Universidad de Sonora
Departamento de Matemáticas
Programa de Maestría en Ciencia de Datos

Proyecto Integrador: Análisis de datos de vinos

Asignatura:

Introducción a la Ciencia de Datos y sus Metodologías

Profesor:

Dr. Juan Martín Preciado Rodríguez

Equipo:

Santiago Francisco Robles Tomayo
Reyna Yanet Hernández Mada
María Elena Martínez Manzanares
Martín José Vega Noriega

Hermosillo, Sonora, a 30 de septiembre del 2022

Resumen ejecutivo

A partir de la base de datos sobre trece variables que influyeron en la calidad del vino tipo vinho verde —"*una variedad de vino de Portugal elaborado en la región de Entre Douro e Minho*" (Wikipedia, 2022b)—, se realizó un análisis exploratorio de datos e implementó la metodología CRISP-DM para construir un modelo Random Forest para discernir la calidad del vino (variable dependiente) a través de sus características (variables independientes), y así brindar una herramiento de toma de decisiones a los productores sobre en qué variables enfocarse para aumentar la calidad de la bebida. Además, se eligió la aplicación *Streamlit* para presentar los resultados, por ser un framework que dispone de herramientas de implantación gratuita, en línea y de programación amigable, con la finalidad de compartir la información en un formato intuitivo y de fácil acceso para los productores. El código fuente del tablero se puede encontrar en el siguiente repositorio de GitHub, para usarse en la siguiente página web.

Índice

1. Comprensión del negocio

1.1. Contexto

Un precedente del uso de datos vinculados a la producción de vino para esta investigación es el trabajo de Cortez et al (2009), quien realizó un experimento de minería de datos usando datos recolectados de mayo 2004 a febrero 2007 con valores atribuidos a vinos tintos y blancos, aunque en su estudio sólo se analizan los datos de blancos. Los registros corresponden con muestras denominadas de *origen protegido*, las cuales fueron analizadas por la Entidad Oficial de Certificación (CVRVV). Los datos se registraron mediante un sistema informático (iLab), que administra los análisis de las muestras desde la solicitud hasta análisis de laboratorio y sensoriales.

En el caso de este proyecto, la base de datos utilizada se obtuvo del sitio web *Machine Learning Repository*, de la Universidad de California en Irvine. Ésta se compone de

> "[...] *dos conjuntos de datos, relacionados con las variantes tintas y blancas del vino portugués "Vinho Verde". En este caso, solo se analiza el dataset de vino blanco. Debido a cuestiones de privacidad y logística, sólo se dispone de variables fisicoquímicas (entradas) y sensoriales (la salida). Por ejemplo, no hay datos sobre tipos de uva, marca de vino, precio de venta del vino, etc*." (Dua & Grph, 2019).

1.2. Objetivos

Objetivos generales:

- Determinar la clasificación del vino a través de sus características.

Objetivos específicos:

- Realizar un análisis multivariado para identificar relaciones de dependencia entre las características del vino.

- Efectuar una implantación en un servicio de host gratuito a través de framework de alto nivel programable en Python.
- Identificar el algoritmo de clasificación o regresión adecuado para discernir la calidad del vino a través de sus características.

1.3. Inventario de recursos

UCI Machine learning repository es un portal creado por la escuela de IT y ciencias computacionales de la Universidad de Irvine, en California. Este repositorio contiene múltiples bases de datos y tablas que son utilizados para ejercicios y análisis de aprendizaje automático. La base de datos de calidad de vinos está registrada como una donación por parte de Paulo Cortez de la Universidad de Minho, Guimarães, Portugal, el 07 de Octubre del 2009, la cual se compone de dos tablas de datos: una correspondiente a los datos de vinos rojos y otra a la de vinos tintos. Para efectos de este proyecto solo se considera la tabla correspondiente a los datos de vino tinto.

1.4. Terminología

- **Variable**: de acuerdo all glosario de términos estadísticos de la OECD (2004), una variable es una característica de una unidad observada que puede asumir más un valor en un conjunto de valores, a los que se puede asignar una medida numérica o una categoría de una clasificación.
- **Característica**: la OECD (2004) define a este concepto como una abstracción de una propiedad de un objeto o de un conjunto de objetos. Las características se utilizan para describir conceptos. También pueden considerarse como los atributos físicos y económicos de un producto que sirven para identificarlo y permiten ubicarlo en alguna partida de una clasificación de productos; los parámetros técnicos y las propiedades determinantes de un producto.

Las palabras característica, variable o columna se utilizan indistintamente en el resto de este trabajo.

1.5. Plan de proyecto

Dadas las características de la base de datos, se considera que un modelo clasificatorio es el adecuado para cumplir el objetivo del presente trabajo, clasificando como variable dependiente la característica “quality” y como variables independientes el resto de las características de la muestra. Es decir, la calidad de la bebida está en función de un conjunto de variables, en su mayoría químicas, relacionadas al proceso de crecimiento y fermentación de la uva.

