udp FACULTAD DE
INGENIERÍA Y CIENCIAS

CONTROL 5 – DATA MINING

AGRUPACIÓN DE CARACTERÍSTICAS QUE INFLUYEN EN LOS PRECIOS DE CASAS EN AMEN, IOWA, EE. UU.

Integrante: María Paz Santibáñez

Profesor: Pablo Lemus

Fecha: 10 de noviembre, 2023

INTRODUCCIÓN

La base de datos Kaggle "House Price" es un conjunto de datos que proporciona 1460 observaciones por cada una de las 81 variables que describen casi todos los aspectos de las viviendas residenciales en Ames, Iowa, EE. UU. Estas variables abarcan una amplia gama de características, incluidas las características físicas de la propiedad, la condición y calidad de varios aspectos de la casa, el tipo y calidad de servicios públicos y las características del vecindario y la zonificación, y el precio de la vivienda.

OBJETIVO

El objetivo de este informe es aplicar algoritmos de agrupación usando técnicas de data mining para identificar patrones y relaciones dentro de los datos, lo que proporcionará una compresión de los factores que influyen en los diferentes rangos de precios de la vivienda en Ames, Iowa, EE. UU.

METODOLOGIA

En primer, se observó los tipos de variables y se rectificó el formato que corresponde según la descripción de cada una, luego se realizó un análisis exploratorio de las variables numéricas y se identificó el porcentaje de datos faltantes para cada variable. Para realizar la imputación de datos faltantes las variables numéricas se utilizó la librería *mice* y se aplicó árboles de decisión con el método *CART* (Classification and Regression Tree) y la eliminación de las observaciones.

A continuación, mediante el uso de la función *prcomp* se realizó el análisis de componentes principales para reducir la dimensión del conjunto de datos *df_features* que considera solo las variables numéricas. Luego, se obtiene el grafico de la reducción de la varianza y de las dos componentes principales.

Por ultimo, se realiza la aplicación de los algoritmos de agrupamiento *k-means* junto a métodos para determinar el número óptimo de clusters, el método del codo indica gráficamente que el numero optimo es el numero donde el total dentro de la suma del cuadrado interrumpe su disminución. Por otro lado, el coeficiente de silueta muestra gráficamente que el numero optimo es el punto que maximiza la anchura media de la silueta.

Finalmente, se aplica el algoritmo *hierarchical clustering* de la librería *clúster.*

RESULTADOS

Para comenzar, se identifica que 32 variables son del tipo numérica y 48 son categóricas o factor, donde la mayoría contiene mas de 3 niveles. Se decide particionar el conjunto de datos en datos tipo numéricos y tipo factor ya que facilita la aplicación de los algoritmos mencionados en la metodología.

Las variables numéricas LotFrontage y MasVnrArea contienen un 17,74% y 0.55% de datos faltantes respectivamente.

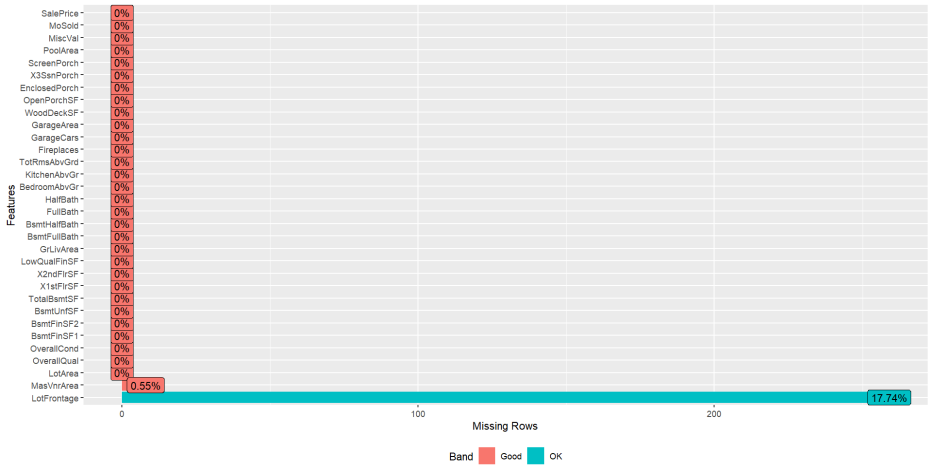

Imagen 1: Porcentaje de valores perdidos subconjunto variables numéricas

Se observa en la imagen 1 que el porcentaje de datos faltantes de la variable MasVnrArea es menor al 5%, por lo tanto, se eliminan 8 observaciones. Para el caso de la variable LotFrontage se imputan las 254 observaciones mediante el método *CART,* utilizando solo las variables numéricas para realizar la “regresión”.

