MODELO PREDICCIÓN DEL PRECIO!

Objetivo

El objetivo de esta práctica consiste en predecir el precio de venta final de cada vivienda de la lista, dado un conjunto de datos de 79 variables explicativas que describen aspectos de las viviendas residenciales en Ames, Iowa.

Aplicando técnicas de regresión y habilidades de análisis exploratorio de datos.

LIMPIEZA DATOS

Después de analizar los campos con valor null y nan y comprobar que son los mismos campos, estudiamos qué variables nos van a ser útiles o no dentro del conjuntos de datos.

Vamos a calcular el porcentaje entre las columnas vacías para compararlas con la correlación de los campos, viendo si las variables son necesarias o se podrían descartar.

En la imagen podemos ver las columnas con mayor cantidad de filas nulas/nan

```
PoolQC          99.520548
MiscFeature     96.301370
Alley           93.767123
Fence           80.753425
FireplaceQu     47.260274
LotFrontage     17.739726
GarageCond       5.547945
GarageType       5.547945
GarageYrBlt      5.547945
GarageFinish     5.547945
GarageQual       5.547945
BsmtExposure     2.602740
BsmtFinType2     2.602740
BsmtFinType1     2.534247
BsmtCond         2.534247
BsmtQual         2.534247
MasVnrArea       0.547945
MasVnrType       0.547945
Electrical       0.068493
Utilities        0.000000
Name: NULL, dtype: float64
```

Correlación entre variables

El cálculo de la correlación es una técnica estadística usada para determinar la relación entre dos o más variables, el coeficiente de correlación puede variar desde -1.00 hasta 1.00.

En nuestro caso, podemos ver que donde mayor correlación entre variables tenemos, es entre el campo GarageCars, que nos indica el tamaño del garaje en capacidad de automóvil y GarageArea, que nos indica el tamaño del garaje en pies cuadrados.

Pero en las variables que nos fijamos, son las que tengan mayor correlación con nuestra variables objetivo, SalePrice, que es el campo OverallQual, y esta nos indica la calidad general del material y del acabado.

Otra que parece importante también y tiene buena correlación es la variable GrLivArea, que nos indica los pies cuadrados de área habitable sobre el nivel del suelo.

```
SalePrice     SalePrice      1.000000
GarageCars    GarageArea     0.882475
GarageYrBlt   YearBuilt      0.825667
GrLivArea     TotRmsAbvGrd   0.825489
1stFlrSF      TotalBsmtSF    0.819530
SalePrice     OverallQual    0.790982
GrLivArea     SalePrice      0.708624
              2ndFlrSF       0.687501
BedroomAbvGr  TotRmsAbvGrd   0.676620
BsmtFinSF1    BsmtFullBath   0.649212
GarageYrBlt   YearRemodAdd   0.642277
SalePrice     GarageCars     0.640409
FullBath      GrLivArea      0.630012
GarageArea    SalePrice      0.623431
TotRmsAbvGrd  2ndFlrSF       0.616423
TotalBsmtSF   SalePrice      0.613581
2ndFlrSF      HalfBath       0.609707
1stFlrSF      SalePrice      0.605852
GarageCars    OverallQual    0.600671
GrLivArea     OverallQual    0.593007
dtype: float64
```

Selección de variables

Una correlación fuerte se acerca al 1.00, por eso en el caso de que el porcentaje en las columnas null/nan no aparezcan en el cálculo de una correlación superando el 0.5%, descartamos esas variables, como en las variables mostradas en la foto, por ejemplo la variable fence, que mide la calidad de la valla o la variable FireplaceQu que mide la calidad de chimenea.

Y en caso de que si que aparezcan dentro del umbral, como por ejemplo el tipo o la ubicación del garaje, que son más relevantes para poder predecir el cálculo del precio de la casa, rellenaremos los datos vacíos con la media del resto de datos de su columna.
Y en caso de que sea una variable categórica utilizaremos la moda de la columna.

