Taller 2: Clasificación de vinos según sus características físico-químicas usando Random Forest

Franco Cornejo Gonzalez (4 horas)
Departamento de Ingeniería Informática
Universidad de Santiago de Chile, Santiago, Chile
franco.cornejo.g@usach.cl

***Resumen*—En este trabajo se clasifica la calidad de los vinos según sus características fisicoquímicas, utilizando el algoritmo Random Forest. Se implementó un rebalanceo de las clases para mejorar la precisión del modelo. Random Forest también permitió identificar la importancia de las variables clave en el proceso de clasificación permitiendo reducir las dimensiones del problema.**

***Palabras claves*—Random Forest, Rebalanceo**

1. Introducción

El vino, que antes era visto como un lujo, ahora lo disfruta un público mucho más amplio. La industria vitivinícola está invirtiendo en nuevas tecnologías para mejorar tanto la producción como la venta, donde la certificación y evaluación de calidad son clave. Estas prácticas ayudan a prevenir la adulteración y aseguran que se mantengan altos estándares de calidad, además de facilitar la producción y la fijación de precios (Cortez, Cerdeira, Almeida, Matos, and Reis, 2009).

La certificación incluye pruebas físico-químicas y sensoriales, aunque clasificar el vino es un desafío, ya que depende del trabajo de los enólogos. Pero gracias a los avances en tecnología de la información, se pueden analizar grandes volúmenes de datos, lo que mejora la toma de decisiones en la industria. Las técnicas de minería de datos, como **Random Forest** pueden ser de utilidad para reconocer patrones sobre la calidad de los vinos. Este estudio aplica Random Forest para clasificar vinos.

Se utiliza el dataset ”Wine Qualityçuenta con 1499 observaciones de distintos vinos con 11 variables numéricas sobre su composición físico-químicas . El dataset fue creado en la Universidad de Minho en Portugal para facilitar la investigación en el área de modelado de preferencias del vino basado en sus propiedades físico-químicas (Cortez et al., 2009).

1.1. Objetivo

El objetivo del presente trabajo es obtener una buena clasificación de los vinos y determinar cuales son las características físico-químicas mas importantes para determinar la calidad de un vino.

2. Metodo y datos

2.1. Descripción de la base de datos

La base de datos ”Wine Quality”fue creada para facilitar la investigación en modelado de preferencias del vino basado en sus propiedades físico-químicas. Este dataset busca servir para aplicar técnicas de minería de datos para predecir la calidad del vino utilizando pruebas analíticas fácilmente accesibles. Incluye información sobre vinos tinto, incorporando tanto pruebas físico-químicas como sensoriales, lo que la convierte en una herramienta útil en la investigación de la enología.

2.2. Descripción de las variables

A continuación se describen algunas de las variables más relevantes del dataset:

- **Acidez Fija (fixed acidity)**: Ácidos naturales presentes en el vino por ejemplo tartárico, málico, cítrico, succínico y láctico, medidos en gramos por litro.
- **Acidez Volátil (volatile acidity)**: Producida durante la fermentación, depende de la actividad de las bacterias lácticas y puede afectar el sabor y estabilidad del vino.
- **Ácido Cítrico (citric acid)**: Ácido presente en las uvas.
- **Azúcar Residual (residual sugar)**: Cantidad de azúcar que queda después de la fermentación.
- **Cloruros (chlorides)**: Anión natural presente en diversas fuentes de agua.
- **Dióxido de Azufre Libre (free sulfur dioxide)**: Actúa como conservante en el vino, protegiéndolo.
- **Dióxido de Azufre Total (total sulfur dioxide)**: Suma del azufre libre y ligado, el segundo no disponible para actividad antimicrobiana o antioxidante.
- **Densidad (density)**
- **pH**: Indica el nivel de acidez o alcalinidad del vino. Un pH de 7 es neutro.
- **Sulfatos (sulphates)**: Aditivo que contribuye a los niveles de dióxido de azufre (SO2), actuando como antimicrobiano y antioxidante.
- **Alcohol**: Grado alcohólico volumétrico del vino.
- **Calidad (quality)**: Variable que mide la calidad del vino, con valores de 0 (muy mala calidad) a 10 (muy buena calidad).

Nombre de Variable	Tipo	Valores
Fixed Acidity	Numérico	4.6 - 15.9 g/dm^3
Volatile Acidity	Numérico	0.12 - 1.58 g/dm^3
Citric Acid	Numérico	0 - 1 g/dm^3
Residual Sugar	Numérico	0.9 - 15.5 g/dm^3
Chlorides	Numérico	0.012 - 0.611 g/dm^3
Free Sulfur Dioxide	Numérico	1 - 68 mg/dm^3
Total Sulfur Dioxide	Numérico	6 - 289 mg/dm^3
Density	Numérico	0.99007 - 1.00369 g/cm^3
pH	Numérico	2.74 - 4.01
Sulphates	Numérico	0.33 - 2.00 g/dm^3
Alcohol	Numérico	8.4 - 14.9 % vol
Quality	Categórico	0 - 10

Tabla 1: Descripción de variables del Dataset de Calidad de Vino

Notar la Tabla 2 para comprender la frecuencia por calidad.

