Comparación de modelos de aprendizaje para la clasificación de vinos

Manuel Espinoza, Jorge Puentes
Universidad de la Frontera
Departamento de Ciencias de la Computación e Informática
Temuco, Chile
{m.espinoza11, j.puentes02}@ufromail.cl

***Abstract*—Wine Quality Data Set es un conjunto de datos entregado por UCI, un repositorio de Machine Learning. Este dataset contiene características de vinos rojos y blancos provenientes del vino portugués "Vinho Verde", donde a cada conjunto de características se le asigna un valor de calidad entre cero y diez. A partir de esto se pretende realizar una comparativa entre distintos modelos de clasificación para obtener el que mejor prediga la calidad de cierto vino dado un set de características.**

Keywords—Análisis de datos, modelos, métricas.

I. Introducción y descripción de los datos.

El dataset que se utilizó en el presente reporte [1] son una serie de datos correspondientes a las propiedades de vinos rojos y blancos (por separado) pertenecientes a una marca llamada "Vinho Verde", los cuales se clasifican con una calidad que oscila entre 0 y 10 (para el detalle de los datos, ver tabla I).

TABLA I Tipos de datos

Columna	Tipo de dato		Desviación estándar	
	Vino tinto	***Vino blanco***	***Vino tinto***	***Vino blanco***
fixed acidity	float64	float64	1.740552	0.843782
volatile acidity	float64	float64	0.179004	0.100784
citric acid	float64	float64	0.194740	0.121007
residual sugar	float64	float64	1.409487	5.071540
chlorides	float64	float64	0.047051	0.021846
free sulfur dioxide	float64	float64	10.456886	17.005401
total sulfur dioxide	float64	float64	32.885037	42.493726
density	float64	float64	0.001887	0.002991
pH	float64	float64	0.154338	0.150985
sulphates	float64	float64	0.169454	0.114114
alcohol	float64	float64	1.065334	1.230495
quality	int64	int64	0.807317	0.885548

Según la descripción de los datos, el problema a tratar corresponde a uno de clasificación, teniendo 11 clases, sin embargo, al visualizar la columna "quality" (ver figura I), se puede notar que en la práctica, los valores realmente oscilan entre 3 y 9, siendo 5, 6 y 7 los valores más recurrentes. Por otra parte, es oportuno recalcar que existen más registros de vino blanco que de vino rojo (ver figura II). Una opción puede ser juntar los datasets y trabajar como gran conjunto, se descarta debido a que los vinos tintos y blancos tienen distintas definiciones de calidad.

FIGURA I Frecuencia de las calidades de vinos rojos (izquierda) y blancos (derecha)

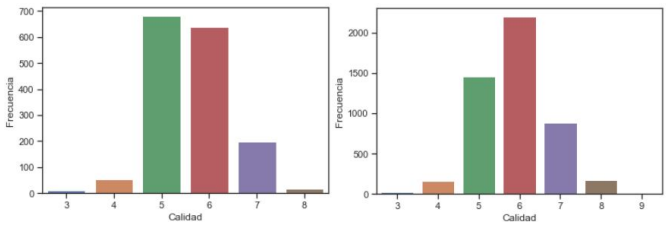
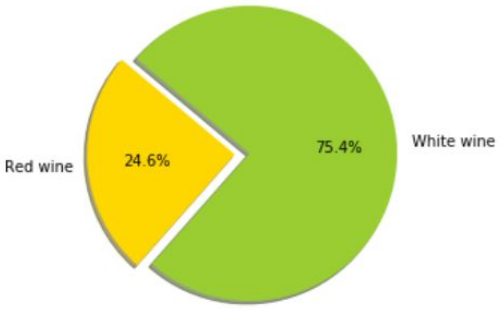


FIGURA II PROPORCIÓN DEL DATASET, HABIENDO UN TOTAL DE 6497 REGISTROS



La importancia de tener la capacidad de clasificar la calidad de cierto vino según sus características con una precisión aceptable, radica principalmente en el provecho que los productores de vinos le pueden sacar, es decir, tener la posibilidad de predecir la calidad de un vino antes de venderlo puede tener un impacto significativo en el negocio de quienes producen.

II. MODELAMIENTO DEL PROBLEMA

A. *Estrategia para abordar el problema: modelo base.*

En primer lugar, para el problema que se presentó anteriormente es necesario escoger un set de modelos de aprendizaje para lograr clasificar la calidad de un vino, estos modelos fueron extraídos de la biblioteca de Python: Scki-learn [2] y se seleccionó vecinos cercanos debido al conocimiento de los parámetros a ajustar.

Luego de escoger el modelo a ajustar, ¿cómo se evalúa el rendimiento para elegir al mejor modelo? Usualmente la estrategia consiste en dividir el conjunto de datos en una parte para entrenar el modelo y otro para testear y así evitar el clásico problema del Overfitting [3], obteniendo el rendimiento con la resta del valor real (la predicción esperada) con el que el modelo predice. Esta técnica es de gran utilidad, sin embargo, existe una incertidumbre al escoger de manera arbitraria la forma en que se dividirá el dataset. Para lo anterior, se utilizará grid search cross-validation, dividiendo el conjunto de datos en diez (k=10), el cual generalmente se usa como alternativa a k=n (siendo n el tamaño del dataset) por el costo computacional que implica [4], teniendo aún en consideración que los conjuntos a trabajar son de tamaño pequeño.

