Universidad de Granada

TRABAJO FIN DE GRADO

INGENIERÍA EN INFORMÁTICA

Desarrollo de un sistema de alerta temprana financiera basada en aprendizaje automático avanzado

Autor
Rafael Vázquez Conejo

Director
Salvador García López

Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación

Granada, julio 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.

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 KNN (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).

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.

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 antiguedad 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.

urlhttps://github.com/sonarsushant/California-House-Price-Prediction/blob/master/EDA
Con longitude un valor más negativo es más al oeste Con latitude un valor más alto está más al norte

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 missing values dentro de los datos, concluyendo en que este dataset no posee ningún Missing value. También se determina que no hay ninguna muestra duplicada.

	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

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**

  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 distribucion 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:

  ??? Imagen boxplot

  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

  ??? Imagen histo

- **Latitude**:

  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.

	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

	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

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

??? Imagen boxplot ??? Imagen histo

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

- **HousingMedianAge**:

  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

  ??? Imagen boxplot ??? Imagen histo

  Efectivamente podemos observar una distribución medianamente uniforme.

- **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

	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

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

??? Imagen boxplot ??? Imagen histo

- **TotalBedrooms**:

  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

  ??? Imagen boxplot ??? Imagen histo

- **Population**:

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

	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

	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

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

??? Imagen boxplot ??? Imagen histo

- **Households**:

  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

  ??? Imagen boxplot ??? Imagen histo

- **MedianIncome**:

  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 existemica de outliers.

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

  ??? Imagen boxplot ??? Imagen histo

	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

	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 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.

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

??? Imagen boxplot ??? Imagen histo

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: ??? Imagen truncado

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 %).

??? Imagen post- truncado

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

2.3.2. Análisis de las relaciones entre variables

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

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 tratrse 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.

???Imagen diagram correla

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 dependiento si la media de las casas de esa zona es de un alto o bajo precio.

??? Scatterplot

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.

?? Con el mapa

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.

?? Scatter ?? los dos

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.

?? scatter

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.

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.

??doble scatter

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

Bibliografía