CURSO DE ANÁLISIS DE DATOS CON PYTHON 1

TRABAJO GRUPAL

INGENIERO

ROTTA SAAVEDRA, PEDRO

INTEGRANTES

ARNIE JEAN PIERE REQUENA MORAN
SEBASTIAN CASTRO PACAHUALA
ALLISON MALDONADO VARGAS
SEBASTIÁN MARCELO PÉREZ
RODRIGO ARIAS VILLANUEVA
GUILLERMO ANTONIO VARGAS MENDIZÁBAL

06 /02 /2022

INDICE

I. INTRODUCCIÓN

La tecnología ha estado en constante cambio en los últimos años, se han logrado avances que, en el siglo pasado, eran inimaginables para el hombre, con estos progresos también se han ido creando nuevas disciplinas, entre ellas es el aprendizaje automático o también llamado Machine Learning, que en simple palabras extrae conocimiento de los datos para poder determinar algoritmos que más tarde se usaran a favor para poder predecir un conjunto de datos. Este tipo de aprendizaje se ve mucho en la vida cotidiana; como las recomendaciones automáticas de YouTube, Facebook o Google; el reconocimiento facial de tus amigos en tus fotos; y aunque es un avance relativamente nuevo, los autos que se manejan solos.

El proyecto se centrará en el de la creación de un programa para el propietario de un viñedo llamado Juan Alberto quien quiere saber si el vino añejo que tiene es de calidad. Por lo general, él lo mide con un proceso bastante largo que le toma mucho tiempo. En consecuencia, busca un programa capaz de predecir si su vino es de calidad. Nosotros ofrecemos a señor Juan un programa capaz de calcular con cierta precisión si su vino es de calidad. Para ello desarrollaremos un programa de Machine Learning, el cual partiremos de una base de datos gigantesca para así poder entrenar al modelo y realizar una predicción más exacta. El programa será desarrollado con el lenguaje de programación Python uno de los lenguajes que es más usado en análisis de datos.

¿Por qué debería Juan tener este programa para su viñedo? Debido a que ofrece:

- Una mayor organización de datos subidos al programa, además de brindar seguridad al no perderlos.

- Un ahorro de tiempo bastante notable, ya que hacerlo manualmente llega a ser tedioso.

- La oportunidad de hacer informes estadísticos de la calidad de sus vinos a través del tiempo con las diferentes librerías de Python.

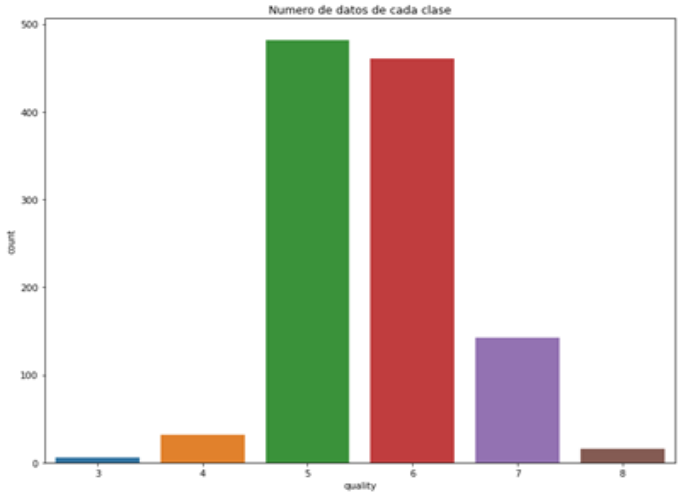
II. ANÁLISIS DEL PROBLEMA

El gran problema aquí es poder obtener la calificación del vino añejado, pero para poder obtener esta calificación, se requiere de usar distintas máquinas, personal, recursos y tiempo; así que se decidió trabajar un código que pueda estudiar distintos vinos y sus calificaciones en distintas áreas para que con solo ingresar algunas características de un nuevo vino, se pueda aproximar a una calificación que no se aleje mucho de la realidad siendo una mejora en el tiempo de evaluación y pronosticar ciertos eventos. Se obtuvieron varios problemas en la comprensión, balance y graficas de los datos.

