Análisis de métodos de clasificación

Modelado de preferencias de vinos

Alexander Acosta y Julián Arango
Universidad EAFIT
Marzo de 2019

Introducción

El siguiente es un trabajo recopila un análisis realizado a partir de un dataset de ejemplo de características de vinos, el cual mide según dichos atributos la calidad de un vino en una escala de 1 a 10, este análisis busca realizar predicciones utilizando el dataset con técnicas de machine learning clásico como regresiones y árboles de decisión, analizando los diferentes hiperparametros para escoger el modelo que tenga un mejor desempeño en la predicción la calidad de los vinos ya sea binaria o en la escala numérica.

Entendimiento del problema

Los vinos pueden ser vistos actualmente como un bien de lujo, Además cuenta con una cantidad de consumidores que va en aumento, se toma como referencia [Cortez et al., 2009] quienes hicieron un estudio de la calidad de los vinos producidos en Portugal, que es uno de los mayores productores de vino.

Se busca invertir en nuevas tecnologías como podría ser el Machine Learning para lograr predecir y perfeccionar las variables tomadas en cuenta a la hora de producir un bueno de excelente calidad que los lleve a lograr la certificación de calidad, de forma que esto permite unas mejores ventas, rentabilidad de los viñedos y vinos de excelente calidad.

Relación entre variables de entrada y salida

Las variables de entrada nos dan un conjunto de 11 que son continuas, y en la salida se otorgó una clasificación binaria de la calidad, dependiendo de algunas de estas 11 variables (para este ejemplo en particular se toman las 5 de más importancia)

- Alcohol
- Volatile acidity
- Density
- Free sulfur dioxide
- Citric acid

Y en la salida se obtiene la clasificación de el vino (bueno ó malo) y una calificación cuantitativa que va de 0 a 10.