Tanto el análisis exploratorio de los datos y entrenamiento del modelo se realizaron a través del lenguaje de programación Python, haciendo uso de la librería *Scikit-learn*. La implantación del modelo entrenado se presentará a través de un tablero programado por medio del framework *Streamlit* —disponible en Python— y será puesto a disposición de uso (host) en la nube comunitaria de *Streamlit*.

De esta manera se presentarán los resultados finales bajo un formato didáctico, en línea y de libre acceso para la comodidad de la empresa.

2. Comprensión de los datos

2.1. Descripción de los datos

Las doce variables que componen la tabla son, básicamente, componentes que definen la calidad, sabor y estructura de un vino. De acuerdo a Cortez (2009), las entradas del dataset incluyen pruebas objetivas (por ejemplo, valores de PH), mientras que las salidas se basan en datos sensoriales (mediana de al menos tres evaluaciones realizadas por expertos en vino). Cada experto calificó la calidad del vino entre 0 (muy malo) y 10 (excelente).

Presentamos a continuación el esquema de la tabla.

1. **Fixed acidity** (*acidez fija*): representa la composición de 6 principales ácidos orgánicos presentes en la uva, como son tartárico, málico, acético, láctico, succínico y cítrico. Se mide en miligramos por litro.
2. **Volatile acidity** (*acidez volátil*): grado de gases ácidos en un vino, medido en gramos por litro.
3. **Citric acid** (*ácido cítrico*): los tres principales ácidos que se encuentran en un vino son los tartáricos, málicos y cítricos; este último suele encontrarse en pequeñas cantidades en la uva. Acorde al portal de la compañía Randox Food Diagnostics:

   > "*Se usa como un suplemento ácido durante el proceso de fermentación para ayudar a los productores de vino a incrementar la acidez del producto, especialmente las uvas cosechadas en climas cálidos (...) En la Unión Europea, el ácido cítrico sólo puede ser usado para propósitos de estabilización y el contenido ácido final no debe exceder 1grs/L*" (Mooney, 2019).

   En este sentido, como en el caso de la acidez volátil, puede medirse como gramos por litro.

4. **Residual sugar** (*azúcar residual*): es una forma de azúcares naturales de la uva que se obtiene posterior a la fermentación; entre más azúcar residual, más dulce el vino. Se mide en gramos por litro.
5. **Chlorides** (*cloruros*): es el grado de sales que contiene un vino, medido en gramos por litro.

6. **Free sulfur dioxide** (*dióxido de azufre libre*): es un gas, cuya composición química es SO2, usado en la fermentación de vino como antioxidante, mantener el vino fresco, evitar levadura, entre otros motivos. Suele medirse en miligramos por litro.
7. **Total sulfur dioxide** (*dióxido de azufre total*): "*es una porción de SO2 aislada en el vino más la porción del gas que está unida a otros químicos, como los aldehídos, pigmentos o azúcar*" (Iowa State Unviersity, 2018). Se mide también en miligramos por litro.
8. **Density** (*densidad*): concentración de vino respecto al agua en el producto final, medido en porcentaje.
9. **pH**: dentro de cualquier líquido, "*el pH es el grado relativo de acidez contra el grado relativo de alcalinidad. Suele medirse en una escala de 0 a 14, donde 7 es neutral. Los productores de vino lo utilizan como una variable para determinar la madurez en relación con la acidez*" (WineSpectator, 2009)
10. **Sulphates** (*sulfatos*): son un tipo de conservador, utilizado en los vinos desde el siglo XIX para "*evitar oxidación, prevenir el crecimiento de microorganismos, preservar color, promover el crecimiento de levadura para mejor fermentación y promover el crecimiento de componentes necesarios en uvas*" (Wb, 2019). Suele medirse en miligramos por litro.
11. **Alcohol**: es el alcohol por volumen (ABV, por sus siglas en inglés). Es decir, la cantidad de etanol por volumen del vino, usualmente medido en porcentaje
12. **Quality** (*calidad*): es la calidad del vino determinada a partir del resto de las características. Esta variable está contenida en el rango entero de 0 a 10.

2.2. Análisis exploratorio de los datos

Se realizó el cálculo de las medidas de tendencia central de los datos (ver Tabla 1), donde fue posible notar que, con excepción de las características "free sulfur dioxide" y "total sulfur dioxide", las columnas de datos presentan una desviación estándar pequeña y con rangos de valores no muy amplios . Esto nos indica un comportamiento en cierto sentido regular de la mayor parte de las variables que han sido muestreadas.

