ugr | Universidad de Granada

INTRODUCCIÓN A LA CIENCIA DE DATOS

MÁSTER DE CIENCIA DE DATOS E INGENIERÍA DE COMPUTADORES

# Trabajo teórico/práctico integrador

**Autor**
Rafael Vázquez Conejo

—

Granada, diciembre de 2022

# Índice general

# Capítulo 1

# Introducción

Este documento recoge el informe obtenido con la realización del trabajo final de la asignatura *Introducción a la Ciencia de Datos*. Este consiste en la realización de un análisis exploratorio de los datos pertenecientes a dos *dataset* distintos, con la intención de comprender y determinar información importante sobre las variables que forman a cada conjunto de datos. Tras ello, dependiendo del dataset, se generarán diversos modelos de clasificación o regresión.[1]

El primer dataset tratado es el **California**, sobre el que se plantea un problema de regresión y la búsqueda de desarrollar diversos modelos óptimos de regresión lineal simple y múltiple, modelos de regresión basados en K-NN (K Nearest Neighbour) y modelos de regresión no lineales. **Bupa** es el segundo dataset, el cual plantea un problema de clasificación que será abordado con modelos de clasificación basados en KNN, modelos basados en LDA y modelos basados en QDA.

# Capítulo 2

# Análisis Exploratorio de los Datos (EDA). Dataset de regresión: California

## 2.1. Descripción del dataset: California

El dataset **California** contiene información relativa a viviendas de California, formada por datos extraídos del censo de Estados Unidos del 1990, cuya variable objetivo es el valor medio de la vivienda para diferentes distritos de California, información expresada en cientos de miles de dólares ($100,000). Cada fila de este conjunto de datos representa un grupo de bloques censales, siendo esta la unidad geográfica más pequeña utilizada por La Oficina del Censo de EE.UU. para publicar sus datos (generalmente un grupo de bloques lo forma una población de 600 a 3.000 personas). [4]

Este dataset consta de 8 variables independientes y una variable dependiente, siendo todos los datos numéricos. Se recoge en total 20640 muestras que representan diferentes bloques censales.
Se presentan a continuación las variables independientes:

- **Longitude**: Variable numérica real que refleja la longitud geográfica del grupo de bloques censal.
- **Latitude**: Variable numérica real que refleja la latitud geográfica del grupo de bloques censal.
- **HousingMedianAge**: Variable numérica entera que refleja la media de antigüedad de una casa en un grupo de bloques.
- **TotalRooms**: Variable numérica entera que refleja la cantidad total de habitaciones por hogar.
- **TotalBedrooms**: Variable numérica entera que refleja la cantidad total de dormitorios por hogar.

- **Population**: Variable numérica entera que refleja la cantidad de personas que residen en un grupo de bloques.

- **Households**: Variable numérica entera que refleja la cantidad de hogares en un grupo de bloques.

- **MedianIncome**: Variable numérica real que refleja el valor medio de los ingresos de cada hogar de un grupo de bloques (medido en decenas de miles de dolares estadounidenses).

La variable dependiente es **MedianHouseValue**, un valor numérico entero que representa el precio medio asociado al valor de una casa en cada distrito de California. [5]

## 2.2. Planteamiento de hipótesis

- California posee una amplia zona de costa permitiendo las variables de latitud y longitud determinar la cercanía de cada grupo de viviendas al mar. ¿Cómo de importante es la localización de la vivienda a la hora de determinar su precio?

- En ocasiones zonas con menos densidad de población suele estar relacionado con poblaciones más privilegiadas, ¿el precio de la vivienda tendrá una alta relación con la densidad de la población?

- La variable MedianIncome indica los ingresos medios de los hogares, ¿posee esta una fuerte relación positiva con el valor de la vivienda?

- Un factor interesante de estudio es la antigüedad de la vivienda, lo esperable sería que una casa nueva tuviera mayor precio que una antigua. Sin embargo, puede que la localización de la vivienda afecte a esta variable y esta premisa no se cumpla.

## 2.3. Procesamiento de los datos

Introducidos los diferentes atributos que constituyen este dataset, el siguiente paso pretende un análisis en detalle de la información contenida en dicho dataset con el objetivo de conocer y determinar cualquier característica de los datos que facilite el posterior desarrollo de modelos predictivos.

El primer paso es la búsqueda de valores perdidos dentro de los datos, concluyendo en que este dataset no posee ningún Missing value. También se determina que no hay ninguna muestra duplicada.

### 2.3.1. Análisis de las variables

Para un mejor análisis del comportamientos de las diferentes variables, se calculan diversas medidas de posición: la media aritmética, mediana, primer y tercer cuartil, valores máximos y mínimos de cada variable. También se estudia la dispersión de las distribuciones mediante el calculo de la desviación típica, mientras que la normalidad de los datos se estudia con los coeficientes de Skewness y Kurtosis.

Se estudia cada variable en detalle mediante el calculo de las medidas previamente mencionadas. Dicho estudio es acompañado con una serie de representaciones gráficas que facilite la comprensión de los datos:

- **Longitude**

| | Longitude |
|---|---|
| Valor mínimo | -124.3 |
| Primer cuartil | -121.8 |
| Mediana | -118.5 |
| Media | -119.6 |
| Tercer cuartil | -118.0 |
| Valor máximo | -114.3 |
| Desviación estandar | 2.003532 |
| Coeficiente de skewness | -0.29777956 |
| Coeficiente de Kurtosis | 1.669879 |

Los estadísticos resultantes nos muestran una distribución con una medida de asimetria muy cercana al valor cero, lo que nos indica que la distribución de esta se aproxima a una distribución normal, pero con cierta dispersión de los datos a la derecha, hecho que claramente se refleja con un valor de la media muy cercano al valor del tercer cuartil. Un coeficiente de Kurtosis tan cercano a cero nos indica que no se presenta dispersión de los valores respecto al centro de densidad de la distribución.

Se estudia de forma gráfica estos resultados por medio de un diagrama de cajas:

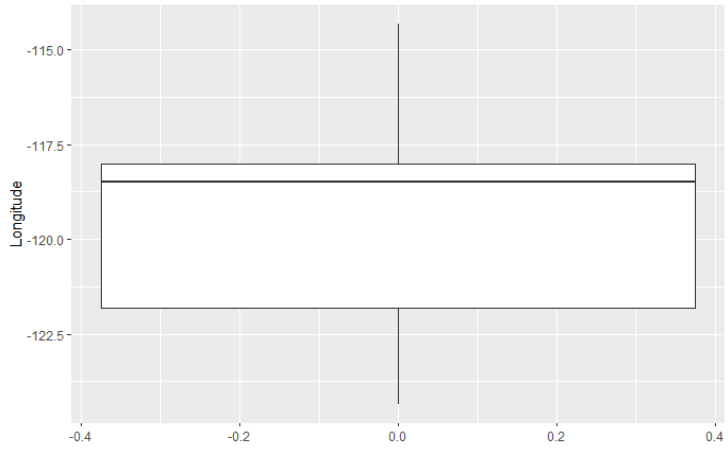



Figura 2.1: Diagrama de cajas Longitude

Efectivamente el diagrama de cajas nos muestra que los datos se concentran en una región central densa, sin presencia de outliers.

Observemos la distribución de los valores en más detalle en un histograma:

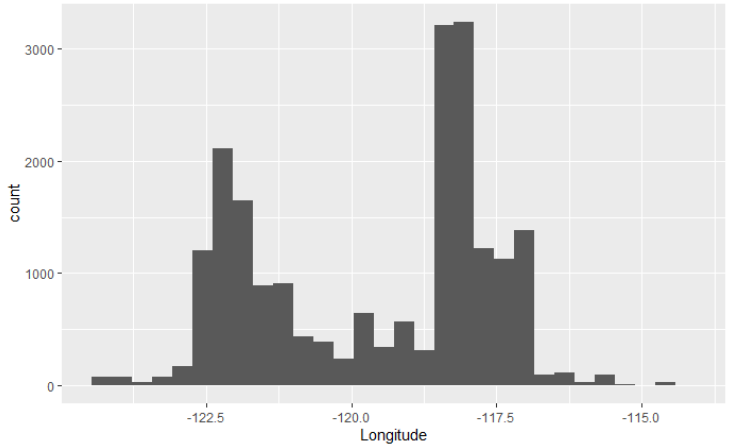



Figura 2.2: Histograma Longitude

- **Latitude**:

| | Latitude |
|---|---|
| Valor mínimo | 32.54 |
| Primer cuartil | 33.93 |
| Mediana | 34.26 |
| Media | 35.63 |
| Tercer cuartil | 37.71 |
| Valor máximo | 41.95 |
| Desviación estandar | 2.135952 |
| Coeficiente de skewness | 0.46591914 |
| Coeficiente de Kurtosis | 1.882220 |

En este caso los datos presentan una leve dispersión a la derecha, detalle que se confirma al observar una media cercana al valor del primer cuartil. De nuevo no se presenta dispersion de los datos respecto de su centro, por lo que será poco probable la existencia de outliers.

Se complementa este estudio con la representación gráfica de esta variable mediante un diagrama de cajas y un histograma

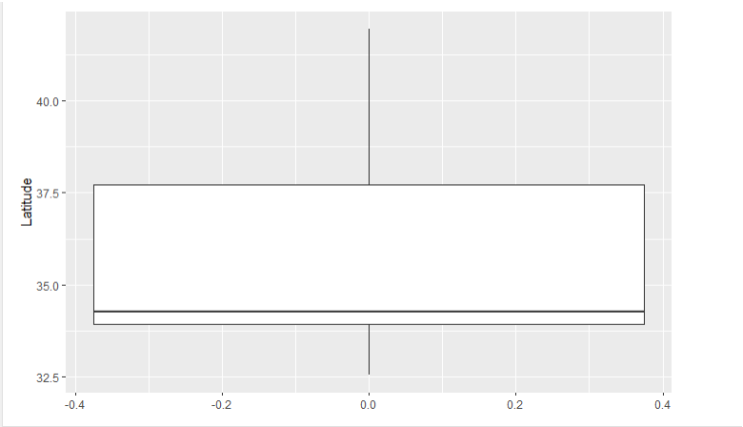



Figura 2.3: Diagrama de cajas Latitude

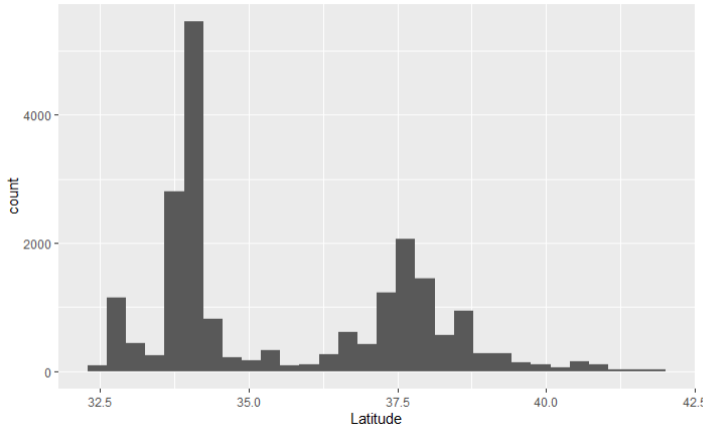



Figura 2.4: Histograma Latitude

Se confirma una distribución concentrada de los datos levemente desplazada a valores situados a la izquierda.

- **HousingMedianAge**:

| | HousingMedianAge |
|---|---|
| Valor mínimo | 1.00 |
| Primer cuartil | 18.00 |
| Mediana | 29.00 |
| Media | 28.64 |
| Tercer cuartil | 37.00 |
| Valor máximo | 52.00 |
| Desviación estandar | 12.585558 |
| Coeficiente de skewness | 0.06032625 |
| Coeficiente de Kurtosis | 2.199274 |

Los resultados estadísticos describen una variable con una región central muy densa y con una muy leve dispersión de los datos respecto a esta región central, descartando nuevamente la presencia de outliers.

Se complementa este estudio con la representación gráfica de esta variable mediante un diagrama de cajas y un histograma:

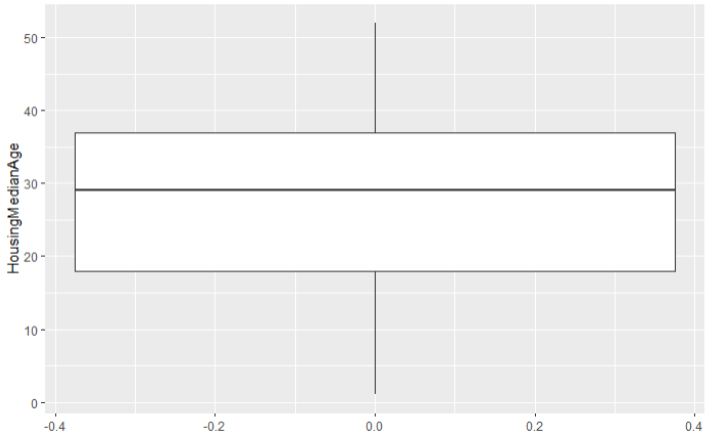



Figura 2.5: Diagrama de cajas HousingMedianAge

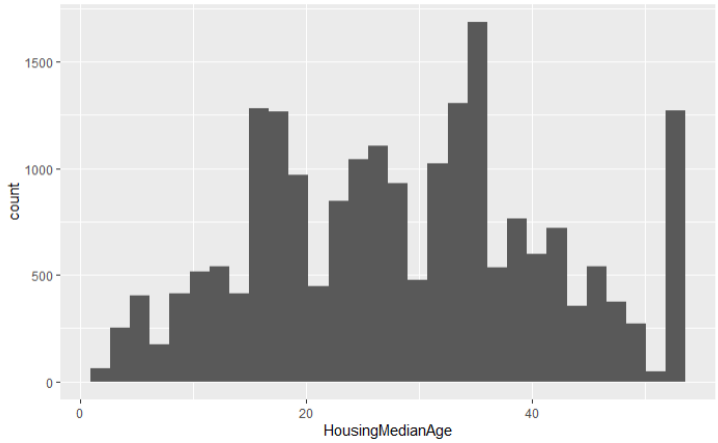



Figura 2.6: Histograma HousingMedianAge

Se emplea el test de Shapiro-Wilk para confirmar que no se supera el nivel mínimo de significancia, y por tanto no es una distribución normal.

- **TotalRooms**:

| | TotalRooms |
|---|---|
| Valor mínimo | 2 |
| Primer cuartil | 1448 |
| Mediana | 2127 |
| Media | 2636 |
| Tercer cuartil | 3148 |
| Valor máximo | 39320 |
| Desviación estandar | 2181.615252 |
| Coeficiente de skewness | 4.14704204 |
| Coeficiente de Kurtosis | 35.622732 |

Los resultados estadísticos hacen referencia a una distribución desplazada a la izquierda. El coeficiente de Kurtosis obtenido se traduce en una amplia dispersión de los datos respecto al centro de distribución de estos, siendo en este caso, a diferencia de los anteriores, muy probable la existencia de outliers situados a la derecha.

Se complementa este estudio con la representación gráfica de esta variable mediante un diagrama de cajas y un histograma

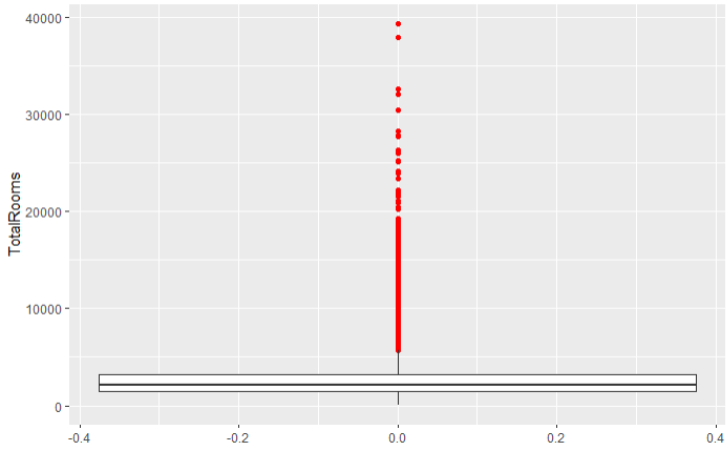



Figura 2.7: Diagrama de cajas TotalRooms

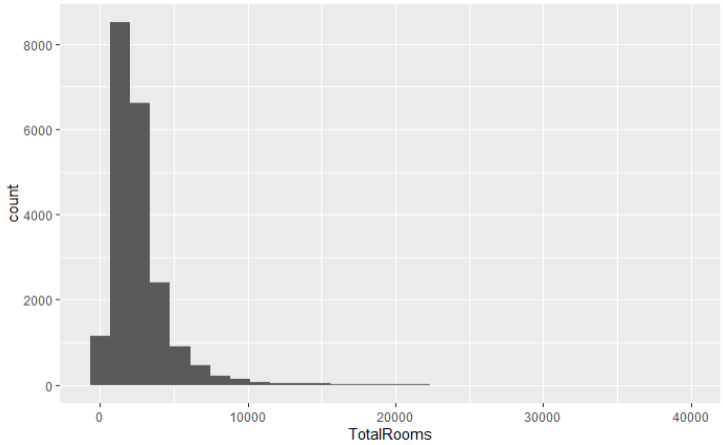



Figura 2.8: Histograma TotalRooms

- **TotalBedrooms**:

| | TotalBedrooms |
|---|---|
| Valor mínimo | 1.0 |
| Primer cuartil | 295.0 |
| Mediana | 435.0 |
| Media | 537.9 |
| Tercer cuartil | 647.0 |
| Valor máximo | 6445.0 |
| Desviación estandar | 421.247906 |
| Coeficiente de skewness | 3.45282180 |
| Coeficiente de Kurtosis | 24.917894 |

Se presenta un caso similar al anterior, en el que de nuevo el centro de la distribución se desplaza a la derecha, a la vez que se revela una amplia dispersión de los datos.

Se complementa este estudio con la representación gráfica de esta variable mediante un diagrama de cajas y un histograma

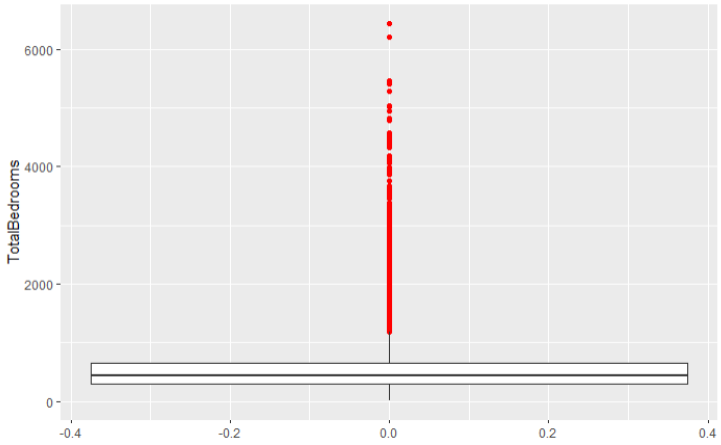



Figura 2.9: Diagrama de cajas TotalBedrooms

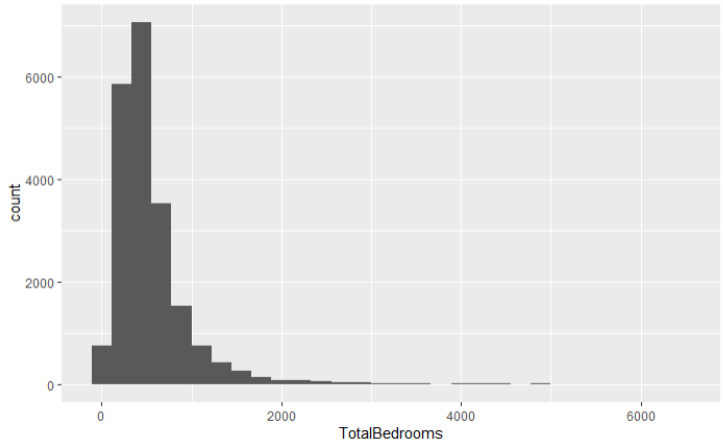



Figura 2.10: Histograma TotalBedrooms

- **Population**:

| | Population |
|---|---|
| Valor mínimo | 3 |
| Primer cuartil | 787 |
| Mediana | 1166 |
| Media | 1425 |
| Tercer cuartil | 1725 |
| Valor máximo | 35682 |
| Desviación estandar | 1132.462122 |
| Coeficiente de skewness | 4.93549951 |
| Coeficiente de Kurtosis | 76.535009 |

Los estadísticos muestra una distribución muy estrecha, desplazada a la izquierda y con dispersión de los datos respecto de su centro, planteando la existencia de outliers.

