Clasificación de Calidad del Vino Rojo por Medio de Random Forest

Luis Orellana Altamirano.

Universidad de Santiago de Chile, Chile
Avenida Ecuador #3659. Estación Central, Santiago de Chile., Chile
luis.orellana.a@usach.cl
fernando.cabrera.ga@usach.cl

Abstracto. El presente trabajo de investigación está centrado principalmente en la calidad del vino en su variedad tinto (rojo). La "calidad" del vino, la cual es típicamente medida, tiene componentes subjetivos, los cuales afectan de mayor o menor medida al sabor. Sin embargo, la principal pregunta planteada en nuestro estudio, ¿es posible ver alguna concordancia entre la calidad acusada por catadores y los índices tomados por laboratorios fisioquimicos? . Para tal efecto, es propuesto el uso de Random Forest, el cual permite realizar clasificaciones según el etiquetado de cada uno de los vinos, de esta manera, buscar semejanza entre las medidas fisioquimicas, y la anticipación de los etiquetados según cada agrupación.

Palabras Clave: Medidas de Calidad, Vino Tinto, Random Forest, Clasificacion.

1 Introducción

Los inicios de la producción de vino se remonta a épocas anteriores de los romanos. En Grecia el vino era loado por poetas y artistas. Incluso hoy en día es considerado un producto infaltable en las mesas. A través de los años, debido a su creciente demanda, el vino se ha industrializado, llevando consigo, medidas de calidad. Por consiguiente, se vuelve imperativo conocer más a fondo las variables que determinan la calidad de este. Si se va más a fondo en este tema, para el presente estudio se aborda la problemática utilizando un dataset de índole público, el cual es facilitado por el repositorio dispuesto por UCI [1]. Este dataset está compuesto por una muestra total de 1599 instancias, los cuales corresponden a impresiones subjetivas que obtuvieron catadores de distintos vinos de Portugal, pertenecientes a la variedad roja (vino tinto). Las variables que son presentadas están en el dataset se pueden ver en la tabla 1. Con la finalidad de determinar si es posible clasificar vinos según su calidad percibida versus cualidades fisioquimicas con un mínimo de error, se ha optado por clasificación de vinos de tipo tinto (rojo) por medio del modelo random forest, el cual permite anticipar etiquetas

(atributo experto) de nuevas instancias por medio del aprendizaje efectuado sobre muestras obtenidas del dataset.

Tabla 1. Medidas fisicoquímicas para vino tinto.

Atributos	Min	Max	Media
Fixed acidity (g(tartaric acid)/dm3)	4.6	15.9	8.3
Volatile acidity (g(acetic acid)/dm3)	0.1	1.6	0.5
Citric acid (g/dm3)	0.0	1.0	0.3
Residual sugar (g/dm3)	0.9	15.5	2.5
Chlorides (g(sodium chloride)/dm3)	0.01	0.61	0.08
Free sulfur dioxide (mg/dm3)	1	72	14
Total sulfur dioxide (mg/dm3)	6	289	46
Density (g/cm3)	0.990	1.004	0.996
pH	2.7	4.0	3.3
Sulphates (g(potassium sulphate)/dm3)	0.3	2.0	0.7
Alcohol (vol.%)	8.4	14.9	10.4

2 Método y Herramientas

2.1 Procesamiento de los Datos

Antes de enfrentar el problema y contestar la pregunta de investigación, se utilizará una serie de acciones de procesamiento de los datos, con la finalidad de sacar más provecho al método random forest.

Eliminar Outliers

Debido a que los datos presentan una gran cantidad de outliers, se han eliminado con la finalidad de obtener distribuciones más homogéneas y similares a distribuciones normales. De esta manera se evita el gran sesgo en el clasificador.

Normalizar Datos

Es necesario considerar que los arboles de decisión (de los cuales random forest está compuesto) necesitan la normalización de sus datos para realizar cálculos de máxima entropía, por lo cual después de la normalización de estos, todos los atributos estudiantes resultan en cifras entre cero y uno.

