



## ***Conclusiones generales Dataset Ames Housing***

### **Equipo de trabajo:**

José Nardulli

Germán Daniel Hilgert

Iván González Seguezzo

### **Profesor:**

Luca Città Giordano

### **Tutor:**

Juan Felipe Gonzalez Sanmiguel

Comisión 22740 Data Science

***Fecha 24/02/2022***

# Índice

|  |           |
|--|-----------|
| <b>Problemática</b>  | <b>3</b>  |
| <b>Objetivos Generales</b>   | <b>3</b>  |
| Objetivos de la investigación:   | 3         |
| Objetivo principal:  | 3         |
| Objetivos secundarios:   | 3         |
| <b>Adquisición de datos (Data Acquisition)</b>                                       | <b>4</b>  |
| <b>Limpieza de datos (Data Wrangling)</b>  | <b>5</b>  |
| Características generales del dataset:   | 5         |
| <b>Conclusiones obtenidas por la limpieza y manejo de datos:</b>                     | <b>6</b>  |
| Manejo de data perdida en el dataset:  | 6         |
| <b>Conclusiones obtenidas en el EDA:</b>   | <b>7</b>  |
| Las correlaciones más fuertes son:   | 7         |
| <b>Análisis Univariado</b>   | <b>8</b>  |
| Conclusiones obtenidas en el análisis Univariado:                                    | 8         |
| Precio de venta:   | 8         |
| <b>Análisis Bivariado</b>  | <b>9</b>  |
| <b>Conclusiones obtenidas en el análisis Bivariado:</b>                              | <b>11</b> |
| Tipos de viviendas:  | 11        |
| Condiciones de venta:  | 11        |
| Barrios:   | 12        |
| Tipos de calle:  | 12        |
| Tipo de Zona:  | 13        |
| <b>Análisis Multivariado</b>   | <b>14</b> |
| <b>Conclusiones obtenidas del análisis Multivariado:</b>                             | <b>15</b> |
| Precio de venta año de construcción y calidad general:                               | 15        |
| Tipo de Zona:  | 15        |
| Estilo de vivienda:  | 15        |
| <b>Algoritmos de Machine Learning</b>  | <b>16</b> |
| <b>Conclusiones de los algoritmos de Machine Learning</b>                            | <b>17</b> |
| <b>Métricas de efectividad para los distintos algoritmos aplicados en el modelo.</b> | <b>20</b> |

## Problemática

Inmobiliaria de primera línea requiere la creación de nuevas campañas publicitarias con el fin de aumentar ventas, para esto solicita al equipo de Data Science una solución para obtener palabras clave y atributos de una vivienda que puedan ser determinantes en la atención del cliente, y sucesiva compra del inmueble.

Para esto se provee la base de datos de la compañía donde se lista el historial de los últimos años.

## Objetivos Generales

En este trabajo, se desarrolla un análisis sobre los datos correspondientes a las viviendas pertenecientes a la ciudad de Ames situada en el condado de Story del estado de Iowa en Estados Unidos. Su población es de 66.427 habitantes según el censo de 2020 y tiene una longitud de 55,90 kilómetros cuadrados.

El "Dataset" contiene información acerca de características muy variadas sobre viviendas y donde cada una fue valuada a un determinado precio. Se entiende que las diversas cualidades como las dimensiones de terreno, la cantidad de habitaciones o la ubicación de cada casa influyen en el proceso de valuación. Teniendo esta premisa en claro, este trabajo tendrá como finalidad entender cómo se comportan las distintas variables y en qué medida influyen en el precio de venta final..

La información extraída puede ser de gran utilidad para agencias inmobiliarias y para potenciales compradores de viviendas en la zona bajo análisis.

### Objetivos de la investigación:

#### Objetivo principal:

Realizar un estudio complejo acerca de las distintas variables que podrían afectar o modificar el precio de una vivienda proyectar las relaciones entre las mismas y realizar modelos de ML con posibilidad de análisis predictivo

#### Objetivos secundarios:

Limpiar y transformar el dataset elegido para poder manipular la información de forma eficiente.

Extraer \*insights\* que permitan un mejor entendimiento de la determinación de los precios de los inmuebles.

Encontrar las correlaciones más importantes entre variables, ya sean directas o inversas.

Determinar qué variables categóricas pueden diferenciar distintos rangos de precio.

## Adquisición de datos (Data Acquisition)

Fuente del dataset:

Realizamos una investigación por diferentes páginas afines a la ciencia de datos realizando una primera selección de datasets con características necesarias para realizar el trabajo de este curso. En esta instancia cada integrante tuvo vía libre a su imaginación.

Fundamentos de la elección:

De todos los encontrados realizamos una segunda selección teniendo en cuenta diferentes parámetros como:

Tema a tratar: Debe ser interesante para los 3 integrantes del grupo.

Objetivo y resultados: El Dataset debe permitir realizar un análisis completo.

Datos suficientes: El dataset debe contener muchos atributos y registros para facilitar el aprendizaje de nuestros algoritmos de Machine Learning.

Originalidad: El tema a tratar debe ser original en la comisión.

Calidad: El trabajo finalizado debe ser de calidad. Para poder formar parte de nuestro Portfolio.

Nos decidimos a trabajar sobre el tema venta y comparación de viviendas. El dataset cuenta con 2930 registros y 80 atributos. Cada registro corresponde a un inmueble. Los atributos corresponden a diferentes características de la vivienda. A continuación se comparte el link de la publicación:

<https://www.kaggle.com/prevek18/ames-housing-dataset>

## Limpieza de datos (Data Wrangling)

Para el considerar un Data Wrangling completo se tiene en cuenta realizar verificar los datos fuera de rango (outliers), los datos faltantes (missing data) y los errores categoriales (categorical data).

Características generales del dataset:

- 2930 registros de viviendas.
- 80 columnas con atributos de cada vivienda. 1 columna adicional PID con número de identificación único de vivienda.
- Tipos de dato:
  - o 43 del tipo Object.
  - o 11 del tipo Float.
  - o 26 del tipo Int64.
- Valores nulos:
  - o 53 columnas con datos completos (2930 not null).
  - o 21 columnas con datos completos <95% (2771 a 2929 not null).
  - o 1 columna con la mitad de datos faltantes (1508 not null).
  - o 4 columnas con muchos datos null. Porcentaje de datos faltantes 80% (13 a 572 not null).

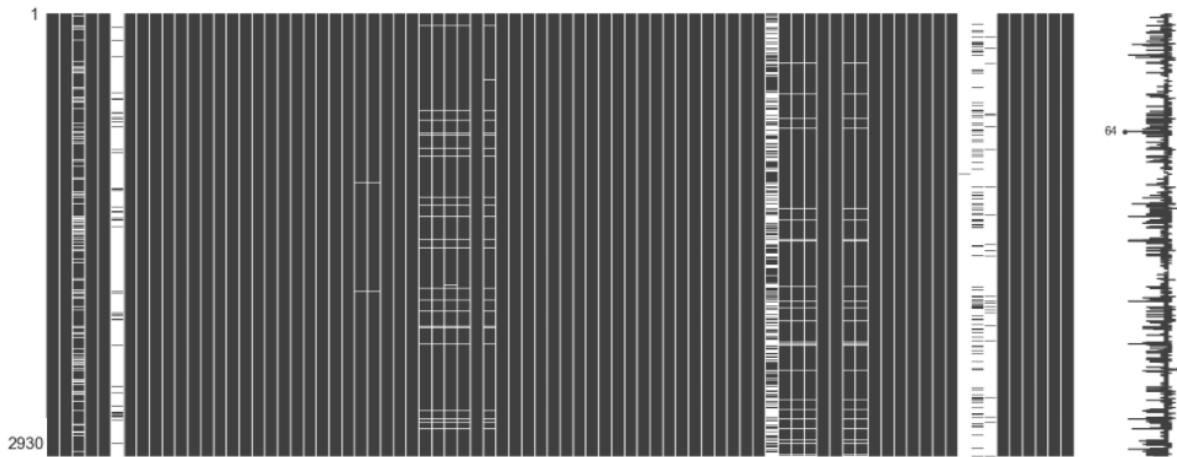


Fig. 1. Gráfico Missingno. Datos faltantes en dataset.