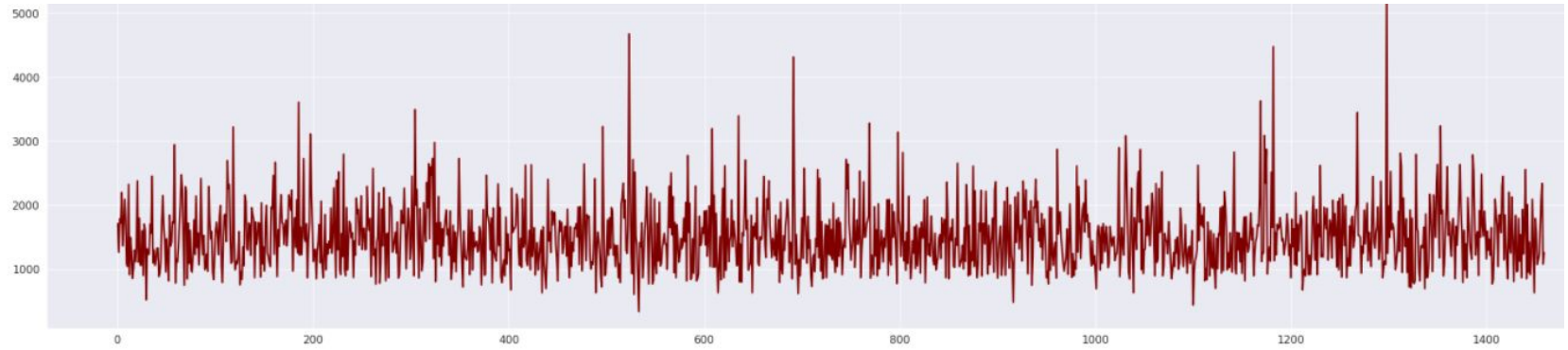
* PoolQC	99.520548
* MiscFeature	96.301370
* Alley	93.767123
* Fence	80.753425
* FireplaceQu	47.260274
* LotFrontage	17.739726

OUTLIERS

Los outliers son los valores que se escapan al rango en donde se concentran la mayoría de muestras, desviaciones que pueden desbalancear o empeorar un entrenamiento.

Uno de los principales outliers que hemos seleccionado ha sido el campo GrLivArea, esta mide los pies cuadrados de área habitable sobre el nivel del suelo, que es una de las variables que más relación tienen para poder predecir el precio de las casas, por lo que cuanto menos desbalance tengamos y más concentradas están las muestras, mayor será el acierto de nuestro modelo.

En un primer análisis hemos decidido deshacernos de todos aquellos que su valor esté por encima de los 1500 pies cuadrados, concentrando así más los datos y quitando algunos datos punteros y otras excepciones que meten ruido al conjunto de datos para el entrenamiento.



Algoritmos de regresión

Se han utilizado cuatro diferentes algoritmos de regresión, el Random Forest Regressor, ExtraTreesRegressor, DecisionTreeRegressor y el GradientBoostingRegressor

En cada uno de ellos se ha utilizado la clase de SelectKBest, que selecciona las características de acuerdo con las k puntuaciones más altas y elimina todas las demás características.

En nuestro caso hemos realizado dos pruebas por cada algoritmo, una con las 20k mejores y otra con las 10k mejores para poder probar diferentes resultados con diferentes configuraciones al entrenar.

También se ha utilizado la clase Pipeline de Scikit-learn, que está diseñada como una forma manejable de aplicar una serie de transformaciones de datos seguidas por la aplicación de un estimador, en nuestro caso la hemos utilizado para cargar el seleccionador de las k mejores variables y aplicar cada algoritmo.

Por último se ha calculado de cada algoritmo el error medio absoluto, y el resultado de la predicción sobre la variable objetivo, el precio de la casa.