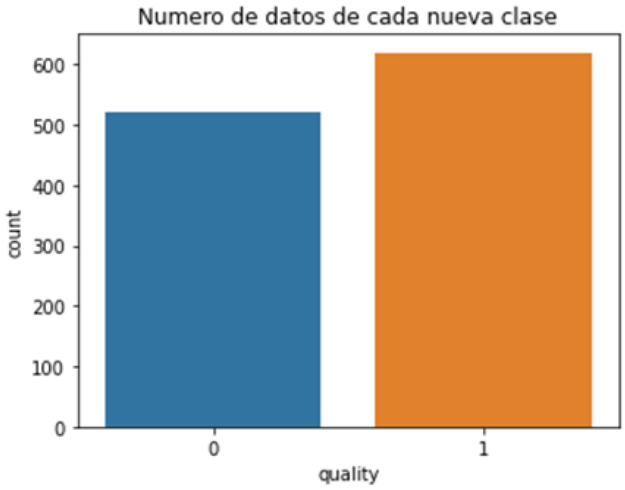
Se nos presentaron algunos problemas a la hora de testear los modelos planteados.

1. Los datos de la columna quality no están bien balanceados, como se ve en esta gráfica:

   La idea principal fue dividir la data en 3 grupos [Malo, Normal, Bueno] pero la data quedaba muy desbalanceada ya que el '5' y '6' tenían mayoría de datos. Ese problema nos produjo gran overfitting.

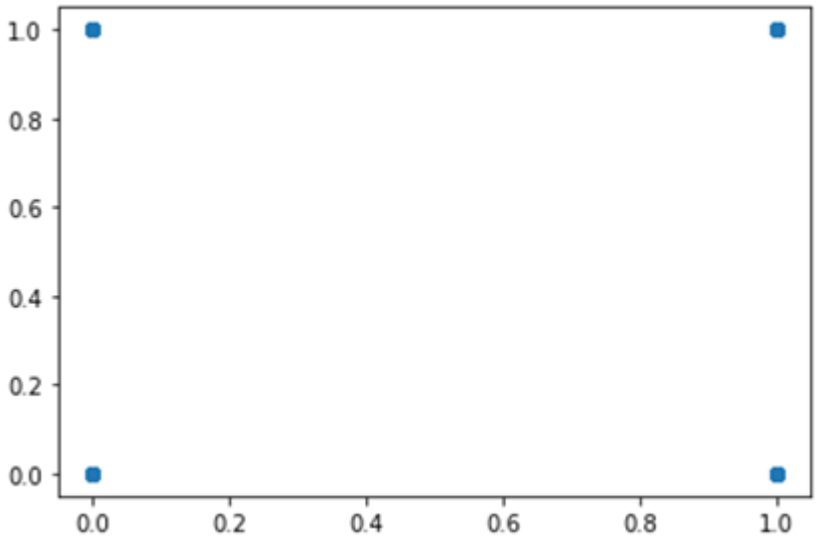


   En conclusión, la data al final se dividió en 2 grupos desde el [0,5] era el grupo del Vino malo y del <5,10] era grupo bueno. Quedando así la columna quality balanceada, el siguiente gráfico:



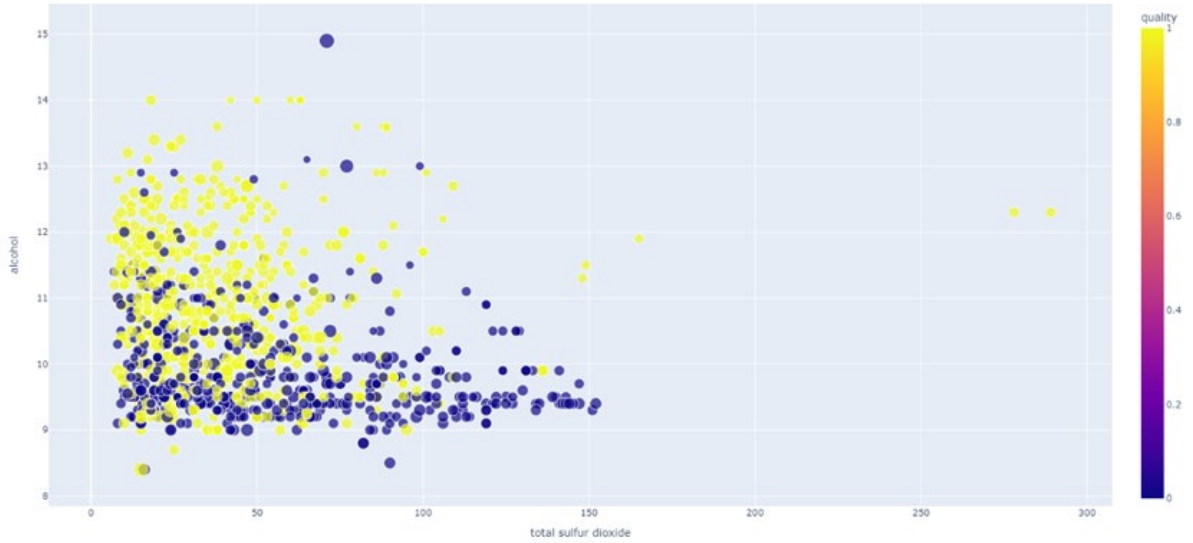
2. Problema de los gráficos.