## Conclusiones obtenidas por la limpieza y manejo de datos:

### Conclusiones obtenidas en el manejo de Outliers

Existe casos que podrían contrariar la lógica inicial de la propuesta en la relación de precio y área y el precio y tamaño del garaje:

- 5 Viviendas con demasiada área de terreno y poco valor.
- 5 garages gigantes y con poco valor
- 4 viviendas con mucho valor y sin área de garaje.
- 2 viviendas extremadamente económicas en comparación con su grupo.

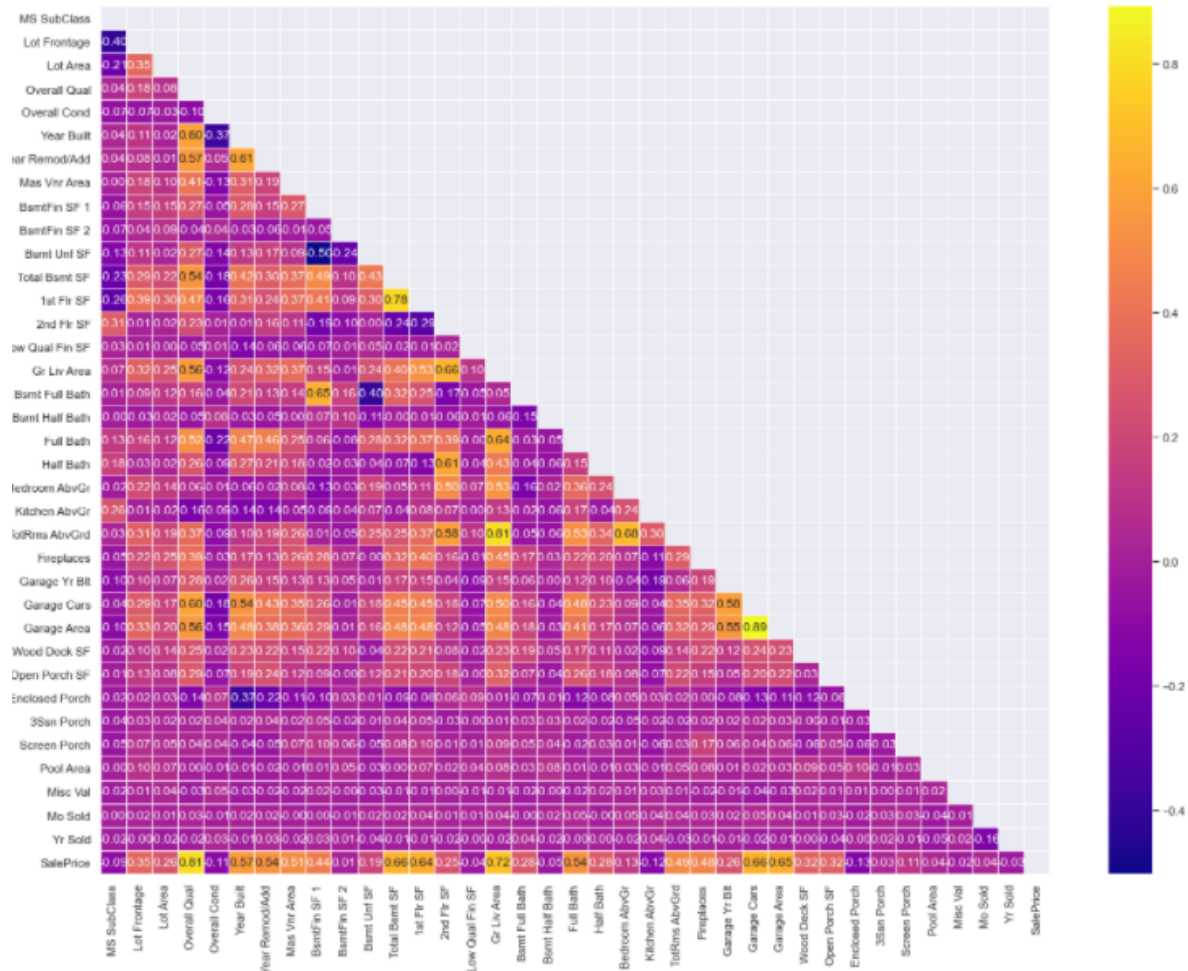
### Manejo de data perdida en el dataset:

- Realizamos análisis exhaustivos de los datos faltantes.
- Dependiendo cada caso con los datos faltantes podemos:
  - o Mantenerlos.
  - o Borrarlos.
  - o Reemplazarlos por otro valor.
- Comenzamos analizando los missing <1%. Dado que son pocos los miramos en detalle para dar una solución individual.
- Se continua con los missing que comprenden de >2% y <17%. Se verifican casos particulares.
- Se finaliza con los missing >80%. Columnas ['Fence', 'Alley', 'Misc Feature', 'Pool QC'] se dropean.
- Luego de verificar los PID y la documentación, podemos estimar que los valores nulos corresponden a que la vivienda no cuenta con *sótano*. Por lo que se procede a rellenar con 0 los valores numéricos y None los valores String.
- Observando la documentación deducimos que los valores faltantes de *Mas Vnr Type* corresponden a que la vivienda no cuenta con revestimiento de mampostería. Por lo que rellenamos los valores por 0 y *None* respectivamente.

Al finalizar la limpieza obtenemos un dataset con 2916 registros y 76 columnas.

## Conclusiones obtenidas en el EDA:

A partir de un análisis de correlación de variables y un mapa Heatmap verificamos qué variables tienen mejor relación entre sí. Nuestro caso particular es ver cómo reaccionan todas las variables respecto al output Precio de venta.



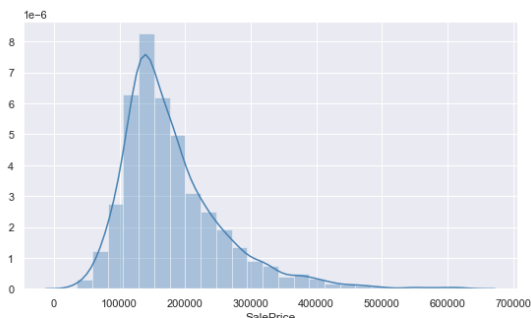
Las correlaciones más fuertes son:

- SalePrice:
  - o Overall Qual. Calidad general de la vivienda.
  - o Gr Liv Area. Pies cuadrados de superficie habitable sobre el nivel del suelo (no cuenta sótano).
  - o Garage Cars. Cantidad de vehículos en Garage.
  - o Garage Area. Pies cuadrados de superficie de Garage.
  - o Total Bsmt SF. Pies cuadrados totales del área del sótano.
  - o Year Built. Año de construcción.
  - o Mas Vnr area. Área de pared con revestimiento de mampostería.
- Year Built:
  - o Garage Yr Built. Año de construcción de garage respecto de la vivienda. Debido a que las construcciones actuales incluyen al Garage en sus planos y ambos se construyen el mismo año.
  - o Overall Qual. A vivienda más nueva mejor calidad de construcción.
- TotRms AbvGr & Gr LivArea. Cantidad de habitaciones respecto al área habitable de la vivienda.
- Garage Cars & Garage Area. Cantidad de vehículos en Garage respecto al área del garage.

