Lunes 09 de enero de 2023

Reporte de investigación de calidad del Vino

Francisco Nuñez Hernandez, Gabriel Díaz Cerón

Coderhouse Comisión 29805

Abstracto

El mundo del vino es complejo y variado, con una gran cantidad de viñas, bodegas, fabricantes y marcas, por lo que tiene una gran complejidad la calidad de una copa de vino. Este trabajo se encargará de analizar las principales variables que afectan la calidad del vino, como lo son la acidez, los azúcares residuales, dióxido de sulfato, entre otros. Será de gran interés para aquellos fabricantes y/o bodegas que buscan la excelencia en sus productos y maximizar su producción.

1. Introducción

El vino**. ¿Qué es lo que hace a un vino un "buen vino"?** El mundo del Vino es un tema muy complejo y entendido solo para algunos. Los **Sommelier's** quienes se consideran los máximos expertos en este tema, nos dicen que cierto vino es bueno o malo. Pero no divulgan el trasfondo de cómo llegaron a esta conclusión. Pareciera una suerte de magia. Sin embargo, como casi todo en la vida, la ciencia nos da la respuesta. Es por eso por lo que con este proyecto nos adentraremos a descubrir, los secretos de lo que convierten a un vino, en un **buen vino**.

2. DataSet

2.1 Visión General

Nuestro dataset llamado **"winequalityN.csv",** consta de 6497 Filas y 13 columnas, para un total de 84423 entradas de datos. Consistentes en type *(White / Red), fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, quality.*

Este fue se obtuvo desde la página de Kaggle. Este contiene suficiente data para realizar el proyecto y aunque existe una variedad más amplia de tipos de Vino, los más populares son el Rojo/tinto y el Blanco. Ambos forman parte del Dataset.

2.2 Variables

Hay 13 variables en el dataset, las cuales son:

1. **Type (Tipo):** *Pudiendo ser Blanco o Rojo/Tinto*
2. **Fixed acidity (Acidez Fija):** *La acidez Influye en el sabor y el Ph, en general la vida útil del Vino.*
3. ***Volatile acidity (Acidez Volatil):*** *Referente a los ácidos destilables al vapor.*
4. ***Citric acid (Ácido Cítrico):*** *Proporciona cierta frescura al vino.*
5. ***Residual sugar (Azúcar Residual):*** *Es la cantidad de azúcar que no pudo ser fermentada.*
6. ***Chlorides (Cloruros):*** *Sales minerales del Vino.*
7. ***Free sulfur dioxide (Dióxido de Azufre Libre):*** *Un potente Antimicrobiano.*
8. ***Total sulfur dioxide (Dióxido de Azufre Total):*** *Se utiliza como conservante. Aunque también como antioxidante y antiséptico.*

9. ***Density (Densidad):*** *También conocido como "peso específico" es un parámetro analítico del vino.*
10. ***Ph (ph):*** *Medida de acidez o alcalinidad de una sustancia acuosa.*
11. ***Sulphates (Sulfatos):*** *Otro conservador, antimicrobiano y antioxidante.*
12. ***Alcohol (Alcohol):*** *Al ser una bebida fermentada, cuenta con graduación de alcohol.*
13. ***Quality. (Calidad):*** *Nota que hace referencia a la calidad del Vino.*

2.3 Análisis

Una pequeña revisión del dataset, nos deja ver que las variables en cada tipo de vino, (blanco o Tinto) interferían de diferente manera en la calidad de este, por lo que era importante dividir el dataset en dos partes, los datos del vino blanco y los del vino tinto.

wine

	type	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	white	7.0	0.270	0.36	20.7	0.045	45.0	170.0	1.00100	3.00	0.45	8.8	6
1	white	6.3	0.300	0.34	1.6	0.049	14.0	132.0	0.99400	3.30	0.49	9.5	6
2	white	8.1	0.280	0.40	6.9	0.050	30.0	97.0	0.99510	3.26	0.44	10.1	6
3	white	7.2	0.230	0.32	8.5	0.058	47.0	186.0	0.99560	3.19	0.40	9.9	6
4	white	7.2	0.230	0.32	8.5	0.058	47.0	186.0	0.99560	3.19	0.40	9.9	6
...	...	...	...	...	...	...	...	...	...	...	...	...	...
6492	red	6.2	0.600	0.08	2.0	0.090	32.0	44.0	0.99490	3.45	0.58	10.5	5
6493	red	5.9	0.550	0.10	2.2	0.062	39.0	51.0	0.99512	3.52	NaN	11.2	6
6494	red	6.3	0.510	0.13	2.3	0.076	29.0	40.0	0.99574	3.42	0.75	11.0	6
6495	red	5.9	0.645	0.12	2.0	0.075	32.0	44.0	0.99547	3.57	0.71	10.2	5
6496	red	6.0	0.310	0.47	3.6	0.067	18.0	42.0	0.99549	3.39	0.66	11.0	6

6497 rows × 13 columns

3. Métodos

Antes que nada, decidimos preprocesar el **Dataset**, haciendo lo siguiente:

3.1 Separamos el en dos los datos y eliminamos nulos.

Como mencionamos anteriormente, debido a que había una notoria diferencia en variables que intervenían en la calidad del vino decidimos separar el data set en dos **dataframes** por tipo de vino. (Blanco y Tinto). Posteriormente obtuvimos la proporción de valores nulos por tipo de vino y el total. Luego, debido a que los valores nulos representaban un **0.15%** decidimos eliminar estas filas nulas.

3.2 Exploratory Data Analysis

Empezando a comprender la magnitud de los datos, nos dimos cuenta de que de los dos **dataframes** que obtuvimos de nuestro **dataset**, (El del vino blanco y el del vino tinto) el que tiene más información, es el **dataframe** del vino Blanco, con un **75.4%** del total de los datos.