Se complementa este estudio con la representación gráfica de esta variable mediante un diagrama de cajas y un histograma

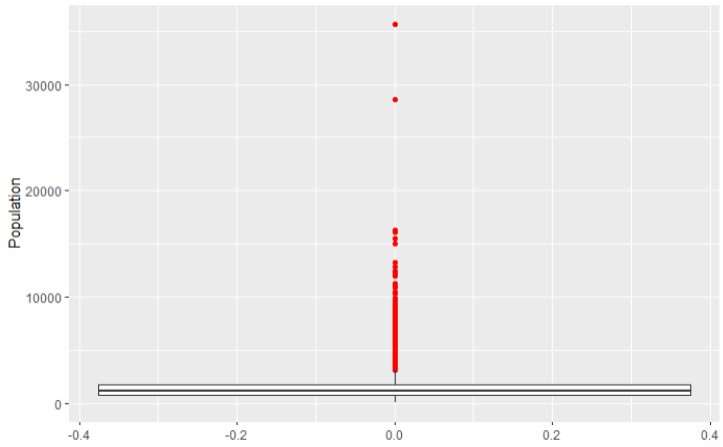



Figura 2.11: Diagrama de cajas Population

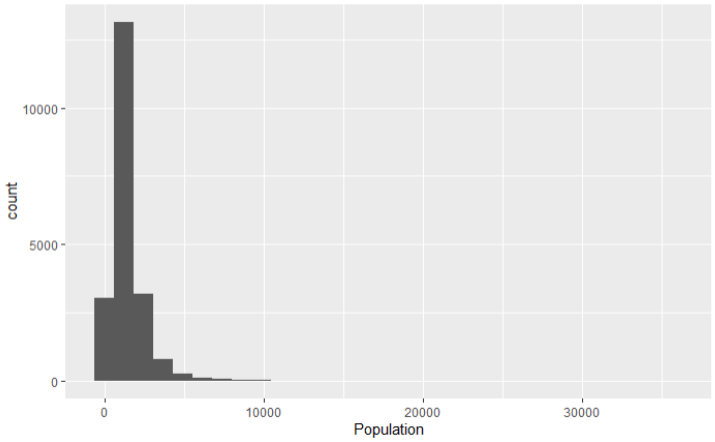



Figura 2.12: Histograma Population

- **Households**:

| | Households |
|---|---|
| Valor mínimo | 1.0 |
| Primer cuartil | 280.0 |
| Mediana | 409.0 |
| Media | 499.5 |
| Tercer cuartil | 605.0 |
| Valor máximo | 6082.0 |
| Desviación estandar | 382.329753 |
| Coeficiente de skewness | 3.41018986 |
| Coeficiente de Kurtosis | 25.052354 |

Los resultados estadísticos hacen referencia a una distribución desplazada a la izquierda con una amplia dispersión de los datos respecto al centro

de distribución de estos, muy probable la existencia de outliers situados a la derecha.

Se complementa este estudio con la representación gráfica de esta variable mediante un diagrama de cajas y un histograma

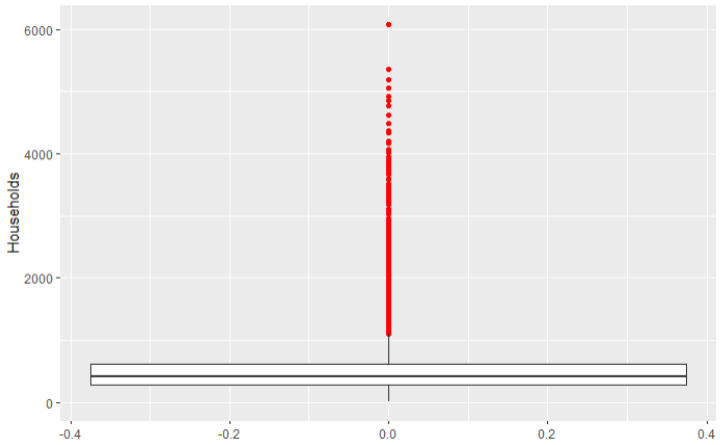



Figura 2.13: Diagrama de cajas Households

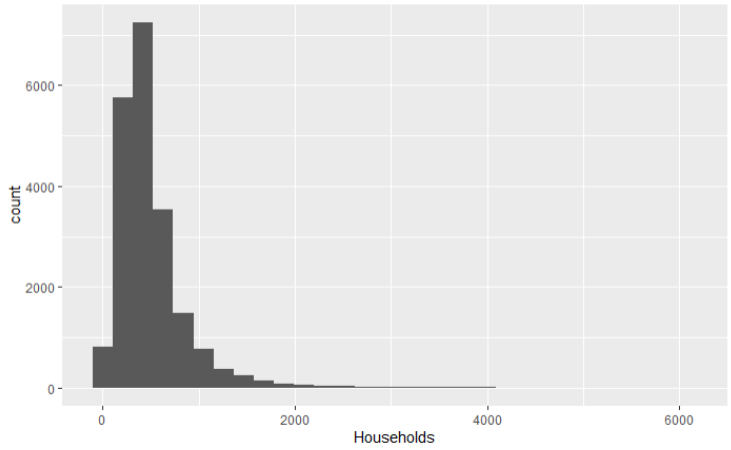



Figura 2.14: Histograma Households

- **MedianIncome**:

| | MedianIncome |
|---|---|
| Valor mínimo | 0.4999 |
| Primer cuartil | 2.5634 |
| Mediana | 3.5348 |
| Media | 3.8707 |
| Tercer cuartil | 4.7432 |
| Valor máximo | 15.0001 |
| Desviación estandar | 1.899822 |
| Coeficiente de skewness | 1.64653703 |
| Coeficiente de Kurtosis | 7.951034 |

De nuevo los datos se encuentran ligeramente desplazados a la izquierda, siendo en este caso leve la dispersión de valores respecto al centro de la distribución, revelando la existencia de outliers.

Se complementa este estudio con la representación gráfica de esta variable mediante un diagrama de cajas y un histograma

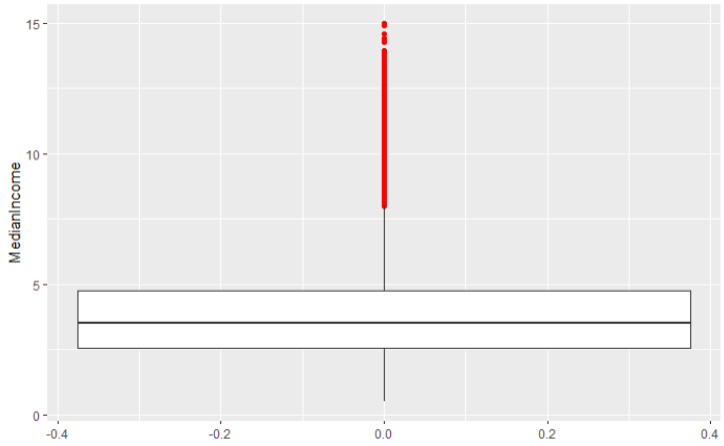



Figura 2.15: Diagrama de cajas MedianIncome

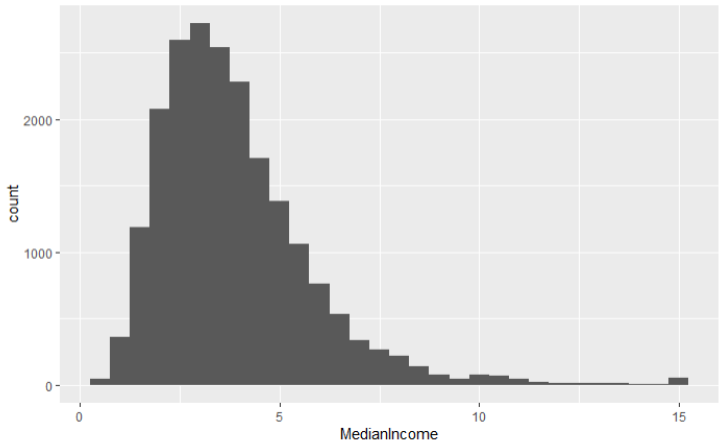



Figura 2.16: Histograma MedianIncome

Se observa que el salario de las personas posee una distribución más o menos normal, pero con la existencia de personas con un salario más elevado que el resto, originando outliers.

Mencionar que aunque se ha indicado en cada variable que no se posee una distribución normal, esta suposición se ha confirmado realizando el test Shapiro-Wilk, el cual efectivamente no ha superado un valor por encima del nivel de significancia para ninguna de las variables, todas poseen un p-value menor a 2.2e-16.

Finalicemos analizando la variable dependiente, **MedianHouseValue**:

| | MedianHouseValue |
|---|---|
| Valor mínimo | 14999 |
| Primer cuartil | 119600 |
| Mediana | 179700 |
| Media | 206856 |
| Tercer cuartil | 264725 |
| Valor máximo | 500001 |
| Desviación estandar | 115395.615874 |
| Coeficiente de skewness | 0.97769221 |
| Coeficiente de Kurtosis | 3.327500 |

Se complementa este estudio con la representación gráfica de esta variable mediante un diagrama de cajas y un histograma:

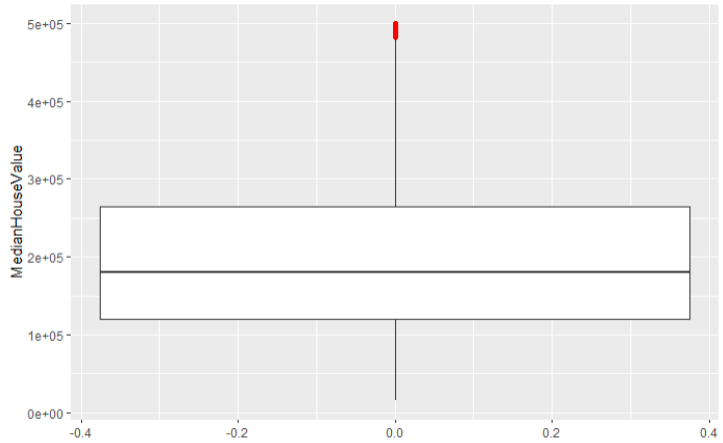



Figura 2.17: Diagrama de cajas MedianHouseValue

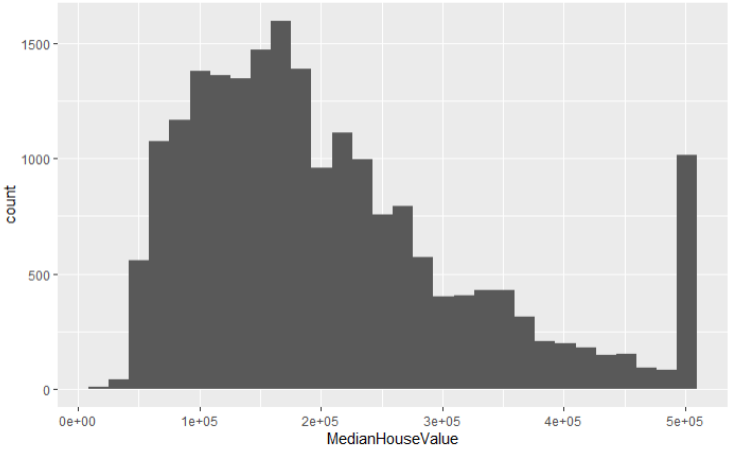



Figura 2.18: Histograma MedianHouseValue

Se observa que la distribución de la variable dependiente posee una región central muy densa y con una leve dispersión de los datos respecto a esta. Se observa un efecto de umbral para aquellas viviendas con un valor muy elevado, ya que todas las viviendas con un valor superior a 500000 han sido fijadas en esta cantidad.

Comparando el valor medio de la vivienda con la media de ingresos se observa con claridad este problema:

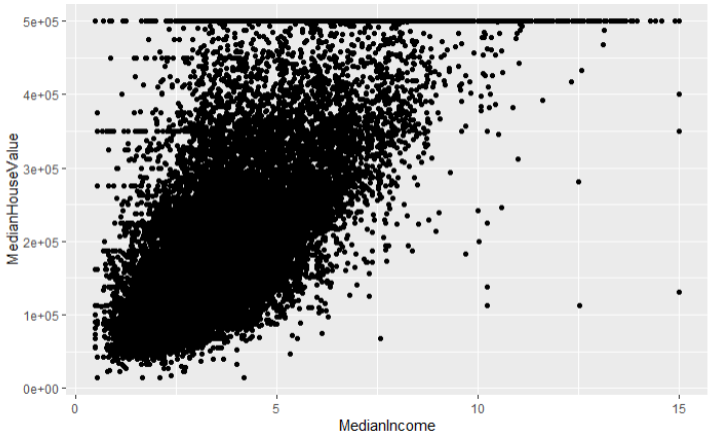



Figura 2.19

Todos estos casos poseen un valor asignado incorrecto y por ello se procede a su eliminación, con el objetivo de evitar que estos afecten en la posterior generación de modelos. Por suerte estos casos representan un porcentaje muy bajo de los datos (sobre 1 %).

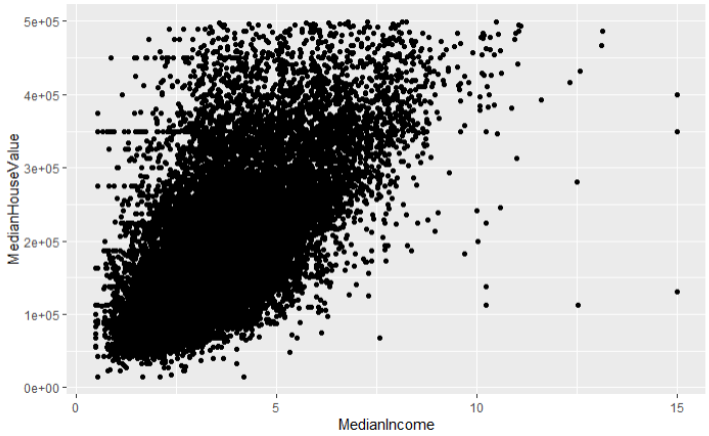



Figura 2.20

Este paso reduce el número de outliers pero aún así se estudiaran en el siguiente apartado en aquellas variables que he considerado de estudio.

### 2.3.2. Análisis de valores anómalos

Nos fijamos en los atributos anteriores en los que se presentaron valores outliers:

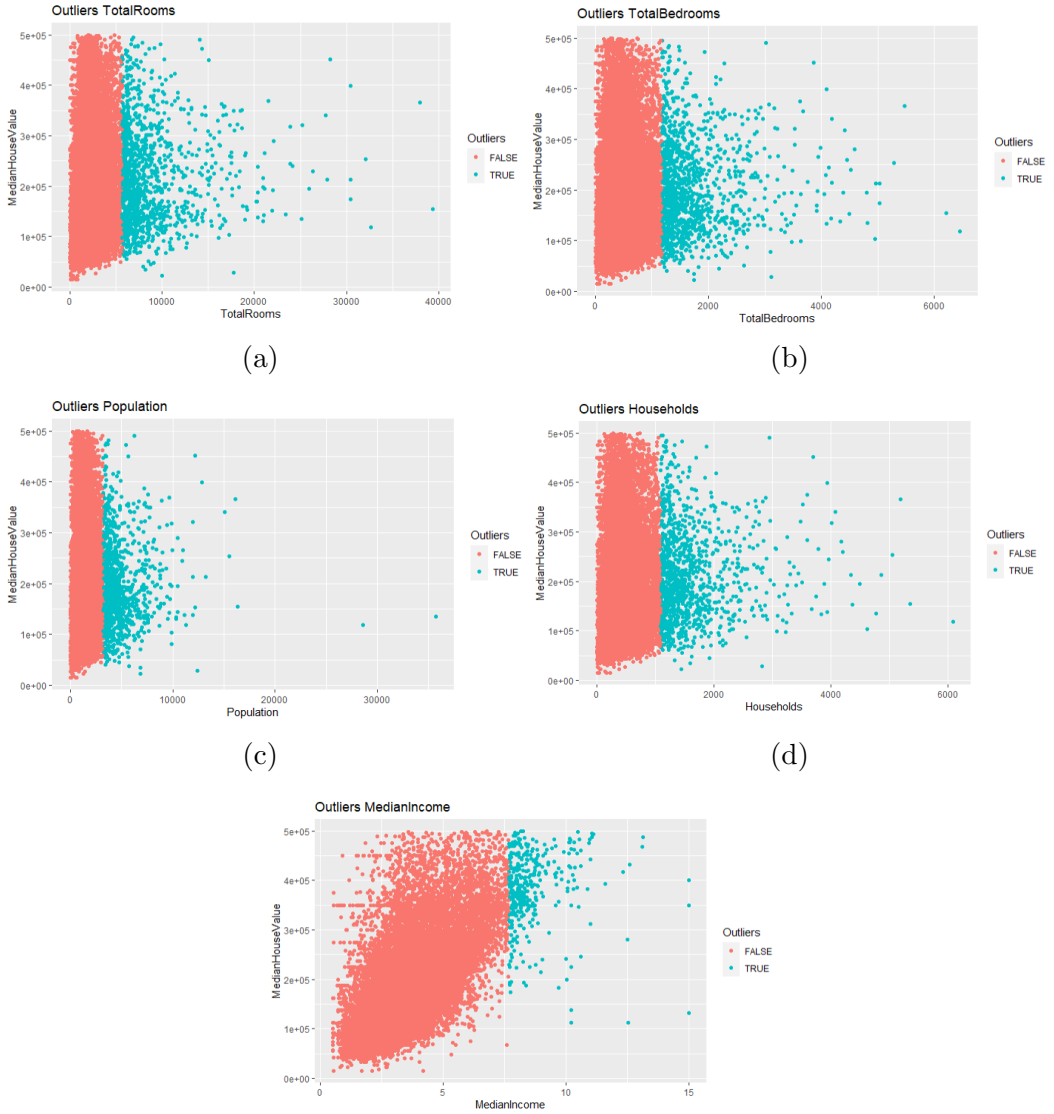



Figura 2.21: Outliers California

Nos fijamos solo en los casos cuyo valor del tercer cuartil más 1.5 veces su rango intercuartil, casos que son considerados como outliers extremos. Efectivamente se detectan numerosos casos que influirán de manera negativa en la realización de los modelos. Al tratarse de un 6 % del total de los datos, un porcentaje muy bajo, pues tenemos un dataset muy denso, he decidido que su eliminación será una ventaja.

## 2.4. Análisis de las relaciones entre variables

Conocidas en profundidad cada una de las variables, el próximo paso es el estudio de las posibles relaciones existentes entre ellas. Este estudio tendrá como objetivo determinar si existe un alto nivel de dependencia entre algunas de las variables, detalle que debe ser tenido en cuando en el posterior proceso de elaboración de modelos.

El siguiente diagrama de correlaciones permite efectuar este estudio entre cada par de variables que forman el dataset. Las correlación es entre cada una de ellas s calcula gracias al test de Kendall.

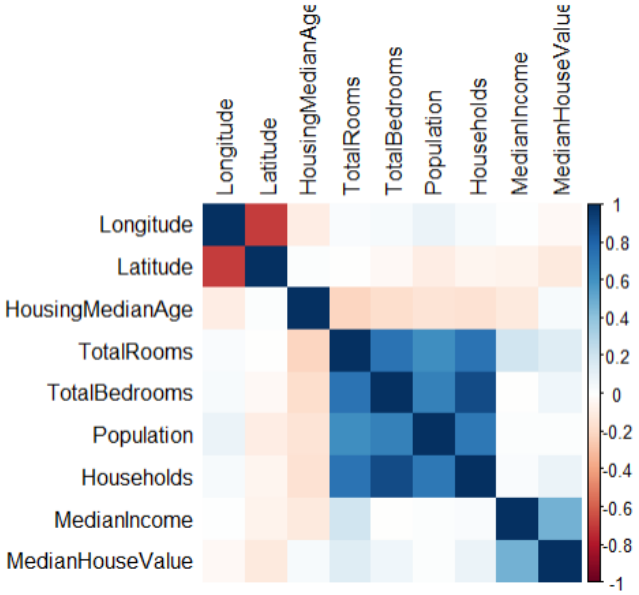



Figura 2.22: Correlaciones California

Observando la matriz de correlación se confirman relaciones entre variables obvias y fáciles de predecir. Algunas de estas relaciones confirman que cuantos más hogares hay un bloque mayor es el número de dormitorios; de igual manera una mayor población también se vincula con un mayor número de dormitorios. Otra relación menos clara y en la que se profundiza en el siguiente apartado es en la existente relación entre la media de ingresos en el hogar con el valor medio de la vivienda.

## 2.5. Comprobación de hipótesis planteadas

**California posee una amplia zona de costa permitiendo las variables de latitud y longitud determinar la cercanía de cada grupo de viviendas al mar. ¿Cómo de importante es la localización de la vivienda a la hora de determinar su precio?**

Para comprobarlo se crea un scatterplot con las variables de latitud y longitud, que utilice la variable MedianHouseValue para asignarle color a cada punto dependiendo si la media de las casas de esa zona es de un alto o bajo precio.

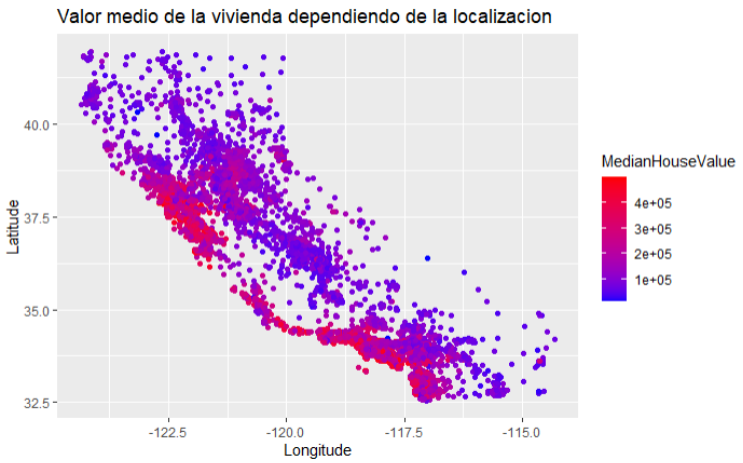



Figura 2.23

Se observa que los puntos representados pertenecen a una representación gráfica del estado de California, por ello para facilitar la compresión de los resultados se haya un mapa de este estado.