Calidad	Frecuencia
1	0
2	0
3	10
4	53
5	681
6	638
7	199
8	18
9	0
10	0
Total:	1599

Tabla 2: Distribución de los datos por *calidad*

2.3. *Preprocesamiento de datos*

En primero lugar se eliminaron los valores atípicos usando el rango intercuartílico y teniendo cuidado de no eliminar los mejores y peores vinos que por su propia naturaleza atípica tendían a ser erradicados. En segundo lugar se balancearon las clases reclasificando en vinos de **Alta** y **Baja** calidad. Para esto se asignaron las clases originales 5, 4 y 3 a la clase **Baja** y las 6, 7 y 8 a la clase **Alta** véase la Tabla3.

Calidad	Frecuencia
Baja	566
Alta	635
Total:	1201

Tabla 3: Distribución de los datos reclasificados *calidad*

2.4. *Análisis estadístico exploratorio*

Para determinar que variables son mas importantes en la clasificación se usaron diagramas de cajas, un diagrama por variable y una caja por categoría Figura 1. Mirando atentamente cada variable resaltan las variables de alcohol y sulfatos en las cuales las cajas se notan mas separadas a diferencia de las demás variables donde las cajas están sobrepuestas. Otra cosa a notar es que las variables con menos dispersión son la de cloruros y azúcar residual.

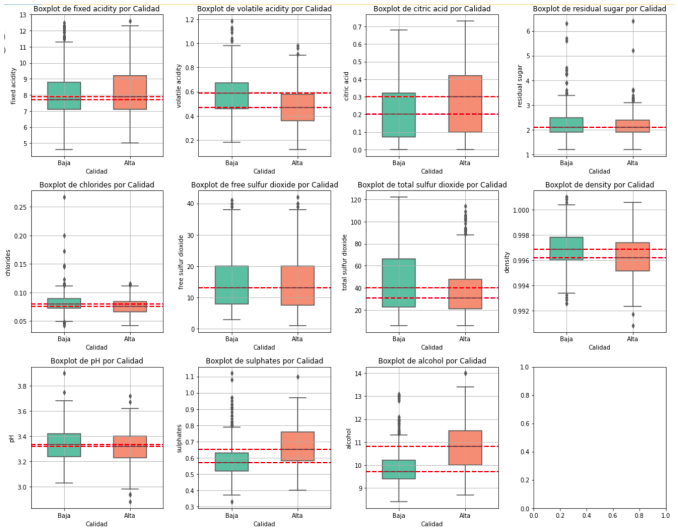

Figura 1: Variables por Calidad

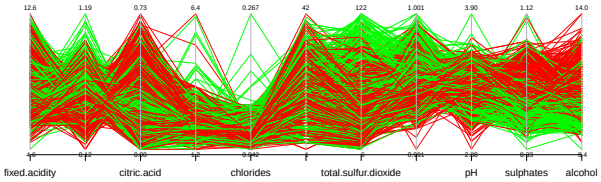

Figura 2: Gráfico coordenadas paralelas

2.5. *Análisis de coordenadas paralelas*

Los gráficos de coordenadas paralelas son una herramienta útil para visualizar datos multivariados. Cada eje representa una variable, y las observaciones se dibujan como líneas que conectan los valores en cada uno de estos ejes.

En la Figura 2 se presenta un gráfico de coordenadas paralelas donde las variables están ordenadas tal como se describe en la sección de Descripción de variables.

Al igual que en los gráficos de cajas, se observa una mayor concentración de vinos de baja calidad con bajos niveles de alcohol y sulfatos. Además, el gráfico revela una diferencia más clara en la acidez volátil, siendo los vinos de alta calidad aquellos con menor acidez volátil. Por último, se aprecia una distinción en el dióxido de azufre total: muchos vinos de baja calidad presentan niveles altos de este compuesto, lo que no era evidente en los gráficos de cajas.

2.6. *Método usado*

2.6.1. ***Arboles de decisión****:* Los árboles de decisión son métodos de aprendizaje automático que realizan clasificaciones de datos a través de la selección recursiva de variables. Su estructura es fácil de entender y permite derivar reglas, lo facilita su interpretación. Sin embargo, tienden a sobreajustarse a los datos de entrenamiento y al centrarse en una sola variable a la vez, pueden tener una capacidad limitada para clasificar y separar en comparación con enfoques más sofisticados, como las redes neuronales o las

máquinas de vectores de soporte (SVM). Por estas razones, a menudo se les considera clasificadores menos robustos (Chacón, 2017).