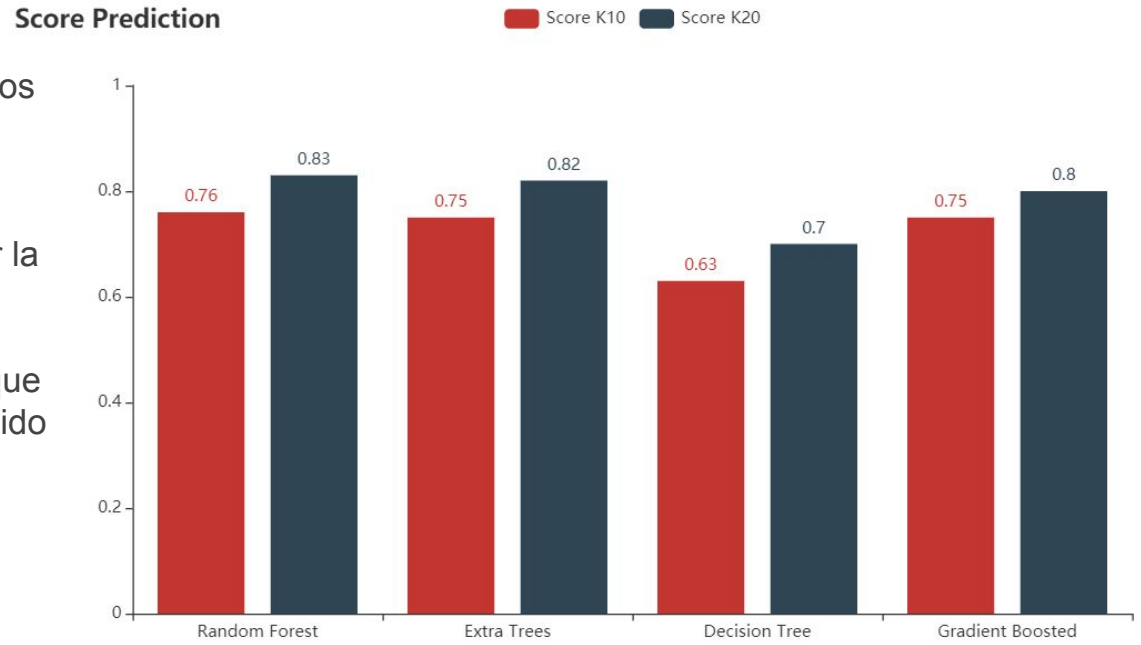
COMPARACIÓN GRÁFICOS

En la gráfica se pueden ver los resultados de los cuatro diferentes algoritmo utilizados para la predicción del precio.

Se ha utilizado la función .r2_score para medir la precisión que va de 1.0 como máximo.

Podemos ver que de los cuatro algoritmos el que mejor resultado da es el Random Forest, seguido por Extra trees y GradientBoosting.

Otra cosa que salta a la vista es que en los cuatro algoritmos el resultado mejora cuando utilizas el k20 que el k10.