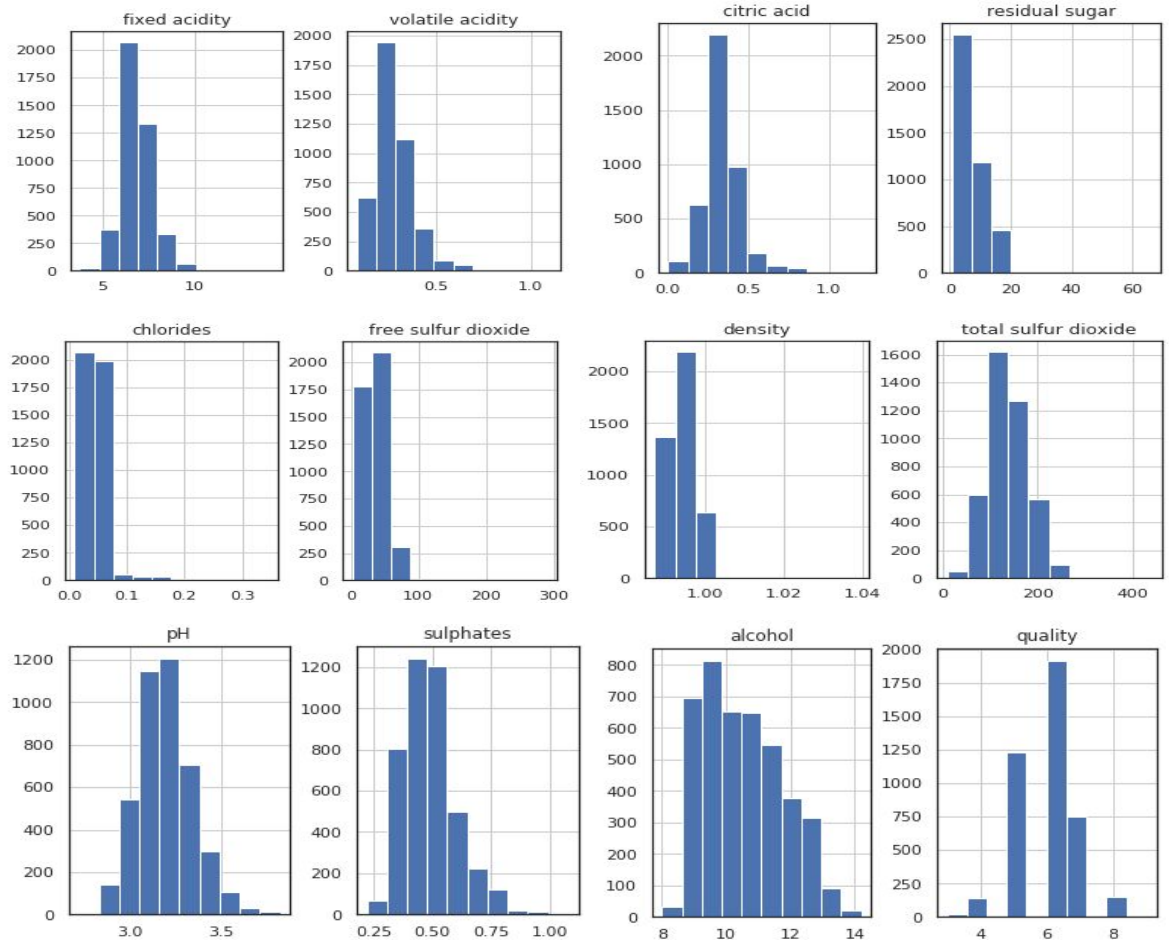
Revisión de la calidad de los datos

Table 1: Descripción de los datos

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
count	4200.000000	4200.000000	4200.000000	4200.000000	4200.000000	4200.000000	4200.000000	4200.000000	4200.000000	4200.000000	4200.000000
mean	6.856381	0.278389	0.333557	6.413357	0.045673	35.219048	137.758333	0.994024	3.187786	0.489826	10.520622
std	0.843218	0.101356	0.120082	5.103230	0.021899	16.952787	42.657233	0.003006	0.150757	0.114469	1.231589
min	3.800000	0.080000	0.000000	0.600000	0.009000	2.000000	9.000000	0.987110	2.720000	0.220000	8.000000
25%	6.300000	0.210000	0.270000	1.700000	0.036000	23.000000	108.000000	0.991720	3.080000	0.410000	9.500000
50%	6.800000	0.260000	0.320000	5.200000	0.043000	34.000000	134.000000	0.993700	3.180000	0.470000	10.400000
75%	7.300000	0.320000	0.390000	9.900000	0.050000	46.000000	167.000000	0.996060	3.280000	0.550000	11.400000
max	14.200000	1.100000	1.230000	65.800000	0.346000	289.000000	440.000000	1.039000	3.820000	1.080000	14.200000

Cómo se puede observar en la tabla anterior, los datos del dataset utilizado no presentan valores faltantes o nulos en ninguno de sus campos, es decir, se tienen 4200 registros con 11 variables numéricas/continuas cada uno. Se hizo un análisis que incluye media, desviación estándar, mínimos, máximos y además los cuartiles para tener una visión general de la distribución de los datos. Por lo observado y como se muestra a continuación se llega a la conclusión de que los datos se encuentran muy desiguales y que por lo tanto sería buena práctica escalarlos o normalizarlos para unos resultados más acordes.