2.6.2. ***Métodos de ensamble:*** Dos enfoques son el Bagging y Boosting. El Bagging selecciona muestras de datos con la misma probabilidad y opera de manera paralelo para entrenar múltiples modelos que "votan" para determinar la respuesta, mientras que Boosting asigna pesos a las observaciones y se enfoca secuencialmente en generar modelos que corrigen el error del modelo anterior (Chacón, 2017).

2.7. Random Forest

Los Bosques Aleatorios o Random Forest son una implementación de Bagging usando Arboles de decisión. Este método combina la predicción de múltiples árboles entrenados en subconjuntos aleatorios del conjunto de datos y en subconjuntos aleatorios de características. Cada árbol en el bosque es entrenado de manera independiente, utilizando muestreo con reemplazo (bootstrap) para generar diferentes conjuntos de entrenamiento. Al hacer predicciones, cada árbol emite un voto y la clase más votada se adopta como la predicción final del modelo. Además, Random Forest ofrece medidas de importancia de las variables, lo que ayuda a identificar cuáles son las más influyentes en la toma de decisiones del modelo. Utiliza el error del conjunto Out-of-Bag (OOB) que son las observaciones no utilizadas para entrenar cada árbol, lo que permite estimar el rendimiento del modelo (Chacón, 2017).

Hiperparámetros principales de Random Forest:

- **ntree:** Número de arboles.
- **mtry:** Número de variables aleatorias consideradas en cada división de los árboles. Se calcula como la raíz cuadrada del número de variables, más o menos uno, es decir, $mtry = \sqrt{p} \pm 1$, donde p es el número total de variables (Chacón, 2017).

3. Resultados

3.1. Aplicación de Random Forest y selección de características

Se aplica Random Forest sobre el conjunto de datos usando diferentes semillas y tras un largo proceso de búsqueda de hiperparámetros que tenia por fin encontrar estabilidad y un bajo OOB se llego a los siguientes: **ntree = 38000**, **mtry = 3** y una **semilla = 1089**. Con estos se llego a un modelo con un **OOB = 17.82 %** y su estabilidad puede verse en la Figura 3. Analizando el gráfico de estabilidad puede notares que ambas calidades comienzan con un error grande y requiere muchos arboles estabilizar ambas a la vez. Los vinos de alta calidad presentan un error del 18.7 % y los de calidad baja un error del 16.7 %.

Para escoger las características se procedió a eliminar de una en una las variables menos importantes(según la Figura 4) notando que al eliminar las 5 menos importantes (azúcar residual, dióxido de azufre libre, ácido cítrico, pH y acidez fija) el OBB aumento aproximadamente 1 %, se continuo eliminando las siguientes 3 variables menos importantes

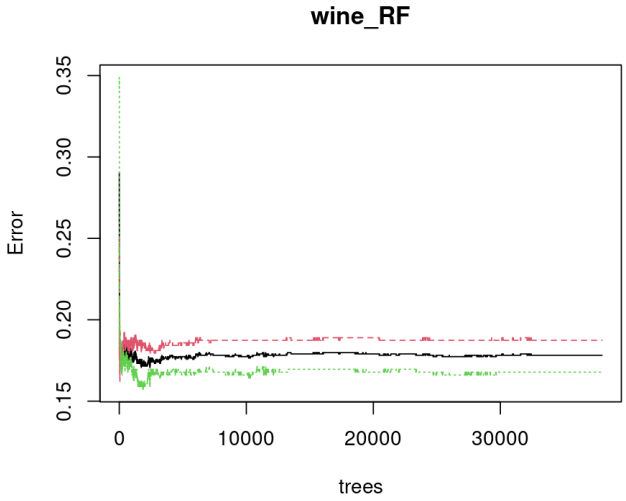

Figura 3: Gráfico de estabilización: ntree = 38000, mtry = 3 y semilla = 1089

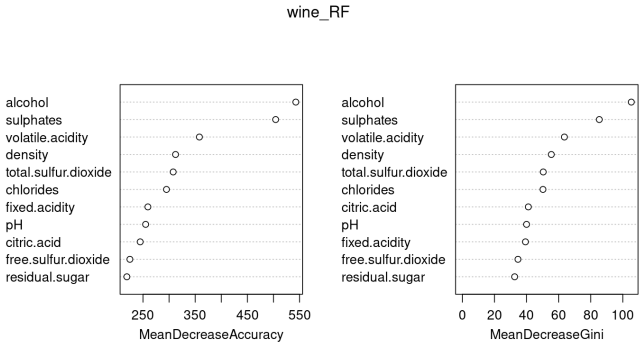

Figura 4: Gráfico Accuracy e Gini, importancia de las variables.

