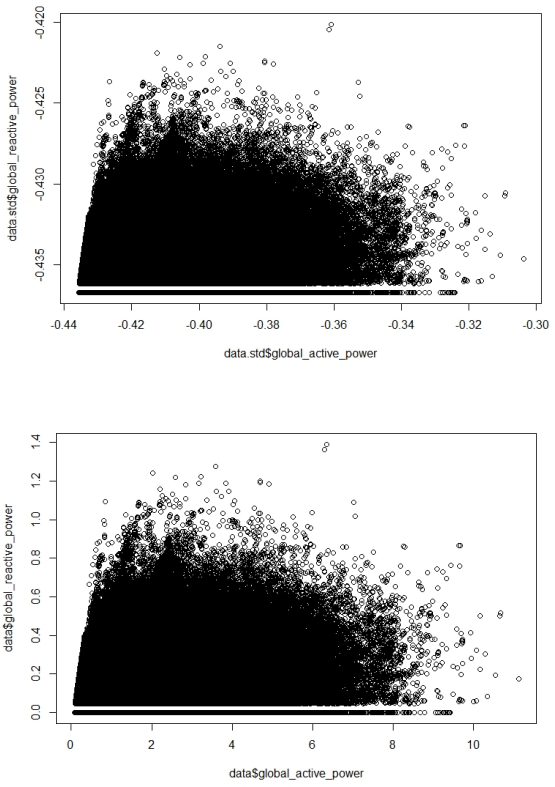

Figura 12: Diagrama de dispersión de los datos estandarizados y sin estandarizar tomando como ejemplo la corriente global activa y la corriente global reactiva

Clusters Jerárquicos

A continuación nos dispusimos a realizar el clúster jerárquico, sin embargo debido a la gran cantidad de datos fue imposible procesar una matriz de distancia tan grande. A pesar de esto, con el objetivo de encontrar un k adecuado para el algoritmo de k-means se realizó el clúster jerárquico con una muestra de los registros. Distintas formas de obtener una muestra fueron probadas (cogiendo los registros de un día determinado, filas al azar) y distintos tamaños, ninguno de los cuales dio una diferencia significativa en la cantidad de dendograma del clúster jerárquico. Este se muestra en la Figura 13 a pesar de que no es posible analizarlo adecuadamente por la gran cantidad de datos:

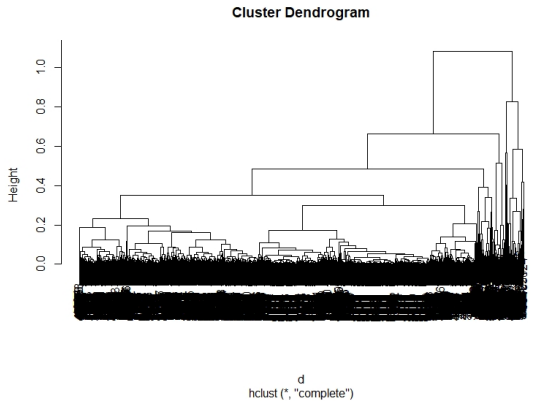

Figura 13: Dendograma de Clúster Jerárquico. Ajuste Completo

Tomando como altura 0.4 tenemos 8 clusters, por lo que ese es el k que utilizaremos en el algoritmo de K-Means.

K-means

Habiendo realizado la clasificación por el método jerárquico comenzamos eligiendo el k a partir del número de clusters encontrado previamente. Por lo tanto, comenzamos con k=8. Se realizó el experimento con distintos valores de k, sin embargo, con este número fue como se obtuvieron mejores resultados, teniendo una similitud de 87.5 % entre los elementos de cada clúster como se puede apreciar en la Figura 14

Por la gran cantidad de datos no se puede analizar correctamente en un gráfico el resultado de k-means, por lo que para analizar las relaciones de forma visual entre variables debemos hacerlo de dos a dos, por ejemplo, en la Figura 15 se muestran los 8 clusters de la corriente global activa y la corriente global reactiva.