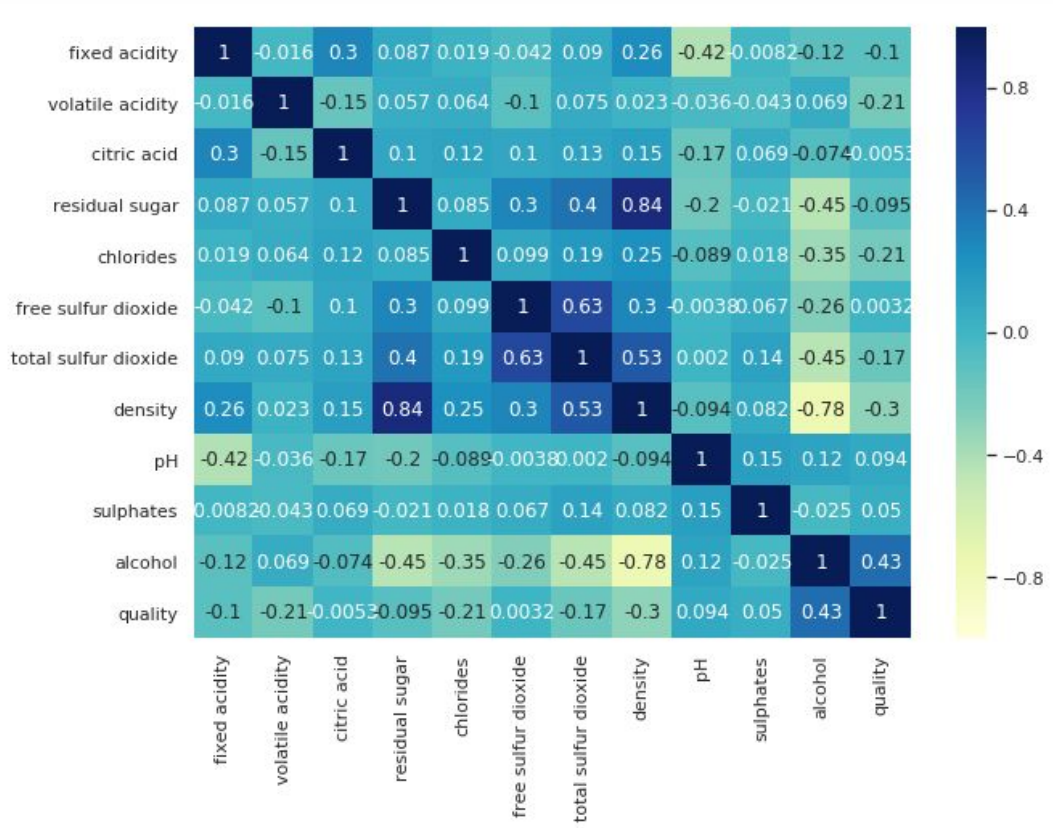
Distribución de variables



Correlación de variables

Se muestra a continuación el mapa de calor con la correlación entre las diferentes variables donde se puede observar correlación mayor a |0.5| entre las variables:

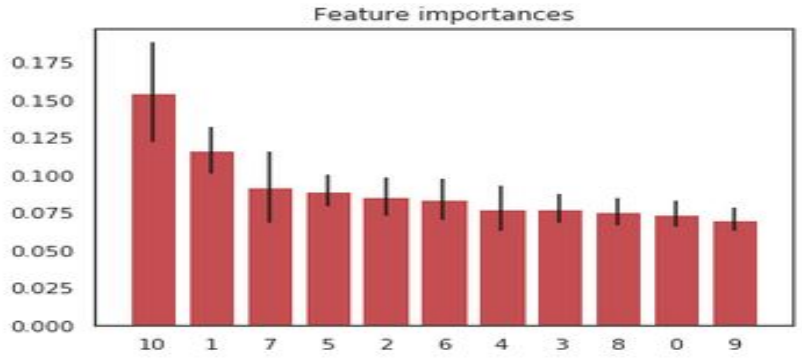
- Total Sulfur Dioxide y Free Sulfur Dioxide
- Density y Residual Sugar
- Density y Total Sulfur Dioxide
- Alcohol y Density



Ranking de variables

Se utilizó una agrupación de árboles de decisión llamada *Extra Tree Classifier* para calcular la importancia relativa de cada una de las variables y realizar así una selección basados en un ranking, evaluando la importancia de cada una de las features con una cantidad de 250 estimadores, usando *la importancia de Gini* para calcular la selección de features, calculada como la reducción total (normalizada) del criterio presentado por esa característica.

```
Feature ranking:
1. alcohol (0.155246)
2. volatile acidity (0.116624)
3. density (0.092000)
4. free sulfur dioxide (0.089839)
5. citric acid (0.085906)
6. total sulfur dioxide (0.083914)
7. chlorides (0.077930)
8. residual sugar (0.077721)
9. pH (0.075916)
10. fixed acidity (0.074325)
11. sulphates (0.070579)
```



Para el caso de clasificación binaria se utiliza como límite de calidad el 5, menor o igual a este valor se considera como un vino de mala calidad (calidad = 0), si es mayor a 5 se considera un vino de buena calidad (calidad = 1).

El porcentaje de vinos malos en los datos de entrenamiento sería de 33%, mientras que los vinos de buena calidad representan el 67% de la muestra, lo que significa que los datos tienen cierto desequilibrio.

