Evaluación de la calidad de Vinos con Algoritmos de Clasificación.

Muñoz, Jeanette Elizabeth, D'Angiolo, Federico Gabriel, Fernando Asteasuain

Universidad Nacional de Avellaneda, Ingeniería en informática

Resumen

El objetivo de este trabajo es aplicar distintas herramientas de Machine Learning para predecir la calidad del vino. Para esto, se toma un conjunto de datos de "Calidad de vinos", el cual contiene información sobre diversas propiedades fisicoquímicas de cada uno de estos. Para realizar este estudio, se utilizarán distintas herramientas, como por ejemplo: Regresión Lineal, Naive Bayes, Decision tree, KNN (k-nearest neighbors) y Redes Neuronales.

Palabras Clave

Regresión Lineal, Naive Bayes, Decision tree, KNN (k-nearest neighbors), Redes Neuronales, precisión, calidad, vinos.

Introducción

Hoy en día, muchas actividades industriales basan sus desarrollos adoptando el avance que se viene dando en el área del Aprendizaje Automático (Machine Learning) **[1-2]**. El fundamento de este trabajo se basa en la exploración y estudio de distintas técnicas trabajadas en el área de Machine Learning.

El objetivo principal de este trabajo es analizar la relación de los componentes químicos que permiten definir si un vino es de calidad o no. Para esto se realizará la comparación de diferentes conceptos de Aprendizaje Supervisado (una de las ramas del Aprendizaje Automático), donde entran en juego algoritmos de Clasificación y Regresión.

Para este estudio, en primer lugar, se analizan los datos obtenidos, luego se los visualizan y a continuación se desarrollan modelos de datos que permitan tomar decisiones. Algunos de estos modelos se obtienen mediante técnicas de Regresión y de Clasificación. Por último, se podrá estudiar mediante distintas herramientas, la calidad de cada uno de estos modelos.

Cada uno de los pasos mencionados anteriormente, contempla el análisis de correlación y sus correspondientes gráficos de distribución. Esto permite a priori, estudiar cuáles variables pueden ser significativas para hacer análisis de Regresión y de Clasificación, según sea el caso.

Los algoritmos a utilizarse seran: Regresión Lineal, Naives Bayes, Árboles de Decisión, KNN y Redes Neuronales. El dataset seleccionado posee las siguientes variables: *pH, Quality, Fixed acidity, Residual sugar, Total sulfur dioxide, Volatile acidity, Citric acid, Chlorides, Free sulfur dioxide, Density, Sulphates, Alcohol sulfur dioxide.*

De dichas variables se seleccionan las que posean características lineales para los algoritmos de regresión y luego para los demás se clasificaran de manera binaria.

el cual lo podrán encontrar en el siguiente link: DATASET WINE QUALITY

Podrán acceder al código del proyecto ingresando al siguiente enlace:

Repositorio de todo el Proyecto

Desarrollo

El Aprendizaje supervisado se divide en dos ramas:

Regresión: predice respuestas continuas. La Regresión tiene dos significados: uno surge de la distribución conjunta de probabilidades de dos variables aleatorias mientras que el otro significado es empírico y nace de la necesidad de ajustar alguna función a un conjunto de datos **[3]**. Un modelo lineal se basa en la suposición de que es posible

aproximar los valores de salida a través de un proceso de regresión basado en la regla:

$$f(x_i) = \alpha_0 + \sum_{i=1}^{m} \alpha_i x_i \quad donde\ A = \{\alpha_0, \alpha_1, ..., \alpha_m\}$$

[1]
Siendo:
f(x) = vector de salida
x = vector de entrada
$\boldsymbol{\alpha}_0$, $\boldsymbol{\alpha}_1$ = coeficientes obtenidos mediante de los datos. **[1]**

Clasificación: puede predecir respuestas discretas y se utiliza cuando los datos pueden ser etiquetados o categorizados en grupos o clases. Los algoritmos utilizados en este trabajo son: Árboles de decisión, K vecinos más cercanos (KNN), Naive Bayes y Redes Neuronales.

Los **árboles de decisión** o de clasificación son un modelo surgido en el ámbito del aprendizaje automático (Machine Learning) y de la Inteligencia Artificial que, partiendo de una base de datos, crea diagramas de construcciones lógicas que nos ayudan a resolver problemas.

