# Implementacion Machine Learning

*DataSet Diamonds

1st J. Jurado
*Ciencias de la Computación e IA*
*Universidad Sergio Arboleda*
Bogotá D.C., Colombia
juan.jurado01@correo.usa.edu.co

## I. Contexto

A lo largo del tiempo, la toma de decisiones siempre ha sido algo de vital importancia en el futuro lejano o cercano para las personas o compañias a nivel mundial, ante esta necesidad surge para el año 2004 de la mano de Google con un nuevo paradigma de procesamiento llamado ”Google Fyle System”, una nueva rama de la inteligencia artifical llamada: ”Machine Learning”, esta nos ofrece la posibilidad de que un sistema aprenda de los datos que le son suministrados mediante la programacion explicita, esto es una gran ayuda para las compañias principalmente, debido a que sus decisiones son basadas en recopilaciones de datos que en su mayoria son muy grandes, asi que mediante ”Manchine Learning” podemos tomar todos estos datos mediante modelos predictivos, con ello lo que realizaremos es dividirlos en dos grupos que son entrenamiento ”Train” y prueba ”Test”, para finalmente poder predecir el valor del dato que desea la compañia y en base a este valor ellos puedan tomar decisiones para la compañia.
Para este trabajo utilizamos el DataSet (conjunto grande de datos) llamado ”Diamonds.csv”, con estos datos crearemos el grupo mencionado anteriormente ”Train”, aplicaremos un analisis exploratorio de los datos, realizaremos una matriz de correlacion para verificar la afinadad que tienen los datos con el dato que deseamos predecir, para esta oportunidad el valor seria el precio ”price”, seleccionamos el modelo de Machine Learning que deseamos para nuestros datos, los entrenamos y mostramos lo datos predictivos que nos genera la programacion, despues hacemos todo esto con el grupo de datos ”Test”, que seran los valores mas importantes en la prediccion, cuando tengamos estos datos, si queremos tener seguridad de que lo que estamos mostrandole a la compañia este desarrollado de buena manera, aplicamos una metrica de rendimiento ”R2” y con esto tendremos un valor generado entre 0 y 1, donde entre mas cerca nos de el valor a 1 mejor sera calificado nuestro modelo predictivo y mas confiabilidad tendra este, con esto ya podremos generar un analisis y conclusion final de los resultados predictivos obtenidos y comunicarle a la compañia la desicion mas factible para ellos segun nuestro modelos y su DataSet.
Ademas en el desarrollo de este trabajo, ademas de desarrollarlo con Sklearn, utilizamos un modelo con Qt, esto para tener una comparativa entre estos dos modelos y poder ver si existe algun cambio significativo en los valores de las predicciones o si los dos modelos son aplicados de manera exitosa y tenemos los mismo resultados en ambos

***Index Terms*—Machine Learning, Regresion Lineal, DataSet, Train, Test, Toma de Desiciones, Modelos Predictivos, Qt**

## II. Modelos

Los modelos predictivos para Machine Learning se pueden clasificar de dos formas:

- Modelos de Clasificacion: Este modelo nos permite, como su nombre lo indica, predecir la pertenencia del dato a una clase, por ejemplo, si tratamos de clasificar de nuestros clientes ”quien es mas probable que nos abonde”, el resultado de este modelo es binario, ya que es un si o un no (1 o 0) con su grado de probabilidad, entonces una salida que podriamos esperar para este ejemplo seria que ”un cliente nos abandonaria con una probabilidad del 0.89”.