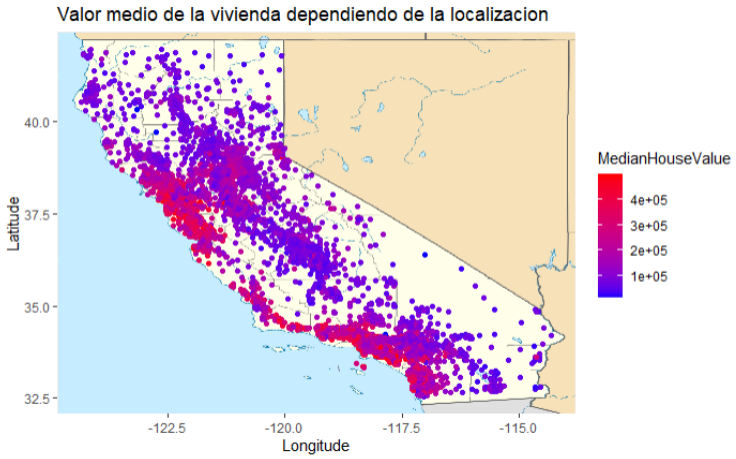
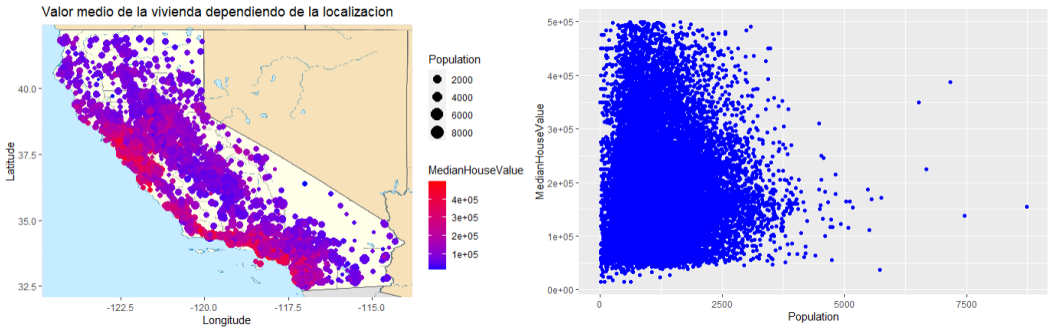



Figura 2.24

Efectivamente observamos que prácticamente todas las casas de alto valor se encuentran relativamente cerca del mar, resaltando la importancia de la proximidad a la costa en el análisis. Pero existe una excepción y se trata de la zona norte de California, donde a pesar de la cercanía al océano las viviendas no poseen un valor elevado. Tras investigar las diferentes ciudades de California he determinado que esto puede ser causado porque las principales ciudades de california se sitúan en la zona de costa central y sur. En la siguiente hipótesis se estudiará en más detalle esta observación.

**En ocasiones zonas con menos densidad de población suele estar relacionado con poblaciones más privilegiadas, ¿el precio de la vivienda tendrá una alta relación con la densidad de la población?**

De nuevo para facilitar la compresión de los resultados se efectuarán las gráficas sobre un mapa de California.




(a) (b)

Observamos que a pesar de que la capital de Callifornia sea Sacramento, situada en aproximadamente Latitud 38 y longitud -122, siendo una de las zonas con mayor población, la mayoría de los bloques situados en ella contienen casas de bajo presupuesto. Además se observa que cuanto más nos acercamos a regiones de montañas (centro/oeste) las poblaciones son más pequeñas al igual que el precio de la vivienda.

Investigando la geografía de California determino que aquellas concentraciones de viviendas de gran precio situadas en la costa corresponden a las grandes ciudades de este Estado, como son San Francisco, Los Ángeles, San Diego o San Jose.

**La variable MedianIncome indica los ingresos medios de los hogares, ¿posee esta una fuerte relación positiva con el valor de la vivienda?**

Nos apoyamos en un scatterplot para observar la relación entre estas dos variables.

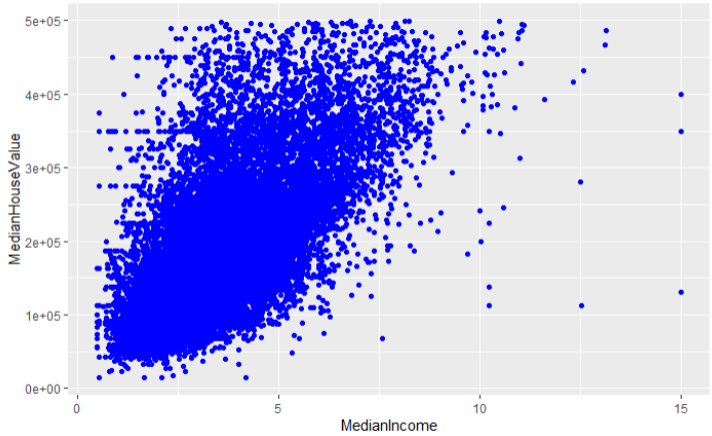
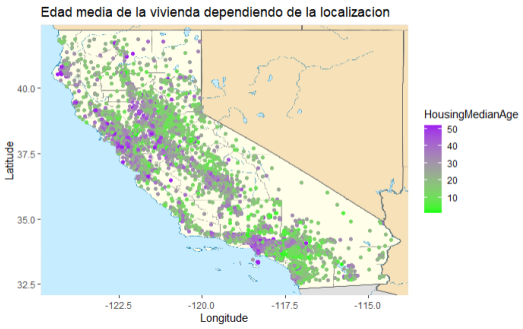


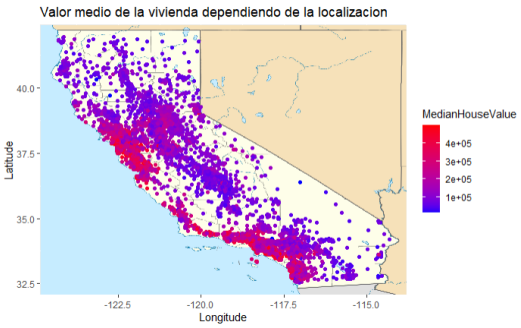

Figura 2.26

Se observa una fuerte relación entre estas variables, en la que efectivamente una media de ingresos elevados suele ir vinculada con viviendas de mayor precio. Esta relación será importante en la posterior creación de modelos de regresión.

**Un factor interesante de estudio es la antigüedad de la vivienda, lo esperable sería que una casa nueva tuviera mayor precio que una antigua. Lo lógico es pensar que una casa nueva tenga mayor valor que una vieja.**




(a)




(b)

Comparando los dos mapas llegamos a la conclusión de que cuanto más vieja la casa más cara es esta, algo que de primeras no parece tener sentido. Sin embargo, como ya hemos descubierto previamente, las viviendas en la costa tienen mayor precio, siendo de nuevo este el factor de mayor impacto en el dataset, junto con el ingreso de dinero en cada domicilio.

Podemos concluir en que cuanto más nueva sea la casa mayor será la probabilidad de que esta se ubique en una zona no costera, en el interior.

# Capítulo 3

# Análisis Exploratorio de los Datos (EDA). Dataset de clasificación: Bupa

## 3.1. Descripción del dataset: Bupa

El dataset **Bupa** contiene los resultados de diferentes tipos de análisis de sangre sensibles a los trastornos hepáticos que pueden surgir con un consumo excesivo de alcohol. Datos donados en el 1990. Cada fila de este conjunto de datos contiene el registro de un solo individuo masculino. [6]

Un dato importante de este conjunto de datos, es que la séptima variable solía malinterpretarse como una variable que indica la presencia o ausencia de un trastorno hepático, pero esto es incorrecto, esta variable fue creada por investigadores para separar los datos en un conjunto de entrenamiento y otro de test.

Teniendo en cuenta la función de esta séptima variable, este dataset está formado por 5 variables independientes, las cuales corresponden a los resultados de diferentes análisis de sangre y una variable dependiente, todas variables numéricas. Se recogen un total de 345 muestras que representan el registro de cada individuo masculino. Se presentan a continuación las variables independientes:

- **mcv**: Variable numérica real que refleja el volumen corpuscular medio
- **alkphos**: Variable numérica real que refleja la fosfatasa alcalina, una enzima responsable de eliminar grupos de fosfatos de varios tipos de moléculas como nucleótidos, proteínas y otros compuestos fosforilados. Los niveles de fosfatasa alcalina elevados podrían ser signo de daño en el hígado.

- **sgpt**: Variable numérica real que refleja la alanina aminotransferasa, enzima que se encuentra principalmente en las células del hígado. Niveles altos de esta puede indicar que tiene algún tipo de daño en el hígado.
- **sgot**: Variable numérica real que refleja la aspartato aminotransferasa, otra enzima del hígado. Los niveles elevados de esta en la sangre pueden indicar hepatitis, cirrosis, mononucleosis u otras enfermedades del hígado
- **gammagt**: Variable numérica real que refleja la gamma-glutamil transpeptidasa, es una enzima hepática. Se mide su nivel en sangre siendo un marcador de laboratorio de enfermedad hepática (mala en altos niveles).
- **selector**: Variable numérica entera creada por los investigadores para dividir los datos en el conjunto de train y test.

La variable dependiente utilizada para la clasificación es **drinks**, un valor numérico real que refleja el número de medias pintas equivalentes a la cantidad de bebidas alcohólicas que se beben por día.

## 3.2. Planteamiento de hipótesis

- **Una mayor masa corporal puede dar lugar a casos en los que un alto consumo de alcohol no se vincule con altos resultados en los análisis de sangre.**
- **¿Un valor alto en uno de los test de análisis de sangre equivale a un valor alto en el resto de test?**

## 3.3. Procesamiento de los datos

Siguiendo la misma idea que con el dataset anterior se pretende profundizar en los diferentes atributos que componen este conjunto de datos para ampliar el conocimiento sobre este y determinar cualquier característica que facilite el posterior desarrollo de modelos.

El primer paso es la búsqueda de valores perdidos dentro de los datos, concluyendo en que este dataset no posee ningún Missing value. Sin embargo se detectan 4 filas duplicadas, las cuales procedemos a eliminar, reduciéndose el número de muestras a 341.

### 3.3.1. Análisis de las variables independientes

Se analiza el comportamiento de las diferentes variables mediante el calculo de diversas medidas de posición: la media aritmética, mediana, primer y tercer cuartil, valores máximos y mínimos de cada variable. También se estudia la dispersión de las distribuciones mediante el calculo de la desviación típica, mientras que la normalidad de los datos se estudia con los coeficientes de Skewness y Kurtosis.

Se estudia cada variable en detalle mediante el calculo de las medidas previamente mencionadas. Dicho estudio es acompañado con una serie de representaciones gráficas que facilite la comprensión de los resultados:

- **mcv**

| | mcv |
|---|---|
| Valor mínimo | 65.00 |
| Primer cuartil | 87.00 |
| Mediana | 90.00 |
| Media | 90.12 |
| Tercer cuartil | 92.00 |
| Valor máximo | 103.00 |
| Desviación estandar | 4.4523855 |
| Coeficiente de skewness | -0.3765269 |
| Coeficiente de Kurtosis | 5.542126 |

Se presenta una variable con un rango de valores definidos entre [65, 103]. Con un centro de distribución muy levemente desplazado a la derecha, se presenta una leve dispersión de los datos respecto a este centro de dispersión. Se profunda la información mediante un diagrama boxplot y un histograma:

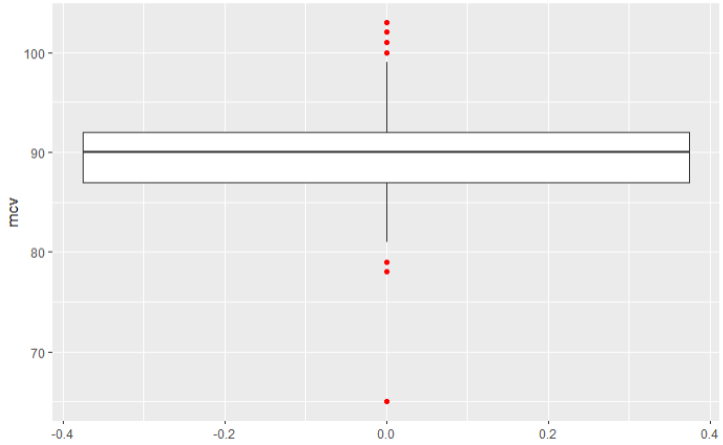



Figura 3.1: Diagrama de cajas mcv

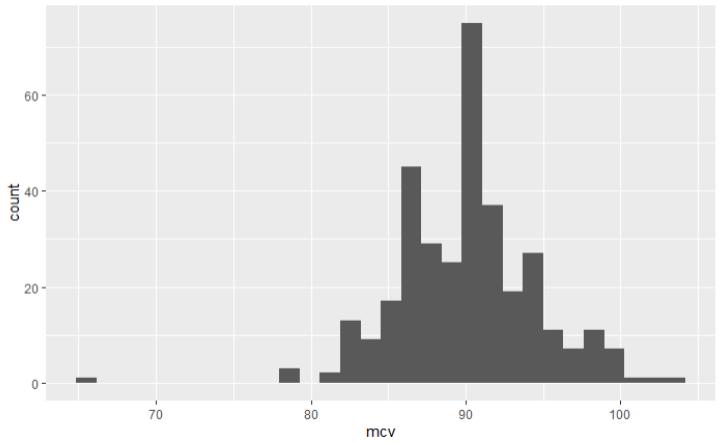



Figura 3.2: Histograma mcv

Se estudia la posibilidad de que los datos sigan una distribución normal a través del test Shapiro-Wilk, no obstante el p-value resultante no supera el nivel de significancia definido en 0.05, por lo que rechazamos la hipótesis nula y negamos una distribución Normal.

- **alkphos**:

| | alkphos |
|---|---|
| Valor mínimo | 23.00 |
| Primer cuartil | 57.00 |
| Mediana | 67.00 |
| Media | 69.89 |
| Tercer cuartil | 80.00 |
| Valor máximo | 138.00 |
| Desviación estandar | 18.4319883 |
| Coeficiente de skewness | 0.7457300 |
| Coeficiente de Kurtosis | 3.690844 |

El dominio de esta variable se halla en el intervalo [23, 138]. Se aprecia una leve desplazamiento del centro de la distribución hacia la izquierda presentando cierta dispersión de los datos respecto este centro.

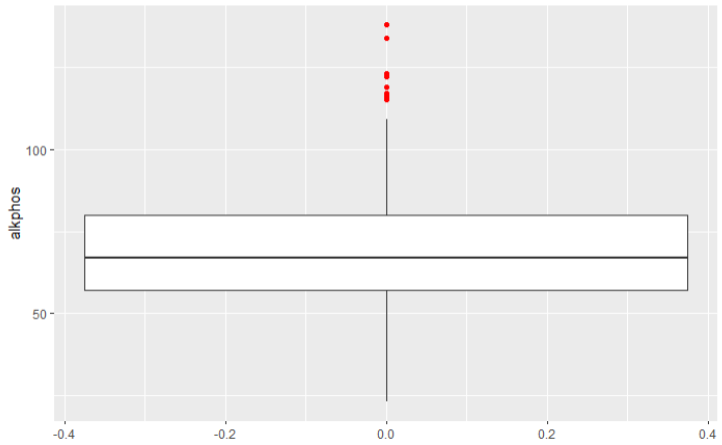



Figura 3.3: Diagrama de cajas alkphos

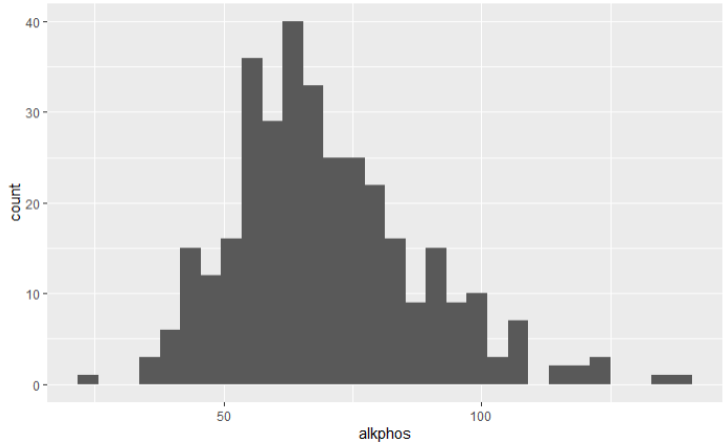



Figura 3.4: Histograma alkphos

De nuevo la distribución se asemeja a una distribución normal, por ello una vez más con el uso del test de Shapiro-Wilk confirmamos un rechazo de la hipótesis nulas, por lo que se niega que la distribución sea normal.

- **sgpt**:

| | sgpt |
|---|---|
| Valor mínimo | 4.00 |
| Primer cuartil | 19.00 |
| Mediana | 26.00 |
| Media | 30.51 |
| Tercer cuartil | 34.00 |
| Valor máximo | 155.00 |
| Desviación estandar | 19.5862490 |
| Coeficiente de skewness | 3.0385262 |
| Coeficiente de Kurtosis | 16.470290 |

La variable tratada posee un rango de valores dentro del intervalo [4, 155]. Diferente a los casos anteriores, esta vez se observa una destacable distribución de los datos respecto el centro de la distribución, el cual se sitúa desplazado a la izquierda. Podemos asumir una elevada presencia de outliers a la derecha.

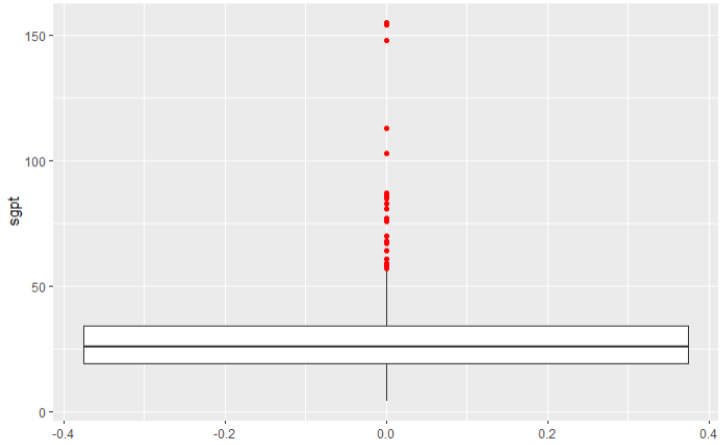



Figura 3.5: Diagrama de cajas sgpt

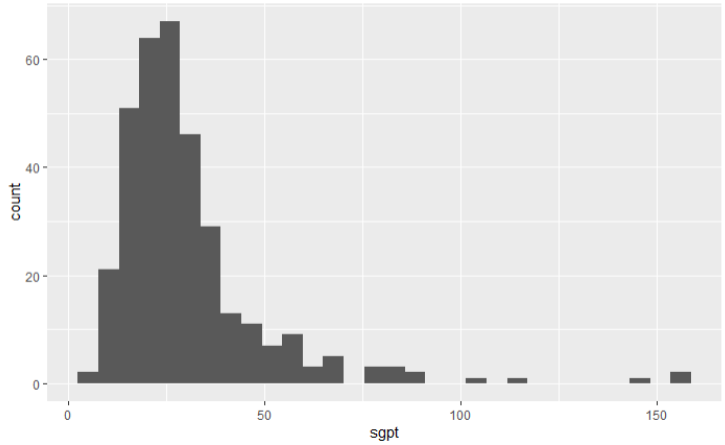



Figura 3.6: Histograma sgpt

En este caso a simple vista podemos confirmar que los datos no presenta una distribución normal, lo cual es confirmado por el test Shapiro-Wilk.

- **sgot**:

| | sgot |
|---|---|
| Valor mínimo | 5.00 |
| Primer cuartil | 19.00 |
| Mediana | 23.00 |
| Media | 24.66 |
| Tercer cuartil | 27.00 |
| Valor máximo | 82.00 |
| Desviación estandar | 10.1155409 |
| Coeficiente de skewness | 2.2703414 |
| Coeficiente de Kurtosis | 10.894188 |

Con un dominio dentro del intervalo [5, 82], esta variable presenta un comportamiento similar a la anterior. Un centro de la distribución desplazado ligeramente a la izquierda con una alta dispersión en los datos, lo cual de nuevo dará lugar a una presencia de outliers a la derecha del centro de la distribución.