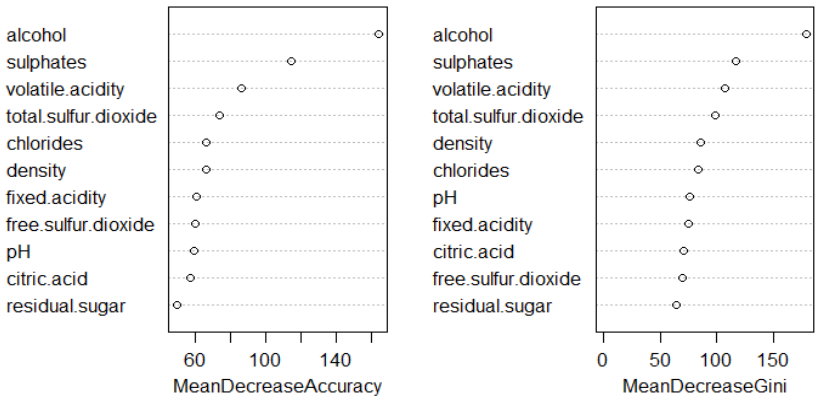
2.2 Método de Minería de Datos

Con la finalidad de obtener conocimiento por medio de técnicas de minería de datos, se utilizará el método Random Forest proveniente de un paquete del software R. Este software es de código abierto, multiplataforma y de programación de alto nivel, el cual trata como caja negra operaciones vectoriales y matriciales. Específicamente, para el análisis de este paper, se utilizaron las librerías "party" y principalmente la librería "randomForest" la cual contiene el método de clasificación. Este método hace uso de árboles de decisión, los cuales, por medio de la máxima entropía, clasifican nuevas instancias según los atributos estudiantes versus atributos expertos. Por otra parte, bosque aleatorio utiliza muchos árboles de decisión a la vez y efectúa muestreos de los datos para cada uno de los árboles. De esta manera, cada una de las instancias tienen chances de ser incluidas en cada muestreo de un 63,2%. Por otra parte, el aprendizaje de los arboles es efectuado por medio de las instancias utilizadas versus las no utilizadas (37%), a esto se le llama "Out of Bag", con las cuales se valida cada modelo generado por el bosque.

2.3 Selección de Parámetros

Haciendo uso del método "Random Forest", se obtuvieron 9 modelos clasificadores, donde por cada uno de ellos, se realizaron 5 pruebas. Posteriormente a ello, fueron seleccionados los mejores valores de cada uno. Los parámetros de entrada para el método fueron 1.000 árboles aleatorios y el número de variables a escoger en el sub-espacio aleatorio de cada nodo estuvieron comprendidas entre los valores 3 a las 11, dando como resultado, la medida de menor error de "Out of Bag" con 9 variables.

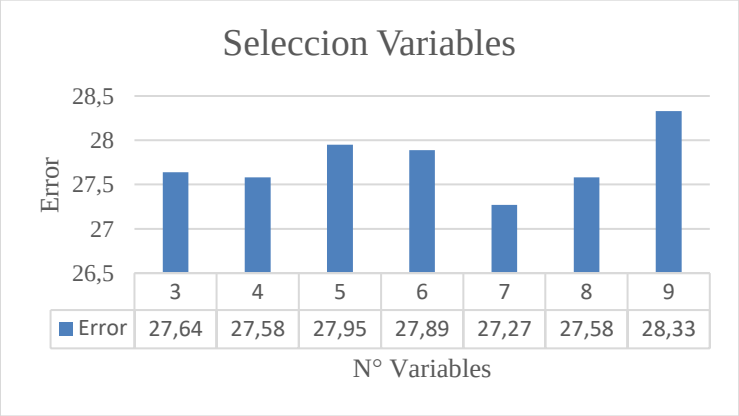
Imagen 1. Importancia de las Variables.