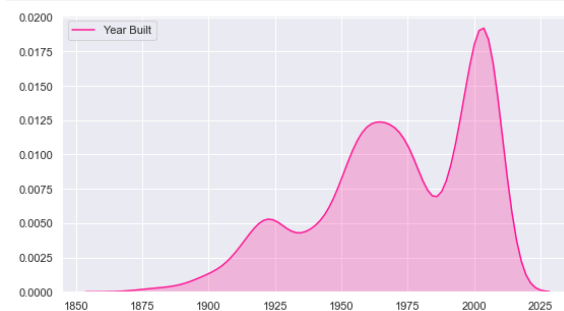
## Análisis Univariado

Se analizan las variables individualmente. Se grafica la distribución de precios de venta, año de construcción, área de garage, mes de venta, entre otros.

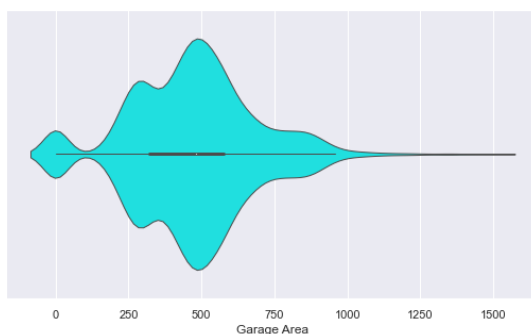
```
#Distribución de la variable PRECIO DE VENTA
sns.distplot(df['SalePrice'], bins=25, color="steelblue", kde=True)
plt.show()
```



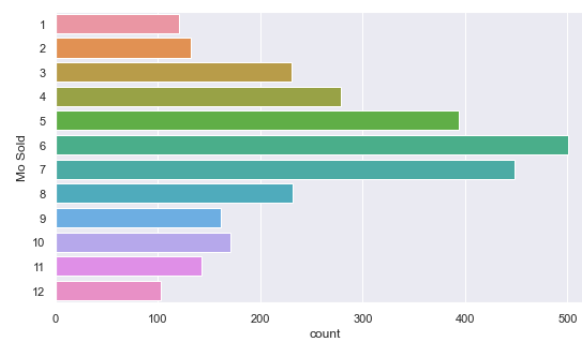
```
plt.figure(figsize=(9,5))
sns.kdeplot(data=df['Year Built'], color="deeppink", shade=True)
plt.show()
```



```
plt.figure(figsize=(9,5))
violinplot(df['Garage Area'],linewidth=1, color="cyan")
plt.show()
```



```
plt.figure(figsize=(9,5))
sns.countplot(y=df['Mo Sold'])
plt.show()
```



### Conclusiones obtenidas en el análisis Univariado:

- El frente de lote se encuentra mayormente entre los valores 50 a 80 metros de longitud.
- En su mayoría son viviendas consideradas con calidad y condiciones en un nivel medio.
- Sólo hay algunas pocas casas con dos cocinas, más del 90% de las casas vendidas poseen una sola cocina.
- Las casas que tienen capacidad para dos autos en su garage supera el doble de las viviendas que poseen sólo uno.
- La cantidad de habitaciones en la totalidad de las casas es mayormente de 6.
- Las casas que poseen 3 habitaciones son mayores en cantidad a las viviendas que poseen 2 y 4 habitaciones juntas.
- La mayoría de las transacciones de venta se concretaron a mitad de año, en los meses de mayo, junio y julio.

### Precio de venta:

- los precios generales de las propiedades examinadas rondan entre los 120000\$ y los 755000\$ con la mayor parte de las propiedades ubicándose en el rango de los 213500\$



## Análisis Bivariado

Se analiza el output SalePrice respecto a columnas de interes ["Garage Area", "Lot Area", "Year Built", "Gr Liv Area", "Full Bath", "TotRms AbvGrd", "Yr Sold", "Mo Sold"], entre otras.

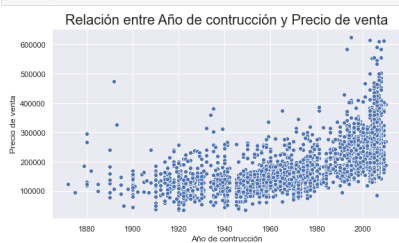


Se realizan gráficas dobles con atributos de interés.

Las pruebas comprenden estas selecciones:

- 1)\_ YearBuilt vs SalePrice.
- 2)\_ Garage Cars vs SalePrice.
- 3)\_ OverallQual vs SalePrice.
- 4)\_ MS SubClass vs Lot FrontAge.
- 5)\_ Full Bath vs SalePrice.
- 6)\_ Garage Cars vs Garage Area.
- 7)\_ TotRms AbvGrv vs Gr Liv Grid.
- 8)\_ Sale condition vs SalePrice.
- 9)\_ Neighborhood vs SalePrice.
- 10)\_ Street vs SalePrice.
- 11)\_ MS Zoning vs SalePrice.
- 12)\_ Overall Cond vs YearBuilt.

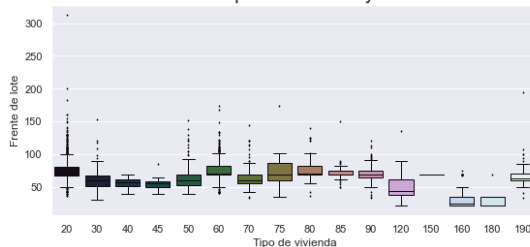
```
sns.scatterplot(x = df['Year Built'], y = df['SalePrice'], palette = 'cubehelix')
VariableX="Año de construcción"
VariableY="Precio de venta"
plt.title(f'Relación entre (VariableX) y (VariableY)', fontsize = 20)
plt.xlabel(VariableX)
plt.ylabel(VariableY)
plt.show()
```



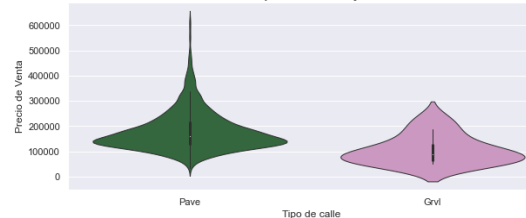
```
plt.figure(figsize=(9,4))
sns.scatterplot(data=df,x="SalePrice", y="Garage Cars")
plt.show()
```



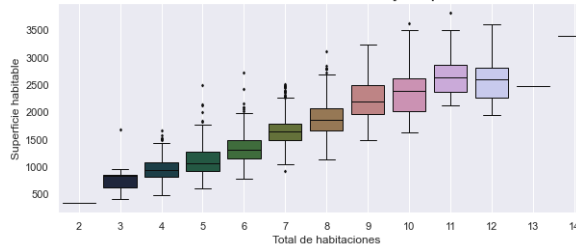
Relación entre Tipo de vivienda y Frente de lote



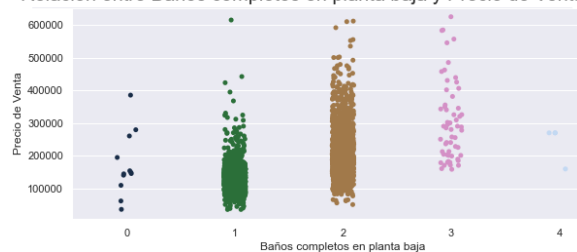
Relación entre Tipo de calle y Precio de Venta



Relación entre Total de habitaciones y Superficie habitable



Relación entre Baños completos en planta baja y Precio de Venta



## Conclusiones obtenidas en el análisis Bivariado:

- se observan que hay tendencias obvias que afectan principalmente el precio de venta, el tamaño del lote la cantidad de comodatos (piscina garajes), aumentan el precio total solapando a otro tipo de data como la ubicación y se aprecie más el tamaño de áreas cubiertas que de áreas verdes y los comodatos más valorados sería la piscina seguida por el garaje
- propiedades más nuevas por regla general presentan un costo superior a propiedades antiguas con marcadas excepciones
- podría marcarse como una obviedad, pero la calidad general impacta directamente sobre el precio del lote, a mayor calidad mayor precio