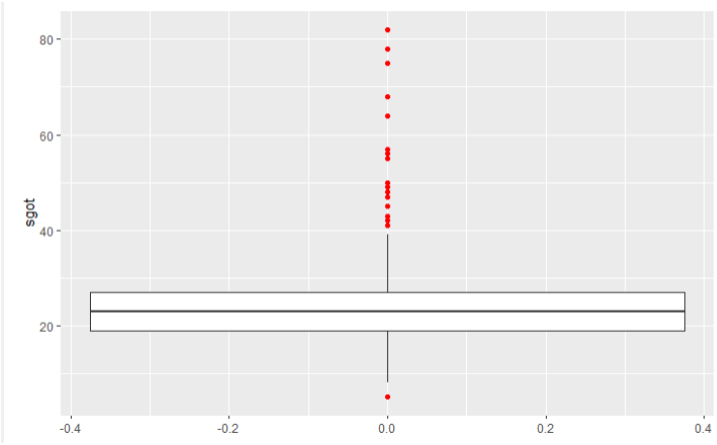



Figura 3.7: Diagrama de cajas sgot

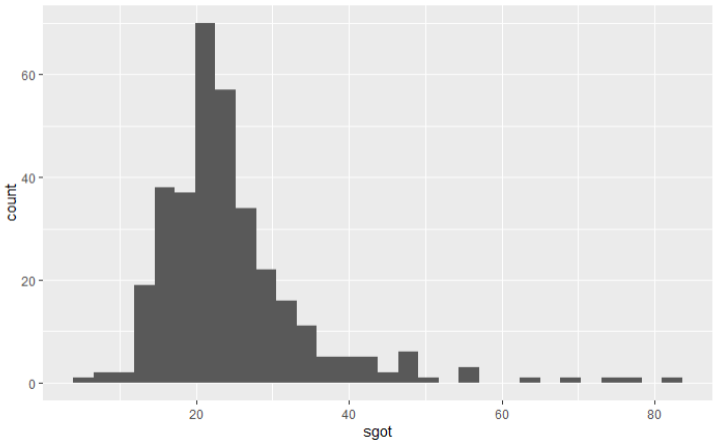



Figura 3.8: Histograma sgot

De nuevo nos alejamos de una distribución normal de los datos.

- **gammagt**:

| | gammagt |
|---|---|
| Valor mínimo | 5.0 |
| Primer cuartil | 15.0 |
| Mediana | 25.0 |
| Media | 38.4 |
| Tercer cuartil | 46.0 |
| Valor máximo | 297.0 |
| Desviación estandar | 39.4393786 |
| Coeficiente de skewness | 2.8387489 |
| Coeficiente de Kurtosis | 13.180100 |

La variable tratada posee los datos dentro del rango [5, 297]. De nuevo un ligero desplazamiento del centro de la distribución a la izquierda junto con una dispersión de estos valores. Se asume una vez más la presencia de outliers a la derecha del centro.

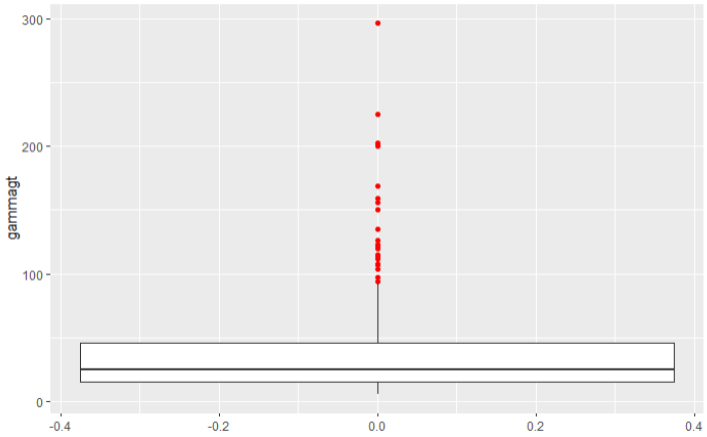



Figura 3.9: Diagrama de cajas gammagt

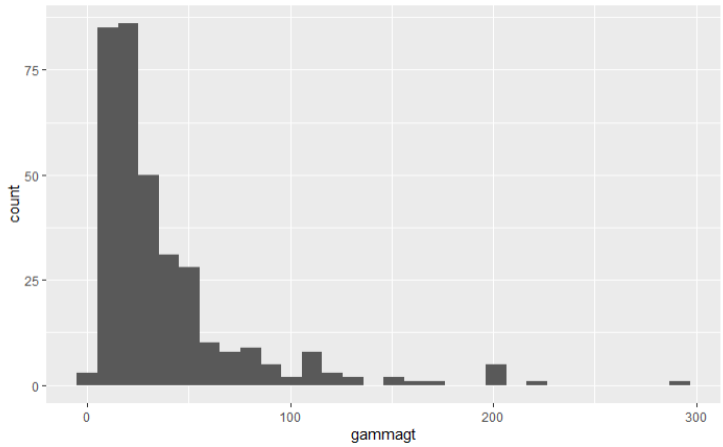



Figura 3.10: Histograma gammagt

Se descarta una distribución normal.

- **selector**: Esta variable no merece de estudio profundo pues valdrá 1 o 2 según el dato pertenezca al conjunto de entrenamiento o test.

### 3.3.2. Análisis de las variable dependiente

Finalicemos analizando la variable dependiente, **drinks**:

| | drinks |
|---|---|
| Valor mínimo | 0.000 |
| Primer cuartil | 0.500 |
| Mediana | 3.000 |
| Media | 3.431 |
| Tercer cuartil | 5.000 |
| Valor máximo | 20.000 |
| Desviación estandar | 3.3416404 |
| Coeficiente de skewness | 1.5616997 |
| Coeficiente de Kurtosis | 6.673044 |

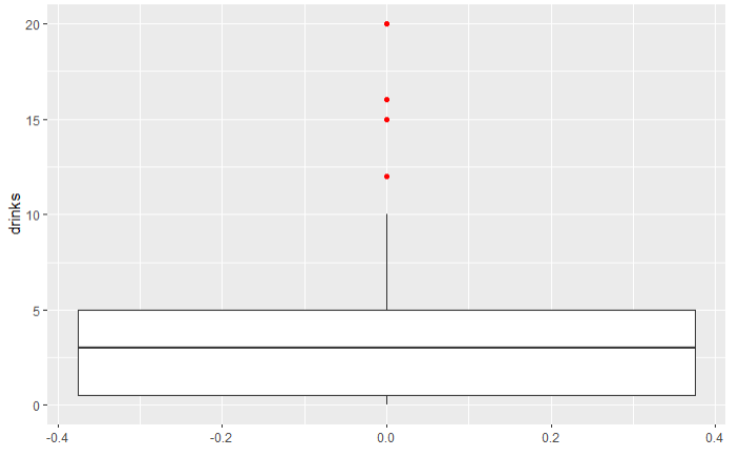



Figura 3.11: Diagrama de cajas drinks

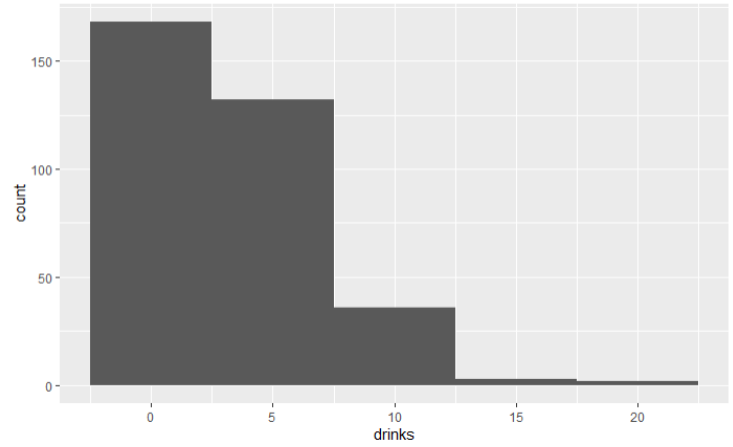



Figura 3.12: Histograma drinks

Se presenta un problema con la variable dependiente, pues existen 16 posibles clases diferentes. Para un modelo de clasificación sobre un dataset este es un valor de clases muy elevado, por ello, tras un proceso de documentación, varios artículos coinciden en separar la variable dependiente en dos posibles valores dependiendo si el valor de drinks $> 5$. [2] Siguiendo esta idea y separando los datos según la variable 'selector', la nueva variable drinks contiene:

| | selector == 1 | selector == 2 |
|---|---|---|
| drinks <= 5 | 100 | 157 |
| drinks >5 | 45 | 43 |

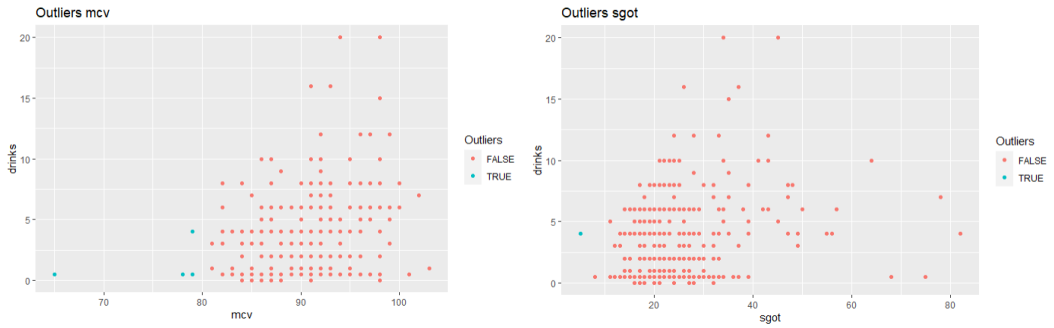
Se observa un desnivel de casos, algo que se podría solucionar si en lugar de drinks $> 5$, se toman aquellas drinks $> 3$ como punto de separación. Estos nos permitiría obtener dos clases con un número similar de casos, sin embargo

sería una separación artificial. Según la información encontrada, tomar menos de 5.5 pintas de medía al día tiene efectos muy bajos en la posibilidad de problemas en el hígado. Mientras que beber por encima de esta cantidad si suele estar vinculado a problemas en el hígado. [3] Por este motivo considero mejor opción tomar drinks $> 5$ como separación en dos posibles clases. De esta forma se ha reducido el número de clases de la variable drinks a 2, tomando un valor de 0 si drinks $\leq 5$ y un valor de 1 si drinks $> 5$.

### 3.3.3. Análisis de valores anómalos

Durante el estudio de las variables se han detectado varios casos en los que la presencia de valores anómalos debe ser considerada y estudiada.

Respecto a los outliers situados a la izquierda:




(a) (b)

Respecto a los outliers situados a la derecha:

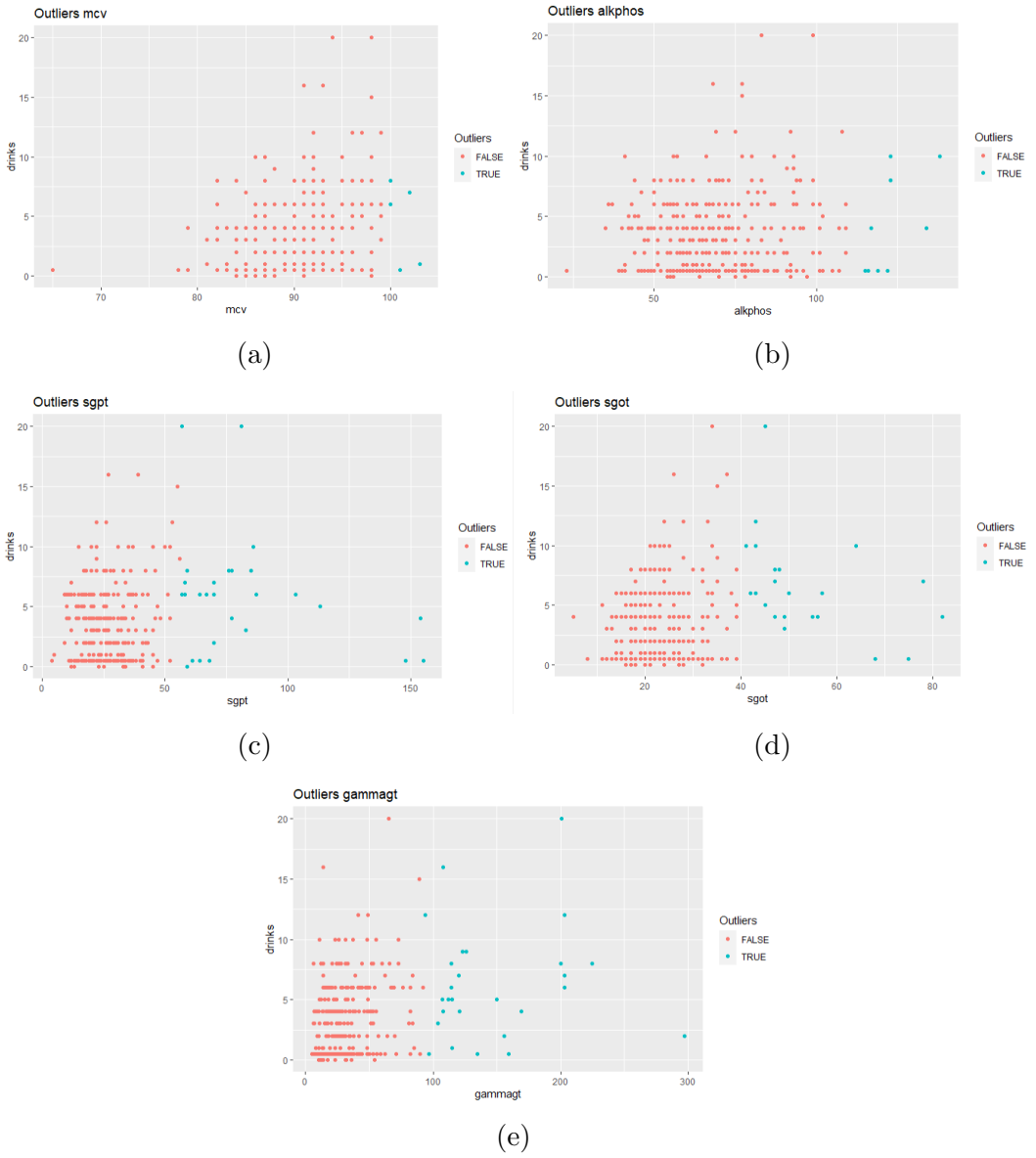



Figura 3.14: Outliers Bupa

En casos como la variable 'sgpt' se observa que la presencia de outliers se debe a que los datos poseen una alta concentración en una zona en concreto. Pese a la existencia de estos casos con valores extremos, considero que su eliminación sería un error. En primer lugar tratamos con un dataset muy pequeño, por lo tanto la eliminación de casos será notable. Por otra parte, el estudio de casos donde los diferentes test de análisis de sangre han obtenido resultados elevados puede ser muy importante para definir casos en los que el consumo de alcohol es elevado y por tanto las posibilidades de una enfermedad de hígado son mayores.

## 3.4. Análisis de las relaciones entre variables

Analizada las distribuciones de las diferentes variables que forman este dataset, se procede en esat sección al análisis de las posibles relaciones existentes entre las mismas, con el objetivo de determinar la posible existencia de relaciones entre variables, detalle a tener en cuenta en el posterior proceso de elaboración de modelos.

El siguiente diagrama de correlaciones permite efectuar este estudio entre cada par de variables que forman el dataset, realizando el calculo de las correlaciones mediante el test de Kendall. Se han excluido la variable 'selector' y 'drinks' para esta representación, pues es evidente que no aportarán información útil.

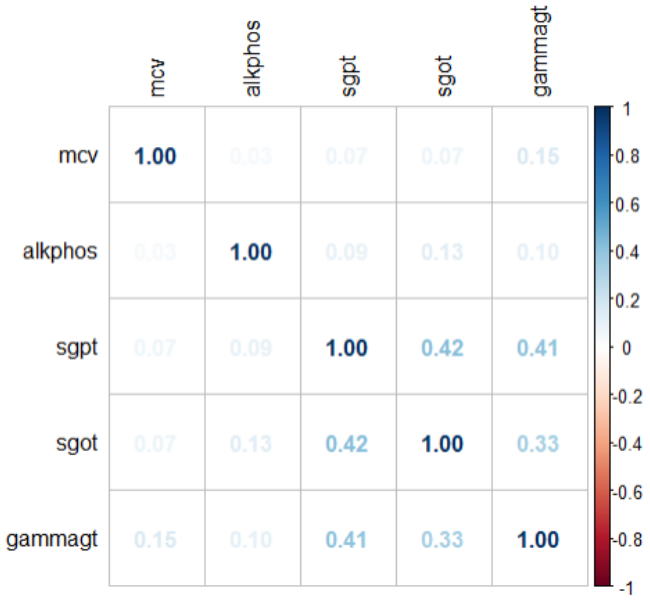



Figura 3.15

Observamos que unicamente destacan dos relaciones, la correlación positiva entre 'sgpt' - 'sgot' y 'sgpt' - 'gammagt'. Esta relación nos indica que estos test de análisis de sangre poseen unas métricas similares, y que por lo tanto, los resultados de estos poseen valores similares. Pese a ello la relación no es lo suficientemente fuerte como para ser altamente considerable. Esta posible relación será tratada en la siguiente sección.

## 3.5. Comprobación de hipótesis planteadas

**Una mayor masa corporal puede dar lugar a casos en los que un alto consumo de alcohol no se vincule con altos resultados en los análisis de sangre.**

Recordando que un valor drinks de 1 significa que el consumo está por encima del recomendado, y por tanto se clasifica como muy posible problema en el hígado, mientras que un valor de drinks de 0 indica un bajo consumo y menor posibilidad de enfermedad, nos apoyamos en las siguientes gráficas:

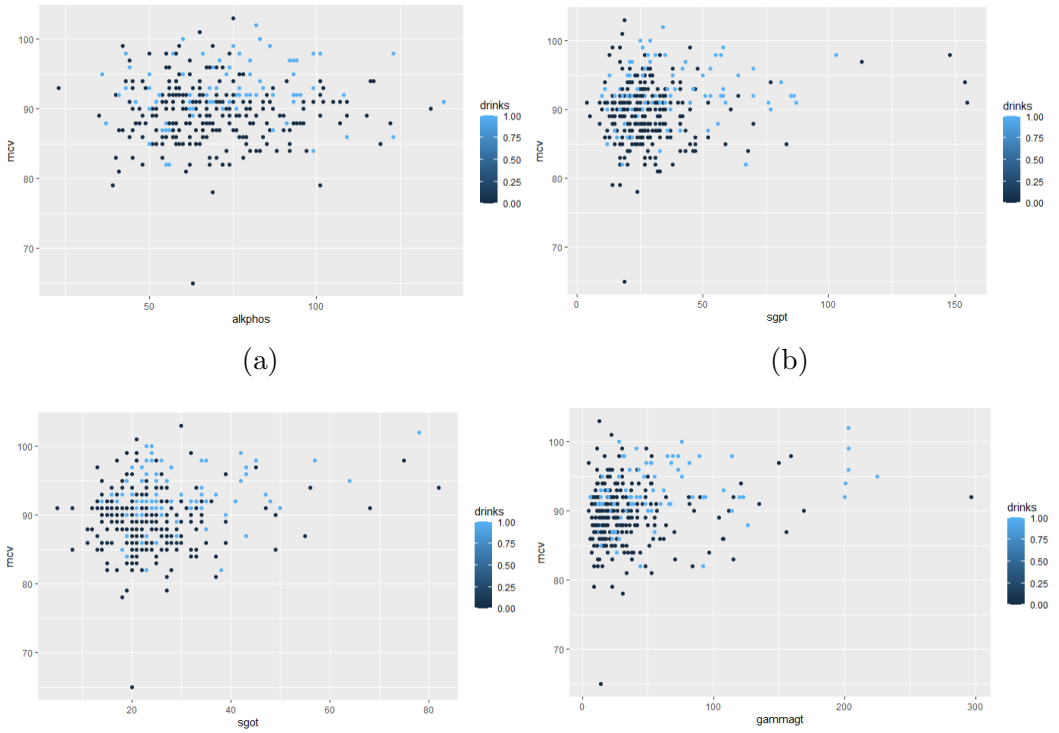



Figura 3.16: Hipótesis 1

Se observa que en varios casos con alta masa corporal (mcv), un considerado alto consumo de alcohol (drinks=1) no se vincula con un alto valor en los resultados del test de análisis de sangre, confirmando la hipótesis planteada.

De este punto se pueden sacar varias conclusiones. La primera es que la masa corporal será de importancia en el desarrollo de correctos modelos de clasificación. Por otra parte, la detección de un alto consumo de alcohol no será posible con un único análisis de sangre, siendo importante la combinación de los cuatro diferentes análisis presentes en el dataset.

**¿Un valor alto en uno de los test de análisis de sangre equivale a un valor alto en el resto de test?**

Es cierto que para todas las variables que representan os resultados de un test de sangre, un valor elevado significa una mayor probabilidad de tener una enfermedad de hígado. Sin embargo, observando los diferentes dominios presentes en cada una de estas variables se observa que cada una tiene su propia métrica.

