Universidad Andrés Bello

Facultad de Ingeniería

27-11-2022

# Análisis de Datos

Diamantes

## MINERIA DE DATOS

Profesor: Jean Paul Maidana Gonzalez

Integrantes:

Benjamín Suárez

Yerko Farfán

Angelo Cancino

Joaquín Mujica

# Índice

## Contenido

Índice ..... 1
Índice de Tablas ..... 1
Introducción ..... 2
Desarrollo ..... 3
Tiempos de procesamiento ..... 6
Predicciones ..... 7
Conclusión ..... 9

## Índice de Tablas

Figura 1 K-vecinos más cercanos ..... **¡Error! Marcador no definido.**

Tabla 1 Tiempos de procesamiento 6

Gráfico 1 Árbol de Decisión **¡Error! Marcador no definido.**
Gráfico 2 Modelo de red neuronal **¡Error! Marcador no definido.**
Gráfico 3 Modelo de regresión lineal múltiple **¡Error! Marcador no definido.**
Gráfico 4 MAE, RMSE y R^2 6
Gráfico 5 Predicción de la Red Neuronal 7
Gráfico 6 Predicción de la regresión Lineal Múltiple 7
Gráfico 7 Predicción KNN 8

# Introducción

El desarrollo de este trabajo fue motivado por la predicción del precio de diamantes, de los cuales tenemos una base de datos con alrededor de 54000 entradas, cada una de las cuales contiene información importante de estos diamantes, tales como sus medidas, proporción de sus medidas, color, claridad, etc. Durante este informe utilizaremos una aproximación al problema con métodos de aprendizaje supervisado, es decir, utilizamos una fracción de toda la información que tenemos disponible y con esta, hacemos que el ordenador pueda encontrar relaciones entre los datos otorgados y el precio de cada diamante, para luego, comprobar con la fracción de datos restante cuán acertadas y confiables son estas relaciones encontradas. Estos métodos son útiles cuando tenemos el poder de procesamiento apropiado para analizar todas estas variables, debido a que se consume una gran cantidad de este cuando tenemos muchas variables y entradas. A pesar de esto, estos algoritmos son tremendamente útiles, identifican patrones y relaciones de formas analíticas que no serían posibles para un ser humano, o quizás, podríamos decir que replican el proceso de pensamiento humano, al analizar conjuntamente muchas variables y emitiendo un juicio al respecto.

Primero que todo, tenemos que identificar que la variable que queremos predecir es continua, por lo cual los modelos están divididos en 2 grupos, El primer grupo que tenemos identificado los modelos pertenecen al modelo supervisado que son regresión lineal múltiple, árbol de decisión, redes neuronales, y el segundo grupo es el modelo no supervisado k-vecinos más cercanos.

Para el caso del grupo de modelo supervisado, intentamos crear modelos que pudieran predecir los precios basados en las variables que tenemos en nuestra base de datos, eliminando también algunas que creemos que no son las mejores para la predicción, en cambio utilizando el modelo no supervisado de los k-vecinos más cercano identifica patrones que tengan en común los datos para poder clasificarlos y poder predecir el precio.

Por último, considerando el alto volumen de datos y atributos que tenemos, utilizaremos un subconjunto de datos de 5000 entradas, debido a que la capacidad de procesamiento a la cual teníamos acceso no es la ideal para trabajar con tantos datos, junto al aumento de imprecisiones al ser más grande el tamaño de datos, lo que empeoraría nuestras predicciones.

# Desarrollo

## Modelo de regresión lineal múltiple

En este modelo, utilizamos dos iteraciones, debido a que nos dimos cuenta que la escala en la que se presentaban algunos datos era muy grande, bordeando los miles, en oposición a la escala de algunos datos que estaban alrededor de 10, por lo que nuestra segunda iteración fue realizada con una normalización logarítmica.