Modelos y evaluación de desempeño

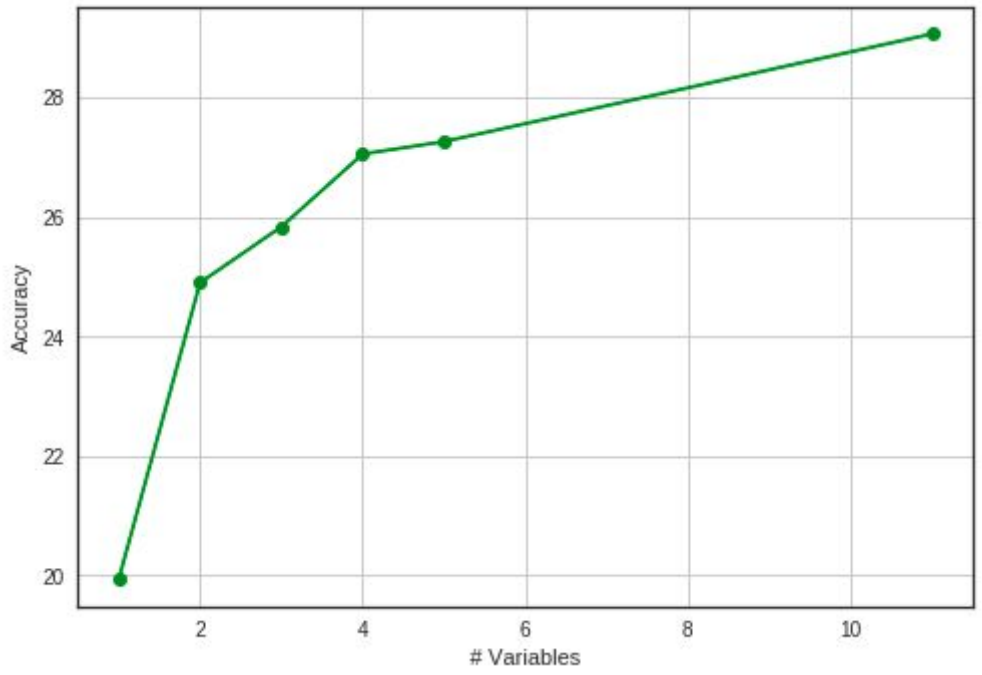
En cada uno de los modelos el particionamiento de los datos de entrenamiento y pruebas fue de 70% - 30% respectivamente. Además de esto se escalaron los datos para que tuvieran una distribución normal. Para esto se utilizó StandardScaler el cual se encarga de modificar los datos para que tengan una distribución de la forma $z = (x - u) / s$, donde x son los datos que se desean escalar, u es el promedio de los datos y s la desviación estándar de los datos.

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
scaler.fit(X)
X_train, X_test, y_train, y_test = train_test_split(scaler.transform(X), y_bin, test_size=0.30)
```

Los modelos utilizados para el análisis y la resolución de este problema fueron:

Regresión lineal:

Se analizaron los resultados del accuracy modificando las variables tenidas en cuenta para el modelo, en este caso con 1, 2, 3, 4, 5 o todas las variables, tomándolas en el orden de importancia obtenido anteriormente mediante análisis. Posteriormente se realizó un análisis del RMSE con el modelo que mejor se comportó.

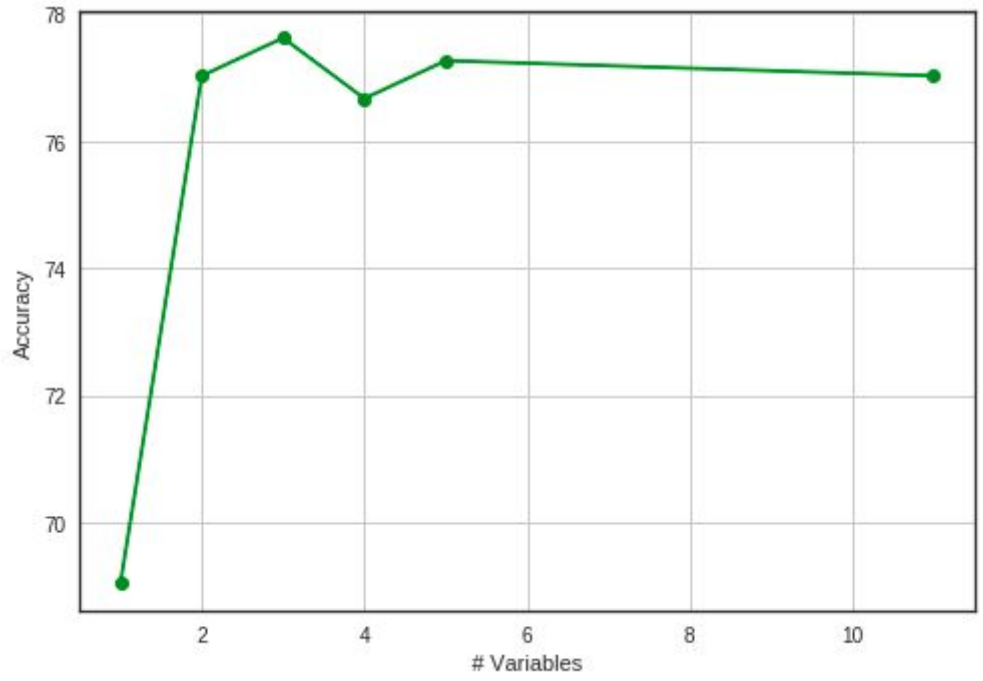