Otro punto importante, es que los vinos del tipo blanco tienen en promedio mejores notas (calidad) que los del tipo tinto.

Para entender mejor toda esta información decidimos graficar cada variable relacionándola con otra, en conjunto con una matriz de correlación.

	Column	Wine_w: 4898	NaN_w_%	Wine_r: 1599	NaN_r_%	Wine_t: 6497	NaN_t_%
0	type	0	0.000000	0	0.000000	0	0.000000
1	fixed acidity	8	0.163332	2	0.125078	10	0.153917
2	volatile acidity	7	0.142915	1	0.062539	8	0.123134
3	citric acid	2	0.040833	1	0.062539	3	0.046175
4	residual sugar	2	0.040833	0	0.000000	2	0.030783
5	chlorides	2	0.040833	0	0.000000	2	0.030783
6	free sulfur dioxide	0	0.000000	0	0.000000	0	0.000000
7	total sulfur dioxide	0	0.000000	0	0.000000	0	0.000000
8	density	0	0.000000	0	0.000000	0	0.000000
9	pH	7	0.142915	2	0.125078	9	0.138525
10	sulphates	2	0.040833	2	0.125078	4	0.061567
11	alcohol	0	0.000000	0	0.000000	0	0.000000
12	quality	0	0.000000	0	0.000000	0	0.000000

3.3 Red Neuronal (Keras)

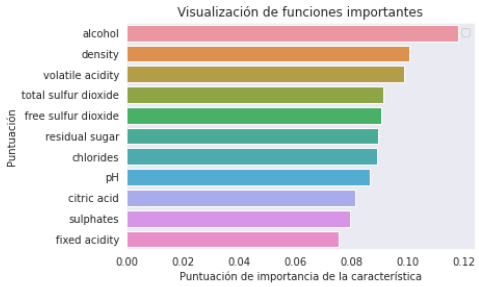
Con el modelo secuencial, entrenamos una red neuronal, con el objetivo, que con la data de la que disponíamos podríamos determinar cuál es la relación de una buena nota (calidad) del vino y las variables que intervienen.

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 batch_normalization_4 (Batc  (None, 11)               44
 hNormalization)

 dense_4 (Dense)             (None, 128)               1536

 dropout_2 (Dropout)         (None, 128)               0

 batch_normalization_5 (Batc  (None, 128)              512
 hNormalization)

 dense_5 (Dense)             (None, 413)               53277

 dropout_3 (Dropout)         (None, 413)               0

 batch_normalization_6 (Batc  (None, 413)              1652
 hNormalization)

 dense_6 (Dense)             (None, 952)               394128

 batch_normalization_7 (Batc  (None, 952)              3808
 hNormalization)

 dense_7 (Dense)             (None, 11)                10483

=================================================================
Total params: 465,440
Trainable params: 462,432
Non-trainable params: 3,008
_________________________________________________________________
```

3.4 Random Forest

Este modelo el que alcanzo más del **90%** de precisión. Por lo que perfila para ser el modelo elegido para el análisis.

```
Model Performance
Average Error: 0.3576 degrees.
Accuracy = 93.65%.
```



Visualización de Características importantes.

3.5 PCA

Debido a que la cantidad de Variables, que usando el modelo PCA, se describen un conjunto de datos, mediante nuevas variables no correlacionadas. Y con el resultado de este, se aplico un modelo de **Gradient Boosting Trees**. El cual esta formado por un conjunto de árboles de decisión individuales, entrenados de forma secuencial.

	precision	recall	f1-score	support
3	0.00	0.00	0.00	2
4	0.00	0.00	0.00	17
5	0.63	0.66	0.65	171
6	0.51	0.61	0.56	148
7	0.49	0.38	0.43	53
8	0.00	0.00	0.00	8
accuracy			0.56	399
macro avg	0.27	0.27	0.27	399
weighted avg	0.53	0.56	0.54	399

4. Resultados de modelos ML

Se realizaron tres modelos distintos de Machine Learning para definir cuál tiene mayor exactitud con el fin de predecir la calidad de un vino en base a sus características mencionadas anteriormente.

El resultado es el siguiente:

Keras: Cerca del 58.36%

Random Forest: 93.25%

PCA: Cerca del 64.41%

Como resultado se concluye que se utilizará el **Modelo Random Forest**, aunque antes de la **Grid Search with Cross Validation**, dado que posterior a este análisis la veracidad final es **91.92%,** que aun así es un resultado por sobre el 90%, teniendo un modelo de machine learning exitoso.

5. Conclusiones

Los vinos tintos son superiores cuando su acidez es mayor.

En el caso de los tres tipos principales de acidez se puede mencionar lo siguiente:

a) **Acidéz Fija**: aumenta levemente en vino tinto de buena calidad, dado que es conservante y corrector de acidez.
b) **Acidéz cítrica total:** aumenta considerablement e en vino tinto con alta calificación, define la fre scura de vino.
c) **Acidéz Variable:** disminuye en vino tinto de bue na calidad, representa principalmente el deterio ro de vino en relación con el tiempo que este envasado.

Al apreciar estos tres puntos, en primera instancia se puede concluir que un vino tinto es de buena calidad si tiene altos niveles de acidez fija y variable.

En relación con el vino blanco y su acidez es necesario realizar más análisis, dado que en estas variables presenta valores constantes en todas las notas que miden su calidad.

En cuanto al PH y el alcohol, el vino que presenta mayor concentración entre un PH más bajo y un grado alcohólico menor. Pero para el vino blanco la concentración se encuentra bastante dispersa.