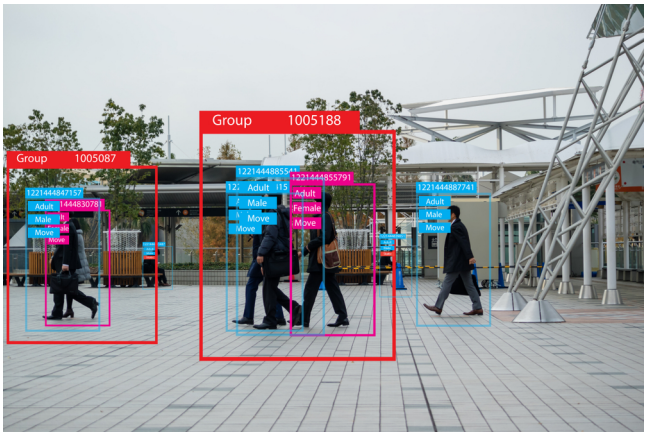



Fig. 1. Representacion grafica de un modelo de clasificacion

- Modelos de Regresion: Este modelo en cambio nos permiten predecir un valor, un ejemplo de este modelo seria, ”cual es el beneficio estimado que obtendremos de un nuevo cliente en la compañia”, una salida que podriamos esperar seria, ”el nuevo cliente nos generara una ganancia total de 100.580 pesos”.

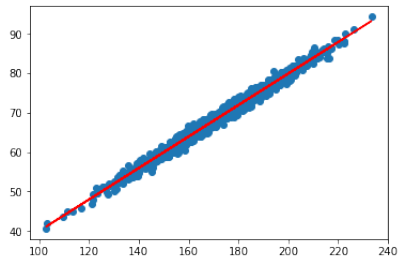



Fig. 2. Representacion grafica de una regresion lineal

Teniendo en cuenta la informacion anterior, podemos ver que el modelo predictivo utilizado en nuestro desarrollo fue el Modelo de Regresion, debido a que este nos permite retornar valores para la variable a predecir.

## III. Metrica

Para el desarrollo de nuestro modelo de prediccion es siempre importante saber si nuestro resultados tienen una confiabilidad baja o alta, esto debido a que para las compañias sera mucho mejor saber el nivel de credibilidad de nuestro modelo, es por ello que se implemento la metrica de rendimiento ”R2”. R2 muestra qué tan bien los términos o datos se ajustan a una curva o línea (Regresion Lineal). El R2 ajustado también indica qué tan bien se ajustan los términos a una curva o línea, pero se ajusta para la cantidad de términos en un modelo. Si agrega más y más inútiles variables a un modelo, el R cuadrado ajustado disminuirá. Si agrega más variables útiles, aumentará R cuadrado ajustado. R2 ajustado siempre será menor o igual a R2, es por esto que la metrica de rendimiento es lo ultimo que le realizamos a nuestro modelo, con esto sabremos si las variables elegidas estan bien y si nuestro modelo es confiable, este valor se dara entre 0 y 1, donde entre mas cerca este del 1, mas confiabilidad o acierto tendra nuestro modelo.

## IV. EDA

EDA (Exploratory Data Analysis) es el analisis exploratorio de los datos, esto es un paso muy importante a la hora de querer manejar los datos debido a que debemos saber que tipos de datos estamos manejando, con esto tambien sabremos de que se trata la salida que nos estan pidiendo y que tipo de dato debemos entregarles, este paso es el primero que debemos realizar en la programacion, ya que con este podremos decidir que modelo de prediccion es el mas acorde a los datos que tenemos y podremos seguir trabajando de ahi en adelante con el modelo seleccionado.
En el EDA realizado para mi DataSet, se pudo observar que los datos eran numericos, donde 6 columnas de nuestro DataSet numerico son de tipo ”float” y 4 coumnas de tipo ”int”, observamos tambien que en nuestro DataSet no hay ningun campo que este vacio y por ultimo observamos la memoria utilizada que nos dio un total de 4.1 MB

```
# A continuación, se realiza el EDA (Análisis exploratorio de los datos)
# 1.- Información general del DataFrame
dfDiamonds.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 53940 entries, 0 to 53939
Data columns (total 10 columns):
 #  Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0  carat   53940 non-null  float64
 1  cut     53940 non-null  int64
 2  color   53940 non-null  int64
 3  clarity 53940 non-null  int64
 4  depth   53940 non-null  float64
 5  table   53940 non-null  float64
 6  price   53940 non-null  int64
 7  x       53940 non-null  float64
 8  y       53940 non-null  float64
 9  z       53940 non-null  float64
dtypes: float64(6), int64(4)
memory usage: 4.1 MB
```

Fig. 3. EDA realizado para el DataSet ”Diamonds”

Dentro de este tambien tenemos una parte muy importante y es poder revisar como es la afinidad de los datos, con esto podremos ubicar que datos tienen mas afinidad con el dato que queremos predecir y asi tener un mejor manejo de los datos, para ello se realiza una matriz de relacion, la cual nos da el siguiente resultado:

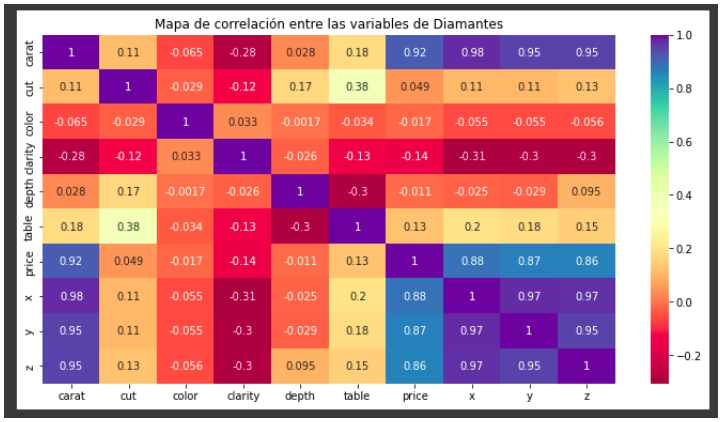



Fig. 4. Matriz de relacion de los datos del DataSet

Como podemos observar el dato que genera mas afinidad con nuestra variable dependiente ”price” es ”carat”, es por esto, que estos valores nos seran de mucha ayuda para poder lograr una prediccion acertada que tenga el mismo porcentaje de afinidad.

para finalizar, debemos observar la distribucion de los datos, con esto sabremos si debemos manejarlos de forma particular o tener algun cuidado con ellos, para esto graficamos el comportamiento de todas las variables, donde nos podemos apreciar lo siguiente:

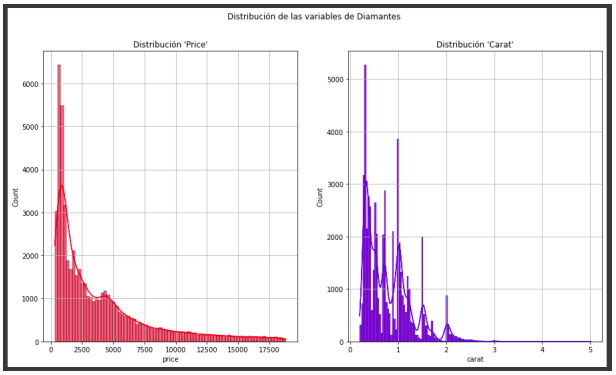



Fig. 5. Grafica distribucion de los datos de ”price” y ”carat”

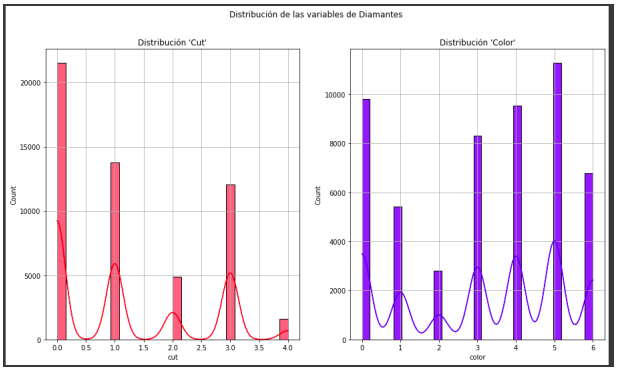



Fig. 6. Grafica distribucion de los datos de ”cut” y ”color”

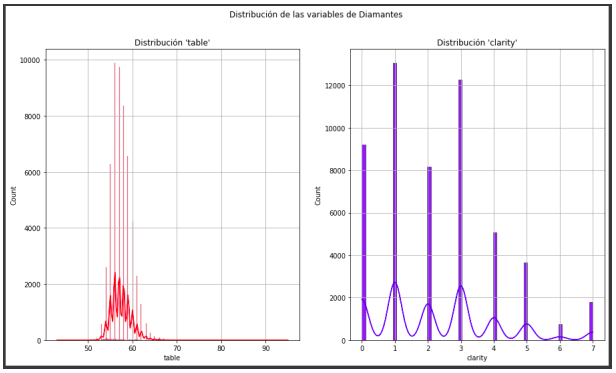



Fig. 7. Grafica distribucion de los datos de ”table” y ”clarity”

