Taller 2: Clasificación de vinos según sus características físico-químicas usando Support vector machine

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

***Resumen*—En este estudio, se clasificaron vinos según sus características físico-químicas utilizando máquinas de vectores de soporte (SVM). Se empleó el dataset "Wine Quality", que contiene 1,599 observaciones y 11 variables numéricas relacionadas con la composición del vino. Tras un preprocesamiento que incluyó la eliminación de valores atípicos y el rebalanceo de las clases, las etiquetas de calidad original se agruparon en dos clases: alta y baja calidad. Se evaluaron modelos de SVM con diferentes configuraciones de parámetros (C y γ) y técnicas de reducción de dimensionalidad para identificar las variables más relevantes. El modelo escogido fue uno de kernel radial cuya precisión fue del 77 % contando con 6 variables.**

***Palabras claves*—Rebalanceo, SVM, Kernel**

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 las **Maquinas de vectores de soporte** pueden ser de utilidad para reconocer patrones sobre la calidad de los vinos. Este estudio aplica SVM 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).

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.

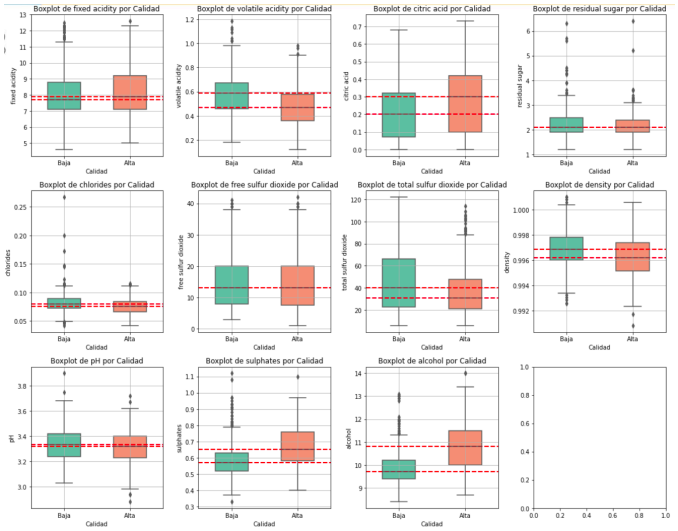

Figura 1: Variables por Calidad

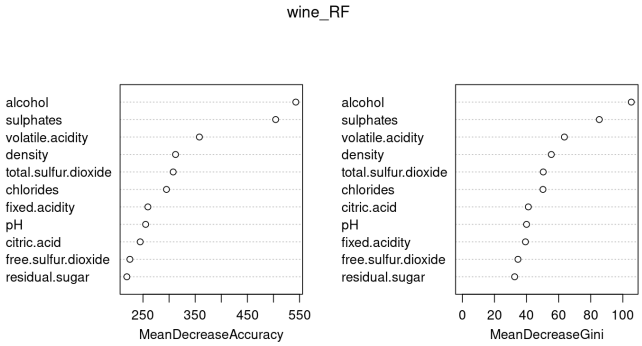

Figura 2: Gráfico Accuracy y Gini, importancia de las variables.

2.5. *Reducción de variables*

Para estudiar la reducción de variables se hizo uso de lo aprendido en el Taller 2, donde el resultado de la importancia de variables de la Figura 2. En dicho trabajo se llego a la conclusión de que eliminando las 6 variables menos importantes tan solo se comprometía un 6 % del error fuera de la bolsa.

Ademas mediante el estudio de la literatura del problema se encontró un gráfico de la importancia de variables echo a partir de SVMs de kernel radial (Figura 3). La principal diferencia entre este análisis y el echo mediante RF fue en las variables de PH, alcohol y densidad. Estando estas 3 variables en puestos muy diferentes y en el caso de la densidad y el pH opuestos.

2.6. *Método usado*

2.6.1. ***Maquinas de vectores de soporte:*** Las Máquinas de Vectores de Soporte (SVM, por sus siglas en inglés) son un poderoso método de clasificación supervisada utilizado en minería de datos y aprendizaje automático. Su objetivo es encontrar un hiperplano óptimo que divida un conjunto de datos en diferentes clases, maximizando la separación entre

