***Estudiante:*** *Briam Pavel Aguirre Moisés*

***CASO DE REGRESIÓN***

**Descripción del problema:**

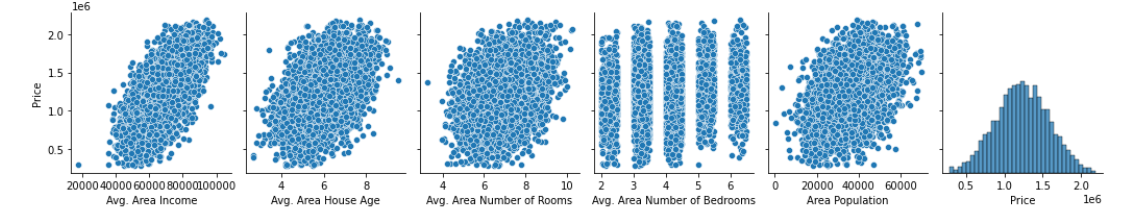
En este caso se quiere predecir el precio de una casa en base a seis variables explicativas: ingreso, antigüedad de la casa, cantidad de habitaciones, cantidad de dormitorios, población y dirección.

|  |  |  |
| --- | --- | --- |
| **Atributo** | **Significado** | **Tipo** |
| **Avg.Area\_Income** | Ingreso promedio por área | Numérica continua |
| **Avg.Area\_House\_Age** | Antigüedad promedio por área | Numérica continua |
| **Avg.Area\_Number\_of\_Rooms** | Habitaciones promedio por área | Numérica continua |
| **Avg.Area\_Number\_of\_Bedrooms** | Dormitorios promedio por área | Numérica continua |
| **Area\_Population** | Población del área | Numérica continua |
| **Address** | Dirección | Categórica nominal |
| **Price** | Precio | Numérica continua |

**Caracterización del Dataset:**

**Análisis Exploratorio de Datos (EDA):**

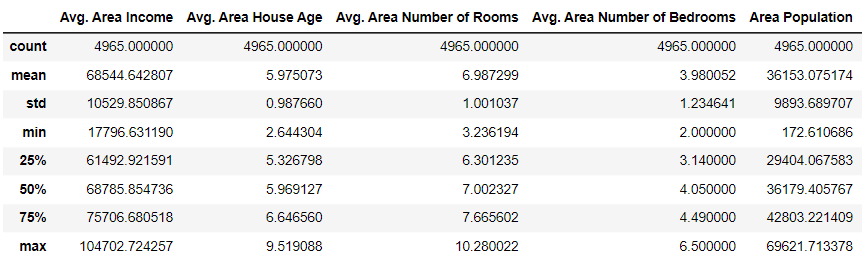
1. La variable Price tiene 35 valores atípicos, entonces se decidió eliminarlas debido a la cantidad ínfima.
2. Las correlaciones entre variables independientes no supera el 70%; por lo tanto, consideramos a todas para ajustar el modelo de regresión lineal.
3. La variable que más explica el precio de las casas es el ingreso promedio (63%) y el que menos es la cantidad de dormitorios promedio (17%).
4. En la gráfica podemos observar que el target Price tiene una posible relación lineal con las demás variables a excepción del número de dormitorios.

****

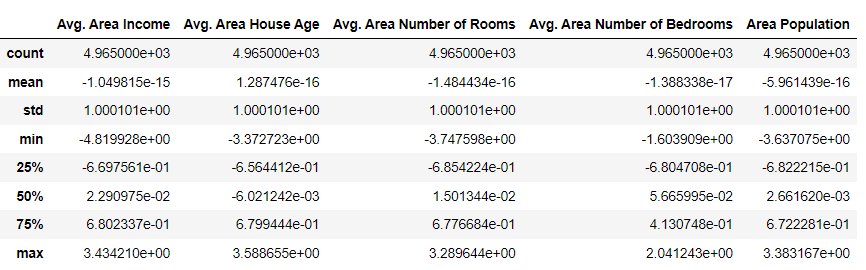
**Preprocesamiento del dataset:**

Debido a que las variables están en diferentes escalas, se estandarizaron.

***Data no estandarizada***



***Data estandarizada***



**Evaluación de los modelos:**

Los dos modelos que se ajustaron son la regresión lineal múltiple y la red neuronal con una capa de entrada (64 neuronas y función de activación relu), dos capas intermedias (32 y 16 neuronas con función de activación relu) y una capa de salida (1 neurona con función de activación lineal)

Para la predicción del precio de la casa es mejor ajustar el modelo de regresión lineal ya que las métricas de RMSE y MAE son menores que el modelo de red neuronal.

|  |  |  |
| --- | --- | --- |
| **Modelo** | **RMSE** | **MAE** |
| Regresión Lineal | 4930560210,35 | 79840,81 |
| Red Neuronal | 5155031496,97 | 81452.35 |

**Discusión de los resultados obtenidos:**

* Se obtuvo menos error en predicción con el modelo de regresión lineal que con la red neuronal. Por lo tanto, el mejor modelo resultó ser el de regresión lineal.
* Se podría mejorar la red neuronal incrementando las capas intermedias o incrementando el número de neuronas por capa.

***CASO DE CLASIFICACIÓN***

**Descripción del problema:**

En este caso el equipo de ventas de una empresa de automóviles ha clasificado a sus clientes en cinco segmentos. La empresa quiere ingresar nuevos productos al mercado y en base a una investigación han deducido que el mercado se mantiene. Han identificado 2627 nuevos clientes potenciales y quieren clasificarlos.