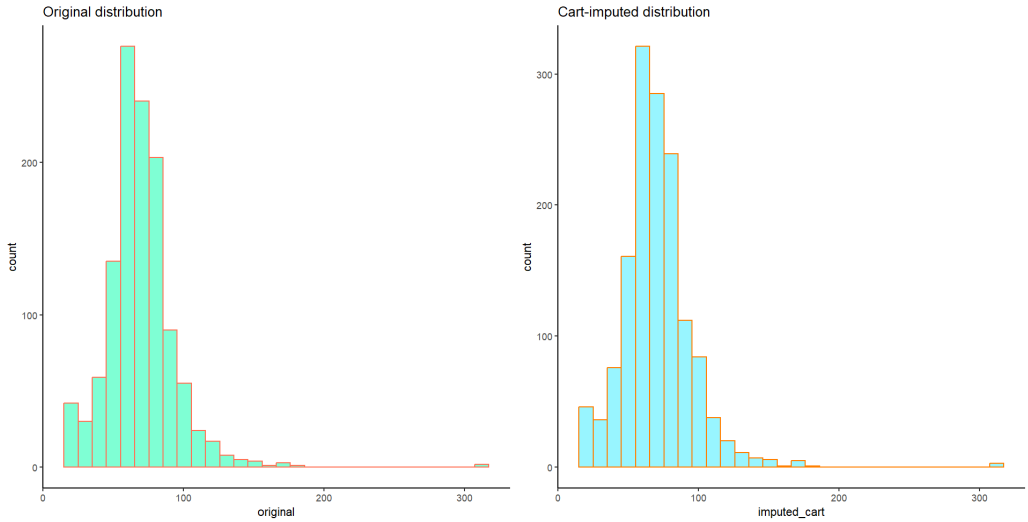

Imagen 2: Distribución original e imputación CART de la variable LotFrontage

A continuación, se realiza el análisis de las componentes principales al conjunto de datos de tipo numérico, ya que la función *prcomp* solo admite entradas numéricas. Las variables categóricas pueden transformarse a formato numérico a través de la codificación one_hot pero aumenta considerablemente la dimensión del conjunto de datos si se tiene variables categóricas con mas de tres niveles, por lo que no serán consideradas para este análisis.

El argumento *scale* de la función *prcomp* debe establecerse como verdadero para que se todas las variables tengan la misma varianza. Los resultados demuestran que las primeras 24 componentes explican aproximadamente el 95% de la varianza de los datos, y así mismo, las variables que más influyen en cada una de las 24 componente son el precio de la casa, los pies cuadrados del segundo piso, los pies cuadrado sin terminar del sótano, en la componente 4 y 23 pies lineales de calle conectados a la propiedad, el numero de cocinas, la cantidad de baños en el sótano, área del porche cubierto en pies cuadrado, en la componente 8 y 20 el área de cubierta de madera en pies cuadrados, en la componente 9 y 13 los pies cuadrados terminados tipo 2, en la componente 10 y 11 el valor de la característica miscelánea, área de porche cerrado en pies cuadrados, área de la piscina en pies cuadrados, e la componente 15, 21 y 22 el tipo de calefacción, pies cuadrados con acabado de baja calidad (todos los pisos), área de porche abierto en pies cuadrados, tamaño del lote en pies cuadrados, área de revestimiento de mampostería en pies cuadrados, baños completos en el sótano y cantidad de baños completos respectivamente.

A continuación, la imagen 3 muestra los gráficos de la varianza y la varianza acumulada según el numero de componentes principales incluidas, lo que demuestra visualmente que considerando 24 componentes se lográ explicar aproximadamente un 95% de la varianza de los datos.