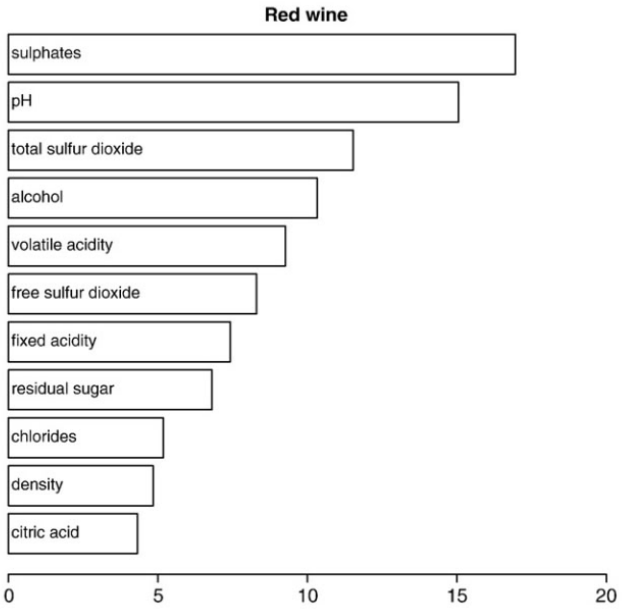

Figura 3: Importancia de las variables según la literatura

ellas. En su forma mas simple permite separa datos linealmente separables, aunque también pude manejar casos mas complejos.

El concepto central de las SVM es el hiperplano separador que divide los datos de tal manera que maximice el margen entre las clases. El margen es la distancia más pequeña entre el hiperplano y los puntos más cercanos de cada clase, denominados vectores soporte. Estos puntos son esenciales porque definen completamente la posición del hiperplano y el resto de los datos no son usados (Chacón, 2017).

En su forma lineal, las SVM buscan un hiperplano que pueda separar las clases de datos de manera óptima cuando estas son linealmente separables. Este enfoque utiliza un parámetro llamado C (costo), que controla la penalización por errores de clasificación en los datos de entrenamiento.

- Un C alto: El modelo intenta clasificar correctamente la mayor cantidad de datos posibles, lo que puede llevar a un sobreajuste al enfocarse en puntos atípicos o ruido.
- Un C bajo: Permite mayor flexibilidad, ignorando ciertos errores para priorizar un margen más amplio, lo que fomenta una mayor generalización del modelo.

Para problemas donde las clases no son separables de manera lineal, las SVM pueden extenderse mediante un kernel radial (RBF, Radial Basis Function). Este kernel somula una transformación a un espacio de características de mayor dimensionalidad donde la separación lineal puede ser posible. El comportamiento del kernel radial se define mediante el parámetro γ. Un bajo γ nos dice que el hiperplano separador tiende a la linealidad y por tanto el problema original es de naturaleza lineal.

Como las SVM son muy sensibles a una mala selección de parámetros, es necesario ajustar cuidadosamente C y γ. Una buena estrategia es realizar una búsqueda en grilla, donde se prueban combinaciones de estos parámetros en un

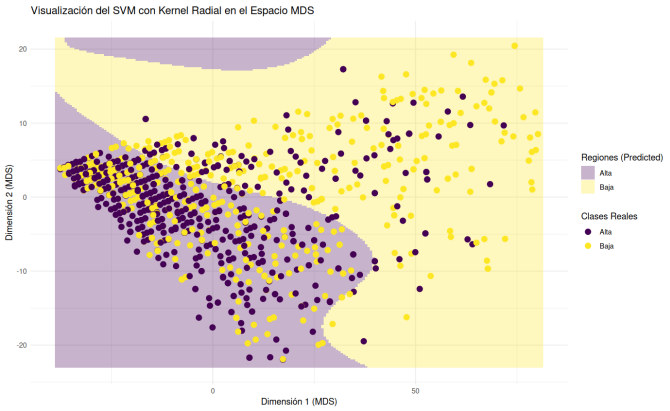

Figura 4: MDS modelo Radial-1

rango definido y se evalúa su desempeño mediante validación cruzada. Este método es computacionalmente demandante, pero si se utiliza correctamente, es decir, ajustando las ventanas de parámetros de manera conveniente, permite una selección de parámetros bastante robusta.

3. Resultados

Para escoger los parámetros se hizo una búsqueda en grilla variando tanto los parámetros de costo como los de γ entres 2^{-10} y $2^{P}10$, después se acotaron estos rangos para afinar la búsqueda obteniendo los parámetros que se ven en la Tabla 4. Ademas para todos los parámetros se hizo una validación cruzada de 5 pliegues.

El mejor modelo encontrado fue el modelo Radial-1 que cuenta con una precisión del 80 % y los parámetros para este modelo fueron de un Costo = 2 y un γ = 0.5, el costo bajo nos indica que se toleraron varios errores y no se tendió al sobre ajuste. Por otro lado el gamma pequeño indica que el modelo no era demasiado no-lineal. Para visualizar los resultados se hizo un escalado multidimensional y se proyecto el hiperplano separador sobre este como se puede ver en la Figura 4.