## Tipos de viviendas:

- 20: 1 PISO - 1946 Y MÁS NUEVOS TODOS LOS ESTILOS
- 30: 1 PISO - 1945 Y MÁS ANTIGUOS
- 40: 1 PISO CON ÁTICO TERMINADO TODAS LAS EDADES
- 45: 1-1/2 HISTORIA - SIN TERMINAR TODAS LAS EDADES
- 50: 1-1/2 HISTORIA TERMINADA TODAS LAS EDADES
- 60: 2 PISOS 1946 Y MÁS NUEVOS
- 70: 2 PISOS 1945 Y MÁS ANTIGUOS
- 75: 2-1/2 HISTORIA TODAS LAS EDADES
- 80: DIVIDIDOS O MULTINIVEL
- 85: VESTÍBULO DIVIDIDO
- 90: DÚPLEX - TODOS LOS ESTILOS Y EDADES
- 120: PUD DE 1 PISO (Desarrollo de unidades planificadas) - 1946 Y MÁS RECIÉN
- 150: 1-1/2 STORY PUD - TODAS LAS EDADES
- 160: PUD DE 2 PISOS - 1946 Y MÁS NUEVOS
- 180: PUD - MULTINIVEL - INCLUYE NIVEL DIVIDIDO/VESTÍBULO
- 190: 2 CONVERSIÓN FAMILIAR - TODOS LOS ESTILOS Y EDADES

Los tipos de casas 160 y 180 son las que tienen menor longitud en el frente de lote. Mientras que los tipos 20, 60, 75 y 80 tienen los frentes de lote más grandes.

Las viviendas con 2 o más baños completos en planta baja conservan precios bastante más elevados que las que tienen 1 o ninguno. A su vez, la mayoría de los datos se encuentran en las casas que tienen uno o dos baños completos.

Cuántos más autos entran en el garaje de cada vivienda, mayor es la superficie que el mismo ocupa.

Se observa una correlación positiva fuerte entre ambas variables. Cuántas más habitaciones tiene la casa, mayor es la superficie que las mismas ocupan.

## Condiciones de venta:

- Normal: Venta normal
- Abnormal: Venta anormal. Comercio, ejecución hipotecaria, venta corta
- AdjLand: Adquisición de terrenos adyacentes
- Alloca: Asignación. Dos propiedades vinculadas con escrituras separadas, generalmente condominio con una unidad de garaje
- Family: Venta entre miembros de la familia
- Partial: La vivienda no se completó cuando se evaluó por última vez (asociada con viviendas nuevas)

Las ventas calificadas como parciales fueron pagadas por las casas con un precio mayor al resto. Mientras que las que se adquirieron como terrenos adyacentes, el precio pagado fue el menor.

#### Barrios:

- Blmngtn: Bloomington Heights
- Blueste: Bluestem
- BrDale: Briardale
- BrkSide: Brookside
- ClearCr: Clear Creek
- CollgCr: College Creek
- Crawfor: Crawford
- Edwards: Edwards
- Gilbert: Gilbert
- IDOTRR: Iowa DOT and Rail Road
- MeadowV: Meadow Village
- Mitchel: Mitchell
- Names: North Ames
- NoRidge: Northridge
- NPkVill: Northpark Villa
- NridgHt: Northridge Heights
- NWAmes: Northwest Ames
- OldTown: Old Town
- SWISU: South & West of Iowa State University
- Sawyer: Sawyer
- SawyerW: Sawyer West
- Somerst: Somerset
- StoneBr: Stone Brook
- Timber: Timberland
- Veenker: Veenker

Las casas más caras se ubican mayormente en: Stone Brook, Northridge Heights y Northridge. Algunos de los barrios más accesibles para adquirir una casa son: Meadow Village, Iowa DOT and Rail Road y Briardale

#### Tipos de calle:

- Pave: Calle pavimentada
- Grvl: Camino de gravilla

Las casas ubicadas en calles pavimentadas tienen en general un precio más elevado que las ubicadas sobre caminos de gravilla.

#### Tipo de Zona:

- A: Agricultura
- C: Comercial
- FV: Residencial de pueblo flotante
- I: Industrial
- RH: Residencial Alta Densidad
- RL: Residencial Baja Densidad
- RP: Parque Residencial de Baja Densidad
- RM: Residencial Media Densidad

Las viviendas ubicadas en zona de Residencial de pueblo flotante son más caras que el resto de las otras zonas. Mientras que las casas en zonas agrícolas, industriales y comerciales son las más baratas. El resto del tipo de zonas residenciales se encuentran en el medio de ambos extremos.

Las viviendas construidas a partir del año 1980 en adelante mayormente tienen una condición media respecto a la totalidad. Por otro lado, las viviendas construidas antes del 1970 tienen condiciones muy heterogéneas. Hay casas en muy buen estado y otras en condiciones muy malas.

Es una tendencia generalizada que las casas con mejor calidad general son aquellas que tienen solo un piso o nivel mientras que en las casas con más pisos la tendencia de calidad general suele ser descendente.

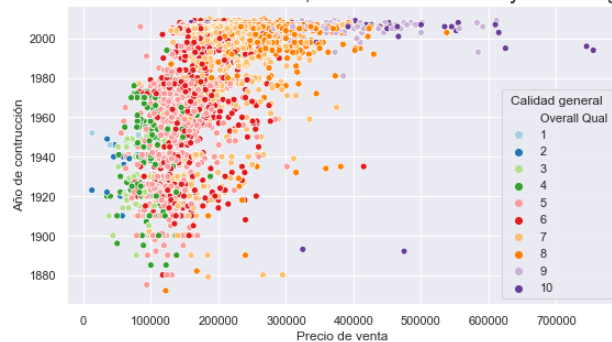
## Análisis Multivariado

Se realizan diferentes gráficas buscando patrones de agrupamiento entre las variables.

Precio de ventas, Año de construcción y Calidad general

```
sns.scatterplot(x = df['SalePrice'], y = df['Year Built'], palette = 'Paired', hue=df['Overall Qual'])
VariableX="Precio de venta"
VariableY="Año de construcción"
VariableCategorica="Calidad general"
plt.title(f'Relación entre {VariableX}, {VariableY} y {VariableCategorica}', fontsize = 20)
plt.xlabel(VariableX)
plt.ylabel(VariableY)
plt.legend(title=VariableCategorica)
plt.show()
```

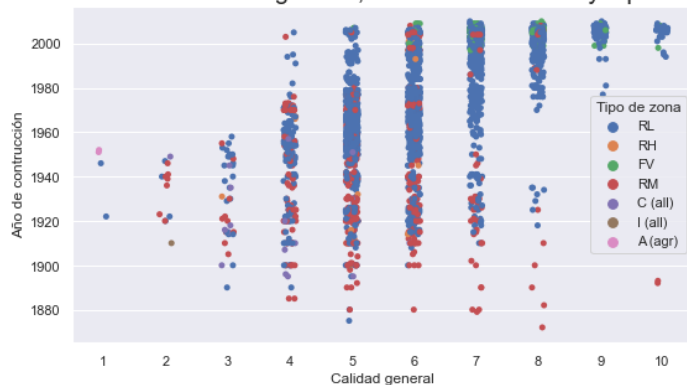
Relación entre Precio de venta, Año de construcción y Calidad general



Calidad general, Año de construcción y Tipo de zona

```
sns.stripplot(x = df['Overall Qual'], y = df['Year Built'], hue=df['MS Zoning'])
VariableX="Calidad general"
VariableY="Año de construcción"
VariableCategorica="Tipo de zona"
plt.title(f'Relación entre {VariableX}, {VariableY} y {VariableCategorica}', fontsize = 20)
plt.xlabel(VariableX)
plt.ylabel(VariableY)
plt.legend(title=VariableCategorica)
plt.show()
```

Relación entre Calidad general, Año de construcción y Tipo de zona



## Conclusiones obtenidas del análisis Multivariado:

la tendencia observada en el análisis bivariado se mantiene:

- Se observan que hay tendencias obvias que afectan principalmente el precio de venta, el tamaño del lote la cantidad de comodatos (piscina garajes), aumentan el precio total solapando a otro tipo de data como la ubicación y se aprecie más el tamaño de áreas cubiertas que de áreas verdes y los comodatos más valorados serían la piscina seguida por el garaje
- Propiedades más nuevas por regla general presentan un costo superior a propiedades antiguas con marcadas excepciones
- Podría marcarse como una obviedad, pero la calidad general impacta directamente sobre el precio del lote, a mayor calidad mayor precio

### Precio de venta año de construcción y calidad general:

Se pueden observar algunas tendencias de agrupamiento de datos en determinadas zonas. Hay una correlación positiva débil entre el precio de venta y el año de construcción. La calidad general está más relacionada con el precio de venta que con el año de construcción, porque se pueden observar más conformaciones de agrupamientos en forma de barras verticales que horizontales. Lo que significa que a medida que el precio crece la calidad va aumentando, mientras que por el lado del año de construcción se pueden ver calidades muy distintas para viviendas construidas durante la misma fecha.