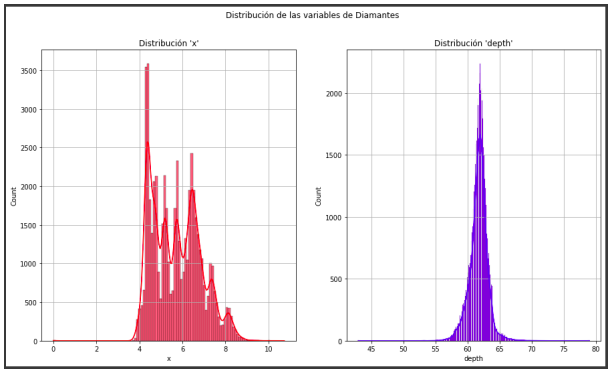



Fig. 8. Grafica distribucion de los datos de ”x” y ”depth”

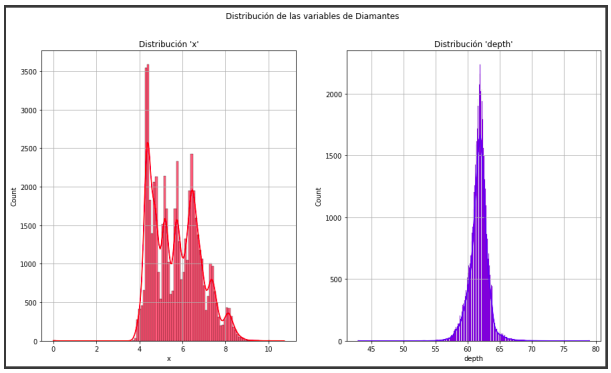



Fig. 9. Grafica distribucion de los datos de ”y” y ”z”

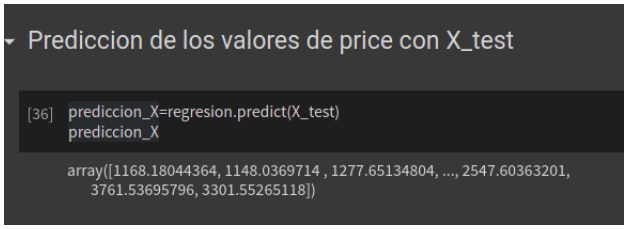
Con estas graficas podemos observar que la distribucion en todos sus datos es una distribucion normal, por lo que no debemos tener ningun tipo de cuidado o excepcion con algun dato al momento de aplicar nuestra regresion para poder llegar a nuestro valores de prediccion.

## V. Apreciaciones

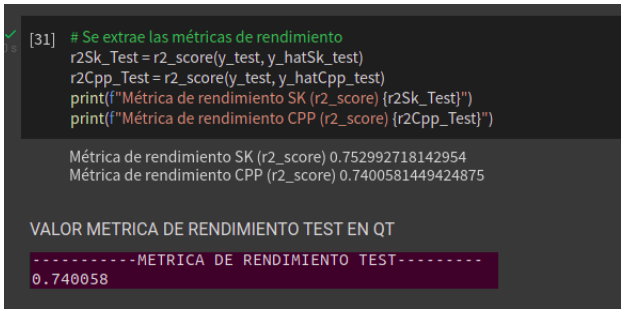
- Apreciamos mendiante el Analisis Exploratorio de Datos que el DataSet suministrado era numerico por lo cual el modelo predictivo mas oportuno para realizar nuestro trabajo era el Modelo de Regresion Lineal.
- En la matriz de correlacion pudimos apreciar que la columna que tiene mas afinidad con nuestra columna de prediccion ”price”, es ”carat” con un valor de 0.92.
- Para el manejo de los datos se debe ser muy cuidadoso, esto ya que en caso de realizar una mala practica con ellos nuestros valores predictivos no seran del todo confiables.
- En la observacion inical de los datos pudimos observar que todos ellos tenian una distribucion normal, esto lo notamos mediante graficas de los valores de las columnas seleccionadas.
- Podemos apreciar que para los datos suministrados, necesitamos la implementacion de la regresion lineal, ya que esta nos permite predecir el comportamiento de una variable dependiente ”price” a partir de una variable independiente ”carat”, en esto es realmente lo que se basa nuestro modelo de prediccion.

