# Un Mecanismo de Clasificación de Calidad del Vino Rojo a través de SVM

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.**

En la industria vitivinícola, la calidad del vino es un aspecto primordial en la producción de este producto. Por tal motivo, indicadores fisicoquímicos son de gran relevancia al momento de medir su calidad. Sin embargo, la habilidad de los catadores cobra real importancia al evaluar la calidad de vino. Es por eso que, en el presente trabajo de investigación, se aborda la pregunta ¿Cuáles son las variables más importantes que inciden en la calidad del vino según índices fisicoquímicos y catadores? Para tal efecto, es propuesto el uso de Support Vector Machine en su modalidad de clasificador. Tras el procesamiento de este método, se obtuvieron resultados que permitieron ver la consonancia que tiene el juicio de catadores y fisicoquímicos luego de realizar selección de importancia de variables por medio del método de selección de variables "Boruta".

**Palabras Clave:** Calidad del Vino, Vino Rojo, SVM, Boruta.

## 1 Introducción

El inicio 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 fisicoquímicas con un mínimo de error, se ha optado por clasificación de vinos de tipo

tinto (rojo) por medio del modelo Support Vector Machine (SVM), el cual permite anticipar etiquetas (clases) de nuevas instancias por medio del aprendizaje efectuado sobre muestras obtenidas del dataset.

| 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 |

### 1.1 Método de Minería de Datos

Con la finalidad de obtener conocimiento a través de las técnicas de minería de datos, es utilizado Support Vector Machine (SVM), el cual, en su modalidad de clasificador, permite para este estudio identificar categorías de vinos (calidad), por medio de las distintas variables en el dataset. La función que permite clasificar nuevas instancias, hace uso de máxima verosimilitud por medio de una función sigmoidal [2] tal como se muestra a continuación:

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

El gran desafío abordado por medio de este método, es la optimización de la ubicación de un hiperplano separador de clases, en donde se toma como referencia los vectores subyacentes a cada una de las clases. Estos vectores son llamados soporte, debido a que permiten la optimización de la distancia de estos vectores de soporte, y el hiperplano que separa las clases. Esto es expresado por medio de la siguiente función de optimización:

$$\min_\theta C \sum_{i=1}^{m} \left[ y^{(i)} cost_1(\theta^T x^{(i)}) + (1 - y^{(i)}) cost_0(\theta^T x^{(i)}) \right] + \frac{1}{2} \sum_{i=1}^{n} \theta_j^2$$

Por otro lado, un supuesto para poder aplicar este método de manera efectiva, es que los datos deben ser linealmente separables. Para lograr este efecto, se utiliza el recurso matemático llamado espacio de Hilbert, el cual permite aumentar la dimensionalidad de las instancias. Con esto se consigue espacio dimensional donde las instancias son linealmente separables y es posible aplicar el producto punto matricial. Para esta investigación se utilizó el kernel de función en base radial [3], la cual es expresada como:

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2),\ \gamma > 0.$$

### 1.2 Procedimiento de Ajuste del Modelo

Con el propósito de obtener un modelo SVM óptimo, se empleó una serie de técnicas encargadas de realizar un tunning de parámetros (función tune de la librería "e1071" de R). Según la investigación realizada por Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin (2010), es necesario localizar los parámetros (C ,γ ) para que el clasificador pueda predecir datos desconocidos con precisión. Para ello, es necesaria la construcción de una grilla de parámetros por cada uno de los kernel y posteriormente localizar los parámetros con el mínimo error por cada kernel. Para cada uno de los kernels, el parámetro de números de conjuntos de validación cruzada utilizado fue 10. En la **Tabla 1**, es posible visualizar los kernels y sus parámetros, los que son comparados con la finalidad de encontrar el mejor tunning posible para la selección del modelo final de SVM.

**Tabla 1.** Resultados para obtención del kernel con parámetros óptimos.

| | **Rango de C** | **Rango de γ** | **Valor óptimo de C** | **Valor óptimo de γ** | **Error de kernel** |
|---|---|---|---|---|---|
| **Lineal** | $2^{-3} a\ 2^{6}$ | - | 2 | | 0,41 |
| **Polinomial** | $2^{-2} a\ 2^{9}$ | - | 32 | | 0,43 |
| **Radial** | $2^{-7} a\ 2^{14}$ | $2^{-7} a\ 2^{12}$ | 1 | 1 | 0,32 |
| **Sigmoidal** | $2^{-7} a\ 2^{14}$ | $2^{-7} a\ 2^{12}$ | 4 | 0.0078 | 0,42 |

Como es posible observar en la Tabla 1, el mejor resultado obtenido fue para el kernel radial con parámetros óptimos (C , γ) de 1 y 1 respectivamente, presentando el mínimo error entre todos los kernels.

### 1.3 Selección e Importancia de las Características

Con el objetivo de disminuir la complejidad del modelo, es posible realizar una selección y verificación de importancia de características a través de la librería Boruta en R. Boruta es un algoritmo envolvente construido alrededor del algoritmo de Random Forest implementado en el paquete R. Es un método conjunto en el cual la clasificación se realiza mediante la votación de múltiples clasificadores débiles imparciales: árboles de decisión [4]. La medida de importancia de un atributo se obtiene como la pérdida de precisión de la clasificación causada por la permutación aleatoria de valores de atributo entre objetos. Aplica una medida de importancia de característica (precisión de disminución media por defecto) para evaluar la importancia de cada característica, donde más alto significa más importante.