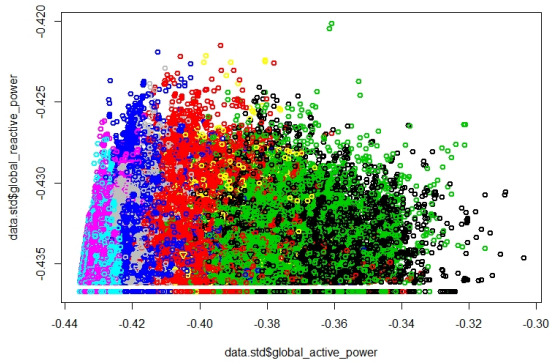

Figura 15: K-Means con 8 graficando los clusters de la Corriente Global Activa y la Corriente Global Reactiva

3.3 Árboles de Decisión

Los métodos basados en árboles de clasificación son poderosas herramientas analíticas para explorar estructuras de datos complejas, y una buena opción si necesitamos generar reglas que puedan entenderse y explicarse fácilmente. Su importancia viene dada porque nos permite encontrar variables claves que identifican los miembros de los grupos actuales, formular reglas para hacer predicciones sobre los miembros de grupos potenciales de nuevos casos y permiten desplegar gráfica y estadísticamente los resultados de los análisis, proporcionando una medida de confianza para ver cuán correcta es la clasificación.

Uno de las mayores dificultades para realizar este análisis fue elegir qué variable intentar predecir ya que todas las variables del data set tienen un valor fraccionario que no los hacen adecuados para una técnica de clasificación, más bien es un problema de regresión. Así que tomamos la decisión de introducir una nueva variable que sería booleana para comprobar si una variable cumple una determinada propiedad. Vimos en la matriz de correlación al analizar los componentes principales (Cuadro 3.1) que esta no es altamente correlacionada, excepto por la intensidad y la corriente global activa, que están bastante relacionadas, por lo que consideramos interesante clasificar una de estas variables. Luego, con los datos numéricos del data set nos propusimos clasificar cuando la corriente global sobrepasa a su media.

Al escoger los conjuntos de entrenamiento y de prueba se tomaron cuatro quintos de la población para entrenar y el quinto restante servirá para probar el árbol y calcular el error de clasificación. Como contamos con un data set grande, la quinta parte de este es lo suficientemente amplio para que los resultados obtenidos en el conjunto de prueba sean estadísticamente válidos.

El árbol que se obtuvo al aplicar el algoritmo clasificando la corriente global activa que sobrepasa una determinada cantidad se muestra en la Figura 16

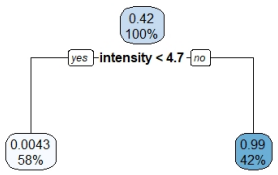

Figura 16: Predicción de valores superiores a la media de la corriente global activa