Tipo	C (costo)	γ	n. de variables	Precisión
Lineal	0.015	–	11	**0.72**
Radial-1	2	0.5	11	**0.8**
Radial-2	64	2	5	**0.77**
Radial-3	2	1	6	**0.77**

Tabla 4: Comparación SVM Precisión

Los modelos Radial-2 y Radial-3 fueron echos a partir de una reducción dimensional basada en reducción de variables. Para encontrar las reducción que reducía mas la complejidad del modelo sin comprometer demasiado la precisión se procedió a ir eliminando las variables de una en una hasta llegar al unto donde la precisión baja notoriamente.

Al analizar el modelo Radial-2 que fue echo a partir de una reducción de dimensiones basado en la literatura (2). notamos que eliminando 6 variables tan solo se perdió un 3 % de la precisión sin embargo puede verse un alto Coste que puede indicar un sobre ajuste. Al gratificar el modelo con MDS (Figura 5) el sobre ajuste puede notarse a simple vista, viendo islas con pocos datos en su interior. A pesar de

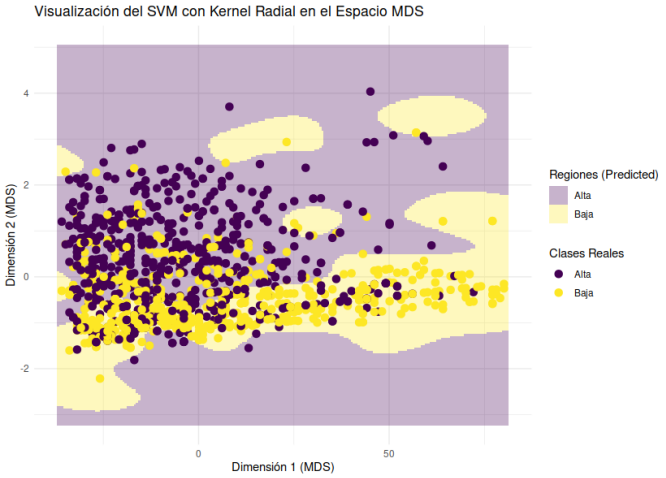

Figura 5: MDS modelo Radia-2

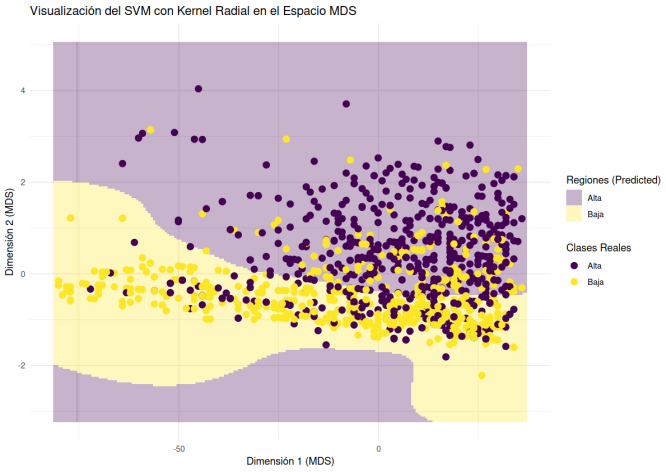

Figura 6: MDS modelo Radial-3

que tan solo viendo la figura no podemos concluir sobre el sobre ajusto dado que es una proyección, si sumamos el dato de los costos podemos determinar que si se sobreajusto el modelo.

Finalmente el modelo que nos pareció mejor fue el Radial-3 que fue confeccionado utilizando el análisis de importancia de las variables echo con RF (Figura 2). Este modelo cuenta con un Costo de 2 y un gamma de 1, parámetros que no parecen acusar ni un sobre ajuste ni una tendencia hacia la linealidad. En este modelo se eliminaron 5 variables sacrificando un 3 % de precisión. Obteniendo una precisión final del 77 %. El gráfico de este modelo puede verse en la Figura 6.

Ademas, notamos que al no relevancia las clases los desempeños de la SVM con dificultad llegaba el 60 % de precisión, si que, no se profundizo en estas cuestiones. Sin embargo fue mucho mejor que el de RF que con todas las clases originales no llegaba al 20 %.

4. Conclusiones

El uso de SVM permitió clasificar eficazmente la calidad del vino a pesar de los desafíos derivados del desbalance y la naturaleza no lineal del problema. El rebalanceo de clases fue crucial para mejorar los resultados, y el ajuste de parámetros mediante búsqueda en grilla optimizó la precisión sin sobreajustar el modelo.

La importancia de las variables fue clave para reducir la complejidad del modelo sin comprometer significativamente la precisión. Se destacó la versatilidad de las SVM y cómo el análisis de sus parámetros puede proporcionar una comprensión más profunda de los datos y su estructura subyacente.

También se noto una mejora considerable frente a RF que no era capas de manejar todas las clases del problema.

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 iv "svm"," in *Taller de minería de datos avanzada PPT*, 2017.