Haciendo uso del método "Boruta" en R, usando los parámetros de maxruns=100 que la librería trae por defecto (este parámetro debe ser aumentado en caso de existir atributos marcados como tentativos) los tres atributos con mayor importancia según el método son presentados en la **Tabla 2**.

Cabe destacar adicionalmente que todos los atributos fueron considerados importantes por el método, sin embargo, fue eliminado el atrivuto "fixed. acidity" el cual fue listado en último lugar según su importancia, sin embargo, los resultados no mejoraron luego de efectuar la eliminación de esa variable, por lo que se descarta la eliminación de alguno de ellos.

**Tabla 2.** Resultados para obtener el kernel con parámetros óptimos.

| Atributo | MeanImp | Consideración |
|---|---|---|
| **Alcohol** | 47.97 | Importante |
| **Sulphates** | 34.86 | Importante |
| **Total Sulfur Dioxide** | 32.04 | Importante |

### 1.4 Matriz de confusión

Posteriormente a la selección de los parámetros óptimos del modelo SVM, es verificado cuáles son los resultados de la clasificación del método obtenidos. En la **Tabla 3** se puede apreciar la matriz de confusión resultante a partir de la ejecución del método SVM. El parámetro de conjuntos usado en la validación cruzada para esta prueba fue de 10.

La exactitud obtenida a partir de este modelo corresponde a un 93,56%.

Para las clases (III, IV y VIII) los valores de especificidad son 1 y sus valores de sensibilidad están entre 0.56 y 0.66 respectivamente.

En el caso de las clases (V, VI y VII) los valores especificidad y sensibilidad están entre los valores 0.91 y 0.99.

**Tabla 3.** Matriz de Confusión.

| | **3** | **4** | **5** | **6** | **7** | **8** |
|---|---|---|---|---|---|---|
| **3** | 6 | 0 | 0 | 0 | 0 | 0 |
| **4** | 0 | 30 | 0 | 0 | 0 | 0 |
| **5** | 3 | 13 | 656 | 27 | 1 | 0 |
| **6** | 1 | 10 | 24 | 609 | 15 | 2 |
| **7** | 0 | 0 | 1 | 2 | 183 | 4 |
| **8** | 0 | 0 | 0 | 0 | 0 | 12 |

### 1.5 Resultados Obtenidos (discusión)

Los parámetros escogidos para el modelo SVM corresponden a un tipo de kernel radial, con valores de parámetros óptimos (C , $\gamma$) de 1 y 1 respectivamente. Esta decisión se basa en que el error presentado por el kernel radial es significativamente inferior comparado con el resto de ellos (alrededor de un 10% en comparación a los otros kernels). Además, es importante considerar que a través de la experimentación se obtuvieron mejores resultados con parámetros más altos en la validación cruzada con la función tune de SVM, tomando valores mayores se logró una disminución del error del kernel radial quedando en 32% aproximadamente.

Para el caso de la selección e importancia de las características es posible corroborar que los atributos más importantes obtenidos por el método Boruta corresponden al alcohol y los sulfatos. Todas las variables del modelo son consideradas como relevantes, por lo que no fue eliminada ninguna de ellas. Por otro lado, como se mencionó anteriormente, fue eliminada la variable “fixed.acidity”, sin embargo, el error del clasificador aumento en un 2%, por lo que tal variable y las demás se mantuvieron en el proceso de clasificación.

El resultado obtenido a partir de los parámetros óptimos refleja precisión del modelo de un 93,56%. Para la mitad de las clases que presentan frecuencias bajas (III, IV y VIII), los valores de especificidad son altos (llegando a 1), por lo que para los casos en que esta clase era detectada además era correctamente clasificada.

Para el caso de las clases (V, VI y VII) presentan valores de especificidad y sensibilidad altos (superiores a 0.9). En la matriz es posible observar que el clasificador para algunos casos le costara distinguir entre las clases V y VI los que a nivel del problema podríamos considerarlos como de vinos rojos de calidad regular.

## 2 Conclusiones

Es posible percatarse de la concordancia entre la información obtenida a partir de la selección de variables de importancia a través de boruta, con lo encontrado en la literatura, obteniendo que el alcohol es la variable más importante tanto para la clasificación como también en los análisis sensoriales que miden la calidad. Por ejemplo, en el trabajo investigativo realizado por King, E. (2012), se destaca la importancia de las concentraciones de alcohol en la percepción del vino tinto en los análisis sensoriales.

Los análisis sensoriales se basan en experiencia y conocimiento de los expertos, que son propensos a factores subjetivos, por lo que contar con un enfoque basado en datos que sea preciso puede ser beneficioso como un soporte de decisión ayudando a una correcta decisión en la evaluación de la calidad del vino.

## 3 Referencias

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

[2] Andrew Ng, CS229 Lecture notes, University of Stanford, 2012 http://cs229.stanford.edu/notes/cs229-notes3.pdf.

[3] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin, A Practical Guide to Support Vector Classification, National Taiwan University, 2010.

[4] M. Kursa, W. Rudnicki, Feature Selection with the Boruta Package, University of Warsaw, 2010.

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