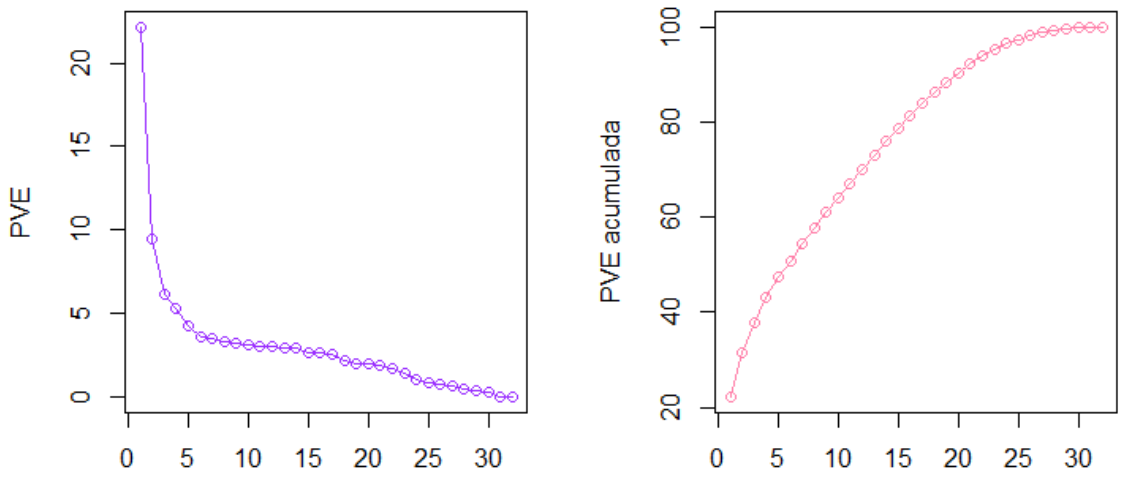

Imagen 3: Varianza y varianza acumulada según número de componentes principales

La imagen 4 muestra gráficamente la relación entre las observaciones y las variables en un conjunto de datos, cada observación está representada por un punto y cada variable está representada por un vector. La longitud y dirección del vector indican la contribución de la variable.

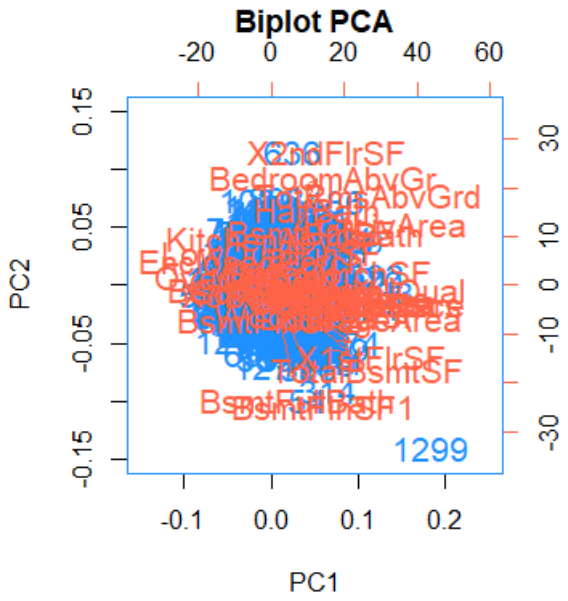

Imagen 4: Biplot PCA del subconjunto de variables numéricas

Por último, se aplica la función *k-means* junto al método del codo *wss* y el coeficiente de silueta *silhouette.* En la imagen 5 se observa que el número óptimo de clúster es dos.

Imagen 5: Clúster plot método *k-means*

Finalmente, la imagen 6 muestra que el método de agrupamiento jerárquico determinó que el número óptimo de clúster es 4.

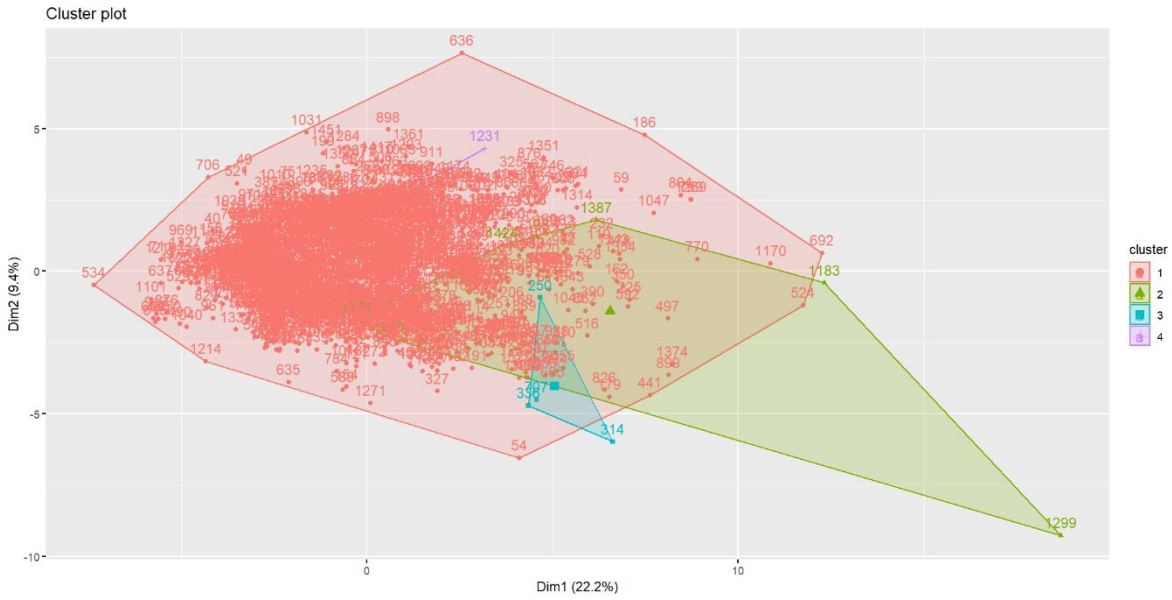

Imagen 6: Clúster plot método *hierarchical clustering*

CONCLUSIÓN

El análisis de componentes principales indica que para reducir la dimensionalidad del subconjunto de variables numéricas se debería no considerar 9 variables. Además, el número óptimo de clúster esta entre 2 y 4 para determinar cual método es mas confiable se debería realizar un análisis detallado de las variables numéricas. Finalmente, para futuros trabajos se debe considerar la inclusión de las variables categóricas, ya que puede contener mucha información significativas para determinar las características que influyen en el precio de las casas en Ames, Iowa, EE. UU.