### Tipo de Zona:

- A: Agricultura
- C: Comercial
- FV: Residencial de pueblo flotante
- I: Industrial
- RH: Residencial Alta Densidad
- RL: Residencial Baja Densidad
- RP: Parque Residencial de Baja Densidad
- RM: Residencial Media Densidad

Hay una correlación muy leve positiva entre la calidad general de las viviendas con su año de construcción. Las viviendas construidas en los últimos años tienden a tener una mayor calidad que las más antiguas. En esta misma zona de la gráfica las viviendas pertenecen al tipo de zona Residencial Baja Densidad. Mientras que algunas pocas casas que son antiguas y son consideradas de muy buena calidad pertenecen a la zona Residencial Media Densidad.

### Estilo de vivienda:

- 1Story: Un piso
- 1.5Fin: Un piso y medio, 2do nivel terminado
- 1.5Unf: Un piso y medio, 2do nivel sin terminar
- 2Story: dos pisos
- 2.5 Fin: Dos pisos y medio, 2do nivel terminado
- 2.5Unf: Dos pisos y medio, 2do nivel sin terminar
- Foyer: Hall dividido
- SLvl: Construido sobre dos niveles

La mayor condición general asignada a las viviendas fue de 5. La mayoría de los datos con esta condición tiene capacidad en el garaje para dos autos, y en menor medida en otros casos, para uno y para tres. Los que tienen capacidad para dos autos son mayormente casas de uno o de dos pisos. Mientras que los que tienen capacidad sólo para un vehículo tienen generalmente un solo piso de vivienda.

# Algoritmos de Machine Learning

Para la fase de algoritmos de prueba Machine Learning se optó por utilizar del tipo supervisado y no supervisado, buscando mejores rendimientos dependiendo las variables involucradas.

Los primeros algoritmos son de regresión lineal. Estos analizan las relaciones que hay entre las diferentes variables y devuelve un coeficiente de relación. Siendo 1 muy relacionado y 0 sin relación aparente. Lo utilizamos para determinar qué variables tienen mejor relación con el output.

El algoritmo del árbol de clasificación encuentra relaciones entre los datos y los clasifica en diferentes grupos, para luego poder predecir el valor de salida u output, en nuestro caso SalePrice o precio de venta.

El algoritmo SVM es del tipo clasificación supervisado. Utiliza los datos de entrada para entrenar hiperplanos que separan clases dependiendo sus atributos.

El algoritmo Random Forest es del tipo supervisado, este crea varios algoritmos de árbol de clasificación y luego genera un promedio de estos.

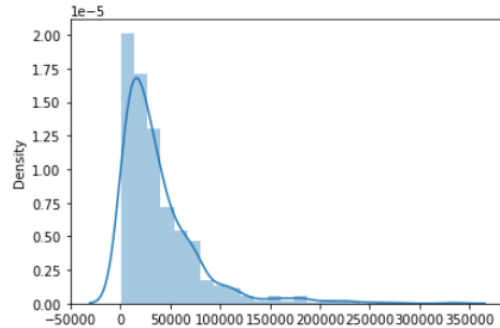
Se listan las variables involucradas para cada tipo de análisis donde encontramos relaciones fuertes:

- Regresión lineal simple
  - Year Built (año de construcción)
  - Sale Price (precio de venta)
- Regresión lineal múltiple
  - Con  $X = df\_final[['Lot Area', 'Year Built', 'Garage Cars', 'Overall Cond', 'SalePrice']]$ .
  - $y = df\_final['SalePrice']$  y también con  $['Overall Qual']$ .
- Árbol de clasificación
  - Con  $X[['Lot Area', 'Year Built', 'Garage Cars', 'Overall Cond', 'SalePrice']]$
  - Con output  $['Overall Qual']$
- Super Vector Regression
  - Con  $X[overall qual, gr liv area]$
  - Con output  $y[sale price]$ .
- Random Forest
  - $X[todos los atributos - central air]$
  - $Y[central air]$



## Conclusiones de los algoritmos de Machine Learning

Usando regresión lineal simple, establecemos el modelo de relación entre las áreas verdes de una propiedad y el precio de venta, para esto usaremos función lineares predictoras lo que nos permitirá estimar el precio de venta de una propiedad.

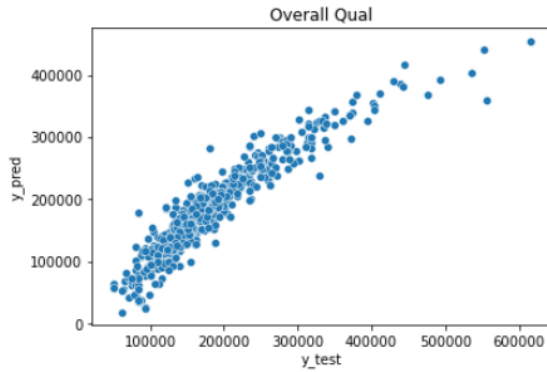


se puede observar como la relación tiene una pendiente acelerada a medida de que el precio general de la propiedad aumenta, lo que nos define que aunque existe cierta relación entre las variables estudiadas (áreas verdes y precio de propiedad) en realidad no es muy significativa a medida de que la propiedad se encarece, lo que nos hace recurrir a un modelo de regresión lineal múltiple para calcular cuál sería el mayor coeficiente de relación a la hora de determinar el precio de venta de una propiedad.

La Regresión lineal múltiple estudia todas las variables numéricas que pudiesen generar un impacto significativo en el precio de la propiedad.

Mostraremos a continuación solo algunas de las de las variables estudiadas con el fin de determinar la enorme diferencia entre los índices de correlación

|                       |              |
|-----------------------|--------------|
| <b>Overall Qual</b>   | 15484.223549 |
| <b>Overall Cond</b>   | 4379.511125  |
| <b>Year Built</b>     | 336.087589   |
| <b>Year Remod/Add</b> | 221.432446   |

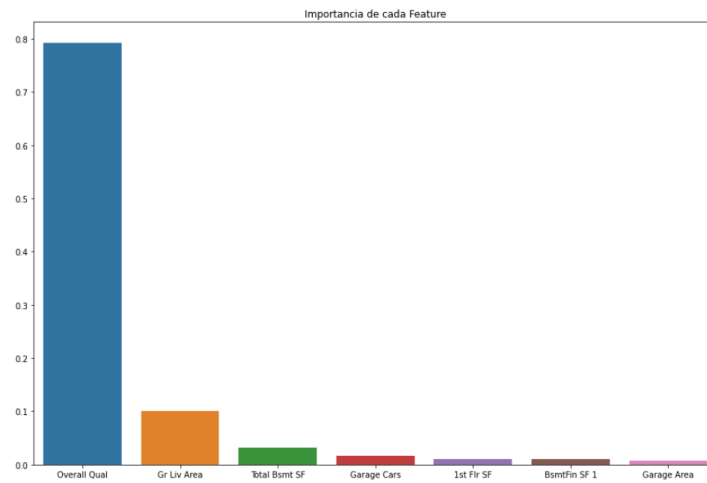


Como se puede observar la calidad general (Overall Qual) tiene un impacto enorme en el precio de venta final.