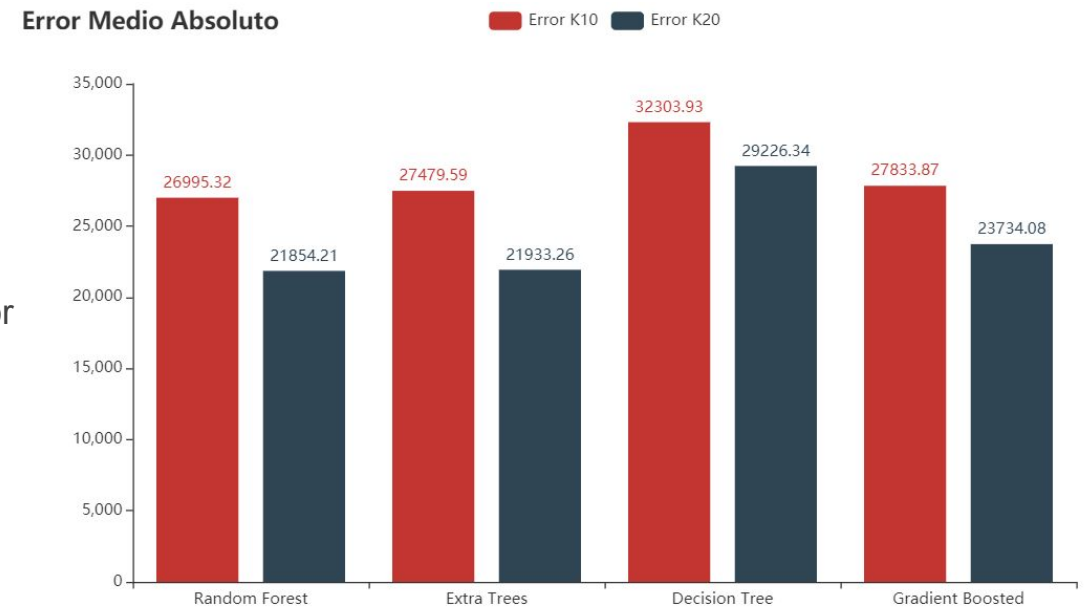


COMPARACIÓN GRÁFICOS

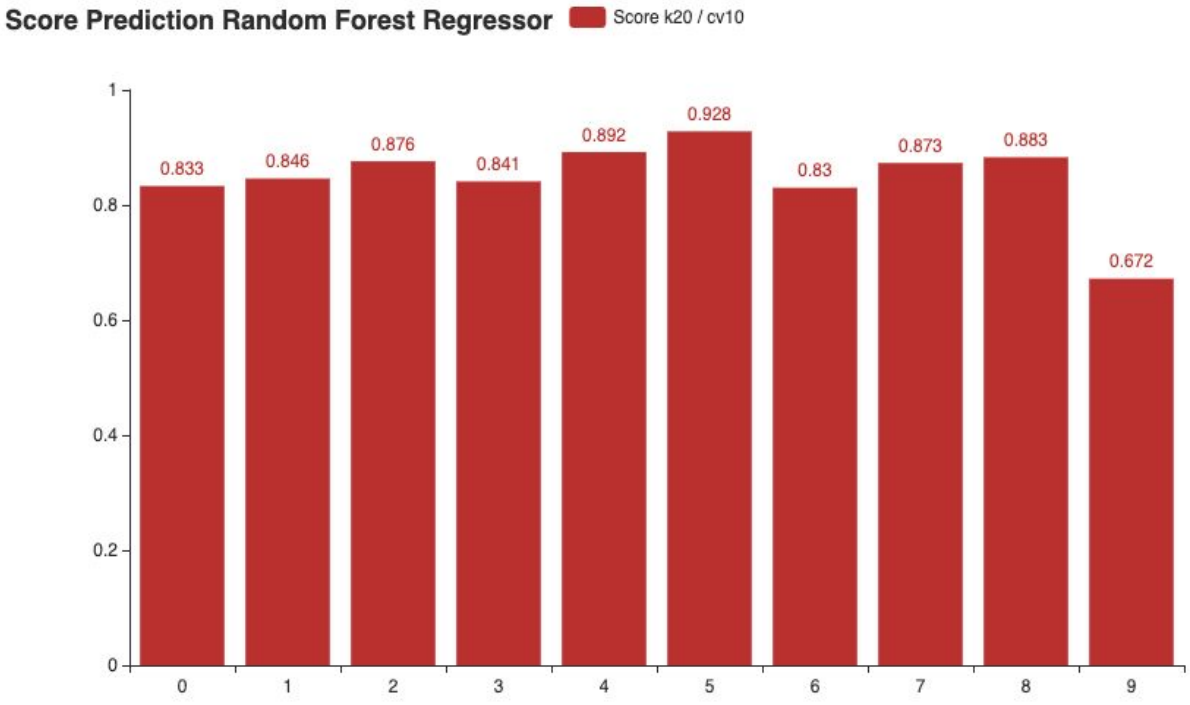
En esta gráfica podemos ver el error medio absoluto de cada algoritmo.

Podemos ver que como en la otra gráfica, se sigue cumpliendo que cuantas más variables selecciones mejor resultado te va a salir y menor error saldrá.