```
from sklearn.metrics import mean_squared_error
clf = linear_model.LinearRegression()
clf = clf.fit(X_train, y_train)
y_predict = clf.predict(X_test)
clf.score(X_test, y_test)
mean_squared_error(y_test, y_predict)
```

```
0.5382587138712316
```

Regresión logística

Se analizaron los resultados del accuracy modificando las variables tenidas en cuenta para el modelo, en este caso con 1, 2, 3, 4, 5 o todas las variables, tomándolas en el orden de importancia obtenido anteriormente mediante análisis. Teniendo como mejor resultado para este caso las 3 variables de más importancia y un accuracy del 78%.




```
clf_3 = linear_model.LogisticRegression(solver="lbfgs")
clf_3 = clf_3.fit(X_train[:,[10,1,7]], y_train)
clf_3.score(X_test[:,[10,1,7]], y_test)*100
```

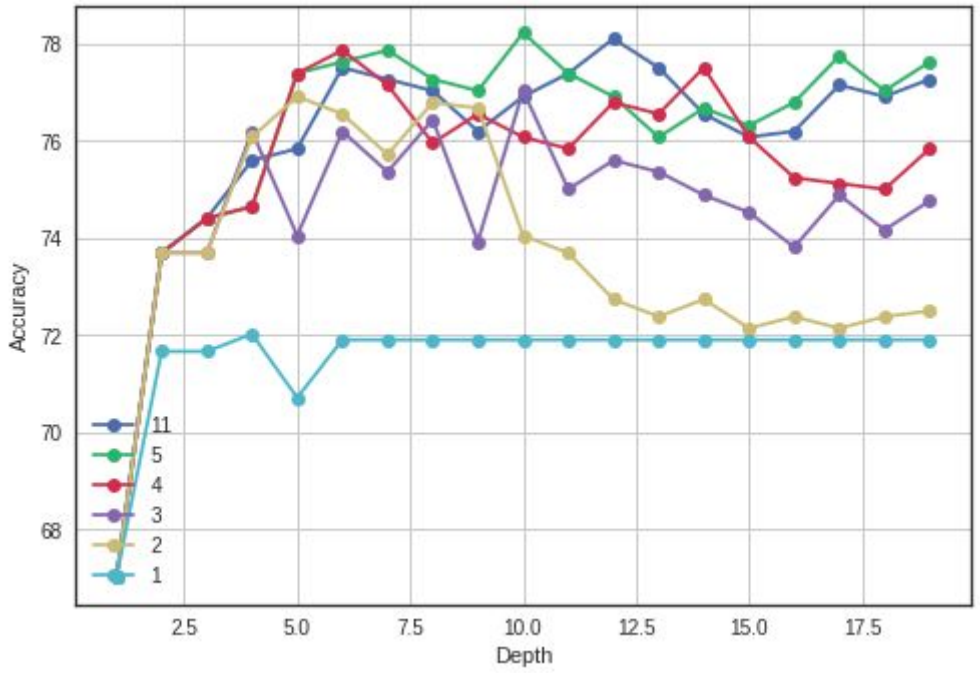
```
77.61904761904762
```

Árboles de decisión:

En este modelo se realizó la predicción modificando el parámetro de profundidad, además de esto se evaluó el modelo modificando el numero de variables, tomandolas en el orden de importancia, para analizar cual tiene un mejor comportamiento a la hora de la predicción. Obteniendo como mejor resultado usar las 5 variables más importante con produndidad 10.

```
clf_5 = tree.DecisionTreeClassifier(max_depth=10)
clf_5 = clf_5.fit(X_train[:,[10,1,7,5,2]], y_train)
clf_5.score(X_test[:,[10,1,7,5,2]], y_test)*100
```

```
78.21428571428571
```



Conclusiones

Se obtienen resultados similares tanto para arboles de decision como para regresión logistica si se usa el mejor resultado para cada uno, sin embargo sigue siendo levemente mejor el resultado obtenido con arboles de decision, una profundidad de 10 y utilizando sólo las 5 variables más importantes Alcohol, Volatile acidity, Density, Free sulfur dioxide, Citric acid.

Referencias

Sklearn.ensemble.ExtraTreesClassifier — scikit-learn 0.20.3 documentation. (2019). Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html

¿Qué es el error cuadrático medio RMSE? | *El blog de franz*. (2018). *El blog de franz*. Retrieved 30 March 2019, from https://acolita.com/que-es-el-error-cuadratico-medio-rmse/