**Caracterización del Dataset:**

|  |  |  |
| --- | --- | --- |
| **Atributo** | **Significado** | **Tipo** |
| ID | Código único | Categórica nominal |
| Gender | Género | Categórica nominal |
| Ever\_Married | Estado Matrimonial | Categórica nominal |
| Age | Edad (en años) | Numérica discreta |
| Graduated | Graduado o no | Categórica nominal |
| Profession | Profesión | Categórica nominal |
| **Work Experience** | Experiencia trabajando (en años) | Numérica discreta |
| **Spending Score** | Puntaje de gasto | Categórica ordinal |
| **Family Size** | Cantidad de miembros en la familia | Numérica discreta |
| **Var\_1** | Categoría anónima | Categórica nominal |
| **Segmentation** | Segmentación | Categórica ordinal |

**Análisis Exploratorio de Datos (EDA):**

1. No se han encontrado instancias duplicadas.
2. Las variables Ever\_Married, Graduated, Profession, Work\_Experience y Family\_Size presentan valores missings.
3. Se tienen variables de tipo númericas que deberán ser discretizadas.
4. Se tienen variables categóricas nominales que deberán ser dumizadas y en el caso de las variables categóricas ordinales, utilizar el método de One Hot Encoder.
5. No se tiene un problema de desbalanceo de data, todas las clases están debidamente representadas.

**Preprocesamiento del dataset:**

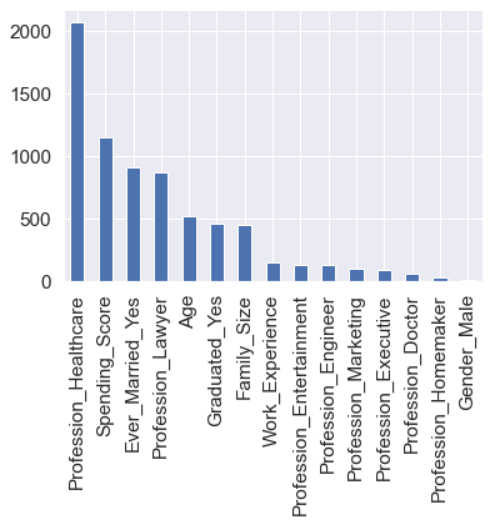
1. Se imputaron los valores correspondientes a la moda y mediana por segmento en cada variable con valores missings.
2. Discretizamos a las variables numéricas de la siguiente manera.

|  |  |
| --- | --- |
| ***Work Experience*** |  |
| [0-2] | Junior |
| <2 – 6] | Semi-Senior |
| <6 – +> | Senior |

|  |  |
| --- | --- |
| ***Family Size*** |  |
| [0-3] | Pequeña |
| <3 – 6] | Mediana |
| <6 – +> | Grande |

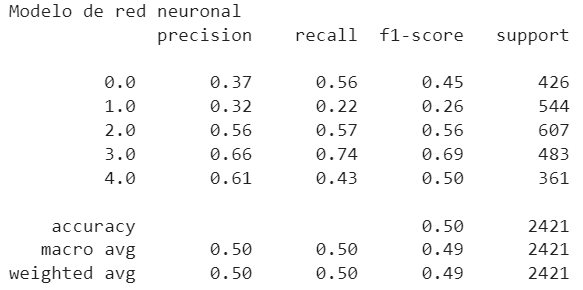
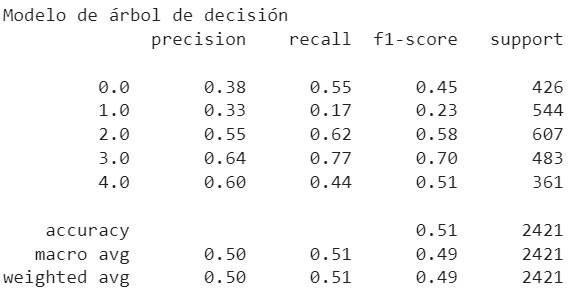
|  |  |
| --- | --- |
| *Age* |  |
| [17 – 18] | Adolescencia |
| <18 – 26] | Juventud |
| <26 – 59] | Adultez |
| <59 - + > | Vejez |

1. Se han dumizado las variables categóricas nominales y se utilizó el método de One Hot Encoder para las variables categóricas ordinales.
2. Se calculó la importancia de variables. En este caso, se decidió utilizar las siete variables más importantes debido a una diferencia significativa con las demás.



**Evaluación de los modelos:**

Los dos modelos que se ajustaron son el árbol de decisión y la red neuronal con una capa de entrada (16 neuronas y función de activación relu), dos capas intermedias (8 y 4 neuronas con función de activación relu) y una capa de salida (5 neuronas con función de activación lineal)

****  ****

**Discusión de los resultados obtenidos:**

* Los modelos de árbol de decisión y red neuronal tienen un poder predictivo muy parecido ya que tienen accuracy, precisión y recall muy cercanos.
* Ambos modelos son muy malos en predicción (No recomendable su uso) ya que tan solo aciertan el 50% de los casos, que prácticamente es lo mismo que lanzar una moneda.
* Los resultados obtenidos se pueden mejorar con una data más grande; es decir, tener más información ya que con la que se tiene no se pueden elaborar reglas que discrimen de mejor manera***.***