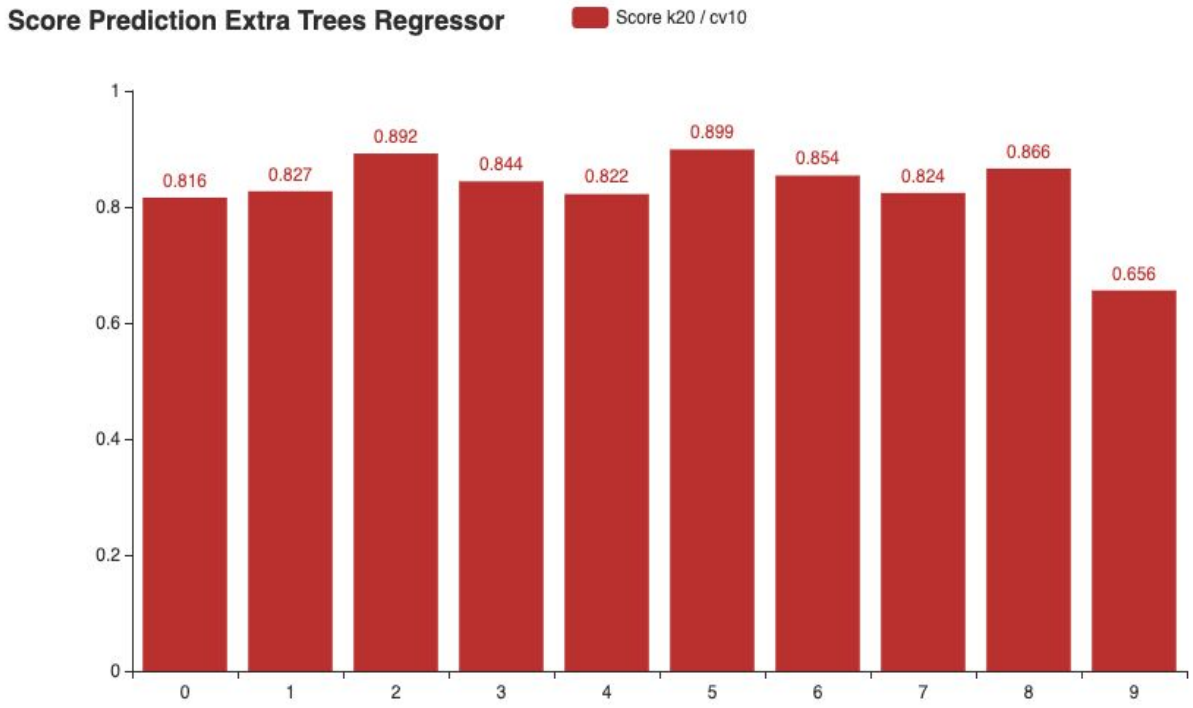
En este caso, cuando mayor error aparece es con el algoritmo Decision tree utilizando las 10k mejores, seguido de GradientBoosting y Extra trees.