Para ello comparemos cada análisis de dos en dos:

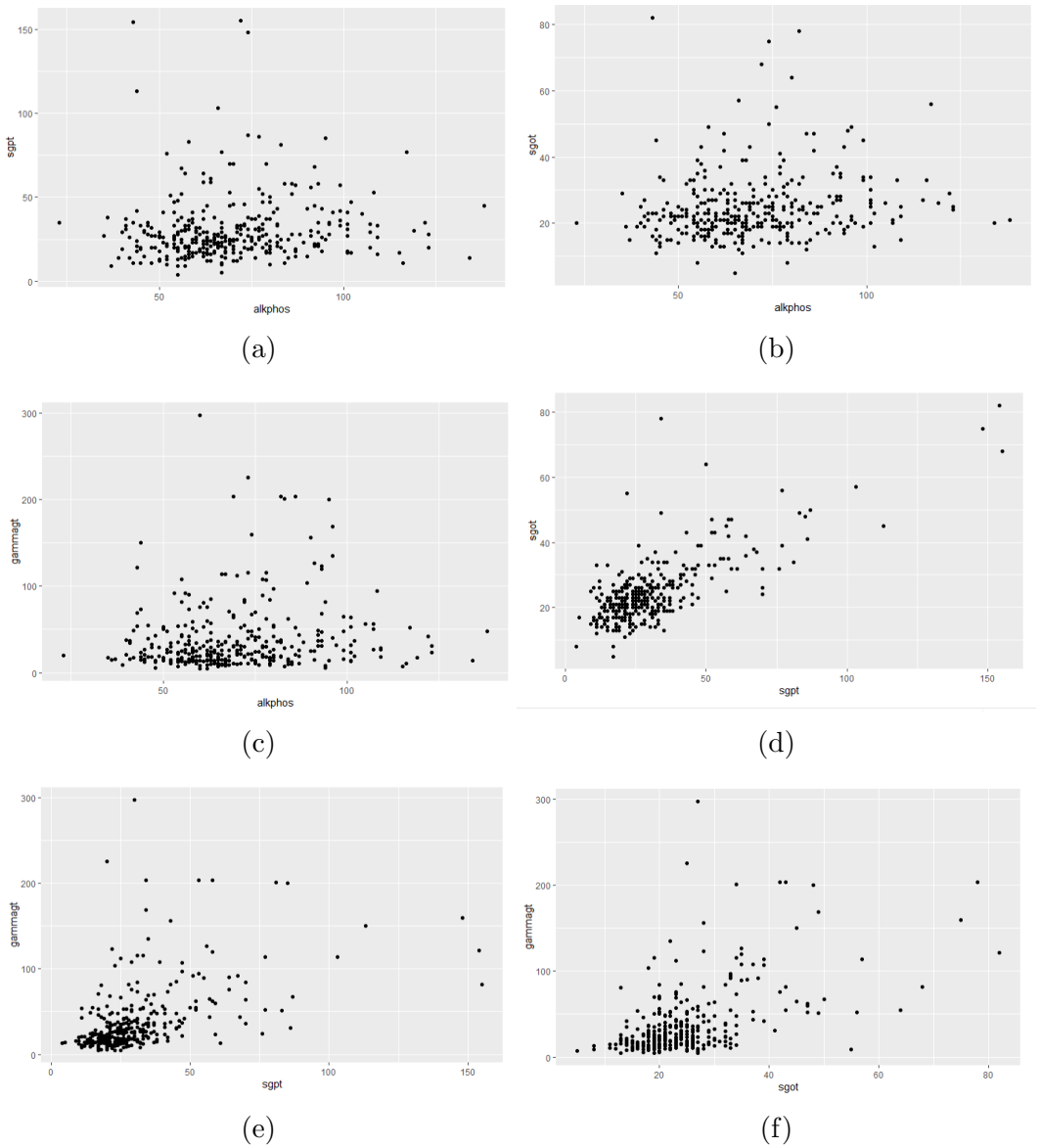



Figura 3.17: Hipótesis 2

En la mayoría de los casos se observa que un resultado elevado en uno de los test de sangre no significa que en los demás se obtenga un valor también

elevado. El único caso en el que un valor alto en ese test de sangre equivale a un valor alto en otro test es en la relación entre 'sgot' y 'sgpt'.

# Capítulo 4

# Regresión

## 4.1. Introducción

Analizado el dataset 'California', en esta sección se desarrollarán diferentes modelos predictivos a partir del mismo.

Comenzando con el estudio y la elaboración de modelos lineales simples, posteriormente se generarán modelos más complejos, modelos no lineales y finalmente modelos basados en K-NN, todos con el objetivo explicar la variable dependiente (MedianHouseValue).

## 4.2. Elaborar modelos lineales simples.

Se presenta un dataset formado por más de 5 variables, por ello es necesario realizar un estudio que permita determinar aquellas 5 más relevantes.

Este estudio consistirá en la generación de los diferentes modelos lineales simples, con el fin de detectar aquellas variable independientes capaces de explicar un considerable porcentaje de la variable dependiente.

Cada modelo generado es evaluado:

| Modelo Lineal | R^2 | R^2 Ajustado | RMSE | p-value |
|---|---|---|---|---|
| MedianHouseValue ~Longitude | 0.002113 | 0.002065 | 115300 | 3.923e-11 |
| MedianHouseValue ~Latitude | 0.02078 | 0.02073 | 114200 | 2.2e-16 |
| MedianHouseValue ~HousingMedianAge | 0.01116 | 0.01111 | 114800 | 2.2e-16 |
| MedianHouseValue ~TotalRooms | 0.018 | 0.01795 | 114400 | 2.2e-16 |
| MedianHouseValue ~TotalBedrooms | 0.00256 | 0.002511 | 115300 | 3.52e-13 |
| MedianHouseValue ~Population | 0.0006076 | 0.0005592 | 115400 | 0.0003976 |
| MedianHouseValue ~Households | 0.004335 | 0.004287 | 115100 | 2.2e-16 |
| MedianHouseValue ~MedianIncome | 0.4734 | 0.4734 | 83740 | 2.2e-16 |

Se representan gráficamente los diferentes modelos generados:

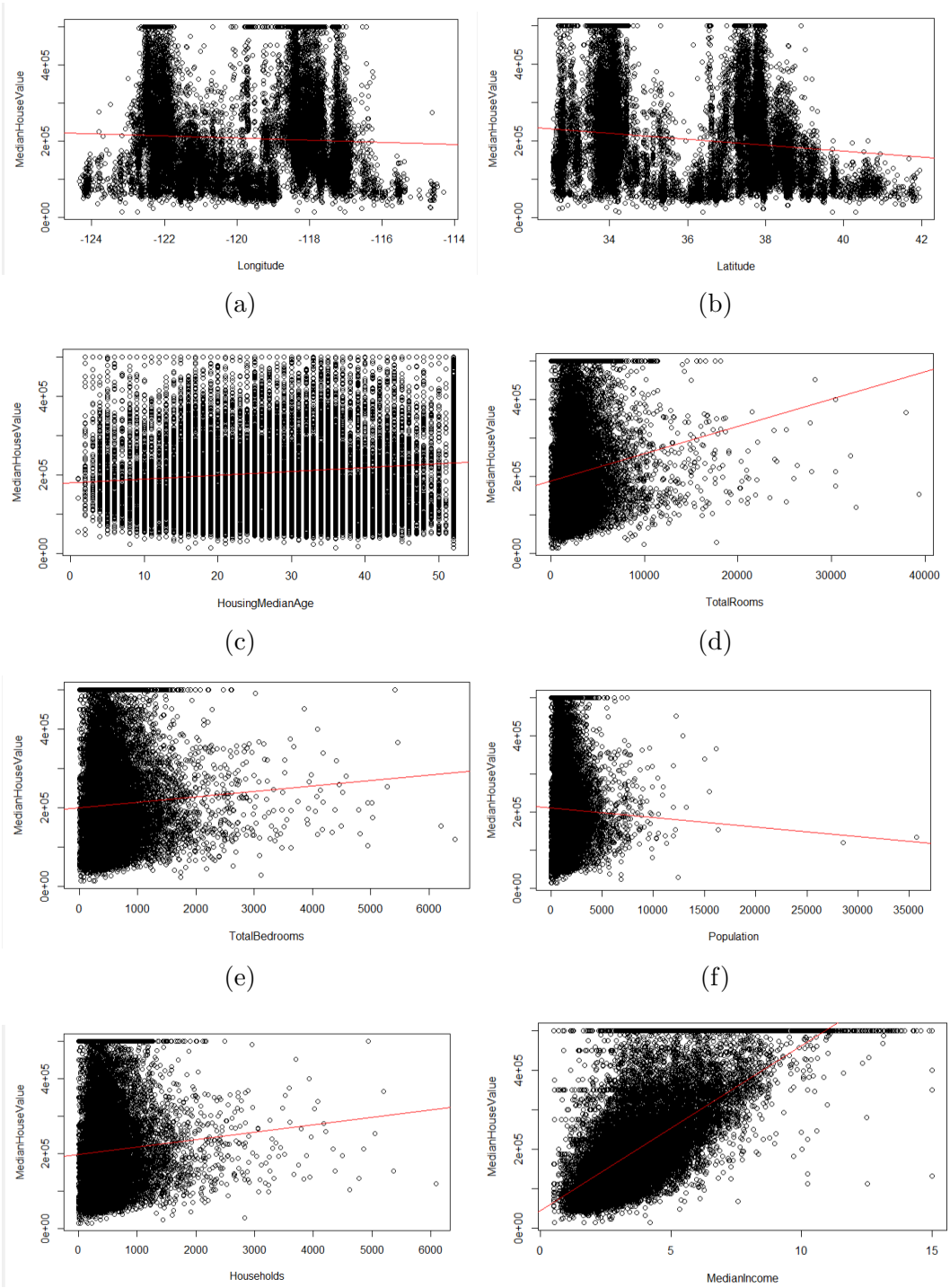



Figura 4.1: Modelos lineales simples

Se observa que todos los modelos lineales generados, exceptuando el que trabaja con la variable independiente 'MedianIncome', presentan un valor de R cuadrado ajustado muy cercano a 0, lo que permite asumir que estas variable por si mismas no pueden llegar a explicar la variable dependiente 'MedianHouseValue'.

Fijando la atención en el caso del modelo con la variable 'MedianIncome', se presenta un valor R cuadrado ajustado próximo al 0.5, lo que se traduce en que esta variable por si sola es capaz de explicar gran parte de la variable dependiente, pero no lo suficiente. La importancia de esta variable será de utilidad para la elaboración de modelos más complejos.

Evaluando las métricas resultantes se observa que las 5 variables que mejor explican la variable dependiente son:

| Modelo Lineal | Rˆ2 | Rˆ2 Ajustado | RMSE | p-value |
|---|---|---|---|---|
| MedianHouseValue ∼MedianIncome | 0.4734 | 0.4734 | 83740 | 2.2e-16 |
| MedianHouseValue ∼Latitude | 0.02078 | 0.02073 | 114200 | 2.2e-16 |
| MedianHouseValue ∼TotalRooms | 0.018 | 0.01795 | 114400 | 2.2e-16 |
| MedianHouseValue ∼HousingMedianAge | 0.01116 | 0.01111 | 114800 | 2.2e-16 |
| MedianHouseValue ∼Households | 0.004335 | 0.004287 | 115100 | 2.2e-16 |

Para finalizar este apartado se determinará cual es el mejor de estos modelos comprobando cual de ellos obtiene el menor error cuadrático medio (RMSE) en los 5 folds propuestos para este dataset.

| Modelo Lineal | Rˆ2 | Rˆ2 Ajustado | RMSE | 5-fold RMSE |
|---|---|---|---|---|
| MedianHouseValue ∼MedianIncome | 0.4734 | 0.4734 | 83740 | 7006888908 |
| MedianHouseValue ∼Latitude | 0.02078 | 0.02073 | 114200 | 13037618917 |
| MedianHouseValue ∼TotalRooms | 0.018 | 0.01795 | 114400 | 13074242952 |
| MedianHouseValue ∼HousingMedianAge | 0.01116 | 0.01111 | 114800 | 13164908193 |
| MedianHouseValue ∼Households | 0.004335 | 0.004287 | 115100 | 13256188343 |

El modelo lineal creado en función de 'MedianIncome' ha obtenido el menor error cuadrático medio en la validación cruzada de 5-folds, confirmando de nuevo que es el mejor modelo de regresión lineal simple posible.

## 4.3. Elaborar modelos regresión lineal múltiple

En este apartado se analizarán y crearán diferentes modelos que consideren varias variables independientes para explicar la variable dependiente. Para ello se procederá mediante un método descendiente, el cual consiste en partir de un modelo lineal que considere todas las posibles variables independientes y, a partir de este, se buscará la obtención de modelos progresivamente mejores mediante la eliminación de variables y/o añadiendo términos no lineales.

El resultado de este primer modelo es el siguiente:

```
Call:
lm(formula = MedianHouseValue ~ ., data = df_california)

Residuals:
    Min      1Q  Median      3Q     Max
-563013  -43592  -11327   30307  803996

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)      -3.594e+06  6.254e+04 -57.468  < 2e-16 ***
Longitude        -4.282e+04  7.130e+02 -60.061  < 2e-16 ***
Latitude         -4.258e+04  6.733e+02 -63.240  < 2e-16 ***
HousingMedianAge  1.156e+03  4.317e+01  26.787  < 2e-16 ***
TotalRooms       -8.182e+00  7.881e-01 -10.381  < 2e-16 ***
TotalBedrooms     1.134e+02  6.902e+00  16.432  < 2e-16 ***
Population       -3.854e+01  1.079e+00 -35.716  < 2e-16 ***
Households        4.831e+01  7.515e+00   6.429 1.32e-10 ***
MedianIncome      4.025e+04  3.351e+02 120.123  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 69530 on 20631 degrees of freedom
Multiple R-squared:  0.6371,    Adjusted R-squared:  0.637
F-statistic:  4528 on 8 and 20631 DF,  p-value: < 2.2e-16
```

Figura 4.2

Media RMSE del modelo lineal múltiple sobre 5-fold: 4844365688

| R^2 | R^2 Ajustado | RMSE | 5-fold RMSE |
|---|---|---|---|
| 0.6371 | 0.637 | 69530 | 4844365688 |

Se observa que el modelo generado ofrece mejor rendimiento frente a cualquiera de los modelos lineales simples generados en el apartado anterior. Por otro lado, los valores p-value del test de Wall no presentan valores significativos en ninguno de los casos, lo que complica la identificación de una variable de la que se pudiera prescindir.

Debido a esta situación, el estudio de nuevos modelos a partir de este mediante la eliminación de variables se realizará comenzando por aquellas variables que no fueron seleccionadas para el estudio de los 5 modelos lineales

simples anteriores.

Por ello se comienza eliminando la variable independiente 'Longitude', generando el siguiente modelo:

```
Call:
lm(formula = MedianHouseValue ~ . - Longitude, data = df_california)

Residuals:
    Min      1Q  Median      3Q     Max
-618282  -47761  -12245   33997  775111

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.247e+05  9.551e+03   13.05   <2e-16 ***
Latitude         -4.637e+03  2.524e+02  -18.37   <2e-16 ***
HousingMedianAge  1.875e+03  4.495e+01   41.72   <2e-16 ***
TotalRooms       -1.695e+01  8.395e-01  -20.19   <2e-16 ***
TotalBedrooms     8.687e+01  7.465e+00   11.64   <2e-16 ***
Population       -3.878e+01  1.169e+00  -33.16   <2e-16 ***
Households        1.329e+02  8.000e+00   16.62   <2e-16 ***
MedianIncome      4.663e+04  3.444e+02  135.38   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 75360 on 20632 degrees of freedom
Multiple R-squared:  0.5737,    Adjusted R-squared:  0.5735
F-statistic:  3966 on 7 and 20632 DF,  p-value: < 2.2e-16
```

Figura 4.3

Media RMSE del modelo lineal múltiple sobre 5-fold: 5688531919

| R^2 | R^2 Ajustado | RMSE | 5-fold RMSE |
|---|---|---|---|
| 0.5737 | 0.5735 | 75360 | 5688531919 |

La eliminación de 'Longitude' da lugar a un modelo con un peor R cuadrado ajustado respecto al anterior, por ello se opta por conservar esta variable y probar eliminando la siguiente no elegida anteriormente: 'Population'.

```
Call:
lm(formula = MedianHouseValue ~ . - Population, data = df_california)

Residuals:
    Min      1Q  Median      3Q     Max
-587504  -45753  -13173   31323  473425

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)      -3.662e+06  6.441e+04  -56.85   <2e-16 ***
Longitude        -4.292e+04  7.347e+02  -58.42   <2e-16 ***
Latitude         -4.139e+04  6.929e+02  -59.74   <2e-16 ***
HousingMedianAge  1.215e+03  4.445e+01   27.33   <2e-16 ***
TotalRooms       -1.604e+01  7.798e-01  -20.57   <2e-16 ***
TotalBedrooms     1.783e+02  6.861e+00   25.99   <2e-16 ***
Households       -8.315e+01  6.751e+00  -12.32   <2e-16 ***
MedianIncome      4.254e+04  3.389e+02  125.53   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 71640 on 20632 degrees of freedom
Multiple R-squared:  0.6147,    Adjusted R-squared:  0.6145
F-statistic:  4702 on 7 and 20632 DF,  p-value: < 2.2e-16
```

Figura 4.4

Media RMSE del modelo lineal múltiple sobre 5-fold: 5130744110

| R^2 | R^2 Ajustado | RMSE | 5-fold RMSE |
|---|---|---|---|
| 0.6147 | 0.6145 | 71640 | 5130744110 |

En este caso los resultados son levemente inferiores. Se prueba por último la eliminación de la variable 'TotalBedrooms', la última de las no elegidas entre las 5 variables más relevantes en la sección anterior.

```
Call:
lm(formula = MedianHouseValue ~ . - TotalBedrooms, data = df_california)

Residuals:
    Min      1Q  Median      3Q     Max
-540492  -44503  -11764   30753  869685

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -3.506e+06  6.271e+04 -55.905   <2e-16 ***
Longitude         -4.207e+04  7.162e+02 -58.748   <2e-16 ***
Latitude          -4.229e+04  6.774e+02 -62.429   <2e-16 ***
HousingMedianAge   1.125e+03  4.341e+01  25.924   <2e-16 ***
TotalRooms        -1.772e+00  6.893e-01  -2.571   0.0102 *
Population        -4.320e+01  1.048e+00 -41.238   <2e-16 ***
Households         1.494e+02  4.343e+00  34.399   <2e-16 ***
MedianIncome       3.835e+04  3.165e+02 121.155   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 69980 on 20632 degrees of freedom
Multiple R-squared:  0.6324,    Adjusted R-squared:  0.6322
F-statistic:  5070 on 7 and 20632 DF,  p-value: < 2.2e-16
```

Figura 4.5

Media RMSE del modelo lineal múltiple sobre 5-fold: 4909112109

| R^2 | R^2 Ajustado | RMSE | 5-fold RMSE |
|---|---|---|---|
| 0.6324 | 0.6322 | 69980 | 4909112109 |

De nuevo los resultados son levemente peores. Si la eliminación de aquellas variables que no fueron seleccionada como relevantes da lugar a peores modelos podemos prácticamente deducir que con la eliminación de cada una de las 5 variables restantes los modelos obtenidos también será peores. Por ello se opta por la realización de modelos que involucren a todas las variables, pero añadiendo términos no lineales.