KNN (k-nearest neighbors) representa un algoritmo de clasificación supervisado que proporcionará nuevos puntos de datos de acuerdo con los datos más cercanos.

El algoritmo que usaremos proviene del modelo bien conocido **Naive Bayes,** que son una clase especial de algoritmos de clasificación de Aprendizaje Automático (Machine Learning), tal y como nos referiremos de ahora en adelante. Se basan en una técnica de clasificación estadística llamada "Teorema de Bayes". Estos modelos son llamados algoritmos "Naive", o "Inocentes" en español. En ellos se asume que las variables predictoras son independientes entre sí. En otras palabras, que la presencia de una cierta característica en un conjunto de datos no está en absoluto relacionada con la presencia de cualquier otra característica.

Una **Red Neuronal**, intenta simular el comportamiento de una red de neuronas de un ser humano. Como tal, cada neurona que resulta ser artificial, debe aprender hasta obtener un modelo de red neuronal completo que sea capaz de clasificar eficientemente.

Cuando se analiza un dataset, conviene realizar un estudio de las variables por eso, a continuación, se estudian todas las variables contenidas en el conjunto de datos (dataset). Para mayor detalle del código asociado a estos gráficos, se deja un enlace a Github:

File in Github with Analysis

Previamente al análisis de Regresión Lineal, conviene estudiar la correlación entre cada una de las variables, para conocer su posible dependencia. A continuación, se observa una Tabla que muestra la correlación entre cada una de las variables.

Análisis de correlación de las variables

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide
fixed acidity	1.00	-0.26	0.67	0.11	0.09	-0.15
volatile acidity	-0.26	1.00	-0.55	0.00	0.06	-0.01
citric acid	0.67	-0.55	1.00	0.14	0.20	-0.06
residual sugar	0.11	0.00	0.14	1.00	0.06	0.19
chlorides	0.09	0.06	0.20	0.06	1.00	0.01
free sulfur dioxide	-0.15	-0.01	-0.06	0.19	0.01	1.00
total sulfur dioxide	-0.11	0.08	0.04	0.20	0.05	0.67
density	0.67	0.02	0.36	0.36	0.20	-0.02
pH	-0.68	0.23	-0.54	-0.09	-0.27	0.07
sulphates	0.18	-0.26	0.31	0.01	0.37	0.05
alcohol	-0.06	-0.20	0.11	0.04	-0.22	-0.07
quality	0.12	-0.39	0.23	0.01	-0.13	-0.05

	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	-0.11	0.67	-0.68	0.18	-0.06	0.12

volatile acidity	0.08	0.02	0.23	-0.26	-0.20	-0.39
citric acid	0.04	0.36	-0.54	0.31	0.11	0.23
residual sugar	0.20	0.36	-0.09	0.01	0.04	0.01
chlorides	0.05	0.20	-0.27	0.37	-0.22	-0.13
free sulfur dioxide	0.67	-0.02	0.07	0.05	-0.07	-0.05
total sulfur dioxide	1.00	0.07	-0.07	0.04	-0.21	-0.19
density	0.07	1.00	-0.34	0.15	-0.50	-0.17
pH	-0.07	-0.34	1.00	-0.20	0.21	-0.06
sulphates	0.04	0.15	-0.20	1.00	0.09	0.25
alcohol	-0.21	-0.50	0.21	0.09	1.00	0.48
quality	-0.19	-0.17	-0.06	0.25	0.48	1.00

Tabla 1. Correlación entre distintas variables

De la Tabla 1, se puede ver que el conjunto de datos no tiene valores nulos en ninguna de las características y etiquetas.

El coeficiente de correlación (**ρ**) se encuentra contenido en el intervalo -1< **ρ** <1. El caso de **ρ** = 0 indica la ausencia de cualquier asociación lineal, mientras que los valores -1 y 1 indican relaciones lineales perfectas. En base a la Tabla 1, donde se muestra el análisis de correlación de todas las columnas del dataset podemos determinar que las variables de 'citric acid', 'fixed acidity', 'density', 'free sulfur dioxide', 'total sulfur dioxide' muestran valores que se aproximan a 1 y el ph de aprox -1 indicando relaciones lineales bastante acertadas.
pH - Fixed acidity = -0.68
Fixed acidity - Citric acid = 0.67
Fixed acidity - Density = 0.67
Free sulfur dioxide - Total sulfur dioxide = 0.67

Regresión Lineal.
Se puede encontrar el código en el siguiente link: **Code Regresión Lineal**
En este trabajo sólo se muestra la Regresión Lineal de las variables:

- *Fixed acidity - Citric acid*
- *PH - Fixed acidity*

Fixed acidity - Citric acid

En la fig. 1 puede observar la regresión lineal de las sustancias de acidity - Citric acid.