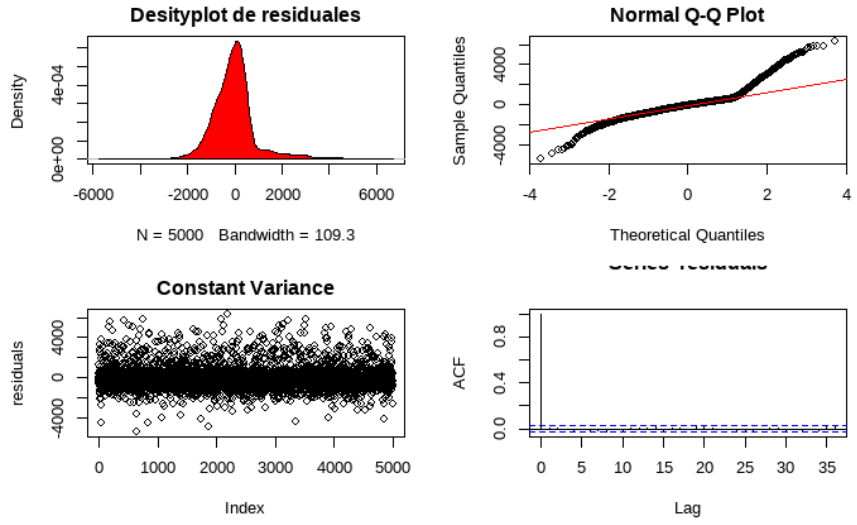


*Gráfico 1 Primer modelo de regresión*

En el siguiente modelo se observa una clara mejoría en la dispersión de los datos, el densityplot mejora y se asemeja más a una distribución normal y el gráfico de la varianza se acumula completamente cerca del 0, salvo por unos pocos outliers.

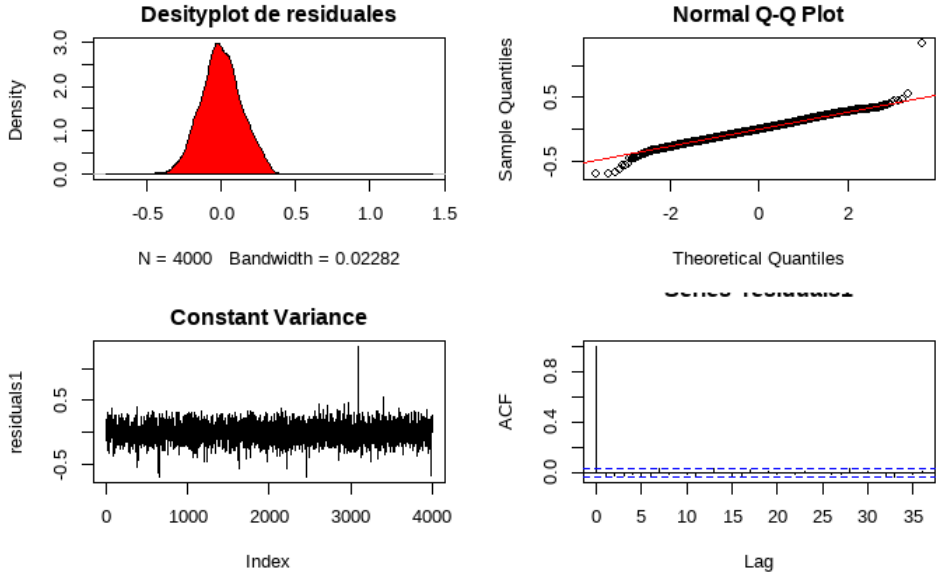


*Gráfico 2 Modelo final de regresión*

## Red neuronal

Este modelo se utilizó con dos capas ocultas debido a que con mayores cantidades simplemente no se ejecutaba correctamente. Este modelo fue el que mejor se comportó con respecto a la medida de R^2, pero también fue el que más tiempo consumió, por lo que hay que notar la relación entre eficiencia y eficacia, pero este análisis se realizará más adelante.