(cloruros, dióxido de azufre total y densidad) y el OBB subió un 2 %. Notar que los porcentajes anteriores dependen de las semillas y los parámetros y son un aproximado de lo que se llego a observar en la experimentación.

Finalmente se seleccionaron las variables acidez volátil, sulfatos y alcohol, lo que reduce la complejidad del problema en gran medida ganando parsimonia e incluso teniendo la posibilidad de graficar en 3 dimensiones.

3.2. Aplicación de Random Forest con las características seleccionadas

Se vuelve a aplicar Random Forest con las siguientes variables **alcohol**, **sulfatos** y **acides volatil**. Con los parámetros **ntree = 14900**, **mtry = 1** y una **semilla = 11**. Con estos parámetros el **OOB** fue de 20.4 % y se logro estabilizar con 14900 (Figura 5) arboles que son menos de la mitad de los que se necesitaba antes.

3.3. Comparación

Al compara los dos modelos de Random Forest es fácil notar que reducir a 3 características se gano mucho en parsimonia

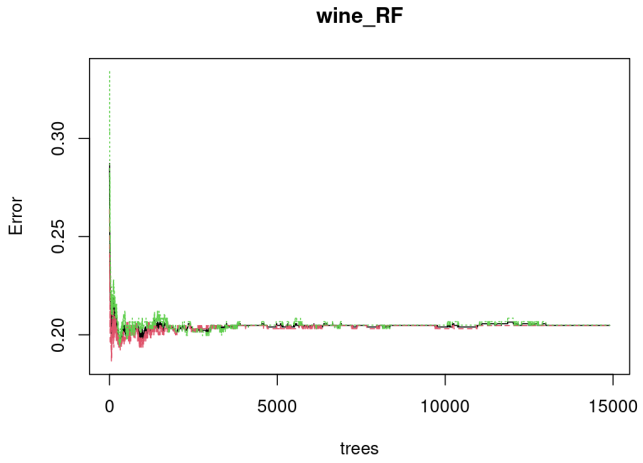

Figura 5: Gráfico de estabilización: ntree = 14900, mtry = 1 y semilla = 11

porque se redujo la cantidad de arboles a menos de la mitad y la cantidad de variables a casi un cuarto de las originales y tan solo sacrificando un 2.58 % del OOB (ver Tabla 4).

Calidad Vino	OOB 11 variables	OOB 3 variables
Baja	16.7	20.1
Alta	18.7	20.6
Promedio	17.82	20.4

Tabla 4: Comparación OOB con y sin selección de características.

A modo de prueba se hizo un modelo Bagging con 400 regresiones logísticas (los resultados pueden verse en el Anexo 6). Lo principal es un F1-score de un 70 % que es peor a lo obtenido con Random Forest. No se hace hincapié en esto ya que no es el objetivo de este trabajo y este Bagging podría mejorarse mucho.

```
Matriz de Confusión:
[[117  41]
 [ 56 147]]

Informe de Clasificación:
{'0': {'precision': 0.6763005780346821, 'recall': 0.740506329113924, 'f1-score': 0.7069486404833837, 'support': 15
8}, '1': {'precision': 0.7819148936170213, 'recall': 0.7241379310344828, 'f1-score': 0.7519181585677749, 'support':
203}, 'accuracy': 0.7313019390581718, 'macro avg': {'precision': 0.7291077358258518, 'recall': 0.7323221300742033,
'f1-score': 0.7294333995255793, 'support': 361}, 'weighted avg': {'precision': 0.7356903455228118, 'recall': 0.73130
19390581718, 'f1-score': 0.7322362088244678, 'support': 361}}
```

Figura 6: Bagging Regresión Logística

4. Conclusiones

En este trabajo se logró una buena clasificación de los vinos según sus características físico-químicas, a pesar de la dificultad inherente del problema. Los resultados consistentes se alcanzaron gracias a un efectivo rebalanceo de las clases, lo cual permitió manejar la desproporción entre las categorías de calidad.

Además, se comprendió la gran utilidad de Random Forest no solo como un clasificador robusto, sino también como un método poderoso para determinar la importancia de las variables. A través de este enfoque, fue posible identificar las variables más relevantes que influyen en la calidad del vino.

Otro aspecto importante es la parametrización que fue clave para mejorar el rendimiento del modelo. La optimización de los hiperparámetros del Random Forest permitió ajustar el modelo de manera eficiente y la búsqueda de buenas semillas, logrando una notable reducción en el error out-of-bag (OOB).

Referencias

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Modeling wine preferences by data mining from physicochemical properties," *Department of Information Systems/R&D Centre Algoritmi, University of Minho*, 2009, viticulture Commission of the Vinho Verde Region (CVRVV), 4050-501 Porto, Portugal.

[2] M. Chacón, "Taller de minería de datos avanzada: Capítulo ii "bosques aleatorios"," in *Taller de minería de datos avanzada PPT*, 2017.