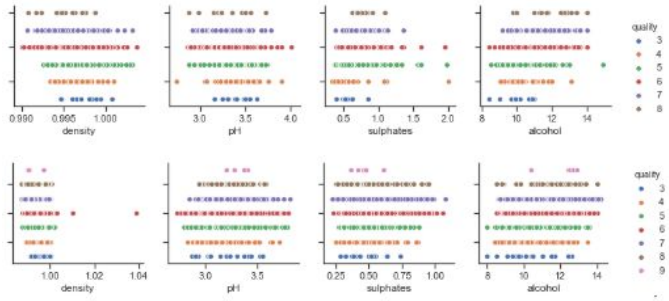
B. *¿Cómo se escogerá el mejor modelo?*

El objetivo que se plantea desde un inicio consiste en identificar el modelo que mejor rendimiento obtenga respecto a la exactitud en las predicciones (accuracy), sin descuidar la sensibilidad (recall), la cual sólo se mostrará debido a que se desconoce cómo lograr un equilibrio entre la exactitud y la sensibilidad utilizando grid search cross-validation.

III. TRAYECTORIA DEL PROYECTO, RESULTADOS E INTERPRETACIÓN

La forma de abordar el problema en cuestión, consistió en primer lugar, en leer la documentación del dataset [1] para obtener un contexto de lo que se quería lograr con la solución. Posteriormente, se procedió a hacer distintas visualizaciones de los datos, como la relación entre variables con la biblioteca de Python, seaborn (ver figura III), para obtener una visión aclarada de la forma en que se distribuían. De lo anterior, fue posible obtener el enfoque para desarrollar el proyecto, que consistió en escoger un modelo de aprendizaje, siendo la exactitud la métrica a maximizar.

FIGURA III RELACIONES ENTRE ALGUNAS PROPIEDADES DEL VINO Y SU CALIDAD (EL DE ARRIBA ES DEL VINO ROJO Y EL INFERIOR, DEL VINO BLANCO)



A. *Resultados e interpretaciones.*

Los resultados obtenidos (ver tabla II), indican que los mejores valores de k para el algoritmo vecinos cercanos son 24 y 59 para vinos tintos y blancos respectivamente, con exactitudes promedio que rondan los 45% y 51%, además de una desviación estándar de las exactitudes cercanas a 0.04, lo cual es indicador de un resultado confiable.

TABLA II RESULTADOS DEL MODELO VECINOS CERCANOS UTILIZANDO AJUSTE DE PARÁMETROS.

Dataset	**Exactitud**		**Sensibilidad**	**Valor de k**
	Valor	***Desviación estándar***		
Vino tinto	0.515957	0.0479794	0.515957	24
Vino blanco	0.453042	0.0299632	0.453042	59

Ciertas columnas del dataset se pueden obviar con tal de mejorar el rendimiento del algoritmo clasificador. El criterio que se utilizó en este caso, fue escoger las dos columnas con mayor desviación estándar (ver tabla I) y eliminarlas. Al realizar esto, los resultados indican que los mejores valores de k para el algoritmo vecinos cercanos son 59 y 48 para vinos tintos y blancos respectivamente, con una mejoría de exactitudes promedio que rondan los 6% y 5%, además de una

desviación estándar baja en ambos casos. Es posible utilizar otros criterios para no considerar ciertas columnas (a través de visualización de la distribución de los datos, por ejemplo) o simplemente probar combinaciones para obtener un mejor rendimiento, sin embargo, aquello no se abordará en este trabajo.

TABLA III RESULTADOS DEL MODELO VECINOS CERCANOS ELIMINANDO DOS COLUMNAS.

Dataset	**Exactitud**		**Sensibilidad**	**Valor de k**
	Valor	***Desviación estándar***		
Vino tinto	0.577861	0.0521376	0.577861	59
Vino blanco	0.4951	0.0459262	0.4951	48

Los resultados obtenidos no son los mejores si se tiene como referencia otros trabajos [1], [5], sin embargo, cumple con el objetivo propuesto que consistía en comparar modelos de aprendizaje para clasificar vinos según la calidad.

CONCLUSIONES

Como se mencionó anteriormente, si bien los resultados no fueron los mejores, es un acercamiento aceptable debido al aprendizaje obtenido sobre cómo comparar distintos modelos según criterios definidos con anterioridad. La principal fortaleza identificada para el presente trabajo, consistió en entender el contexto del problema de forma efectiva mediantes diferentes métodos de visualización, mientras que la debilidad consistió en la falta de conocimiento técnico y experiencia sobre cómo optimizar las métricas que se definan como de interés, además, se carece de entendimiento respecto de los modelos existentes para esta área de estudio.

De forma especulativa, en el futuro, al presente trabajo se le podrían agregar distintos algoritmos de clasificación, tales como “Decision Tree” o “Random Forest”, los cuales se decidieron dejar fuera de este artículo debido a la falta de conocimiento respecto a la forma de optimizar los hiperparámetros. En este artículo se trabajó con los datos de los vinos tinto y blanco de forma separada, siendo que en un trabajo más completo se debería considerar los resultados que se obtendrían al juntar ambos conjuntos de datos. Por último, queda pendiente considerar otros criterios para la obtención de mejores resultados de rendimiento al quitar una o más columnas del dataset, además de probar cross-validation con k=n.

REFERENCIAS

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

[2] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[3] Hawkins, D. M. (2004). The problem of overfitting. Journal of chemical information and computer sciences, 44(1), 1-12.

[4] Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. arXiv preprint arXiv:1811.12808.

[5] V. Kumar. Prediction of quality of Wine. https://www.kaggle.com/vishalyo990/prediction-of-quality-of-wine