A modo de experimentación, se eliminaron tres variables, las cuales son las menos importantes según los resultados de random forest y que tienen menor correlación con

el atributo experto (según muestra la imagen 1), las cuales son “residual.sugar”, “citric.acid” y “fixed.acidity”, obteniendo como resultado la eliminación de sólo las dos primeras anteriormente nombradas considerando la métrica del erro para tomar esta decisión. Pudimos apreciar que el error bajo de un 27,83% a un 27,64%. Nuevamente se realizó la comprobación de ver cuántas variables debe escoger cada modelo en el sub-espacio aleatorio de nodos, obteniendo los resultados de la imagen 2. Cada una de las pruebas se realizó cinco veces con la finalidad de obtener el error más bajo. Con eso, se consiguió reducir el error de un 27,64% a un 27,27%. Y por último se consiguió un clasificador que permitió procesar cada instancia como se muestra en la tabla 2.

Imagen 2. Métrica de OOB para n Variables.



Por último, se comprobó la cantidad de árboles a generar, y los resultados sugieren que al aumentar la cantidad de árboles no existe una mejora en el error OOB. Esto se puede apreciar al observar los resultados generados por medio de 10.000 árboles, en los cuales el error permanece constante.

2.4 Variables Representativas

Es necesario recordar la pregunta de investigación la cual trata de abordar si es posible clasificar la calidad de vinos con respecto a sus medidas fisioquimicas, y cuan relacionadas están estas con las apreciaciones de los catadores y los vinos degustados. Tales métricas van desde el 0 al 10, donde 0 es un vino de muy mala calidad y 10 de muy buena calidad.

Por otra parte, es importante determinar las variables que representan la calidad del vino. Ante esto, se tienen dos aristas importantes. Una es la percepción del sabor, y las otras variables que son recurrentemente utilizadas por laboratorios fisicoquímicos. En este último, las variables que representan mayor sabor al vino según Doug Nierman (2004) son:

- Acidez Fija.
- Acidez Volátil.
- Ácido Cítrico.

- Ph.
- Alcohol.

Por otra parte, el análisis sensorial es el menos comprendido de los sentidos humanos [3], y además no es bastamente conocido ni entendido aún [4]. Esto hace que el problema sea bastante complejo de abordar. Sin embargo, los resultados obtenidos serán analizados para determinar si la calificación de los catadores corresponde a las medidas químicas planteadas por Gavin Sacks y David Jeffery (2016). Ante esto, por medio del método random forest se obtuvieron una serie de variables importantes, las cuales son según su grado de importancia esta ordenada como:

- Alcohol.
- Sulfatos.
- Acidez Volátil.
- Dióxido de Sulfuro total.
- Densidad.

Tabla 2. Matriz de Confusión Según Resultados Finales Obtenidos.

	3	**4**	**5**	**6**	**7**	**8**	**Error OOB**
3	**0**	**0**	**7**	**3**	**0**	**0**	100%
4	1	0	36	16	0	0	100%
5	0	1	559	117	4	0	17,9%
6	0	2	123	481	32	0	24,6%
7	0	0	7	70	121	1	39%
8	0	0	0	8	8	2	88,8%

2.5 Resultados Obtenidos (discusión)

En relación a las variables que fueron consideradas importantes por random forest tenemos que la de mayor relevancia es el Alcohol. En la información del apartado estadístico del primer laboratorio, pudimos observar la correlación entre las variables del dataset. La variable que presenta mayor correlación con la calidad del vino es el alcohol (0.48).Es observable su importancia en Random Forest ya que presenta los valores más altos de las métricas de MeanDecreaseAccuracy y MeanDecreaseGini.
Según la investigación efectuada por King, E (2012), es destacable la importancia de las concentraciones de alcohol en la percepción del vino tinto en los análisis sensoriales.