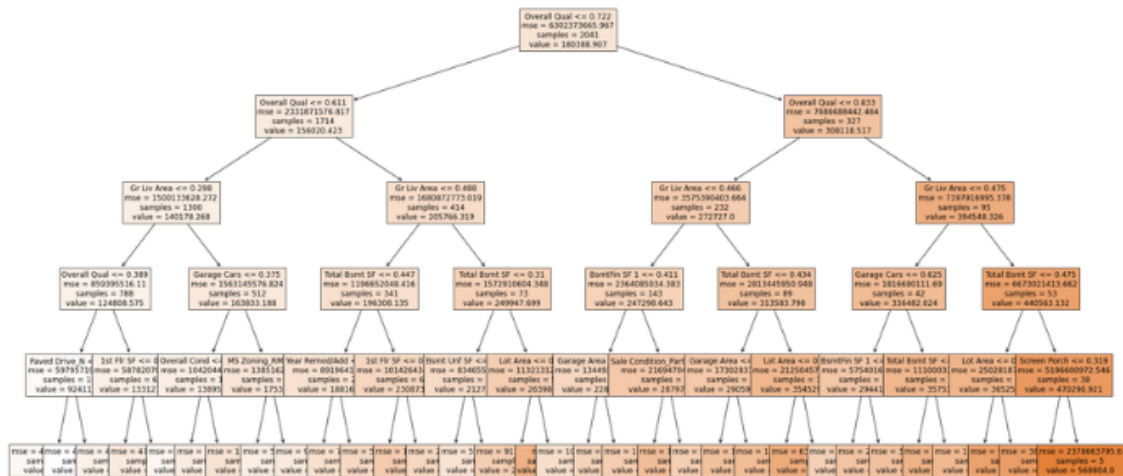
Al realizar el estudio general con árbol de decisión podemos obtener conclusiones más claras acerca de la relación entre las diferentes variables, y al mismo tiempo generar un modelo de predicción más exacto.

Entre las variables estudiadas podemos determinar la importancia de las mismas dentro de nuestro modelo predictivo.

|                    | Overall Qual | Gr Liv Area | Total Bsmt SF | Garage Cars | 1st Flr SF | BsmtFin SF 1 | Garage Area |
|--------------------|--------------|-------------|---------------|-------------|------------|--------------|-------------|
| <b>Importancia</b> | 0.792264     | 0.101342    | 0.031253      | 0.016339    | 0.010929   | 0.010017     | 0.007158    |

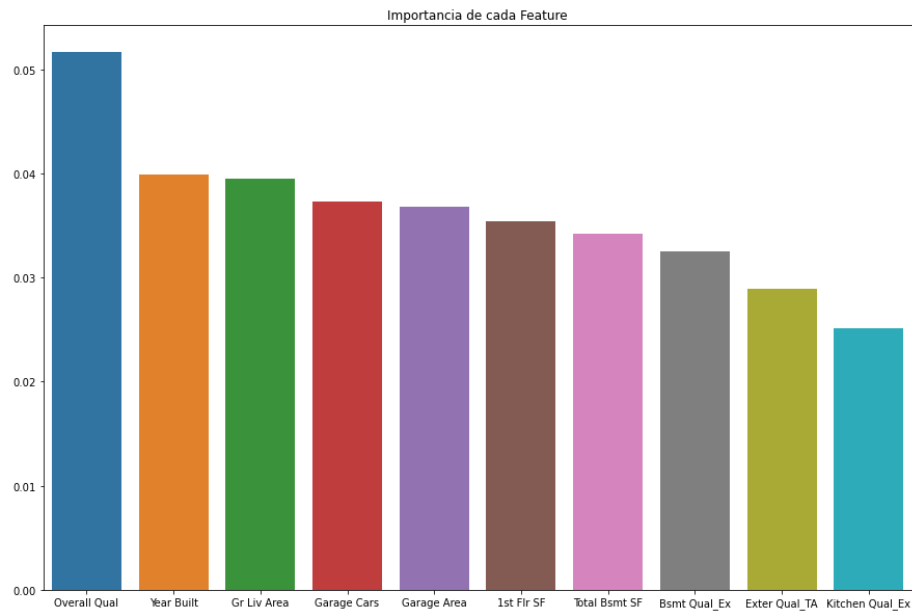


Se observa claramente que de nuevo la calidad general es la variable o feature más importante, y por un amplio margen en cuanto al modelo predictivo tipo árbol observamos también que la calidad general es una variable determinista a la hora de calcular el precio de la propiedad.



Para asegurar nuestras conclusiones anteriormente observadas aplicaremos a nuestro modelo el algoritmo de Random Forest el cual aplica modelos de regresión y clasificatorio aleatorios con el fin de asegurar conclusiones variando métodos de estudio

|             | Overall Qual | Year Built | Gr Liv Area | Garage Cars | Garage Area | 1st Flr SF | Total Bsmt SF | Bsmt Qual_Ex | Exter Qual_TA | Kitchen Qual_Ex |
|-------------|--------------|------------|-------------|-------------|-------------|------------|---------------|--------------|---------------|-----------------|
| Importancia | 0.051639     | 0.039855   | 0.039536    | 0.037329    | 0.036772    | 0.035446   | 0.034249      | 0.032515     | 0.02889       | 0.025081        |



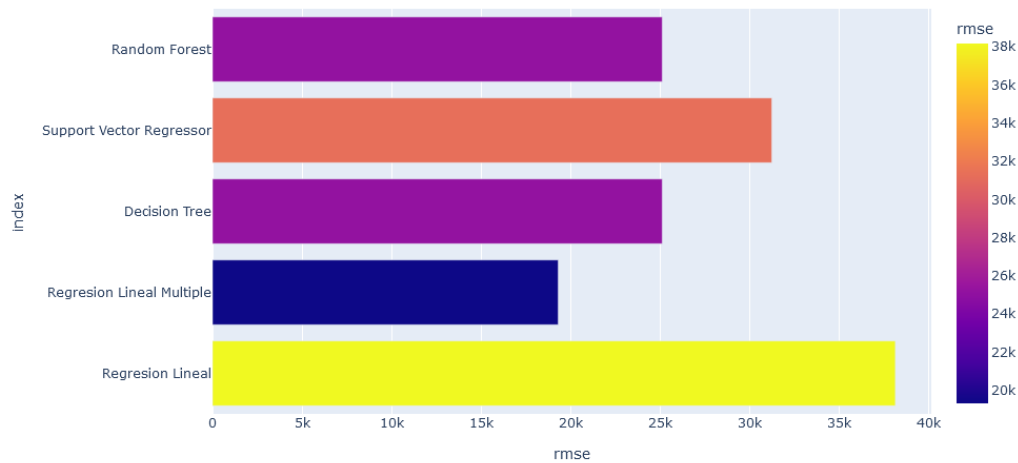
Se observa que la tendencia es similar a los modelos antes aplicados. Calidad general (Overall Quality) sigue siendo la variable más importante sin importar el modelo, la única diferencia es que este caso el resto de las variables tienen un incremento en importancia lo cual es lógico ya que la calidad general es una aglomeración de las otras variables que otorgan valor a una propiedad, el conjunto de features generalmente impactan más que los mismos variables por separado.