Siendo 'MedianIncome' la variable que mejor explica la variable dependiente comencemos añadiéndola elevada al cuadrado:

```
Call:
lm(formula = MedianHouseValue ~ . + I(MedianIncome^2), data = df_california)

Residuals:
    Min      1Q  Median      3Q     Max
-499603  -44139  -11085   30992  783310

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -3.499e+06  6.276e+04 -55.743  < 2e-16 ***
Longitude         -4.138e+04  7.195e+02 -57.503  < 2e-16 ***
Latitude          -4.109e+04  6.811e+02 -60.327  < 2e-16 ***
HousingMedianAge   1.219e+03  4.329e+01  28.165  < 2e-16 ***
TotalRooms        -1.047e+01  8.059e-01 -12.990  < 2e-16 ***
TotalBedrooms      1.274e+02  6.965e+00  18.291  < 2e-16 ***
Population        -3.710e+01  1.081e+00 -34.323  < 2e-16 ***
Households         4.060e+01  7.511e+00   5.405 6.55e-08 ***
MedianIncome       5.030e+04  8.652e+02  58.136  < 2e-16 ***
I(MedianIncome^2) -8.539e+02  6.782e+01 -12.590  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 69260 on 20630 degrees of freedom
Multiple R-squared:  0.6399,    Adjusted R-squared:  0.6397
F-statistic:  4073 on 9 and 20630 DF,  p-value: < 2.2e-16

RMSE de 5-fold sobre Households: 4807345588
```

Figura 4.6

| R^2 | R^2 Ajustado | RMSE | 5-fold RMSE |
|---|---|---|---|
| 0.6399 | 0.6397 | 69260 | 4807345588 |

Se observa la generación de un modelo levemente superior respecto al formado por todas las variables. Se prueba a continuación elevándola a 4:

```
Call:
lm(formula = MedianHouseValue ~ . + I(MedianIncome^4), data = df_california)

Residuals:
    Min      1Q  Median      3Q     Max
-420863  -43736  -10739   30866  773489

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -3.425e+06  6.235e+04 -54.933  < 2e-16 ***
Longitude         -4.049e+04  7.135e+02 -56.753  < 2e-16 ***
Latitude          -4.018e+04  6.752e+02 -59.506  < 2e-16 ***
HousingMedianAge   1.265e+03  4.300e+01  29.427  < 2e-16 ***
TotalRooms        -1.222e+01  8.020e-01 -15.243  < 2e-16 ***
TotalBedrooms      1.373e+02  6.917e+00  19.849  < 2e-16 ***
Population        -3.639e+01  1.072e+00 -33.953  < 2e-16 ***
Households         3.727e+01  7.450e+00   5.002 5.71e-07 ***
MedianIncome       4.673e+04  4.484e+02 104.206  < 2e-16 ***
I(MedianIncome^4) -4.333e+00  2.021e-01 -21.447  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 68770 on 20630 degrees of freedom
Multiple R-squared:  0.645,     Adjusted R-squared:  0.6449
F-statistic:  4165 on 9 and 20630 DF,  p-value: < 2.2e-16

RMSE de 5-fold sobre Households: 4738327923
```

Figura 4.7

| R^2 | R^2 Ajustado | RMSE | 5-fold RMSE |
|---|---|---|---|
| 0.645 | 0.6449 | 68770 | 4738327923 |

Se produce una muy leve mejora, seguir por este camino no dará lugar a mejores resultados. Dado que tener en cuenta una de la variable más representativa origina modelos con mejores resultamos, añadamos el producto de las dos siguientes variables determinada como más relevante, 'Latitude' y 'TotalRooms':

```
Call:
lm(formula = MedianHouseValue ~ . + I(MedianIncome^4) * Latitude *
    TotalRooms, data = df_california)

Residuals:
    Min      1Q  Median      3Q     Max
-359971  -43560  -10597   30665  787456

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                       -3.394e+06  6.382e+04 -53.188  < 2e-16 ***
Longitude                         -4.022e+04  7.159e+02 -56.180  < 2e-16 ***
Latitude                          -4.013e+04  7.300e+02 -54.968  < 2e-16 ***
HousingMedianAge                   1.255e+03  4.289e+01  29.270  < 2e-16 ***
TotalRooms                        -2.147e+01  4.192e+00  -5.122 3.05e-07 ***
TotalBedrooms                      1.513e+02  7.030e+00  21.519  < 2e-16 ***
Population                        -3.566e+01  1.071e+00 -33.297  < 2e-16 ***
Households                         3.269e+01  7.477e+00   4.372 1.23e-05 ***
MedianIncome                       4.728e+04  4.502e+02 105.016  < 2e-16 ***
I(MedianIncome^4)                 -2.010e+00  4.713e+00  -0.426    0.670
Latitude:I(MedianIncome^4)        -1.271e-01  1.317e-01  -0.965    0.335
TotalRooms:I(MedianIncome^4)       1.026e-03  1.474e-03   0.696    0.487
Latitude:TotalRooms                1.723e-01  1.157e-01   1.489    0.136
Latitude:TotalRooms:I(MedianIncome^4) -6.048e-06  4.150e-05  -0.146    0.884
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 68570 on 20626 degrees of freedom
Multiple R-squared:  0.6471,    Adjusted R-squared:  0.6469
F-statistic:  2910 on 13 and 20626 DF,  p-value: < 2.2e-16

RMSE de 5-fold sobre Households: 4715278045
```

Figura 4.8

| R^2 | R^2 Ajustado | RMSE | 5-fold RMSE |
|---|---|---|---|
| 0.6471 | 0.6469 | 68570 | 4715278045 |

Se observa un modelo con unos mejores resultados, sin embargo, al observar los valores 'p-value' se desconfía de algunas variables añadidas en la multiplicación.

Por ello se considera en eliminar la variable 'Latitude' multiplicando.

```
Call:
lm(formula = MedianHouseValue ~ . + I(MedianIncome^4) * TotalRooms,
    data = df_california)

Residuals:
    Min      1Q  Median      3Q     Max
-357380  -43557  -10523   30585  781449

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                -3.399e+06  6.222e+04 -54.620  < 2e-16 ***
Longitude                  -4.013e+04  7.123e+02 -56.337  < 2e-16 ***
Latitude                   -3.970e+04  6.747e+02 -58.832  < 2e-16 ***
HousingMedianAge            1.256e+03  4.288e+01  29.283  < 2e-16 ***
TotalRooms                 -1.540e+01  8.508e-01 -18.099  < 2e-16 ***
TotalBedrooms               1.502e+02  6.998e+00  21.466  < 2e-16 ***
Population                 -3.576e+01  1.070e+00 -33.407  < 2e-16 ***
Households                  3.416e+01  7.434e+00   4.595 4.36e-06 ***
MedianIncome                4.726e+04  4.498e+02 105.073  < 2e-16 ***
I(MedianIncome^4)          -6.525e+00  2.843e-01 -22.951  < 2e-16 ***
TotalRooms:I(MedianIncome^4)  8.308e-04  7.605e-05  10.925  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 68570 on 20629 degrees of freedom
Multiple R-squared:  0.6471,    Adjusted R-squared:  0.6469
F-statistic:  3782 on 10 and 20629 DF,  p-value: < 2.2e-16

RMSE de 5-fold sobre Households: 4713390798
```

Figura 4.9

| R^2 | R^2 Ajustado | RMSE | 5-fold RMSE |
|---|---|---|---|
| 0.6471 | 0.6469 | 68570 | 4713390798 |

Efectivamente la variable 'Latitude' multiplicando no aportaba ninguna mejor.

Se decide continuar agregando los siguientes términos más significativos multiplicando, 'HousingMedianAge' y 'Households'

```
                                                    t value Pr(>|t|)
(Intercept)                                        -54.720  < 2e-16 ***
Longitude                                          -56.983  < 2e-16 ***
Latitude                                           -59.515  < 2e-16 ***
HousingMedianAge                                     1.452  0.14644
TotalRooms                                          -6.090 1.15e-09 ***
TotalBedrooms                                       19.083  < 2e-16 ***
Population                                         -33.494  < 2e-16 ***
Households                                          -5.025 5.07e-07 ***
MedianIncome                                        91.613  < 2e-16 ***
I(MedianIncome^4)                                   -9.634  < 2e-16 ***
TotalRooms:I(MedianIncome^4)                         1.578  0.11459
HousingMedianAge:I(MedianIncome^4)                   0.540  0.58901
HousingMedianAge:TotalRooms                         -2.731  0.00633 **
Households:I(MedianIncome^4)                        -0.380  0.70397
TotalRooms:Households                                6.349 2.22e-10 ***
HousingMedianAge:Households                         12.319  < 2e-16 ***
HousingMedianAge:TotalRooms:I(MedianIncome^4)       -6.353 2.15e-10 ***
TotalRooms:Households:I(MedianIncome^4)              2.117  0.03425 *
HousingMedianAge:Households:I(MedianIncome^4)        6.586 4.64e-11 ***
HousingMedianAge:TotalRooms:Households              -4.436 9.22e-06 ***
HousingMedianAge:TotalRooms:Households:I(MedianIncome^4)  -4.553 5.33e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 67850 on 20619 degrees of freedom
Multiple R-squared:  0.6546,    Adjusted R-squared:  0.6543
F-statistic:  1954 on 20 and 20619 DF,  p-value: < 2.2e-16

RMSE de 5-fold sobre Households: 4641780595
```

Figura 4.10

| R^2 | R^2 Ajustado | RMSE | 5-fold RMSE |
|---|---|---|---|
| 0.6546 | 0.6543 | 67850 | 4641780595 |

De nuevo se mejoran los resultados, pero se observa que en varios casos en los que interviene I($MedianIncome^4$) aparecen valores de 'p-value' que añaden desconfianza a este término, por ello se decide probar sin elevarlo a 4.

```
                                                    Pr(>|t|)
(Intercept)                                          < 2e-16 ***
Longitude                                            < 2e-16 ***
Latitude                                             < 2e-16 ***
HousingMedianAge                                    0.000437 ***
TotalRooms                                           < 2e-16 ***
TotalBedrooms                                        < 2e-16 ***
Population                                           < 2e-16 ***
Households                                          0.037620 *
MedianIncome                                         < 2e-16 ***
TotalRooms:MedianIncome                              < 2e-16 ***
HousingMedianAge:MedianIncome                       0.000298 ***
HousingMedianAge:TotalRooms                         0.012710 *
Households:MedianIncome                              < 2e-16 ***
TotalRooms:Households                               5.21e-06 ***
HousingMedianAge:Households                         1.58e-06 ***
HousingMedianAge:TotalRooms:MedianIncome             < 2e-16 ***
TotalRooms:Households:MedianIncome                  0.045215 *
HousingMedianAge:Households:MedianIncome             < 2e-16 ***
HousingMedianAge:TotalRooms:Households              0.259299
HousingMedianAge:TotalRooms:Households:MedianIncome 3.59e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 67370 on 20620 degrees of freedom
Multiple R-squared:  0.6594,    Adjusted R-squared:  0.6591
F-statistic:  2102 on 19 and 20620 DF,  p-value: < 2.2e-16

RMSE de 5-fold sobre Households: 4577783319
```

Figura 4.11

| R^2 | R^2 Ajustado | RMSE | 5-fold RMSE |
|---|---|---|---|
| 0.6594 | 0.6591 | 67370 | 4577783319 |

La suposición era correcta y de nuevo han mejorado los resultados. Sin embargo, los valores *p-value* indican que no es necesaria la presencia de algunos de los términos del modelo (caso de la multiplicación de HousingMedianAge:TotalRooms:Households).

Finalmente nos quedamos con este último modelo como mejor modelo múltiple generado.

## 4.4. Aplicar el algoritmo k-NN para regresión

En esta sección se elaboran y estudian modelos basados en k-NN. Este estudio se realizará sobre el mejor modelo obtenido en la sección anterior con diferentes valores de k (número de vecinos más cercanos), con el objetivo de determinar el valor de k más cercano al óptimo.

Por tanto la siguiente tabla recoge, para diferentes valores de k, el valor de RMSE obtenido tras realizar la validación cruzada de 5-folds, para cada caso, junto con la media de estos resultados. Utilizando en todos ellos el mejor modelo múltiple conseguido en el apartado anterior: MedianHouseValue . +MedianIncome*TotalRooms*HousingMedianAge*Households.

| k | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Media RMSE |
|---|---|---|---|---|---|---|
| 5 | 4541760347 | 4530471424 | 4565272732 | 4385513226 | 4745290191 | 4553661584 |
| 7 | 4283086269 | 4381302961 | 4329091939 | 4106711363 | 4473274188 | 4314693344 |
| 10 | 4157055323 | 4196558009 | 4173551945 | 4018343901 | 4313895845 | 4171881005 |
| 20 | 4046433324 | 4096849734 | 4069779207 | 3898401005 | 4167615566 | **4051576629** |
| 25 | 4046433324 | 4096849734 | 4069779207 | 3898401005 | 4162774625 | 4054847579 |
| 50 | 4161252917 | 4184933077 | 4160816036 | 4002315397 | 4258437778 | 4153551041 |

Siendo la media del error RMSE obtenida en el modelo de regresión múltiple de 4577783319 sobre 5-folds, determinamos que con un k=5 ya se obtienen mejores resultados con el uso de k-NN en la validación cruzada 5-folds, siendo el valor óptimo de k un valor cercano a 20.

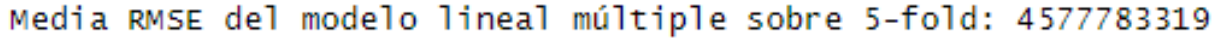



Figura 4.12

## 4.5. Comparar los resultados de los dos algoritmos de regresión múltiple

Finalmente, se comparan los resultados de los algoritmos de regresión múltiple, k-NN y, adicionalmente, el modelo M5', cuyos resultados han sido aportadas por el profesorado, ya que no ha sido tratado en el desarrollo de este proyecto. Estas comparativas se realizan con las mismas condiciones sobre cada algoritmo, no se mantienen consideraciones específicas, se trabaja con todas las variables (MedianHouseValue .).

En primer lugar, se utiliza el test de Wilcoxon, un test no paramétrico que permite determinar si existes relación entre dos muestras. De esta forma, se observa si existen diferencias significativas entre la regresión lineal múltiple y k-NN. La comparativa se realiza utilizando RMSE como medida de precisión para cada modelo. Fijando el nivel de significancia en 0.05, si el p-value resultante está por debajo de este nivel, se rechaza la hipótesis nula del test, lo que indica que uno de los modelos da unos resultados significativamente distintos a los del otro modelo.

Test modelo lineal (R+) vs modelo k-NN:(R-)

| R+ | R- | p-value |
|---|---|---|
| 78 | 93 | 0.7660294 |

Puesto que el p-value resultante posee un valor superior al nivel de significancia, no es posible rechazar que ambos algoritmos ofrecen resultados similares.

A continuación, se utiliza el test de Friedman para determinar si existen diferencias significativas entre los 3 modelos:

Friedman rank sum test

| chi-squared | df | p-value |
|---|---|---|
| 8.4444 | 2 | 0.01467 |

En este caso el valor de p-value si se sitúa por debajo del nivel de significancia, por ello se rechaza la hipótesis nula, es decir, existe al menos diferencias significativas entre un par de los algoritmos evaluados. Finalmente aplicamos el test post-hoc de Holm para averiguar qué par es diferente y cuales se pueden considerar similares

| Modelos Regresión | Lineal Múltiple | k-NN |
|---|---|---|
| k-NN | 0.580 | - |
| M5' | 0.081 | 0.108 |

Se observa la existencia de diferencias significativas entre M5' con la regresión lineal múltiple, pues su valor de p-value no superar el nivel de significancia. Los dos pares restantes poseen similitud, destacando similitud entre k-nn y regresión lineal múltiple.

# Capítulo 5

# Clasificación

## 5.1. Introducción

Analizado el dataset Bupa, en este capítulo se desarrollan diferentes modelos predictivos utilizando la información obtenida previamente. Se construirán modelos basados en k-NN (aplicado a la clasificación), modelos basados en LDA (Lineal Discriminant Analisys) y QDA (Quadratic Discriminant Analysis).

Importante: Tras la realización del EDA se ha optado por aplicar alguno de los cambios efectuados. Se han eliminado aquellas filas detectadas como repetidas, se ha eliminado la columna 'selector', y se han normalizado los datos. El cambio más importante es que la variable dependiente, 'drinks', ha sido separada en dos clases tal y como se justifica en el apartado 3.3.2

## 5.2. Utilizar el algoritmo k-NN probando con diferentes valores de k

En primer lugar, se elaboran y estudian modelos de clasificación bastados en k-NN usando como métrica de rendimiento el Accuracy (precisión): la fracción de predicciones que el modelo realizó correctamente. Se crearán diferentes modelos cada uno con un valor de k diferente.

| | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Fold 6 | Fold 7 | Fold 8 | Fold 9 | Fold 10 | Media |
|---|---|---|---|---|---|---|---|---|---|---|---|
| k=1 | 0.7058824 | 0.7352941 | 0.7000000 | 0.6969697 | 0.7647059 | 0.7058824 | 0.7647059 | 0.5588235 | 0.6060606 | 0.6764706 | 0.6914795 |
| k=3 | 0.7647059 | 0.6764706 | 0.8333333 | 0.6363636 | 0.8529412 | 0.7647059 | 0.7647059 | 0.7647059 | 0.7575758 | 0.7352941 | 0.7550802 |
| k=5 | 0.7647059 | 0.6764706 | 0.9000000 | 0.6060606 | 0.8529412 | 0.7352941 | 0.7941176 | 0.7352941 | 0.6969697 | 0.8235294 | 0.7585383 |
| k=7 | 0.7647059 | 0.7941176 | 0.9000000 | 0.6666667 | 0.8529412 | 0.8235294 | 0.7941176 | 0.7647059 | 0.6969697 | 0.7058824 | 0.7763636 |
| k=10 | 0.7352941 | 0.6764706 | 0.9000000 | 0.6969697 | 0.8529412 | 0.8235294 | 0.7352941 | 0.8823529 | 0.6969697 | 0.7352941 | 0.7735116 |
| k=13 | 0.7647059 | 0.7352941 | 0.8666667 | 0.6363636 | 0.8529412 | 0.7941176 | 0.7352941 | 0.8823529 | 0.6969697 | 0.7941176 | 0.7758824 |
| k=15 | 0.7647059 | 0.7352941 | 0.9333333 | 0.6666667 | 0.8823529 | 0.7941176 | 0.7941176 | 0.8529412 | 0.6969697 | 0.7941176 | **0.7914617** |
| k=17 | 0.7941176 | 0.6764706 | 0.9333333 | 0.6666667 | 0.8823529 | 0.7941176 | 0.7352941 | 0.8529412 | 0.7272727 | 0.7941176 | 0.7856684 |
| k=20 | 0.7647059 | 0.7352941 | 0.9000000 | 0.6363636 | 0.8823529 | 0.7941176 | 0.7352941 | 0.8529412 | 0.7272727 | 0.7941176 | 0.7822460 |
| k=30 | 0.7352941 | 0.7058824 | 0.9000000 | 0.5757576 | 0.8823529 | 0.7941176 | 0.7058824 | 0.8823529 | 0.6969697 | 0.7941176 | 0.7672727 |

Tras la ejecución de todos los casos, se observa que el modelo que ha obtenido el mejor rendimiento es aquel en el que el valor de k = 15, con una precisión del 79,1 %.

## 5.3. Utilizar el algoritmo LDA para clasificar.

En esta sección, se elaboran y estudian modelos de clasificacion bastados en LDA. Este tipo de modelo se realizan en base a una serie de asunciones, las cuales se deben garantizar para asegurar el correcto rendimiento del modelo.

La primera de estas asunciones se centra en la normalidad en las distribuciones de cada atributo. Como ya se observó durante la realización del análisis exploratorio 3.3.1, el test de Shapiro-Wilk confirmó que no existe ninguna distribución normal dentro de los datos.

Otra asunción de este método es que todas las variables deben tener una varianza similar. Observemos en la siguiente tabla si es ese el caso:

| Variables | Varianza |
|---|---|
| mcv | 19.8237364 |
| alkphos | 339.7381922 |
| sgpt | 383.6211489 |
| sgot | 102.3241677 |
| gammagt | 1555.4645851 |

Fácilmente se observa que no hay similitud entre estos valores.

Pese a no cumplir con las asunciones previas, se realiza el modelo LDA sin tener la garantía de correcto rendimiento del modelo. Para ello, se utiliza la función 'lda()' del paquete 'MASS'. Este método requiere especificar la variable a predecir y aquellas que se usarán para la predicción. Tanto en este caso como en el posterior modelo QDA se han utilizado todas las variables independientes para la predicción menos la variable 'selector'.

Los resultados del modelo LDA sobre 10-folds:

| | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Fold 6 | Fold 7 | Fold 8 | Fold 9 | Fold 10 | Media |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Accuracy | 0.7058824 | 0.7352941 | 0.8666667 | 0.6363636 | 0.8235294 | 0.7941176 | 0.7058824 | 0.8823529 | 0.7575758 | 0.7941176 | 0.7701783 |

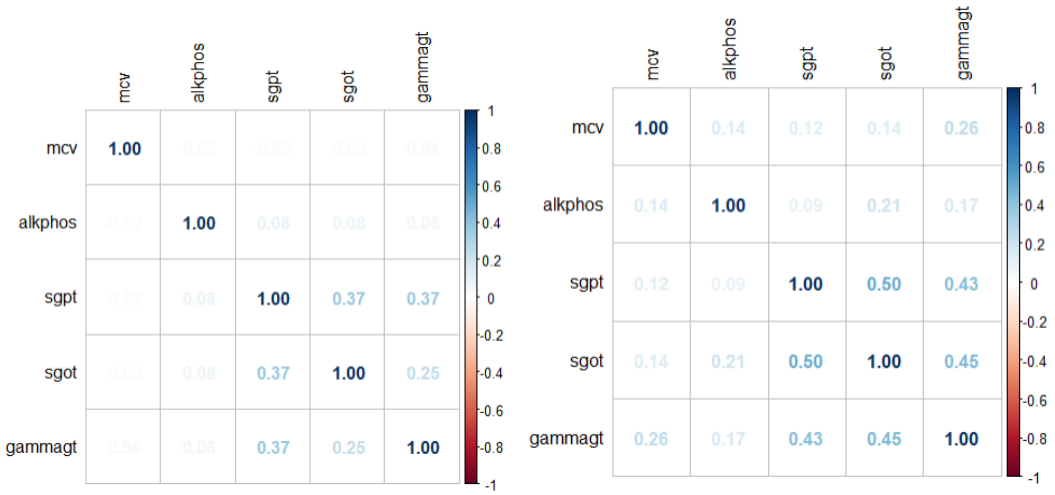
## 5.4. Utilizar el algoritmo QDA para clasificar.