Pendiente: [[6.00613927]]
Término independiente : [6.67786814]

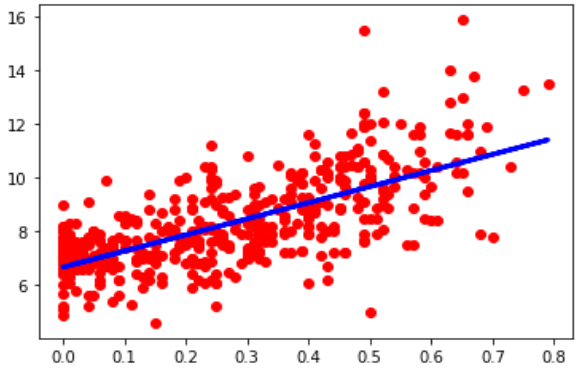


Fig. 1. Análisis de Regresión Lineal Fixed para acidity - Citric acid

MAE (Mean Absolute Error): 0.94
MSE (Mean Square Error): 1.51
RMSE (Root Mean Square Error): 1.23
R2 (Coeficiente de determinación) : 0.48

PH - Fixed acidity

En la fig. 2 puede observar la regresión lineal de las sustancias del PH - Fixed acidity.

Pendiente: [[-7.72375284]]
Término independiente : [33.91255379]

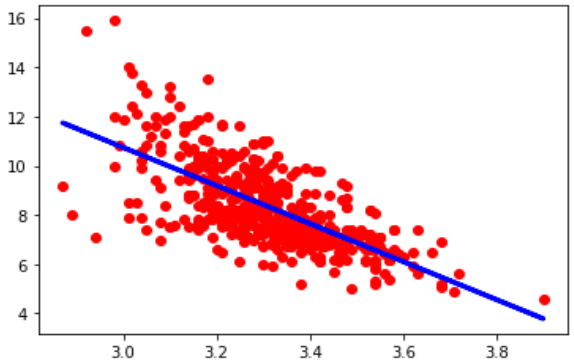


Fig. 2. Análisis de Regresión Lineal Fixed para PH - Fixed acidity

MAE (Mean Absolute Error): 0.99
MSE (Mean Square Error): 1.58
RMSE (Root Mean Square Error): 1.26
R2 (Coeficiente de determinación): 0.45

Para concluir este análisis, a continuación, se expone una tabla donde se observa la comparación entre distintos tipos de métricas:

	MAE	MSE	RMSE	R2
citric acid - fixed acidity	0.94	1.51	1.23	0.48
pH - fixed acidity	0.99	1.58	1.26	0.45

Tabla 2. Comparación de distintas métricas para Regresión Lineal.

Con la tabla 2 podemos comparar los resultados obtenidos en las regresiones Lineales de los compuestos elegidos.

Clasificación. Definición de nuevas etiquetas.

Para los siguientes modelos se re-clasifica a cada categoría de calidad del dataset en: Bueno = 1; Malo (quality <= 5) = 0.

En esta parte del trabajo se realizarán los siguientes algoritmos:

- Naives Bayes
- Arbol de decision
- Red Neuronal
- K-Nearest Neighbor Classifier.

Para evaluarlos, se utilizará la matriz de confusión binaria y su Accuray. En la siguiente imagen podrán observar los datos obtenidos de cada modelo con su respectiva comparación mediante Matriz de Confusión se encuentra conformada de la siguiente manera:

Fig. 3. Matriz de Confusión.

Naive Bayes Classifier

Code Naive Bayes - Accuracy: 0.73

En la figura 4 podemos observar la matriz de confusión obtenida para evaluar el clasificador.

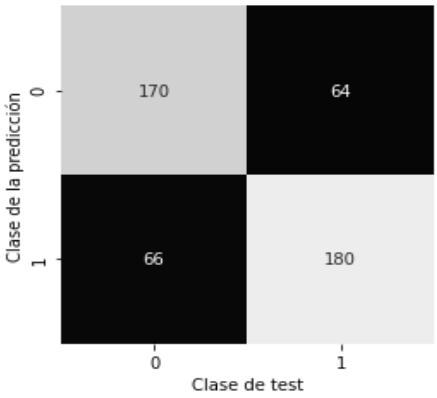


Fig. 4. Matriz de Confusión - Naive Bayes.