## VI. Toma de decisiones

Esta seccion es muy importante para la finalidad y exito de nuestro trabajo, para la toma de desiciones debemos tener muy presente primero los valores que nos arrojo la codificacion, estos valores fueron:




Despues de esto, debemos saber que tan confiable es nuestro modelo de prediccion, para ello utilizamos la metrica de rendimiento anteriormente mencionada y segun el valor que nos muestre sabremos si los valores de la prediccion son de confiar o no, este valor de la metrica es:




Con estos dos valores ya podemos hablar de si la prediccion es confiable o no, viendo la metrica de rendimiento, tenemos un valor de 0.74, este nos representa un valor bueno teniendo en cuenta la gran cantidad de datos que tenemos en el DataSet, por lo que la decision esperada por parte la compañia es que aprueben estos valores predictivos sobre el precio (”price”) y los utilicen en su compañia.

### A. Conclusiones

- Pudimos concluir que para la implementacion de nuestro modelo de prediccion el primer paso y el mas importante es la EDA (Analisis Exploratorio de los Datos), con esto podremos saber como realizar nuestra prediccion y que es lo que debemos hacer.
- Para la utilizacion de la metrica de rendimiento podemos concluir que no siempre se utiliza un modelo de regresion, sino que dependiendo de los datos, tenemos otro modelo de prediccion como lo es el modelo de clasificacio,
- Aprendi a lo largo del desarrollo que el manejo de los datos debe ser lo mas cuidadoso posible, debido a que si se realiza un cambio brusco o drastico en el DataSet, este nos arrojara valores muy poco comunes o no esperados y nuestro modelo que esta bien terminaria mostrando la metrica de rendimiento todo lo contrario.
- También se puede concluir que siempre es bueno para todo modelo de prediccion tener una metrica de rendimiento, esto nos da cierta tranquilidad y confianza en que lo que hicimos se encuentra bien y no tendremos un fallo en el futuro o una prediccion mala.
- También se puede concluir que para el modelo es muy importante separar los datos en los dos grupos ”Train” y ”Test”, debido a que se tendra un mejor manejo de los datos y ademas podremos hacer el entrenamiento del sistema para que al momento de utilizarlo con ”Test”, este sistema ya este entrenado y nos pueda dar unos valores optimos.
- También se puede concluir que tanto en el modelo implementado con Sklearn (Python) y en el modelo Cpp (Qt) los valores arrojados son los mismos, es por esto que podemos concluir que ambos modelos fueron implementados de manera exitosa para el DataSet.
- Por último, se puede decir que el ejercicio propuesto por el profesor para la evaluación del parcial de tercer corte fue muy satisfactorio para nuestra formación como ingenieros, ya que adoptamos los conceptos de Machine Learning, codificacion en Python, codificacion y manejo de Qt, entre otros, de una manera práctica y diferente a la convencional que ayuda a que el conocimiento se adopte de una mejor manera a partir de la práctica.

## VII. Bibliografía


- Machine Learning. (s. f.). IBM Colombia. https://www.ibm.com/co-es/analytics/machine-learning
- http://datascience.recursos.uoc.edu/es/preprocesamiento-de-datos-con-sklearn/
- REGRESIÓN LINEAL – Revista Chilena de Anestesia. (s. f.). Revista Chilena de Asistencia. https://revistachilenadeanestesia.cl/regresion-lineal/
- Las 11 técnicas más utilizadas en el modelado de análisis predictivos - Insight — Keyrus. (s. f.). keyrus. https://keyrus.com/sp/es/insights/las-11-tecnicas-mas-utilizadas-en-el-modelado-de-analisis-predictivos:

## VIII. Biografia

JUAN PABLO JURADO MUÑOZ Nació en Bogotá el 31 de mayo del 2000, en la clínica de la policía nacional. Es una persona alegre, algo timido, que le gusta practicar deporte, pasar tiempo con sus amigos, estar con su familia, crecer en conocimientos y ser mejor persona día tras día.