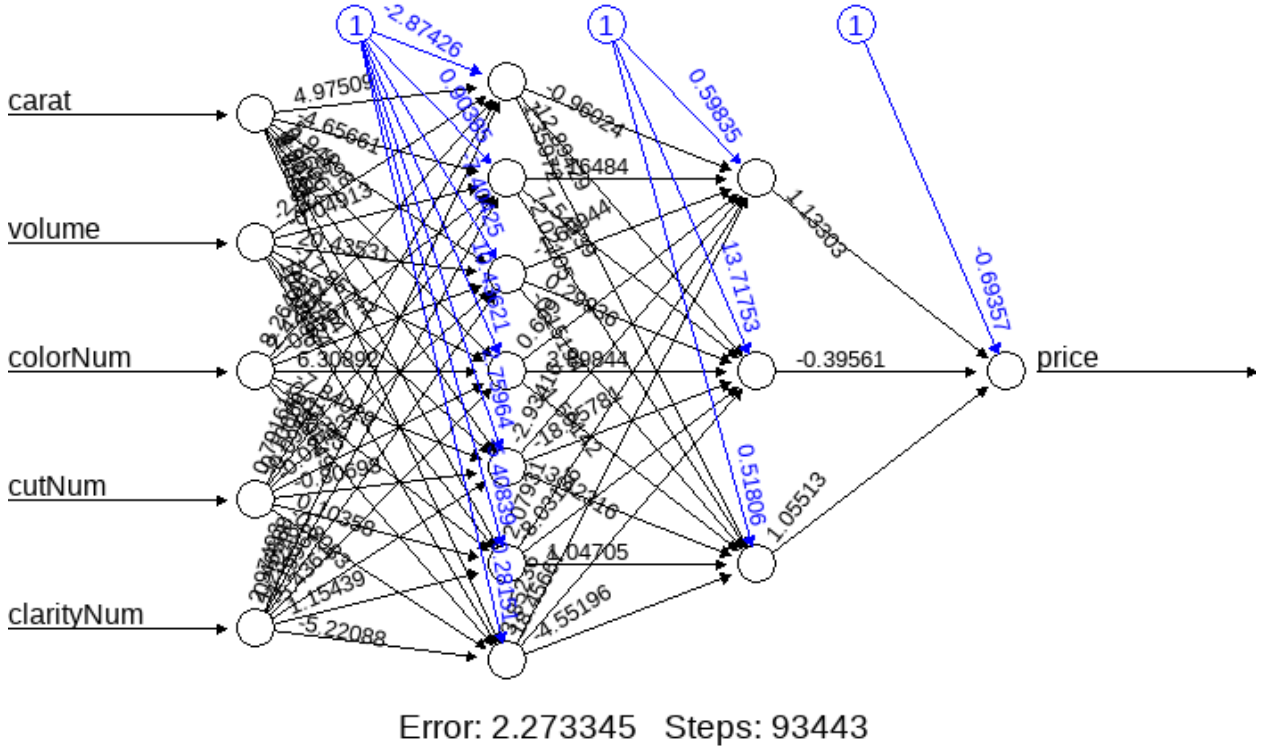


*Gráfico 3 Modelo de red Neuronal*

## Árbol de Decisión

En este caso, se utilizó un modelo de árbol de decisión, el cual encontró que la profundidad máxima óptima era de 3. Estos datos fueron preprocesados por la librería caret, para poder obtener mejores resultados.

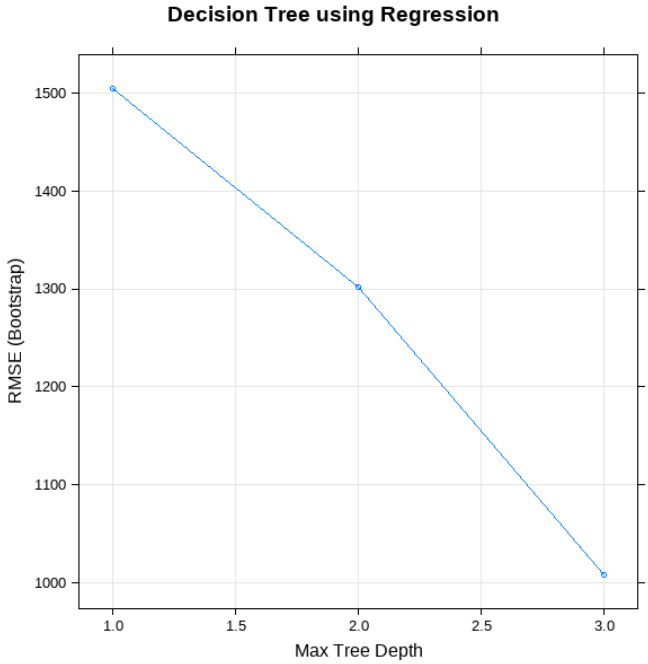


*Gráfico 4 Modelo de regresión árbol de decisión*

## K-Vecinos más cercanos

Este método es el único de aprendizaje no supervisado, este fue uno de los que tuvo un mejor desempeño, a pesar de ser poco intensivo su consumo de recursos. El modelo óptimo que fue encontrado fue de k=9.

```
k-Nearest Neighbors

5000 samples
   5 predictor

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 5000, 5000, 5000, 5000, 5000, 5000, ...
Resampling results across tuning parameters:

  k  RMSE      Rsquared   MAE
  5  639.8008  0.9444321  354.8787
  7  631.0933  0.9458929  348.7853
  9  628.8260  0.9463327  346.0035

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was k = 9.
```

*Figura 1 Modelo de K-NN*

## Tiempos de procesamiento

Respecto al método que más tiempo tomó, este corresponde a la red neuronal, este es el modelo que a cuanto de tiempo de procesamiento se refiere, es el que más tiempo toma probablemente debido a la densidad de datos a procesar como resultado de la compleja combinación entre los nodos entre las capas de la misma red.