Al dividir los ‘y’ target en 0 (‘Malo’) y 1(‘Bueno’) nos surgió un problema que no pudimos resolver. La gráfica que se hizo con el ytest, ytrain solo presentaba 2 clases. Así que quedaba una gráfica así:

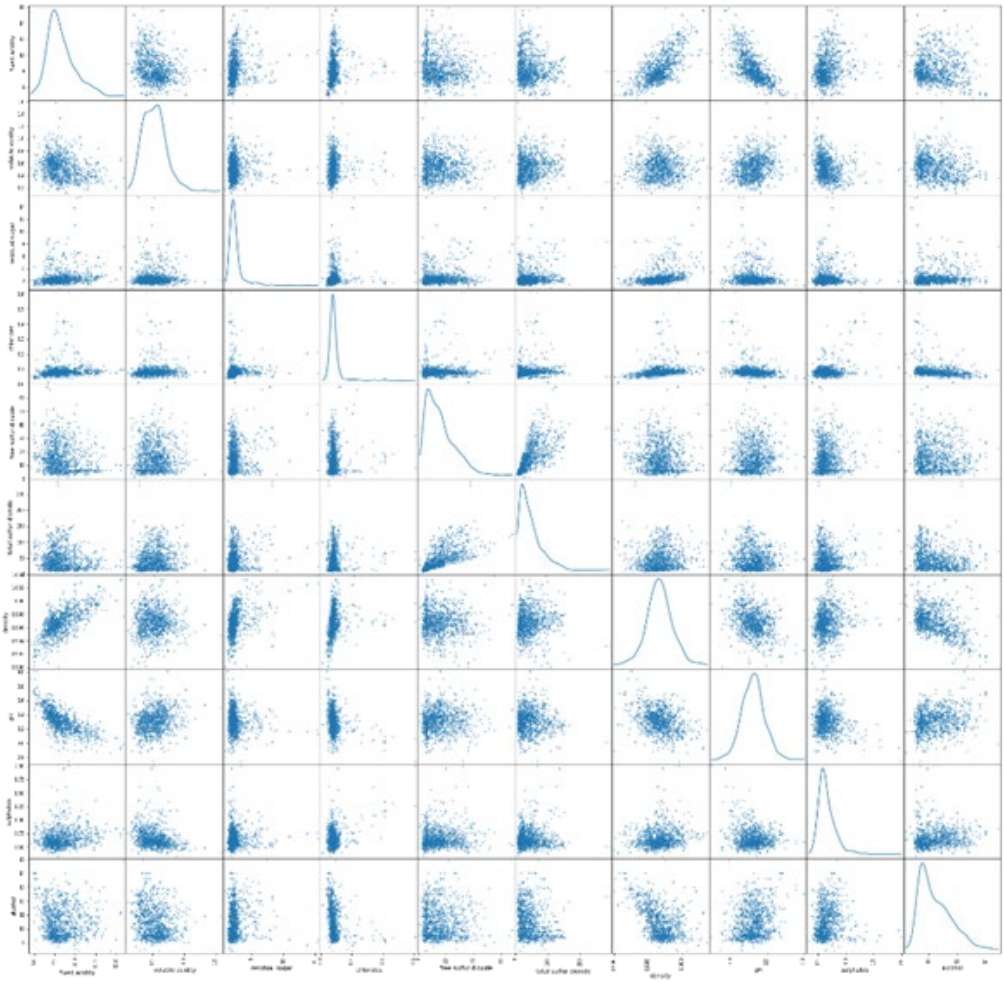


Las dimensiones en X features es muy grande para ser graficada, al tener ese problema se ha optado solo a escoger cuatro X features, los más predominantes en los datos.

- ★ Total sulfur dioxide.
- ★ Alcohol.
- ★ Fixed acidity
- ★ Quality [Por ser la respuesta del sistema]



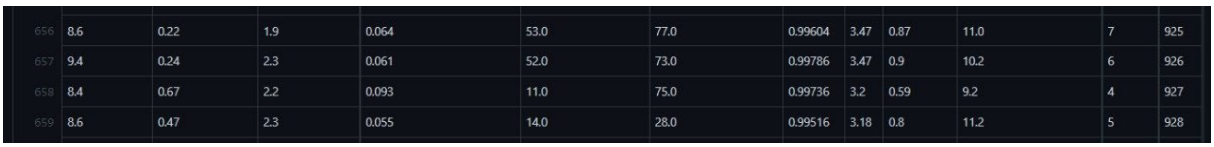
Los gráficos de correlación de datos tenían un pequeño problema, pues se tenían que sacar todos y eran once gráficos. Se resolvió con el módulo “pd.plotting.scatter_matrix” que nos sacaba todos los gráficos en una sola matriz y se podía determinar la forma de los gráficos de cada uno.