Como se puede observar solo se necesita de la intensidad para predecir altos valores de la corriente global activa, lo que corrobora la alta relación entre estas dos

```
K-means clustering with 8 clusters of sizes 653754, 461999, 186066, 352796, 45990, 173616, 55173, 119886

Cluster means:
  global_active_power global_reactive_power  intensity  voltage submetering1 submetering2 submetering3
1          -0.4323152            -0.4353719 -0.4170996 2.450576 -0.436452416  -0.43198788   -0.4308451
2          -0.4185850            -0.4353097 -0.3615856 2.445286 -0.435496200  -0.43206941   -0.2245160
3          -0.4067979            -0.4348005 -0.3100971 2.399321 -0.431168050  -0.42805160   -0.2185950
4          -0.4317764            -0.4355513 -0.4154710 2.495321 -0.436480200  -0.43289591   -0.4324245
5          -0.3922463            -0.4344670 -0.2476580 2.407071 -0.426495275  -0.00706841   -0.3091708
6          -0.4318939            -0.4354319 -0.4150216 2.396842 -0.436313219  -0.43228502   -0.4335653
7          -0.3888607            -0.4343661 -0.2337625 2.400470  0.006657717  -0.40730989   -0.3031096
8          -0.4148507            -0.4348194 -0.3437360 2.429797 -0.431740859  -0.42478425   -0.4324087

Clustering vector:
   [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 5 5 5 5 5 5 5 3 5 3 3 5 3 3 3 3 3 3 3 3 3 3 3
  [64] 3 3 3 3 3 3 3 3 5 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [127] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [190] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [253] 3 3 3 3 3 3 3 3 3 2 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 8 8 8 8 8 8 1 8 8 8 8 8 8 8 8 1 8 8 8 8 1 8 8
 [316] 8 8 1 8 8 8 8 8 1 8 8 8 8 1 8 8 8 8 8 1 8 8 8 8 1 8 8 8 8 8 8 8 8 8 8 8 1 8 8 8 8 8 8 8 8 8 8 8 8 1 8 8 8 8 8 8 8 8 1 8 8 8 4
 [379] 8 8 8 1 8 8 8 8 8 8 8 8 1 8 8 8 8 8 1 8 8 8 8 8 8 8 4 1 8 8 8 8 8 8 1 8 8 8 8 8 1 1 8 8 8 8 8 4 4 8 8 8 8 8 1 8 8 8 8 8 4 1 8
 [442] 8 8 8 8 4 8 8 8 8 8 1 8 8 8 8 8 4 8 8 8 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 8 5 5 5 5 5 5 5 5 5 5 5 8 5 5
 [505] 5 4 8 5 5 8 1 4 8 8 8 8 8 4 4 8 8 8 4 4 8 8 8 4 4 8 8 8 4 4 8 8 8 4 4 8 8 8 4 4 8 8 8 8 4 4 8 8 8 4 4 8 8 8 4 4 8 8 4 4 8 8 8
 [568] 4 4 8 8 8 4 4 8 8 8 8 4 4 8 8 8 4 4 8 8 8 4 4 8 8 8 8 4 4 8 8 8 8 4 4 8 8 8 4 4 8 8 8 8 4 4 8 8 8 4 4 8 8 8 4 4 8 8 8 4 4 8 8
 [631] 8 8 4 8 8 8 8 4 4 8 8 8 4 4 8 8 8 4 4 8 8 8 8 4 4 8 8 8 8 4 4 8 8 8 8 4 4 3 3 3 3 2 2 2 3 3 3 2 2 3 3 3 3 2 2 3 3 3 3 2 2 3 3
 [694] 3 3 2 2 3 3 3 3 2 2 3 3 3 3 2 2 3 3 8 4 4 8 8 8 8 8 4 4 8 8 8 4 4 8 8 8 8 4 4 8 8 8 4 4 8 8 8 4 4 8 8 8 4 8 8 8 4 8 8 8 4 8 8
 [757] 8 4 8 8 4 4 8 8 8 4 8 8 8 4 8 8 8 8 8 8 4 4 8 8 8 8 8 8 4 8 8 8 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
 [820] 4 1 1 1 1 1 1 1 1 1 1 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 1 1 8 8 8 8 8 8 8 8 4 4 8 8 8 8 8 8 8 8 4 4 8 8 8 8 8 8 8 8 1 4 8 8
 [883] 8 8 8 8 4 4 8 8 8 8 8 8 8 1 1 8 8 8 8 8 8 8 4 4 8 8 8 8 8 8 8 8 4 8 8 8 8 8 8 8 1 1 8 8 8 8 8 8 8 8 1 8 8 8 8 8 5 5 5 5 5 5 5
 [946] 5 5 5 1 5 8 8 8 8 8 8 1 8 8 8 8 8 8 8 1 8 8 8 8 8 8 8 1 1 8 8 8 8 8 8 1 1 8 8 8 8 8 8 8 1 1 8 8 8 8 8 8 1 1 3
 [ reached getOption("max.print") -- omitted 2048280 entries ]

Within cluster sum of squares by cluster:
[1]  299.3952  821.1896 1006.8134  235.8802 1809.2241  155.6910 1578.6298  477.3038
 (between_SS / total_SS =  87.5 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
```

Figura 14: Salida del algoritmo de K-means con 8 clusters

variables. Se obtiene solo un error de 0.0045, lo que constituyen muy buenos resultados.