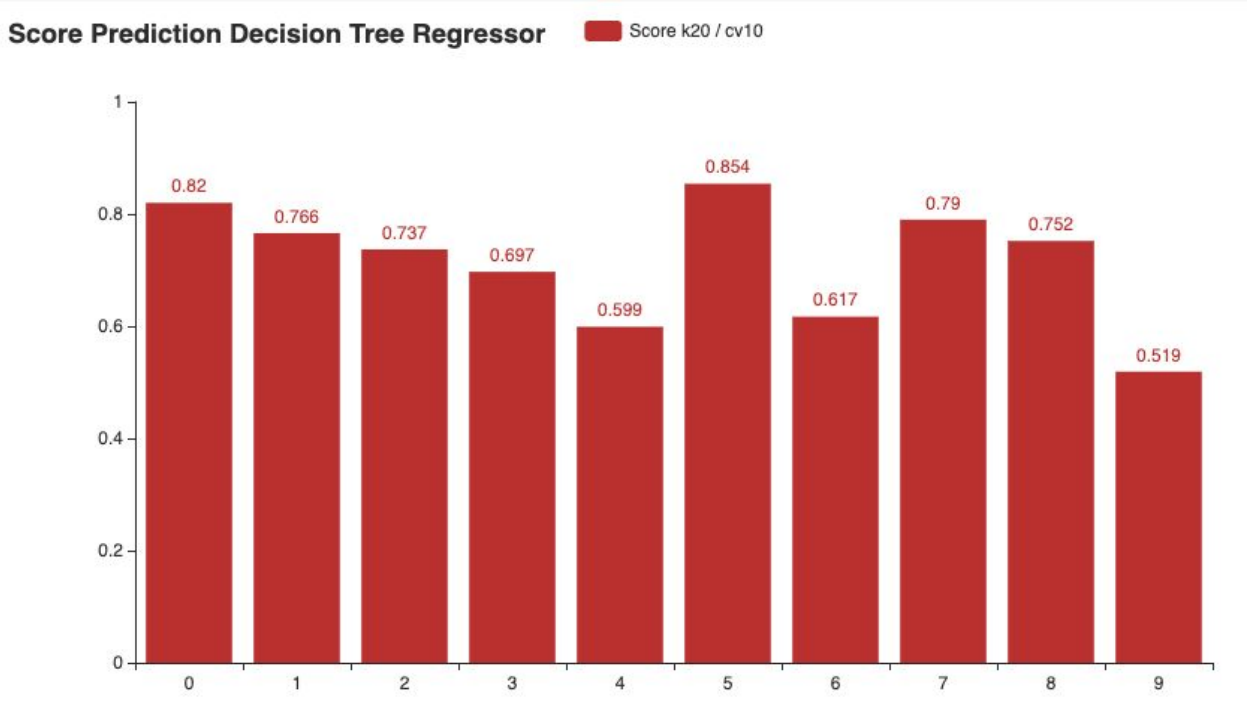
CROSS - VALIDATION



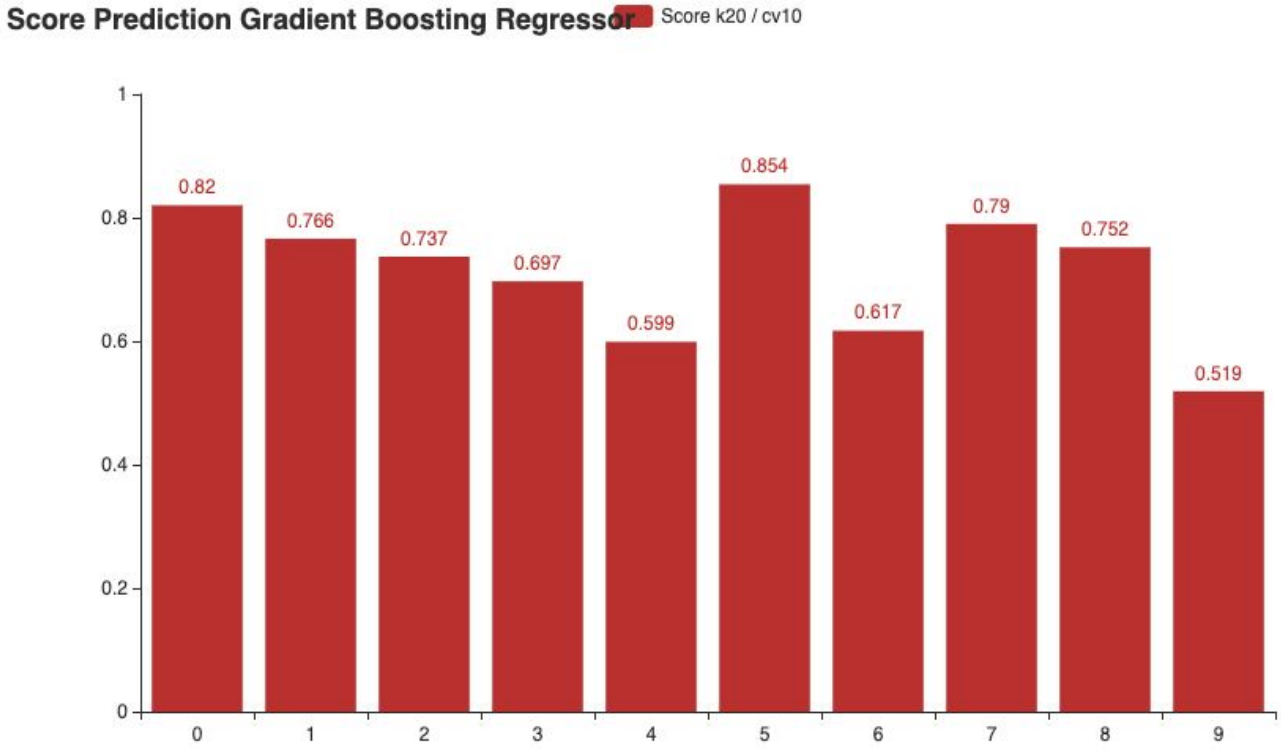
CROSS - VALIDATION



CROSS - VALIDATION



CROSS - VALIDATION



CONCLUSIONES

Dados los resultados obtenidos en el **cross - validation**:

- Random Forest Regressor = 0.8473458049476686
- Extra Trees Regressor = 0.8300519653943574
- Decision Tree Regressor = 0.7151738215845189
- Gradient Boosting Regressor = 0.8638985734434254

Podemos concluir que el algoritmo más correcto para realizar dicha predicción de precio de las casa es el Random Forest Regressor mediante SelectKBest (chi2, k=20) el cuál da un resultado de :

- Mean absolute error: 21854.211
- R2 score report: 0.825