Finalmente, se elaboran y estudian modelos de clasificación bastados en QDA. Respecto a las asunciones de QDA son similares a las de LDA, la única diferencia es que la similitud de las varianzas de cada variable se debe tener en cuenta respecto a cada clase a predecir. Para comprobar si se cumple, se utiliza el test de Levene:

| Variables | Df | F value | Pr(>F) |
|---|---|---|---|
| mcv | 1 | 0.0391 | 0.8434 |
| alkphos | 1 | 2.8435 | 0.09266 |
| sgpt | 1 | 10.646 | 0.001215 |
| sgot | 1 | 4.3484 | 0.03779 |
| gammagt | 1 | 15.386 | 0.0001061 |

Se observa que exceptuando el caso de la variable 'mcv', el resto no llega a un valor de p-value que supere el nivel de significancia (0.05), por lo tanto se rechaza la hipótesis nulas, confirmando que existen diferencia significativas entre las varianzas.

Otra de las asunciones de este modelo es la ausencia de niveles de correlación entre las distintas variables para cada clase. Se observan estos niveles de correlación mediante el coeficiente de Kendall para cada una de las dos clases:




(a) drinks = 0 (b) drinks = 1

En ambos casos se observa que los niveles de correlación son bajos en prácticamente todos los casos, por lo que esta asunción si se cumple.

Para ejecutar QDA se utiliza la función 'qda()' de la librería 'MASS' y una vez más se utilizan todas las variables independientes para predecir la dependiente, menos la variable 'selector'.

Obtenemos los siguientes resultados sobre 10-folds:

| | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Fold 6 | Fold 7 | Fold 8 | Fold 9 | Fold 10 | Media |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Accuracy | 0.7352941 | 0.6764706 | 0.9000000 | 0.4848485 | 0.8235294 | 0.7941176 | 0.7058824 | 0.7941176 | 0.7272727 | 0.7647059 | 0.7406239 |

## 5.5. Comparar los resultados de los tres algoritmos

Los tres algoritmos han dado lugar a resultados bastantes similares, respecto al valor de precisión generado. El mejor resultado ha sido obtenido con el algoritmo knn con un número de k = 15. Justificando unos inferiores resultados con LDA se debe partir de que no se han cumplido las asunciones que acompañan a este algoritmo lo que explica el peor rendimiento de este algoritmo. De forma similar ocurre con QDA, en ambos casos se añade el factor de que no tenemos el suficiente número de casos en cada clase.

| | knn (k=15) | LDA | QDA |
|---|---|---|---|
| Accuracy | 0.7914617 | 0.7701783 | 0.7406239 |

Aunque se haya alcanzado una precisión superior al 70 % en todos los casos, se debe recordar que partimos de un problema balance entre el número de casos para cada una de las dos clases, lo cual se traduce en una peor clasificación la clase con menor número de casos iniciales.

Se finaliza con la aplicación del test de Friedman para determinar si realmente existen diferencias significativas entre los tres modelos creados. Para ello, se toman los valores de precisión resultantes para cada modelo aportado por el profesorado. En ese mismo archivo modificaremos los resultados sobre el dataset tratado, 'Bupa', para que contenga los resultados de los algoritmos efectuados a los largo de este proyecto.

Friedman rank sum test

| chi-squared | df | p-value |
|---|---|---|
| 0.7 | 2 | 0.7047 |

Obteniendo un p-value superior al nivel de significancia no es posible rechazar la hipótesis nula de que los tres modelos evaluados presenten similitud entre ellos. En definitiva, se concluye en que los tres modelos generados ofrecen resultados similares sobre el conjunto de datos trabajado.

# Capítulo 6

# Apéndice

## 6.1. Apéndice A. Código EDA dataset de regresión 'California'

```
#Cargo librerias necesarias
library(moments)
library(ggplot2)
library(corrplot)
library(ggpubr)
library(tidyverse)

#Cargo el dataset
california <- read.csv('california.dat', comment.char = '@',
header = FALSE, stringsAsFactors = TRUE)

#Compruebo dimensiones e info de las variables
dim(california)
str(california)

#Pongo nombre a las variables
names(california) <- c("Longitude", "Latitude", "HousingMedianAge"
    ,
"TotalRooms", "TotalBedrooms", "Population", "Households",
"MedianIncome", "MedianHouseValue")

#Asignacion automatica, facilita el acceso a los campos
#n <- length(names(california)) - 1
#names(california)[1:n] <- paste ("X", 1:n, sep="")
#names(california)[n+1] <- "Y"

#Missing values
any(is.na(california))

#muestras duplicadas
sum(duplicated(california))

#Medidad importantes
```

```
summary(california)
desviacion = apply(california, 2, sd)
format(desviacion, scientific = FALSE)
skev <- apply(california, 2, skewness)
kurt <- apply(california, 2, kurtosis)
format(skev, scientific = FALSE)
format(kurt, scientific = FALSE)

#Representacion grafica de los diagramas de cajas
ggplot(california, aes(y=Longitude)) +
geom_boxplot(outlier.color = "red")
ggplot(california, aes(y=Latitude)) +
geom_boxplot(outlier.color = "red")
ggplot(california, aes(y=HousingMedianAge)) +
geom_boxplot(outlier.color = "red")
ggplot(california, aes(y=TotalRooms)) +
geom_boxplot(outlier.color = "red")
ggplot(california, aes(y=TotalBedrooms)) +
geom_boxplot(outlier.color = "red")
ggplot(california, aes(y=Population)) +
geom_boxplot(outlier.color = "red")
ggplot(california, aes(y=Households)) +
geom_boxplot(outlier.color = "red")
ggplot(california, aes(y=MedianIncome)) +
geom_boxplot(outlier.color = "red")

ggplot(california, aes(y=MedianHouseValue)) +
geom_boxplot(outlier.color = "red")


#Representacion grafica de los histogramas
ggplot(california, aes(x=Longitude)) +
geom_histogram()
ggplot(california, aes(x=Latitude)) +
geom_histogram()
ggplot(california, aes(x=HousingMedianAge)) +
geom_histogram()
ggplot(california, aes(x=TotalRooms)) +
geom_histogram()
ggplot(california, aes(x=TotalBedrooms)) +
geom_histogram()
ggplot(california, aes(x=Population)) +
geom_histogram()
ggplot(california, aes(x=Households)) +
geom_histogram()
ggplot(california, aes(x=MedianIncome)) +
geom_histogram()

ggplot(california, aes(x=MedianHouseValue)) +
geom_histogram()


#Test Shapiro-Wilk para comprobar normalidad en los datos
```

```
shapiro.test(sample(california[ , 'Longitude'], 5000))
shapiro.test(sample(california[ , 'Latitude'], 5000))
shapiro.test(sample(california[ , 'HousingMedianAge'], 5000))
shapiro.test(sample(california[ , 'TotalRooms'], 5000))
shapiro.test(sample(california[ , 'TotalBedrooms'], 5000))
shapiro.test(sample(california[ , 'Population'], 5000))
shapiro.test(sample(california[ , 'Households'], 5000))
shapiro.test(sample(california[ , 'MedianIncome'], 5000))


#Elimino los casos de MedianHouseValue == 500000
ggplot(california , aes(x = MedianIncome , y= MedianHouseValue)) +
geom_point()

california_clean <- drop(california[california$MedianHouseValue <
    500000, ])
#Para resetear los indices
rownames(california_clean) <- NULL

ggplot(california_clean , aes(x = MedianIncome , y=
    MedianHouseValue)) +
geom_point()


# Estudio de los Outliers
ggplot(california_clean , aes(x = TotalRooms , y =
    MedianHouseValue, color = TotalRooms > (quantile(TotalRooms,
    0.75) + IQR(TotalRooms)*1.5) )) +
geom_point() +
labs(title="Outliers TotalRooms", color="Outliers")

ggplot(california_clean , aes(x = TotalBedrooms , y =
    MedianHouseValue, color = TotalBedrooms > (quantile(
    TotalBedrooms, 0.75) + IQR(TotalBedrooms)*1.5) )) +
geom_point() +
labs(title="Outliers TotalBedrooms", color="Outliers")

ggplot(california_clean , aes(x = Population , y =
    MedianHouseValue, color = Population > (quantile(Population,
    0.75) + IQR(Population)*1.5) )) +
geom_point() +
labs(title="Outliers Population", color="Outliers")

ggplot(california_clean , aes(x = Households , y =
    MedianHouseValue, color = Households > (quantile(Households,
    0.75) + IQR(Households)*1.5) )) +
geom_point() +
labs(title="Outliers Households", color="Outliers")

ggplot(california_clean , aes(x = MedianIncome , y =
    MedianHouseValue, color = MedianIncome > (quantile(MedianIncome
    , 0.75) + IQR(MedianIncome)*1.5) )) +
geom_point() +
```

```
labs(title="Outliers MedianIncome", color="Outliers")


#Elimino los outliers
out_TotalRooms <- quantile(california_clean$TotalRooms, 0.75) +
    IQR(california_clean$TotalRooms)*1.5

out_TotalBedrooms <- quantile(california_clean$TotalBedrooms,
    0.75) + IQR(california_clean$TotalBedrooms)*1.5

out_Population <- quantile(california_clean$Population, 0.75) +
    IQR(california_clean$Population)*1.5

out_Households <- quantile(california_clean$Households, 0.75) +
    IQR(california_clean$Households)*1.5



california_new <- drop(california_clean[california_clean$
    TotalRooms < out_TotalRooms ,])
california_new <- drop(california_clean[california_clean$
    TotalBedrooms < out_TotalBedrooms ,])
california_new <- drop(california_clean[california_clean$
    Population < out_Population ,])
california_new <- drop(california_clean[california_clean$
    Households < out_Households ,])
#Para resetear los indices
rownames(california_new) <- NULL

#Se eliminan un 2% solamente, interesa
dim(california_new)


#Matriz de correlaciones
corr_matrix <- cor(california_new, method = "kendall")
corrplot(corr_matrix, method = "color", tl.col = "black")


#Primera hipotesis
ggplot(california_new, aes(x = Longitude, y = Latitude, color =
    MedianHouseValue, hue = MedianHouseValue))+
geom_point() +
labs(title="Valor medio de la vivienda dependiendo de la
    localizacion", color="MedianHouseValue") +
scale_color_gradient(low="blue", high="red")

mapa <- png::readPNG("mapa.png")

ggplot(california_new, aes(x = Longitude, y = Latitude, color =
    MedianHouseValue, hue = MedianHouseValue))+
background_image(mapa)+
geom_point() +
labs(title="Valor medio de la vivienda dependiendo de la
```

```
    localizacion", color="MedianHouseValue") +
scale_color_gradient(low="blue", high="red")


#Segunda hipotesis
ggplot(california_new, aes(x = Longitude, y = Latitude, color =
    MedianHouseValue, size = Population))+
background_image(mapa)+
geom_point() +
labs(title="Valor medio de la vivienda dependiendo de la
    localizacion", color="MedianHouseValue") +
scale_color_gradient(low="blue", high="red")

ggplot(california_new, aes(x = Population, y = MedianHouseValue))+
geom_point(color = "blue")


#Hipotesis 3
ggplot(california_new, aes(x = MedianIncome, y = MedianHouseValue)
    )+
geom_point(color = "blue")


#Hipotesis 4
ggplot(california_new, aes(x = HousingMedianAge, y =
    MedianHouseValue))+
geom_point(color = "blue")

ggplot(california_new, aes(x = Longitude, y = Latitude, color =
    HousingMedianAge))+
background_image(mapa)+
geom_point() +
labs(title="Edad media de la vivienda dependiendo de la
    localizacion", color="HousingMedianAge") +
scale_color_gradient(low="green", high="purple")

ggplot(california_new, aes(x = Longitude, y = Latitude, color =
    MedianHouseValue, hue = MedianHouseValue))+
background_image(mapa)+
geom_point() +
labs(title="Valor medio de la vivienda dependiendo de la
    localizacion", color="MedianHouseValue") +
scale_color_gradient(low="blue", high="red")
```

## 6.2. Apéndice B. Código EDA dataset de clasificación 'Bupa'

```
#Cargo librerias
library(moments)
library(ggplot2)
library(tidyverse)

#Cargo el dataset
bupa <- read.csv('bupa.dat', comment.char = '@',
header = FALSE, stringsAsFactors = TRUE)

dim(bupa)
str(bupa)

#Asigno nombre a las variables
names(bupa) <- c("mcv", "alkphos", "sgpt",
"sgot", "gammagt", "drinks", "selector")


#Missing values
any(is.na(bupa))

#muestras duplicadas
sum(duplicated(bupa))

bupa <- bupa[!duplicated(bupa), ]
dim(bupa)

#Para resetear los indices
rownames(bupa) <- NULL

#Medidadas importantes
summary(bupa)

desviacion = apply(bupa, 2, sd)
format(desviacion, scientific = FALSE)
skev <- apply(bupa, 2, skewness)
kurt <- apply(bupa, 2, kurtosis)
format(skev, scientific = FALSE)
format(kurt, scientific = FALSE)


#Representacion grafica box_plot
ggplot(bupa, aes(y=mcv)) +
geom_boxplot(outlier.color = "red")
ggplot(bupa, aes(y=alkphos)) +
geom_boxplot(outlier.color = "red")
ggplot(bupa, aes(y=sgpt)) +
geom_boxplot(outlier.color = "red")
ggplot(bupa, aes(y=sgot)) +
geom_boxplot(outlier.color = "red")
```

```
ggplot(bupa, aes(y=gammagt)) +
geom_boxplot(outlier.color = "red")


ggplot(bupa, aes(y=drinks)) +
geom_boxplot(outlier.color = "red")



#Representacion grafica histogramas
ggplot(bupa, aes(x=mcv)) +
geom_histogram()
ggplot(bupa, aes(x=alkphos)) +
geom_histogram()
ggplot(bupa, aes(x=sgpt)) +
geom_histogram()
ggplot(bupa, aes(x=sgot)) +
geom_histogram()
ggplot(bupa, aes(x=gammagt)) +
geom_histogram()


ggplot(bupa, aes(x=drinks)) +
geom_histogram(bins = 5)


#Shapiro-wilk test
shapiro.test(sample(bupa[ , 'mcv']))
shapiro.test(sample(bupa[ , 'alkphos']))
shapiro.test(sample(bupa[ , 'sgpt']))
shapiro.test(sample(bupa[ , 'sgot']))
shapiro.test(sample(bupa[ , 'gammagt']))



#Analisis Outliers izquierda
ggplot(bupa, aes(x = mcv   , y = drinks, color = mcv < (quantile(
    mcv, 0.25) - IQR(mcv)*1.5) )) +
geom_point()  +
labs(title="Outliers mcv", color="Outliers")


ggplot(bupa, aes(x = sgot   , y = drinks, color = sgot < (quantile(
    sgot, 0.25) - IQR(sgot)*1.5) )) +
geom_point()  +
labs(title="Outliers sgot", color="Outliers")


#Analisis Outliers izquierda
ggplot(bupa, aes(x = mcv   , y = drinks, color = mcv > (quantile(
    mcv, 0.75) + IQR(mcv)*1.5) )) +
geom_point()  +
labs(title="Outliers mcv", color="Outliers")
```

```
ggplot(bupa, aes(x = alkphos   , y = drinks, color = alkphos > (
    quantile(alkphos, 0.75) + IQR(alkphos)*1.5) )) +
geom_point()  +
labs(title="Outliers alkphos", color="Outliers")

ggplot(bupa, aes(x = sgpt   , y = drinks, color = sgpt > (quantile(
    sgpt, 0.75) + IQR(sgpt)*1.5) )) +
geom_point()  +
labs(title="Outliers sgpt", color="Outliers")

ggplot(bupa, aes(x = sgot   , y = drinks, color = sgot > (quantile(
    sgot, 0.75) + IQR(sgot)*1.5) )) +
geom_point()  +
labs(title="Outliers sgot", color="Outliers")

ggplot(bupa, aes(x = gammagt   , y = drinks, color = gammagt > (
    quantile(gammagt, 0.75) + IQR(gammagt)*1.5) )) +
geom_point()  +
labs(title="Outliers gammagt", color="Outliers")


#Preparo la variable dependiente
unique(bupa$drinks)

bupa$drinks[bupa$drinks <= 5] <- 0
bupa$drinks[bupa$drinks > 5] <- 1

summary(bupa)



#Separo segun la variable selector
bupa_train <- bupa %>% filter(selector == 2)
bupa_test <- bupa %>% filter(selector == 1)

dim(bupa_train)
dim(bupa_test)
bupa_train %>% count(drinks)
bupa_test %>% count(drinks)


#Matriz de correlaciones
no_label_bupa <- bupa %>% select(c(-drinks, -selector))
corr_matrix <- cor(no_label_bupa, method = "kendall")
corrplot(corr_matrix, method = "num", tl.col = "black")



#Hipotesis 1
ggplot(bupa, aes(x = alkphos, y = mcv, color = drinks))+
geom_point()
```

```
ggplot(bupa, aes(x = sgpt, y = mcv, color = drinks))+
geom_point()

ggplot(bupa, aes(x = sgot, y = mcv, color = drinks))+
geom_point()

ggplot(bupa, aes(x = gammagt, y = mcv, color = drinks))+
geom_point()



#hipotesis 2
ggplot(bupa, aes(x = alkphos, y = sgpt))+
geom_point()

ggplot(bupa, aes(x = alkphos, y = sgot))+
geom_point()

ggplot(bupa, aes(x = alkphos, y = gammagt))+
geom_point()


ggplot(bupa, aes(x = sgpt, y = sgot))+
geom_point()

ggplot(bupa, aes(x = sgpt, y = gammagt))+
geom_point()


ggplot(bupa, aes(x = sgot, y = gammagt))+
geom_point()
```