Se probó hacer el mismo análisis con otra variable. Comprobamos si era posible predecir cuando estos valores excedían la media a partir de las otras variables. La más altamente relacionada después de la corriente global activa y la intensidad es la submedida3 (Tabla 3.1), y se obtuvo el árbol mostrado en la Figura 17

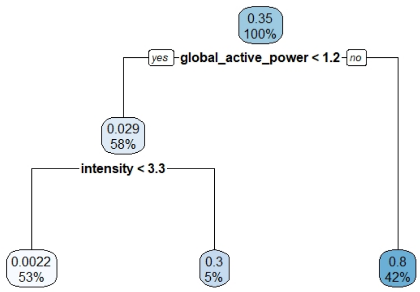

Figura 17: Predicción de valores superiores a la media de la submedida3

De igual manera solo se necesitaron de dos variables para la predicción de la submedida3: la corriente global activa y la intensidad. Sin embargo, se obtuvo un error de 0.45, lo que representa un valor bastante alto.

4. ANOVA

El análisis de varianzas o ANOVA a pesar de su nombre, es una técnica estadística que permite la comparación de las medias de una característica de varias poblaciones. Los ANOVA evalúan la importancia de uno o más factores al comparar las medias de la variable de respuesta en los diferentes niveles de los factores. Hacen una prueba de hipótesis para evaluar la similitud de las medias de dos o más poblaciones. La hipótesis nula establece que todas las medias de la población (medias de los niveles de los factores) son iguales mientras que la hipótesis alternativa establece que al menos una es diferente.

Con el objetivo de cumplir los supuestos necesarios para que el experimento de ANOVA sea válido se utilizaron varios subconjuntos de variables. La relación entre la intensidad y la corriente global activa que fue ampliamente analizada en este trabajo fue una de las combinaciones que se probaron. Al igual que con la regresión lineal no se encontró combinación de variables que cumpliera los supuestos. Al final se decidió elegir la submedida3 (la variable que le corresponde al consumo de un calentador eléctrico y un aire acondicionado) como factor de interés y la corriente global activa y la intensidad como factores de bloque para analizar con mayor profundidad la relación existente entre estas variables.

Al elegir la intensidad y la corriente global activa como factores de bloque y la submedida3 como factor de interés no proponemos responder la interrogante ¿existen diferencias entre la submedida3 para los distintos valores de la corriente global activa y la intensidad?

	Df	Sum Sq	Mean Sq	F value	$Pr(>F)$
Corriente Activa	1	59482639	59482639	1562629	$< 2e-16$ ***
Intensidad	1	8389201	8389201	220387	$< 2e-16$ ***
Residuals	2049277	78007245	38		

—

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Cuadro 3: Salida de anova con dos factores de bloque

Como el valor p es menor que la significación prefijada $\alpha = 0.05$, tanto para la corriente global activa como la intensidad, se rechaza H_0 en ambos casos y se acepta que al menos un par de corrientes globales activas tienen promedio diferente, así como un par de intensidad tienen promedios diferentes con respecto a la submedida3.