| | Intel i7-8750H | AMD Ryzen 5 3550H | Intel i5-5200U | Intel(R) Core(TM) i5-10400F |
|---|---|---|---|---|
| Regresion | 0.352 secs | 0.490s | 0.829s | 0.315846s |
| Red | 2.188405 mins | 2.139m | 1.529m | 1.431405m |
| arbol | 0.648011 secs | 1.065s | 2.892s | 0.5575409 |
| KNN | 1.237999 secs | 1.185s | 2.663s | 0.838016s |

*Tabla 1 Tiempos de procesamiento*

Acá podemos comparar los métodos implementados, como podemos ver MAE y RMSE, tienen valores altos en lo que viene siendo dt y lm, en RMSE también nos dan valores altos en knn y nn, el R Squared, todos los valores son cercanos al 1.

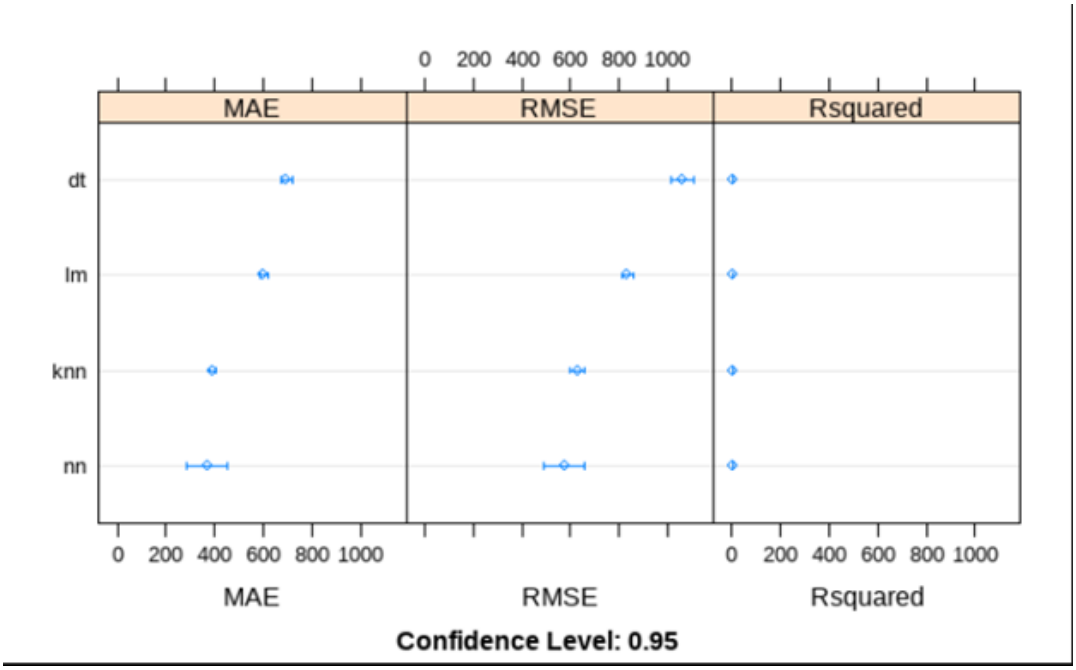


*Gráfico 5 MAE, RMSE y R^2*

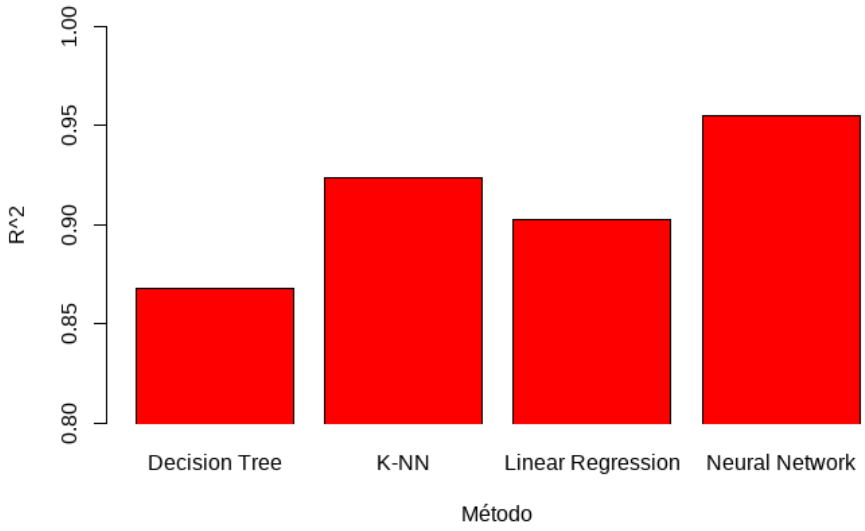


*Gráfico 6 Valores de R^2*

Aquí se puede observar claramente la precisión de cada uno de estos modelos utilizando el R^2, debido a que esta es la mejor medida estadística para medir en nuestro caso. Podemos observar que a pesar de que la red neuronal tuvo el tiempo de ejecución más alto, fue el modelo más preciso, junto a esto, podemos notar que los k-nn fueron muy precisos a pesar de ser marginalmente más demorosos. Tomando en cuenta este comportamiento, la regresión lineal múltiple y el árbol de decisión no serían nuestros modelos preferidos. Considerando estos datos, habría que pensar en equilibrar los recursos que tenemos a