## Métricas de efectividad para los distintos algoritmos aplicados en el modelo.

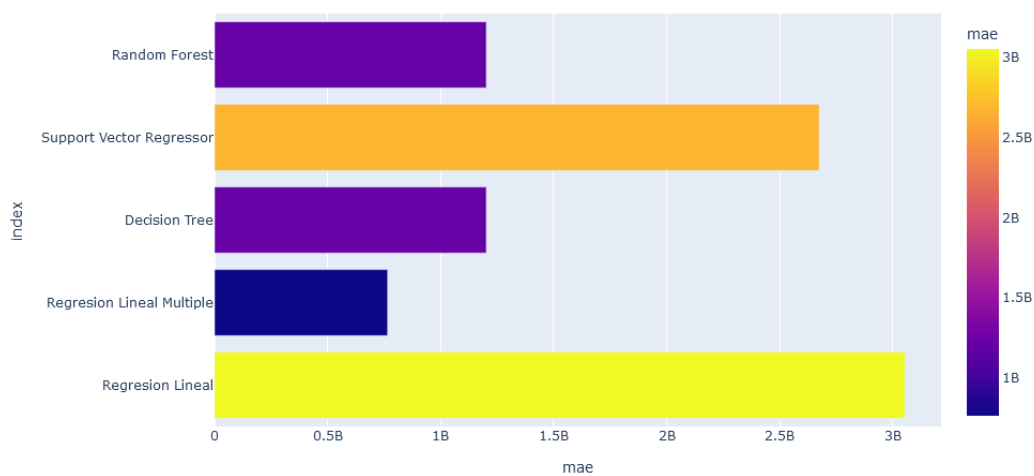
Usando las desviación de la media cuadrada podemos analizar cual de todos los algoritmos aplicados al modelo de datos tuvo una mayor cantidad de errores a la hora de predecir y por cual decantarnos a la hora de hacer los análisis concluyentes

Este gráfico nos indica que la menor precisión fue obtenida aplicando el algoritmo de regresión lineal simple.

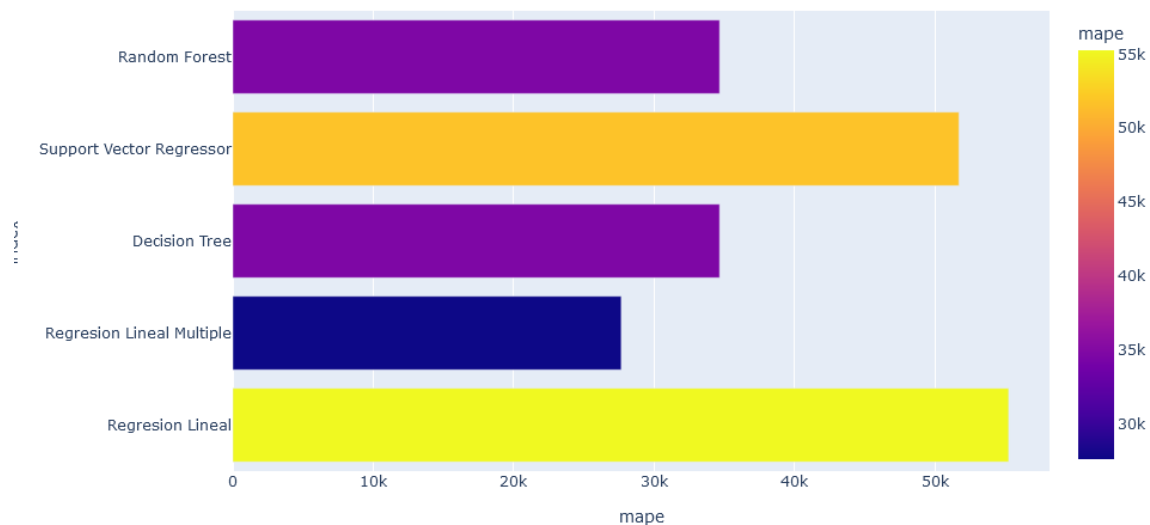
El error absoluto medio (MA



E) nos indica que algoritmo tuvo una mayor cantidad de errores entre la muestra y la predicción actual fue el de regresión lineal simple.



La media de porcentaje absoluto de error (Mape) nos dará la indicación absoluta, de cual algoritmo fue el que tuvo el mayor grado de imprecisión a la hora de hacer análisis predictivos en el modelo, y por tanto que algoritmo es el que debería evitar usarse.



Una vez asegurados que el mejor algoritmo para nuestro modelo sería el de regresión lineal múltiple se proseguirá con la aplicación de algoritmos de optimización dentro del modelo.

## Optimización de algoritmos de Machine Learning

Para lograr una buena optimización de los algoritmos se realizan múltiples pruebas buscando los mejores parámetros para cada modelo aplicado.

Una vez encontrados se utilizan diferentes herramientas para optimizar aún más los algoritmos planteados.

### Validación simple vs cruzada

Se utiliza como validación simple la conocida herramienta de `train_test_split()` de SKlearn, mientras que para la validación cruzada se utiliza `LeaveOneOut()` de SKlearn.

Los resultados obtenidos son similares en rendimiento y tiempos de ejecución por lo que el uso de una validación cruzada no tiene un efecto positivo en este algoritmo.

### Aplicación de PCA

Dado que nuestro Dataset original cuenta con 80 tipos de atributos y algunos no tienen correlación con el output. Suponemos que se genera un ruido por sobrealimentación del algoritmo. Para eso utilizamos la herramienta PCA para la reducción de variables.

Los resultados obtenidos no fueron mejores que con el Dataset original.

## Subdivisión de Dataset

Otra opción para reducir los atributos del Dataset es subdividir de manera manual los atributos con mejor correlación.

Por lo que se seleccionan las 10 variables mejor correlación respecto al output y se ejecutan el algoritmo de prueba Random Forest.

El resultado obtenido es una leve mejora en el rendimiento, con un MAE menor respecto al Dataset original.

## Algoritmos avanzados de Boosting

### Adaboost

Dado que este algoritmo debe utilizar como base un modelo más débil, en este caso se toma Regresión Lineal Múltiple para que pueda entrenarse. Dicho modelo vuelve a entrenarse sobre sí mismo y, a través de la asignación de pesos según la tasa de error, permite generar mejores resultados. Se eligió trabajar con un total de 100 estimadores, dado que con valores más elevados, no producía mejoras significativas. De todas maneras, finalmente no se consiguieron resultados satisfactorios.

### Gradient Boosting

Se trata de un algoritmo que convierte modelos débiles en modelos fuertes a través de la función de pérdida basada en este caso aplicado entre los valores de las viviendas reales y los predichos. A través de su aplicación, se ve una notable mejora en los resultados.

### LightGBM

Este modelo trabaja sobre la mejora de gradientes utilizando algoritmos basados en árboles. El parámetro configurado para su entrenamiento fue el número de estimadores en 60, debido a que se logran resultados muy buenos en un tiempo rápido de ejecución. La particularidad que tiene este algoritmo es que crece el árbol de manera vertical, y no de manera horizontal. Esta característica hace que los rendimientos obtenidos sean notablemente mejores que otros modelos aplicados anteriormente. Una de las posibles razones es la cantidad de registros utilizados en su entrenamiento porque en datasets con pocas filas tendería a un modelo sobreajustado.