Por último, necesitamos verificar los supuestos del modelo:

1. Los e_{ij} siguen una distribución normal con media cero

2. Los e_{ij} son independientes entre sí.

3. Los residuos de cada tratamiento tienen la misma varianza σ^2

Para esto primero inspeccionamos los residuos de forma gráfica:

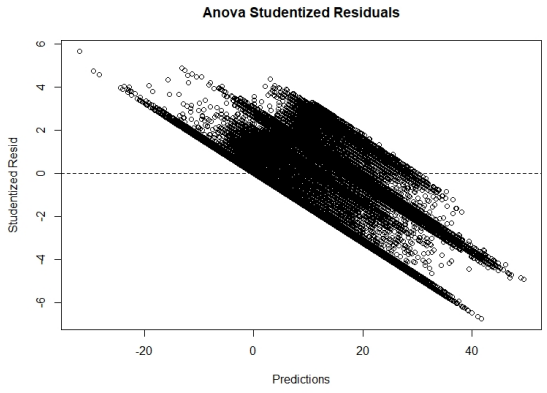

Figura 18: Gráfico estandarizado de los residuos para probar la homogeneidad

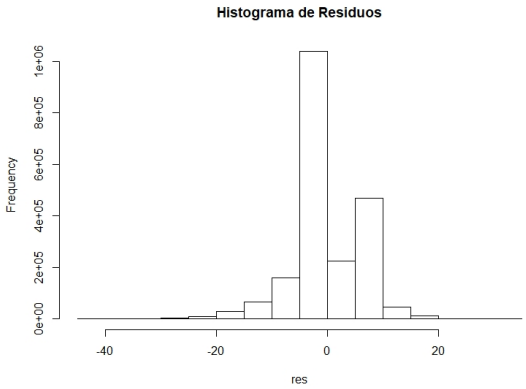

Figura 19: Histograma de los Residuos para probar normalidad

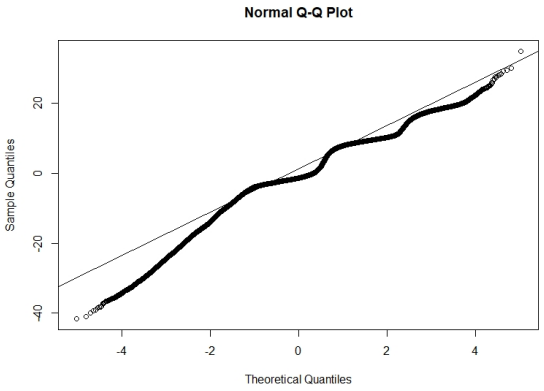

Figura 20: Q-Q Plot para probar normalidad

Como se pueden observar en el gráfico estandarizado de residuos (Figura 18), no parece tener varianza constante, el histograma de los residuos (Figura 19) muestra un poco de comportamiento normal sin embargo, el qq-plot (Figura 20) presenta varias desviaciones. Parece no cumplir todos los supuestos pero probamos con los tests solo para estar seguros:

```
        Shapiro-Wilk normality test

data:  sample(res, 5000)
W = 0.87299, p-value < 2.2e-16
```

Figura 21: Prueba de Normalidad de Shapiro-Wilk

La prueba de Shapiro-Wilk es significativa por tanto podemos rechazar H_0, por lo que no se cumple la hipótesis de la normalidad. Entonces, no tiene validez el experimento. Sin embargo, continuaremos con la validación de los supuestos.

La prueba de Barlett no es aplicable para el análisis de la homogeneidad de las varianzas, pues se necesita que existan replicas para aplicarla. Pero por el gráfico de la Figura 18 no parece que la varianza sea constante.

Por último, veamos que sucede con el test de independencia.

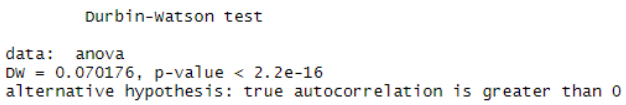

Figura 22: Prueba de Independencia de Durbin-Watson

Se tiene que $p - valor < 0.05$ por lo que se rechaza H_0, lo que significa que las variables no son independientes, entonces el supuesto de independencia no se cumple tampoco.