## 6.3. Apéndice C. Código Regresión sobre California

```
#Cargo librerias
library(kknn)
library(dplyr)
library(ggplot2)
library(corrplot)

#Cargo el dataset
df_california <- read.csv("california.dat", comment.char="@",
    header=F)

#Funcion para poner nombre a las columnas
add_colnames <- function(df) {
  colnames(df) <- c("Longitude", "Latitude", "HousingMedianAge",
  "TotalRooms", "TotalBedrooms", "Population", "Households",
  "MedianIncome", "MedianHouseValue")
  df
}

df_california <- add_colnames(df_california)


# Modelos lineales simples
lm_Longitude <- lm(MedianHouseValue ~ Longitude, data = df_
    california)
lm_Latitude <- lm(MedianHouseValue ~ Latitude, data = df_
    california)
lm_HousingMedianAge <- lm(MedianHouseValue ~ HousingMedianAge,
    data = df_california)
lm_TotalRooms <- lm(MedianHouseValue ~ TotalRooms, data = df_
    california)
lm_TotalBedrooms <- lm(MedianHouseValue ~ TotalBedrooms, data = df
    _california)
lm_Population <- lm(MedianHouseValue ~ Population, data = df_
    california)
lm_Households <- lm(MedianHouseValue ~ Households, data = df_
    california)
lm_MedianIncome <- lm(MedianHouseValue ~ MedianIncome, data = df_
    california)




summary(lm_Longitude)
summary(lm_Latitude)
summary(lm_HousingMedianAge)
summary(lm_TotalRooms)
summary(lm_TotalBedrooms)
summary(lm_Population)
```

```
summary(lm_Households)
summary(lm_MedianIncome)


#Represento graficamente los modelos
plot(MedianHouseValue ~ Longitude, data = df_california)
abline(lm_Longitude,col="red")

plot(MedianHouseValue ~ Latitude, data = df_california)
abline(lm_Latitude,col="red")

plot(MedianHouseValue ~ HousingMedianAge, data = df_california)
abline(lm_HousingMedianAge,col="red")

plot(MedianHouseValue ~ TotalRooms, data = df_california)
abline(lm_TotalRooms,col="red")

plot(MedianHouseValue ~ TotalBedrooms, data = df_california)
abline(lm_TotalBedrooms,col="red")

plot(MedianHouseValue ~ Population, data = df_california)
abline(lm_Population,col="red")

plot(MedianHouseValue ~ Households, data = df_california)
abline(lm_Households,col="red")

plot(MedianHouseValue ~ MedianIncome, data = df_california)
abline(lm_MedianIncome,col="red")



# Funcion para aplicar lm a 5-fold
run_lm_fold <- function(i, x, model, tt = "test") {
  # Cargar conjuntos de train
  file <- paste(x, "-5-", i, "tra.dat", sep="")
  x_tra <- read.csv(file, comment.char="@")

  # Cargar conjuntos de test
  file <- paste(x, "-5-", i, "tst.dat", sep="")
  x_tst <- read.csv(file, comment.char="@")



  x_tra <- add_colnames(x_tra)
  x_tst <- add_colnames(x_tst)


  if (tt == "train") {
    test <- x_tra
  } else {
    test <- x_tst
  }
```

```
  # Entrenar el modelo sobre el conjunto de train
  formula <- terms(model)
  model_eval <- lm(formula=formula, data=x_tra)

  # RMSE sobre test
  yprime <- predict(model_eval, test)
  #MSE
  sum(abs(test$MedianHouseValue - yprime)^2)/length(yprime)
}



#Procedo a evaluar todos los modelos lineales
cat('RMSE de 5-fold sobre MedianIncome:', mean(sapply(1:5, run_lm_
    fold, 'california', lm_MedianIncome), fill=T))

cat('RMSE de 5-fold sobre Latitude:', mean(sapply(1:5, run_lm_fold
    , 'california', lm_Latitude), fill=T))

cat('RMSE de 5-fold sobre TotalRooms:', mean(sapply(1:5, run_lm_
    fold, 'california', lm_TotalRooms), fill=T))

cat('RMSE de 5-fold sobre HousingMedianAge:', mean(sapply(1:5, run
    _lm_fold, 'california', lm_HousingMedianAge), fill=T))

cat('RMSE de 5-fold sobre Households:', mean(sapply(1:5, run_lm_
    fold, 'california', lm_Households), fill=T))



# Modelo lineal multiple con todas las variables
fit_mult1 <- lm(MedianHouseValue ~ ., data=df_california)
summary(fit_mult1)

cat('Media RMSE del modelo lineal multiple sobre 5-fold:', mean(
    sapply(1:5, run_lm_fold, 'california', fit_mult1), fill=T))


#Modelo lineal multiple sin Longitude
fit_mult2 <- lm(MedianHouseValue ~ . -Longitude , data=df_
    california)
summary(fit_mult2)

cat('Media RMSE del modelo lineal multiple sobre 5-fold:', mean(
    sapply(1:5, run_lm_fold, 'california', fit_mult2), fill=T))


#Modelo lineal multiple sin Population
fit_mult3 <- lm(MedianHouseValue ~ . -Population , data=df_
    california)
summary(fit_mult3)

cat('Media RMSE del modelo lineal multiple sobre 5-fold:', mean(
```

```
    sapply(1:5, run_lm_fold, 'california', fit_mult3), fill=T))


#Modelo lineal multiple sin TotalBedrooms
fit_mult4 <- lm(MedianHouseValue ~ . -TotalBedrooms , data=df_
    california)
summary(fit_mult4)

cat('Media RMSE del modelo lineal multiple sobre 5-fold:', mean(
    sapply(1:5, run_lm_fold, 'california', fit_mult4), fill=T))


#Modelo lineal multiple MedianIncome^2
fit_mult5 <- lm(MedianHouseValue ~ . +I(MedianIncome^2) , data=df_
    california)
summary(fit_mult5)

cat('Media RMSE del modelo lineal multiple sobre 5-fold:', mean(
    sapply(1:5, run_lm_fold, 'california', fit_mult5), fill=T))

#Modelo lineal multiple MedianIncome^4
fit_mult6 <- lm(MedianHouseValue ~ . +I(MedianIncome^4) , data=df_
    california)
summary(fit_mult6)

cat('Media RMSE del modelo lineal multiple sobre 5-fold:', mean(
    sapply(1:5, run_lm_fold, 'california', fit_mult6), fill=T))

#Modelo lineal multiple I(MedianIncome^4)*Latitude*TotalRooms
fit_mult7 <- lm(MedianHouseValue ~ . +I(MedianIncome^4)*Latitude*
    TotalRooms , data=df_california)
summary(fit_mult7)

cat('Media RMSE del modelo lineal multiple sobre 5-fold:', mean(
    sapply(1:5, run_lm_fold, 'california', fit_mult7), fill=T))


#Modelo lineal multiple I(MedianIncome^4)*TotalRooms
fit_mult8 <- lm(MedianHouseValue ~ . +I(MedianIncome^4)*TotalRooms
     , data=df_california)
summary(fit_mult8)

cat('Media RMSE del modelo lineal multiple sobre 5-fold:', mean(
    sapply(1:5, run_lm_fold, 'california', fit_mult8), fill=T))


#Modelo lineal multiple I(MedianIncome^4)*TotalRooms*
    HousingMedianAge*Households
fit_mult9 <- lm(MedianHouseValue ~ . +I(MedianIncome^4)*TotalRooms
    *HousingMedianAge*Households , data=df_california)
summary(fit_mult9)

cat('Media RMSE del modelo lineal multiple sobre 5-fold:', mean(
```

```
    sapply(1:5, run_lm_fold, 'california', fit_mult9), fill=T))


#Modelo lineal multiple MedianIncome*TotalRooms*HousingMedianAge*
    Households
fit_mult10 <- lm(MedianHouseValue ~ . +MedianIncome*TotalRooms*
    HousingMedianAge*Households , data=df_california)
summary(fit_mult10)

cat('Media RMSE del modelo lineal multiple sobre 5-fold:', mean(
    sapply(1:5, run_lm_fold, 'california', fit_mult10), fill=T))



#Funcion para aplicar knn sobre 5-folds
run_knn_fold <- function(i, x, formula, k,  tt = "test") {
  # Cargar conjuntos de train
  file <- paste(x, "-5-", i, "tra.dat", sep="")
  x_tra <- read.csv(file, comment.char="@")

  # Cargar conjuntos de test
  file <- paste(x, "-5-", i, "tst.dat", sep="")
  x_tst <- read.csv(file, comment.char="@")



  x_tra <- add_colnames(x_tra)
  x_tst <- add_colnames(x_tst)


  if (tt == "train") {
    test <- x_tra
  } else {
    test <- x_tst
  }

  # Entrenar el modelo sobre el conjunto de train
  model_eval <- kknn(formula=formula, x_tra, test, k=k)

  # RMSE sobre test
  yprime <- model_eval$fitted.values
  #MSE
  sum(abs(test$MedianHouseValue - yprime)^2)/length(yprime)
}


#Pruebo con diferentes valores de k
cat('RMSE del modelo k-NN sobre 5-fold:', sapply(1:5, run_knn_fold
    , 'california', MedianHouseValue ~ . +MedianIncome*TotalRooms*
    HousingMedianAge*Households, k = 5), fill=T)
cat('Meida RMSE del modelo k-NN sobre 5-fold:', mean(sapply(1:5,
    run_knn_fold, 'california', MedianHouseValue ~ . +MedianIncome*
    TotalRooms*HousingMedianAge*Households, k = 5), fill=T))
```

```
cat('RMSE del modelo k-NN sobre 5-fold:', sapply(1:5, run_knn_fold
    , 'california', MedianHouseValue ~ . +MedianIncome*TotalRooms*
    HousingMedianAge*Households, k = 7), fill=T)
cat('Meida RMSE del modelo k-NN sobre 5-fold:', mean(sapply(1:5,
    run_knn_fold, 'california', MedianHouseValue ~ . +MedianIncome*
    TotalRooms*HousingMedianAge*Households, k = 7), fill=T))


cat('RMSE del modelo k-NN sobre 5-fold:', sapply(1:5, run_knn_fold
    , 'california', MedianHouseValue ~ . +MedianIncome*TotalRooms*
    HousingMedianAge*Households, k = 10), fill=T)
cat('Meida RMSE del modelo k-NN sobre 5-fold:', mean(sapply(1:5,
    run_knn_fold, 'california', MedianHouseValue ~ . +MedianIncome*
    TotalRooms*HousingMedianAge*Households, k = 10), fill=T))


cat('RMSE del modelo k-NN sobre 5-fold:', sapply(1:5, run_knn_fold
    , 'california', MedianHouseValue ~ . +MedianIncome*TotalRooms*
    HousingMedianAge*Households, k = 20), fill=T)
cat('Meida RMSE del modelo k-NN sobre 5-fold:', mean(sapply(1:5,
    run_knn_fold, 'california', MedianHouseValue ~ . +MedianIncome*
    TotalRooms*HousingMedianAge*Households, k = 20), fill=T))


cat('RMSE del modelo k-NN sobre 5-fold:', sapply(1:5, run_knn_fold
    , 'california', MedianHouseValue ~ . +MedianIncome*TotalRooms*
    HousingMedianAge*Households, k = 25), fill=T)
cat('Meida RMSE del modelo k-NN sobre 5-fold:', mean(sapply(1:5,
    run_knn_fold, 'california', MedianHouseValue ~ . +MedianIncome*
    TotalRooms*HousingMedianAge*Households, k = 25), fill=T))


cat('RMSE del modelo k-NN sobre 5-fold:', sapply(1:5, run_knn_fold
    , 'california', MedianHouseValue ~ . +MedianIncome*TotalRooms*
    HousingMedianAge*Households, k = 50), fill=T)
cat('Meida RMSE del modelo k-NN sobre 5-fold:', mean(sapply(1:5,
    run_knn_fold, 'california', MedianHouseValue ~ . +MedianIncome*
    TotalRooms*HousingMedianAge*Households, k = 50), fill=T))


#knn sobre modelo con todas las variables
cat('RMSE del modelo k-NN sobre 5-fold:', sapply(1:5, run_knn_fold
    , 'california', MedianHouseValue ~ ., k = 20), fill=T)
cat('Meida RMSE del modelo k-NN sobre 5-fold:', mean(sapply(1:5,
    run_knn_fold, 'california', MedianHouseValue ~ ., k = 20), fill
    =T))



# Comparativa LM, k-NN y M5'
```

```r
#leemos la tabla con los errores medios de test
resultados <- read.csv("regr_test_alumnos.csv")
tablatst <- cbind(resultados[,2:dim(resultados)[2]])
colnames(tablatst) <- names(resultados)[2:dim(resultados)[2]]
rownames(tablatst) <- resultados[,1]


#Calculo mis valores de LM y KNN para aniadirlos en la tabla
media_lm <-mean(sapply(1:5, run_lm_fold, 'california', fit_mult1),
    fill=T)
media_knn <- mean(sapply(1:5, run_knn_fold, 'california',
   MedianHouseValue ~., k = 20), fill=T)

tablatst["california", 1] <- media_lm
tablatst["california", 2] <- media_knn

#Comparar lm con knn con Wilcoxon
# + 0.1 porque wilcox R falla para valores == 0 en la tabla
difs <- (tablatst[,1] - tablatst[,2]) / tablatst[,1]
wilc_1_2 <- cbind(ifelse (difs<0, abs(difs)+0.1, 0+0.1),
ifelse (difs>0, abs(difs)+0.1, 0+0.1))

colnames(wilc_1_2) <- c(colnames(tablatst)[1], colnames(tablatst)
   [2])
head(wilc_1_2)

LMvsKNNtst <- wilcox.test(wilc_1_2[,1], wilc_1_2[,2], alternative
   = "two.sided", paired=TRUE)
Rmas <- LMvsKNNtst$statistic
pvalue <- LMvsKNNtst$p.value

LMvsKNNtst <- wilcox.test(wilc_1_2[,2], wilc_1_2[,1], alternative
   = "two.sided", paired=TRUE)
Rmenos <- LMvsKNNtst$statistic


cat('Test modelo lineal (R+) vs modelo k-NN:(R-)', fill=T)
cat('Valor R+: ',Rmas, fill=T)
cat('Valor R-: ',Rmenos, fill=T)
cat('p-value del test: ',pvalue, fill=T)


test_friedman <- friedman.test(as.matrix(tablatst))
test_friedman

tam <- dim(tablatst)
groups <- rep(1:tam[2], each=tam[1])
pairwise.wilcox.test(as.matrix(tablatst), groups, p.adjust = "holm
   ", paired = TRUE)
```

## 6.4. Apéndice D. Código Clasificación sobre Bupa

```
#Cargo librerias
library(tidyverse)
library(class)
library(caret)
library(car)
library("MASS")
library(corrplot)


#Cargo el dataset
bupa <- read.csv('bupa.dat', comment.char = '@',
header = FALSE, stringsAsFactors = TRUE)

dim(bupa)
str(bupa)

#asigno nombre a las variables
names(bupa) <- c("mcv", "alkphos", "sgpt",
"sgot", "gammagt", "drinks", "selector")


#Elimino casos duplicados
bupa <- bupa[!duplicated(bupa), ]

#Para resetear los indices
rownames(bupa) <- NULL

#Defino unicamente dos clases en la variable dependiente drinks
unique(bupa$drinks)
bupa$drinks[bupa$drinks <= 5] <- 0
bupa$drinks[bupa$drinks > 5] <- 1



#Separo segun la variable selector
bupa_train <- bupa %>% filter(selector == 2)
bupa_test <- bupa %>% filter(selector == 1)

bupa_train <- bupa_train %>% select(-selector)
bupa_test <- bupa_test %>% select(-selector)


summary(bupa_train)


#Normalizo los datos, la variable dependiente no
bupa_train_scale <- as.data.frame(scale(bupa_train))
bupa_train_scale[,'drinks'] <- bupa_train$drinks
bupa_test_scale <- as.data.frame(scale(bupa_test))
```

```
bupa_test_scale[,'drinks'] <- bupa_test$drinks


# Modelo con k=7
knn_pred_1 <- knn(train=bupa_train_scale[,-ncol(bupa_train_scale)
    ],                                                    test=bupa_
    test_scale[,-ncol(bupa_test_scale)],
cl=bupa_train_scale[,ncol(bupa_train_scale)], k=7)

# Evaluo los resultados del modelo creado
result <- table(knn_pred_1, bupa_test_scale[, 'drinks'])

# Precision obtenida
sum(diag(result)) / length(knn_pred_1)



# Modelos basados en k-NN
#Cargo todos los modelos pues debo aplicar los cambios
# Cargar conjuntos de entrenamiento
bupa_train1 <- read.csv("bupa-10-1tra.dat", comment.char="@")
bupa_train2 <- read.csv("bupa-10-2tra.dat", comment.char="@")
bupa_train3 <- read.csv("bupa-10-3tra.dat", comment.char="@")
bupa_train4 <- read.csv("bupa-10-4tra.dat", comment.char="@")
bupa_train5 <- read.csv("bupa-10-5tra.dat", comment.char="@")
bupa_train6 <- read.csv("bupa-10-6tra.dat", comment.char="@")
bupa_train7 <- read.csv("bupa-10-7tra.dat", comment.char="@")
bupa_train8 <- read.csv("bupa-10-8tra.dat", comment.char="@")
bupa_train9 <- read.csv("bupa-10-9tra.dat", comment.char="@")
bupa_train10 <- read.csv("bupa-10-10tra.dat", comment.char="@")

# Cargar conjuntos de test
bupa_test1 <- read.csv("bupa-10-1tst.dat", comment.char="@")
bupa_test2 <- read.csv("bupa-10-2tst.dat", comment.char="@")
bupa_test3 <- read.csv("bupa-10-3tst.dat", comment.char="@")
bupa_test4 <- read.csv("bupa-10-4tst.dat", comment.char="@")
bupa_test5 <- read.csv("bupa-10-5tst.dat", comment.char="@")
bupa_test6 <- read.csv("bupa-10-6tst.dat", comment.char="@")
bupa_test7 <- read.csv("bupa-10-7tst.dat", comment.char="@")
bupa_test8 <- read.csv("bupa-10-8tst.dat", comment.char="@")
bupa_test9 <- read.csv("bupa-10-9tst.dat", comment.char="@")
bupa_test10 <- read.csv("bupa-10-10tst.dat", comment.char="@")


#Funcion para hacer las transformaciones del EDA en todos los
    folds
df_prepare <- function(df){
  names(df) <- c("mcv", "alkphos", "sgpt",
  "sgot", "gammagt", "drinks", "selector")

  #eliminar repetidos
  df <- df[!duplicated(df), ]
  #Para resetear los indices
```

```
  rownames(df) <- NULL

  #Eliminados la columna selector, no es util
  df <- df %>% select(-selector)

  df$drinks[df$drinks <= 5] <- 0
  df$drinks[df$drinks > 5] <- 1

  df_scale <- as.data.frame(scale(df))
  df_scale['drinks'] <- df$drinks

  df_scale
}



# Generamos listas que contienen los folds de Validacion Cruzada.
train_list <- list(bupa_train1, bupa_train2, bupa_train3, bupa_
    train4, bupa_train5, bupa_train6, bupa_train7, bupa_train8,
    bupa_train9, bupa_train10)

test_list <- list(bupa_test1, bupa_test2, bupa_test3, bupa_test4,
    bupa_test5, bupa_test6, bupa_test7, bupa_test8, bupa_test9,
    bupa_test10)


#Aplicamos los cambios en todos
train_list <- lapply(train_list, df_prepare)
test_list <- lapply(test_list, df_prepare)




# Evaluar modelos con validacion cruzada 10-fold
knn_fold_bupa <- function(train_list, test_list, k) {
  sapply(1:length(train_list), function(i) {
    # Aplicamos modelo
    pred <- knn(train = train_list[[i]] %>% select(-drinks),
    test = test_list[[i]] %>% select(-drinks),
    cl = train_list[[i]]$drinks, k=k)
    # Calculamos el accuracy
    sum(pred == test_list[[i]]$drinks) / length(pred)
  })
}

# Evaluar knn train con diferentes k
knn_1 <- knn_fold_bupa(train_list, test_list, 1)
knn_2 <- knn_fold_bupa(train_list, test_list, 3)
knn_3 <- knn_fold_bupa(train_list, test_list, 5)
knn_4 <- knn_fold_bupa(train_list, test_list, 7)
knn_5 <- knn_fold_bupa(train_list, test_list, 10)
knn_6 <- knn_fold_bupa(train_list, test_list, 13)
knn_7 <- knn_fold_bupa(train_list, test_list, 15)
```

```
knn_8 <- knn_fold_bupa(train_list, test_list, 17)
knn_9 <- knn_fold_bupa(train_list, test_list, 20)
knn_10 <- knn_fold_bupa(train_list, test_list, 30)

#Agrupo los resultados en una tabla
l <- rbind(knn_1, knn_2, knn_3, knn_4, knn_5, knn_6, knn_7, knn_8,
    knn_9, knn_10)
l_mean <- cbind(apply(l, 1, mean))
l_all <- cbind(l, apply(l, 1, mean))


# Knn en train del mejor modelo
knn_7 <- knn_fold_bupa(train_list, test_list, 15)
mean(knn_7)



# Comprobamos variabilidad para LDA
var_test <- apply(bupa, 2, var)
var_test


# Funcion para LDA en validacion cruzada
lda_fold_bupa <- function(formula, train_list, test_list) {
  sapply(1:length(train_list), function(i) {
    # modelo lda
    fit <- lda(formula, data = train_list[[i]])
    # Calculamos predicciones
    pred <- predict(fit, test_list[[i]]%>%select(-drinks))

    # Calculamos la precision
    sum(pred$class == test_list[[i]]$drinks) / length(pred$class)
  })
}


#Aplico lda
lda_fold <- lda_fold_bupa(drinks ~ ., train_list, test_list)

mean(lda_fold)



# Comprobamos variabilidad para QDA
#Cambio para poder usar levene test
bupa$drinks[bupa$drinks == 0] <- '0'
bupa$drinks[bupa$drinks == 1] <- '1'

leveneTest(mcv ~ drinks, bupa)
leveneTest(alkphos ~ drinks, bupa)
leveneTest(sgpt ~ drinks, bupa)
leveneTest(sgot ~ drinks, bupa)
leveneTest(gammagt ~ drinks, bupa)
```

```r
bupa$drinks[bupa$drinks == 0] <- 0
bupa$drinks[bupa$drinks == 1] <- 1

#Correlacion por clases
bupa_0 <- bupa %>% filter(drinks == 0) %>% select(-c(selector,
    drinks))

corr_matrix <- cor(bupa_0, method = "kendall")
corrplot(corr_matrix, method = "num", tl.col = "black")

bupa_1 <- bupa %>% filter(drinks == 1) %>% select(-c(selector,
    drinks))

corr_matrix <- cor(bupa_1, method = "kendall")
corrplot(corr_matrix, method = "num", tl.col = "black")



# Funcion para QDA en validacion cruzada
qda_fold_bupa <- function(formula, train_list, test_list) {
  sapply(1:length(train_list), function(i) {
    # modelo lda
    fit <- qda(formula, data = train_list[[i]])
    # Calculamos predicciones
    pred <- predict(fit, test_list[[i]]%>%select(-drinks))

    # Calculamos la precision
    sum(pred$class == test_list[[i]]$drinks) / length(pred$class)
  })
}

qda_fold <- qda_fold_bupa(drinks ~ ., train_list, test_list)
mean(qda_fold)


#Agrupamos los resultados de los 3 modelos
results <- cbind(mean(knn_7), mean(lda_fold), mean(qda_fold))
results

#leemos la tabla con los errores medios de test
resultados <- read.csv("clasif_test_alumos.csv")
tablatst <- cbind(resultados[,2:dim(resultados)[2]])
colnames(tablatst) <- names(resultados)[2:dim(resultados)[2]]
rownames(tablatst) <- resultados[,1]

#Pongo mis resultados en mi dataset bupa
tablatst["bupa", ] <- results

# test friedman
test_friedman <- friedman.test(as.matrix(tablatst))
```

test_friedman

# Bibliografía


[1] J. Han, M. Kamber, and J. Pei, *Data Mining : Concepts and Techniques*.

[2] "Nomograms for Visualizing Linear Support Vector Machines." https://www.researchgate.net/publication/33550053_Nomograms_for_Visualizing_Linear_Support_Vector_Machines

[3] "A Study of Support Vector Machine Algorithm for Liver Disease Diagnosis." http://article.sapub.org/10.5923.j.ajis.20140401.02.html

[4] "The California housing dataset." https://inria.github.io/scikit-learn-mooc/python_scripts/datasets_california_housing.html

[5] "Descripción del conjunto de datos de viviendas de California." https://developers.google.com/machine-learning/crash-course/california-housing-data-description

[6] "Liver Disorders Data Set." https://archive.ics.uci.edu/ml/datasets/liver+disorders