nuestra disposición, tales como el tiempo y poder de procesamiento, y cuánta precisión necesitamos de nuestro modelo. Pero sin lugar a dudas, los más destacados serían el K-NN y la red neuronal, a pesar de ser más complejos de entrenar.

## Predicciones

Como buscamos predecir la variable price si comparamos todos los modelos que desarrollamos a lo largo del trabajo los 2 que nos llegarán a servir son la regresion lineal multiple y el KNN, esto ya que se acercan más a nuestra franja roja, por otro lado, la red neuronal nos ayuda a visualizar que los precios son o muy altos o bajos, y se encuentran poco en el medio.

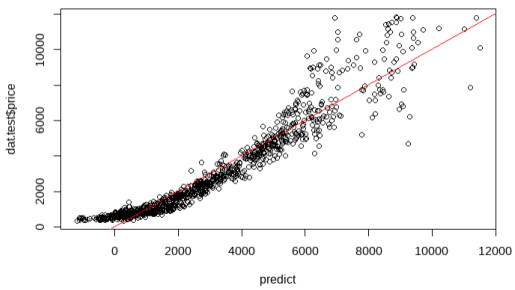


*Gráfico 8 Predicción de la regresión Lineal Múltiple*

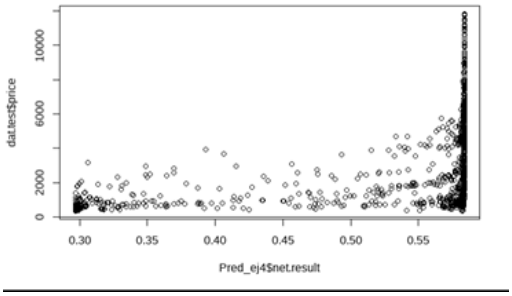


*Gráfico 7 Predicción de la Red Neuronal*

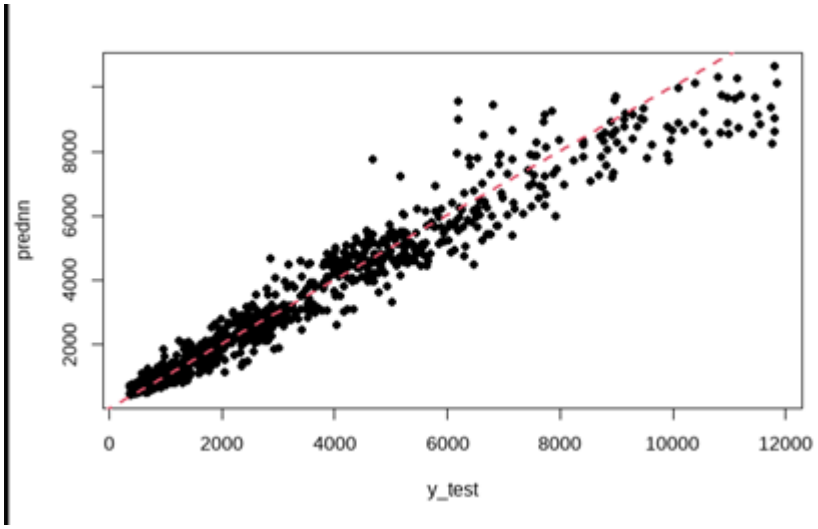


*Gráfico 9 Predicción KNN*

## Conclusión

Podemos concluir que en los tiempos de ejecución sin importar que, la red neuronal siempre es la que más se va a tardar, esto ya que, como tenemos que predecir una variable continua y tenemos una gran cantidad de datos, esto hace que el proceso se dificulte y tarde mucho más, por otra parte, el o los mejores modelos para poder predecir nuestra variable continua serían el de regresion lineal multiple y Knn, esto lo podemos ver en la parte de las predicciones.