ANEXO

Variables numéricas

Nombre variable	Descripción variable
SalePrice	Precio de venta de la propiedad en dólares
LotFrontage	Pies lineales de calle conectados a la propiedad
LotArea	Tamaño del lote en pies cuadrados
OverallQual	Calidad general del material y del acabado
GeneralCond	Calificación de condición general
MasVnrArea	Área de revestimiento de mampostería en pies cuadrados
BsmtExposure	Paredes de sótano a nivel de jardín o de salida
BsmtFinSF1	Pies cuadrados acabados tipo 1
BsmtFinSF2	Pies cuadrados acabados tipo 2
BsmtUnfSF	Pies cuadrados sin terminar de sótano
TotalBsmtSF	Pies cuadrados totales de área del sótano
1stFlrSF	Pies cuadrados del primer piso
2ndFlrSF	Pies cuadrados del segundo piso
LowQualFinSF	Pies cuadrados con acabado de baja calidad (todos los pisos)
GrLivArea	Superficie habitable sobre el nivel del suelo (pies cuadrados)
BsmtFullBath	Baños completos en sótano
BsmtHalfBath	Medios baños del sótano
Baño Completo	Baños completos sobre rasante
HalfBath	Medios baños sobre el nivel del suelo
Dormitorio	Número de dormitorios por encima del nivel del sótano
Cocina	Número de cocinas
TotRmsAbvGrd	Total de habitaciones sobre rasante (no incluye baños)
Chimeneas	Número de chimeneas
GarageCars	Tamaño del garaje en capacidad de coche
Área del garaje	Tamaño del garaje en pies cuadrados
WoodDeckSF	Área de cubierta de madera en pies cuadrados
OpenPorchSF	Área de porche abierto en pies cuadrados
Porche cerrado	Área de porche cerrado en pies cuadrados
3SsnPorch	Área de porche de tres estaciones en pies cuadrados
ScreenPorch	Área del porche cubierto en pies cuadrados
PoolArea	Área de la piscina en pies cuadrados
MiscVal	Valor de la característica miscelánea
MoVendido	Mes vendido

Variables categóricas

Nombre variable	Descripción variable
AñoVendido	Año vendido
SaleType	Tipo de venta
SaleCondition	Condición de venta
MSSubClass	Clase de construcción
MSZoning	Clasificación general de zonificación
PoolQC	Calidad de la piscina
Valla	Calidad de la valla
MiscFeature	Función miscelánea no cubierta en otras categorías
GarageQual	Calidad del garaje
GarageCond	Estado del garaje
PavedDrive	Camino pavimentado
FireplaceQu	Calidad de la chimenea
Tipo de garaje	Ubicación del garaje
GarageYrBlt	Año de construcción del garaje
GarageFinish	Acabado interior del garaje
KitchenQual	Calidad de cocina
Año de construcción	Fecha de construcción original
AñoRemodAdd	Fecha de remodelación
RoofStyle	Tipo de techo
RoofMatl	Material del tejado
Exterior1º	Revestimiento exterior de la casa 1°
Exterior2º	Revestimiento exterior de la casa 2°
MasVnrType	Tipo de revestimiento de mampostería
Calefacción	Tipo de calefacción
HeatingQC	Calidad y estado de la calefacción
CentralAir	Aire acondicionado central
Eléctrico	Sistema eléctrico
Calle	Tipo de acceso vial
Callejón	Tipo de acceso al callejón
LotShape	Forma general de la propiedad
LandContour	Planitud de la propiedad
Utilidades	Tipo de utilidades disponibles
LotConfig	Configuración del lote
LandSlope	Pendiente de la propiedad
Barrio	Ubicaciones físicas dentro de los límites de la ciudad de Ames
Condición 1	Proximidad a la carretera principal o al ferrocarril
Condición 2	Proximidad a una carretera principal o vía férrea
BldgType	Tipo de vivienda
HouseStyle	Estilo de vivienda
ExterQual	Calidad del material exterior
ExterCond	Estado actual del material en el exterior
Cimentación	Tipo de cimentación
BsmtQual	Altura del sótano
BsmtCond	Estado general del sótano
Funcional	Calificación de funcionalidad del hogar
BsmtFinType1	Calidad del área terminada del sótano
BsmtFinType2	Calidad de la segunda área terminada (si está presente)