Las variables que fueron eliminadas por la escasa relevancia fueron: residual sugar y citric acid. En el análisis estadístico se puede observar que el coeficiente de correlación de Spearman es extremadamente bajo para ambas (0,03 y 0,21 respectivamente).

Los parámetros escogidos para el modelo random forest fueron: número de árboles: 1000 y 9 variables escogidas para la generación del sub-espacio aleatorio. Se consideró el árbol con error mínimo de (27,27%) para la selección de estos parámetros.

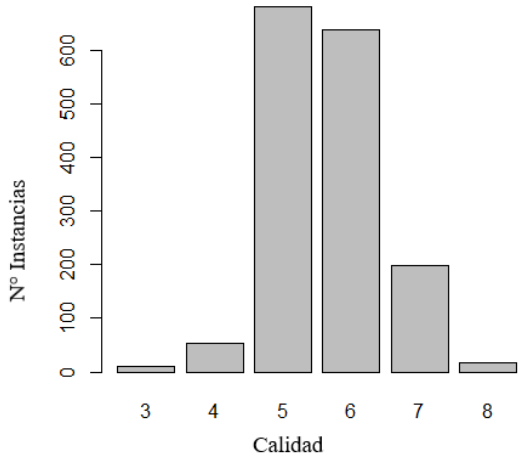
En la Tabla 2 se muestran que los resultados del clasificador para las categorías etiquetadas con valores iguales a 3, 4 y 8, contabilizan una gran cantidad de errores. Los

resultados del clasificador para las categorías 5,6 y 7 contabilizan poca cantidad de errores, fluctuando el error entre los valores 17,9% a 39%.

Como se puede apreciar en la Imagen 3, al observar cómo se distribuyen las instancias con respecto a las clases podemos percatarnos que existe desbalance, por lo que las características no proporcionan mucha información para distinguir entre las clases.

Para este tipo de escenarios, en donde existe un gran desbalance entre las clases pueden ser abordados a través de otras variantes del método random forest[6].

Imagen 3. Volumen de instancias por cada calificación.



3 Conclusiones

Existe concordancia entre la selección de variables de importancia para random forest, con la calidad del vino tinto acusada por los catadores, obteniendo en ambos casos que el alcohol es importante en la clasificación del método como también en los análisis sensoriales realizados por los catadores de vino.

Para el caso de la variable residual sugar tanto en random forest como en el análisis estadístico plantean que tiene poca relevancia. El valor de residual sugar que es considerado aceptable en un vino tinto va entre 2 g/L a 3 g/L e indica que tan dulce es el sabor del vino. La opinión del especialista técnico Tim Gaiser, indica que el olfato es el sentido más importante en la evaluación de la calidad del vino, por lo que no se establece un relación directa entre las azucares residuales con el proceso de evaluación realizado por los catadores del vino tinto [7].

4 Referencias

[1] A. Asuncion, D. Newman, UCI Machine Learning Repository, University of California, Irvine, 2007 http://www.ics.uci.edu/~mlearn/MLRepository.html.

[2] Gavin Sacks and David Jeffery, Understanding Wine Chemistry, University of California, Irvine, 2016 *http://waterhouse.ucdavis.edu/*.

[3] Cortez,P.,Teixeira,J.,Cerdeira,A.,Almeida,F.,Matos,T.,Reis,J. (2009). *Using Data Mining for Wine Quality Assessment*, Springer-Verlag Berlin Heidelberg 66-79.

[4] D. Smith, R. Margolskee, (2006). *Making sense of taste*, Scientific American, Special issue 16 (3) 84–92.

[5] King, E. ,. (2012). The influence of alcohol on the sensory perception of red wines. *ElSEVIER*.

[6] Chen, C. ,Liaw A.,Breiman,L. (2004). Using Random Forest to Learn Imbalanced Data.

[7] Gaiser,T. Using the Deductive Tasting Technique. Obtenido de: http://www.timgaiser.com/how-to-taste-wine.html