Curva ROC (Receiver operating characteristic)

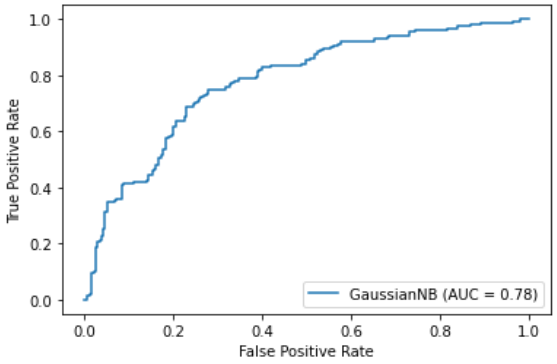


Fig. 5. Curva ROC- Naive Bayes.

Este resultado fue el mejor en la comparación de resultados con:

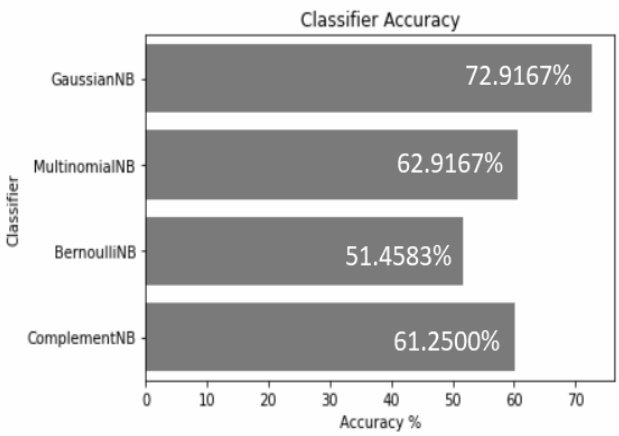


Fig. 6. Comparación con otros algoritmos Naive-Bayes

Árbol de decisión

En la figura 10 podemos observar la matriz de confusión obtenida para evaluar el clasificador y en la figura 11 veremos la ramificación del árbol.

Matriz de confusión binaria

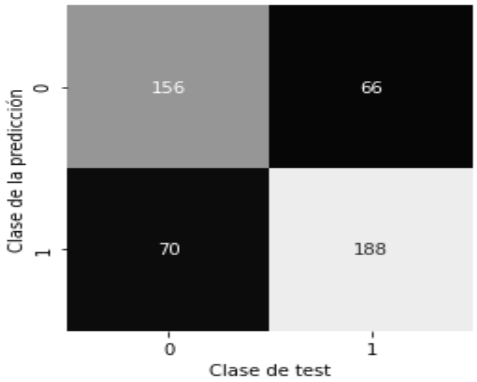


Fig. 7. Matriz de Confusión - Decision Tree.

[Code Decision Tree]

Accuracy: 0.72
Mean Absolute Error: 0.28

Visualizando el árbol

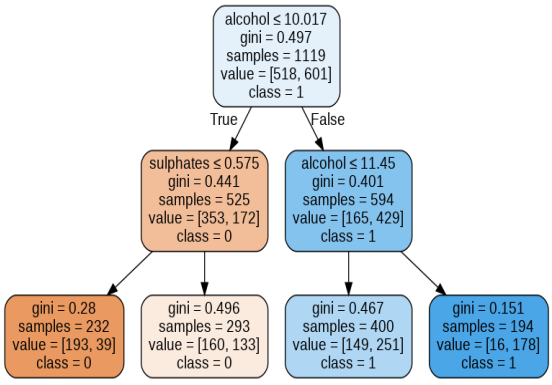


Fig. 8. Diagrama de árbol

Red neuronal

Se generó una red neuronal se utiliza en la capa de salida sigmoid y en las otras tres capas usan la función ReLU. Para poder entrenar el modelo debemos definir la función de pérdida[1], el optimizador y las métricas a usar:

```
model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
```