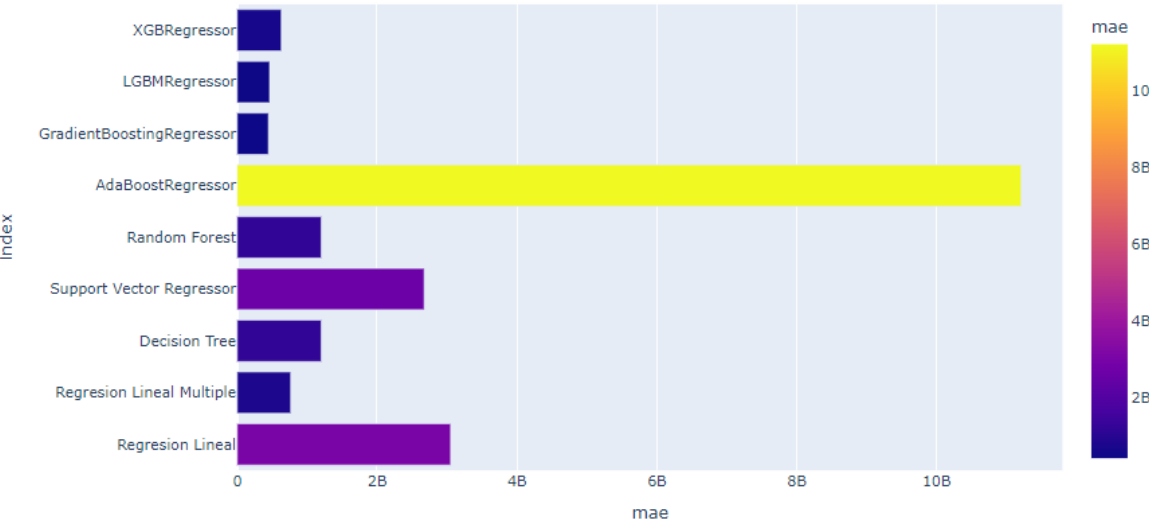
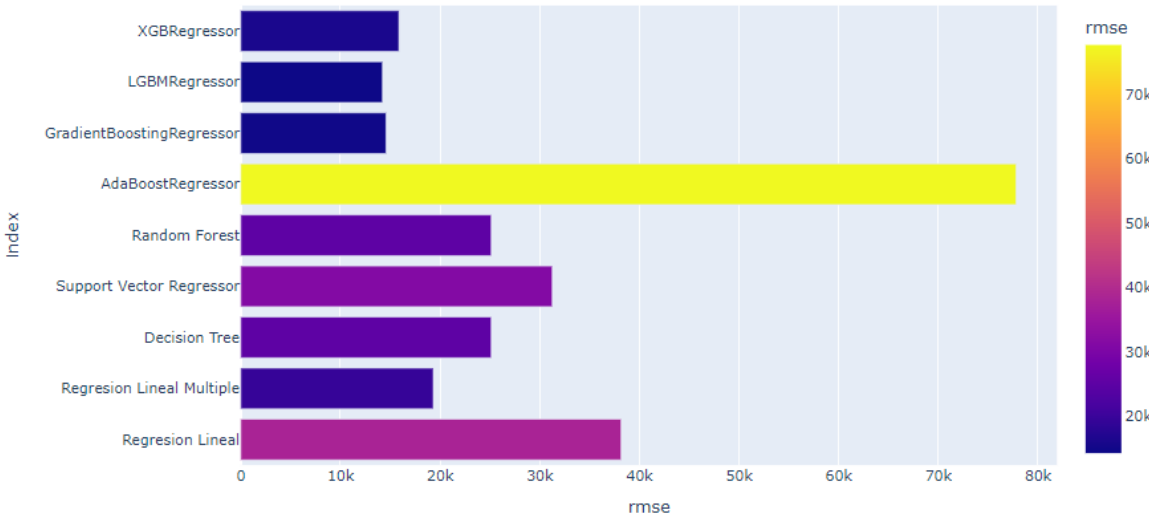
### XGBoost

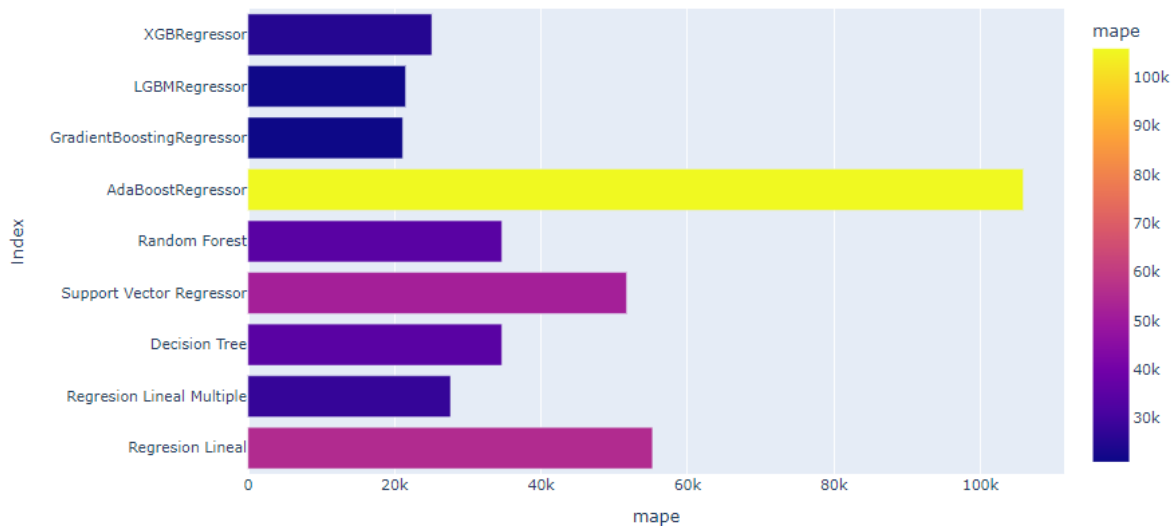
Es otra variante del modelo de Gradient Boosting y lo interesante de ella es que además de tomar como base de entrenamiento los modelos anteriores generados, cada modelo nuevo es comparado con el anterior hasta llegar a un punto en que no hay mejoras. En este caso de uso, los resultados son muy buenos, como todos los modelos de descenso de gradientes.

## Evaluación de las métricas

Se comparan métricas de todos los modelos planteados y se realizan gráficas comparativas.

|                            | rmse        | mae               | mape         |
|----------------------------|-------------|-------------------|--------------|
| Regression Lineal          | 38167.00000 | 3052735783.00000  | 55252.00000  |
| Regression Lineal Multiple | 19321.00000 | 765216326.00000   | 27663.00000  |
| Decision Tree              | 25136.00000 | 1201921164.00000  | 34669.00000  |
| Support Vector Regressor   | 31260.00000 | 2674040438.00000  | 51711.00000  |
| Random Forest              | 25136.00000 | 1201921164.00000  | 34669.00000  |
| AdaBoostRegressor          | 77809.00000 | 11211803625.00000 | 105886.00000 |
| GradientBoostingRegressor  | 14583.00000 | 446378029.00000   | 21128.00000  |
| LGBMRegressor              | 14226.00000 | 463546389.00000   | 21530.00000  |
| XGBRegressor               | 15872.00000 | 630641452.00000   | 25113.00000  |





## Conclusiones finales

Validación simple vs cruzada: Los resultados obtenidos corresponden que la validación simple es sumamente efectiva por mejor rendimiento y menor tiempo de ejecución. En ningún caso planteado la validación cruzada mejoró el rendimiento del algoritmo, en algunos casos se acercó al valor anteriormente probado.

La aplicación de una reducción de variables con PCA no resultó efectiva en nuestros algoritmos planteados. En el mejor de los casos se obtuvieron resultados similares respecto a algoritmos sin PCA. Los tiempos de ejecución utilizando PCA y Random Forest fueron excesivamente elevados, teniendo que reducir parámetros de Random Forest a valores mínimos para que el procesamiento sea menor a 20 minutos.

En cambio la subdivisión del Dataset fue un acierto para nuestro caso de estudio. Dado que nuestro Dataset original cuenta con 80 tipos de atributos y algunos no tienen correlación con el output. Suponemos que generan ruido en el resultado se seleccionan las 10 variables mejor correlacionadas con el output, y utilizando un Random Forest de prueba, se obtuvieron excelentes resultados de MAE, mejorando los obtenidos con PCA y con el Dataset original.

Respecto a los algoritmos avanzados de boosting se encontraron los errores más bajos posibles, mejorando hasta en un 25% respecto al resultado de subdivisión del Dataset. De los 4 algoritmos probados, 3 tuvieron muy buenos resultados (XGBoosting, LightGBM y GradientBoosting) mientras que 1 (AdaBoosting) resultó con errores significativamente altos. Los tiempos de ejecución de estos algoritmos fueron en todos los casos muy reducidos, menores a 3 segundos.