Aunque el gráfico es grande y seguro determina muchas confusiones, es el único que tenemos para poder ver como las variables se correlacionan.

III. ANÁLISIS DEL PROGRAMA

Ejecutando el código ya acabado. Tomamos ciertos valores extraídos del CSV (fila 658), los eliminamos de este para usarlos de prueba y corroborar si el programa funcionaba correctamente.



656	8.6	0.22	1.9	0.064	53.0	77.0	0.99604	3.47	0.87	11.0	7	925
657	9.4	0.24	2.3	0.061	52.0	73.0	0.99786	3.47	0.9	10.2	6	926
658	8.4	0.67	2.2	0.093	11.0	75.0	0.99736	3.2	0.59	9.2	4	927
659	8.6	0.47	2.3	0.055	14.0	28.0	0.99516	3.18	0.8	11.2	5	928

Ahora, el programa nos va a pedir datos específicos del vino de cierta cosecha; tales como:

- Fixed Acidity (acidez fija)
- Volatile Acidity (acidez volátil)
- Residual Sugar (Azúcar residual)
- Chlorides (Cloruros)
- Free sulfur dioxide (dióxido de azufre libre)
- Total sulfur dioxide (dióxido de azufre total)
- Density (densidad)
- pH (pH de la muestra)
- Sulphates (sulfatos)
- Alcohol (alcohol)

El modelo elegido fue Random Forest Classifier ya que presentaba una buena generalización de los datos. Y presentaba mucha más precisión que lo otros modelos con PCA o Kmeans + PCA.

Finalmente, según los valores colocados por el/la cliente para cada uno de los datos solicitados, nos dará el resultado si es que el vino es de buena o mala calidad cómo se logra observar en el final del programa.

```
C:\Users\LENOVO>cd Desktop

C:\Users\LENOVO\Desktop>cd PROGRAMACION

C:\Users\LENOVO\Desktop\PROGRAMACION>python Model_end.py
Bienvenido al programa de prediccion de vinos.
Ingrese "P", para evaluar si su Vino es Bueno o Malo
Enter para salir
: P
Desea ver la precision que tiene el modelo? (S/N): S
La presicion de este modelo es de: 82.895%
Ingresa el valor fixed acidity: 8.4
Ingresa el valor volatile acidity: 0.67
Ingresa el valor residual sugar: 2.2
Ingresa el valor chlorides: 0.093
Ingresa el valor free sulfur dioxide: 11.0
Ingresa el valor total sulfur dioxide: 75.0
Ingresa el valor density: 0.99736
Ingresa el valor pH: 3.2
Ingresa el valor sulphates: 0.59
Ingresa el valor alcohol: 9.2
El vino es de baja calidad

C:\Users\LENOVO\Desktop\PROGRAMACION>
```

Cabe recalcar que nuestro programa tiene un porcentaje de precisión como está señalado al inicio de su ejecución. (82.895%)

IV. CONCLUSIONES

1. Gracias a este programa le hemos facilitado al dueño del viñedo a calificar y testear sus cosechas de vino de una manera mucho más rápida y efectiva, solo anotando valores en los datos requeridos en el programa y a su vez, se ahorra tiempo y dinero que podría invertir para mejorar las cosechas.

2. Después de haber visto el código y cómo funciona, se puede dar por entendido que Machine Learning nos puede ayudar para anticipar cosas antes de que siquiera comience a realizar el proceso como en este caso es la evaluación de un buen vino añejo.

3. Con el problema inspeccionado, nos damos con la tarea de enseñarle a nuestra máquina, ideas de las características de un buen vino desde distintas perspectivas para que así, con los nuevos lotes solo ingresamos algunas calidades y características para poder pronosticar el futuro que le depara al lote que se ingresó.

4. El ahorro de recursos y tiempo lleva a una mejora bastante sustancial al momento de ser trabajado en una empresa, pero a pesar de esto, no muchas empresas usan este tipo de tecnologías, sobre todo en nuestro país, ya que no se fomenta mucho en los medios y solo los que tienen cierto interés cerca de este campo pueden acceder y conocer más de este mundo del Machine Learning.