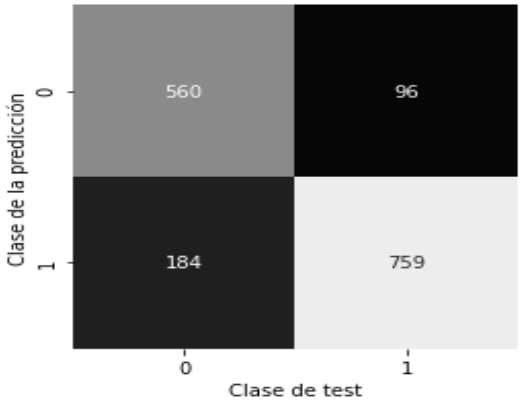
En este caso se probó con estos valores de batch_size = 10, epochs = 1000, para obtener un

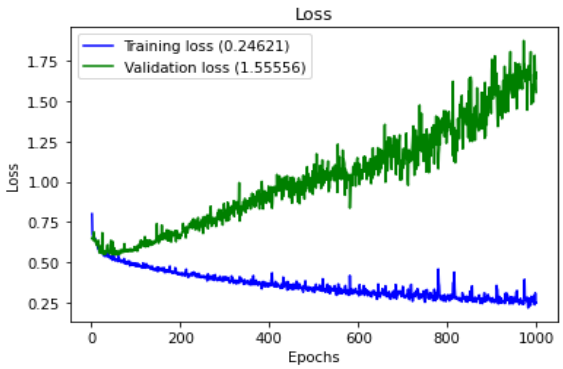
[Code Red Neuronal]

Accuracy: 86.06 - Loss: 27.10

Esto nos permite observar que cuando más iteraciones tenga más disminuye la pérdida.

Matrix de confusion binary





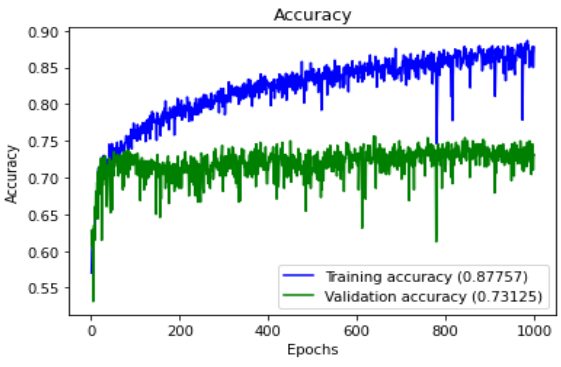


Fig. 9. Matriz de Confusión y Diagramas de Accuracy - Red Neuronal.

[1] https://programmerclick.com/article/67471070714/

KNN: k-Nearest Neighbor Classifier

Con un rango de k = (1, 120) Se obtiene la siguiente gráfica:

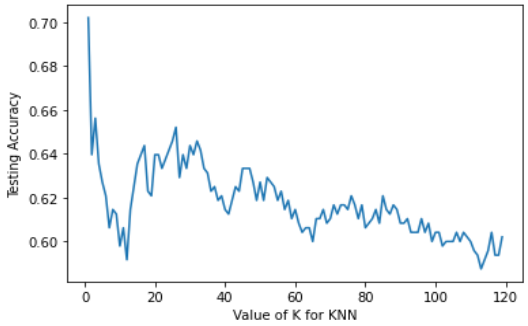


Fig. 10. Accuracy vs “K”

{'algorithm': 'brute', 'metric': 'seuclidean', 'n_neighbors': 101, 'p': 1, 'weights': 'distance'}

Code KNN(k-nearest neighbors)

Accuracy: 0.82 - Loss: 0.18

Matriz de confusión binaria

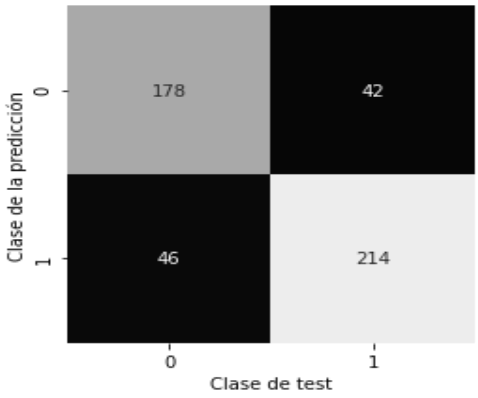


Fig. 11. Matriz de Confusión - KNN.

Curva ROC (Receiver operating characteristic)

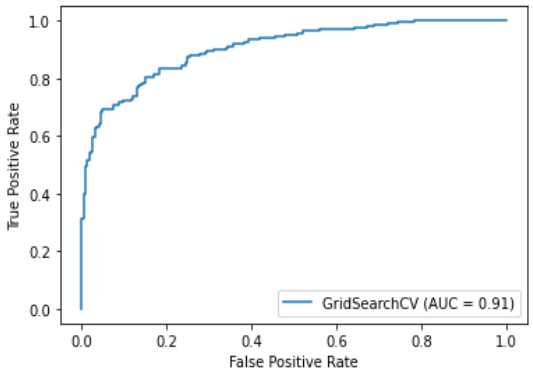


Fig. 12. TPR vs FPR.

Conclusiones.

Haciendo un análisis sobre los algoritmos estudiados, se ve que con Regresión Lineal se pueden comparar las variables:'citric acid', 'fixed acidity', 'density', 'free sulfur dioxide', 'Total sulfur dioxide' dado el factor de correlación encuentra contenido en el intervalo -1< **ρ** <1 indicando relaciones lineales bastante acertadas, es decir, antes de trabajar con algoritmos de Regresión Lineal, conviene estudiar dicho factor.

Por otro lado, en cuanto a los algoritmos de clasificación, se puede ver que mediante Redes Neuronales se obtiene el mejor resultado la exactitud (accuracy) mide el porcentaje de casos que el modelo ha acertado. Es decir, el modelo acierta el 86% de las veces.

Tras reproducir varios algoritmos para el análisis del dataset podemos detectar la variante diferencial en cada resultado. Para cuantificar mejor esto, a continuación se resumen los resultados de accuracy en la Tabla 3.

Algoritmo	Accuracy
KNN	0.82
Red Neuronal	0.86
Arbol de Desicion	0.72
Naive Bayes	0.73

Tabla 3. Resumen de algoritmos

Para concluir este trabajo, en el cual se realizó predicción y clasificación, se destaca que la evaluación no se equipara a la que podrían realizar expertos en el área.

Por ello no se considera establecer predicciones en su totalidad de que un vino sea bueno o no, ya que la experiencia de poder catar un vino y no solo evaluarlo por

su composición, podría darle mayor o menor puntuación.
Además, a la hora de llevar a cabo la cata de un vino, deberemos fijarnos en los denominados factores externos, que sería el espacio físico donde va a realizarse la cata, dentro de los cuales destacamos la copa de vinos o "catavinos", la sala de cata y la temperatura de servicio de los vinos, etc.

Haciendo un análisis sobre los algoritmos estudiados, se ve que con Regresión Lineal se pueden comparar las variables:'citric acid', 'fixed acidity', 'density', 'free sulfur dioxide', dado el factor de correlación estudiado, es decir, antes de trabajar con algoritmos de Regresión Lineal, conviene estudiar dicho factor.

Por otro lado, en cuanto a los algoritmos de clasificación, se puede ver que mediante Redes Neuronales se obtiene el mejor resultado el cual se acerca al 86% de accuracy.

Como trabajo a futuro, se pueden comparar otros algoritmos de clasificación y poder así, obtener conclusiones que puedan extender este estudio.

Repositorio de todos los códigos

Agradecimientos

A la Universidad Nacional de Avellaneda con la carrera de Ingeniería en Informática y a los docentes que guiaron en mi formación académica, les agradezco de todo corazón lo que me transmitieron y este trabajo es una muestra de lo que aprendí y sigo aprendiendo.

Referencias.

[1] Selection of important features and predicting wine quality using machine learning techniques. Yogesh Gupta. GLA University, Mathura, INDIA.

[2] The role of olfaction in the elaboration and use of the Chardonnay wine concept. Jordi Ballester a,*, Catherine Dacremont, Yves Le Fur, Patrick Etievant. Food Quality and Preference 16 (2005) 351–359

[3] Probabilidad y Estadística. Aplicaciones y Métodos. George, Canavos. 1988. ISBN:968-451-856-0

[4] Machine Learning Algorithms. Giuseppe Bonaccorso. July 2017. ISBN 978-1-78588-962-2

[5]https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009/code
https://es.wikipedia.org/wiki/Cata_de_vinos

[6]https://empresas.blogthinkbig.com/ml-a-tu-alcance-matriz-confusion/

[7]https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/