Por medio de un diagrama de caja (ver Gráfico 1) se logra apreciar que las variables "free sulfur dioxide" y "total sulfur dioxide" tienen mínimos y máximos definidos por medio del rango intercuartil considerablemente más chicos que el

mínimo y máximo real, resultando de (1.0, 42) y (6,122) respectivamente. En todos los casos, hay presencia de una cantidad considerable de datos atípicos.

Además, un análisis bivariado y multivariado a la muestra fue realizado, por medio del cálculo de correlaciones y del Factor de Inflación de Varianza (también llamado *VIF*, por sus siglas en inglés), respectivamente. Del cálculo de correlaciones, se tomó como regla que una correlación en valor absoluto mayor a 0.70 es significante. Los cálculos obtenidos son presentados en el Gráfico 2, y se hace notar que la mayoría de las características no presentaban un comportamiento fuertemente relacionado, estando todas las correlaciones calculadas dentro del rango (-6.8,6.8).

Del análisis multivariado (ver Gráfico 2), se consideró que una variable no presentaba multicolinealidad cuando su VIF resultaba menor o igual que 10, obteniendo del análisis que solamente las variables “citric acid”, “residual sugar”, “chlorides”, “free sulfur dioxide” y “total sulfur dioxide” obtuvieron VIF aceptable para el análisis.

Por último, se realizó un análisis de balance de datos (ver Tabla 3), donde se hizo notar que la tabla de datos presenta un desbalance importante con respecto a la variable objetivo, ya que se obtuvo que la mayoría de los registros corresponden con los niveles de calidad 5 y 6, representando el 85% de la muestra, no habiendo registros con calidad de 0 al 2 y, 9 y 10. El imbalance de datos puede comprometer el proceso del entrenamiento del modelo clasificatorio, por lo que resulta necesario efectuar una técnica de remuestreo en el proceso de tratamiento de los datos y, por otro lado, la inexistencia de registros de las calidades más altas y bajas nos indica que el modelo no tendrá la capacidad de identificar calidades dentro de estos rangos extremos.

A partir de lo estudiado, se concluyó que para poder hacer una clasificación de la variable “quality” es necesario considerar múltiples características dentro del modelo, del cual se conjetura factible la eliminación de una cantidad considerable de variables para realizar el modelado final. A su vez se conjetura como necesario en el preprocesamiento del desbalance de datos con respecto a la variable target.

Tabla 1: Datos con medidas de tendencia central básicas de los datos desglosados por columnas.

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
Media	8.31	0.52	0.27	2.53	0.08	15.87	46.46	0.99	3.31	0.65	10.42
Desviación estándar	1.74	0.17	0.19	1.40	0.04	10.46	32.89	0.0018	0.15	0.16	1.06
Mínimo	4.60	0.12	0.0	0.9	0.012	1.0	6.0	0.99	2.74	0.33	8.4
Máximo	15.9	1.58	1.0	15.50	0.611	72.0	289.0	1.003	4.01	2.0	14.9

Fuente: elaboración propia con datos de Dua, D. and Graff, C. (2019)

Tabla 2: Valores de VIF por columna.

Variable	**VIF**
Fixed acidity	74.45
Volatile acidity	17.96
Citric acid	9.19
Residual sugar	4.66
Chlorides	6.64
Free sulfur dioxide	6.46
Total sulfur dioxide	6.60
Density	1528.15
pH	1078.17
Sulphates	22.46
Alcohol	147.61
Quality	77.72

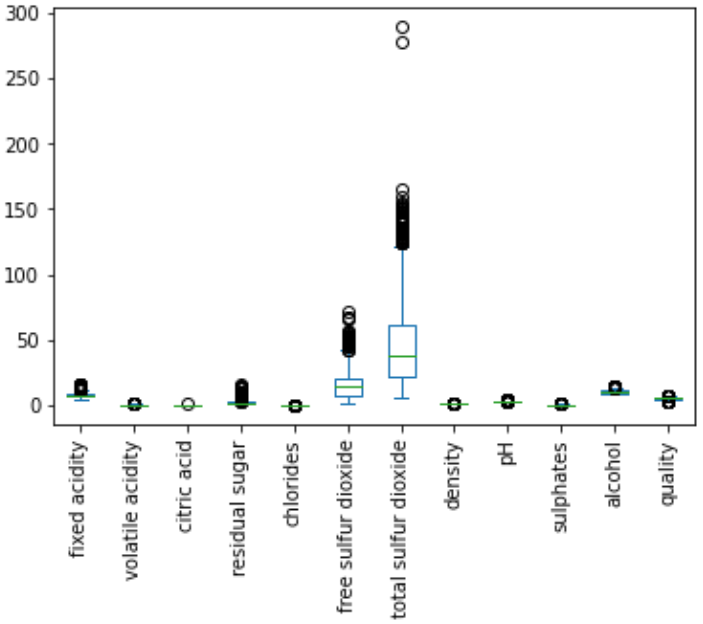
Fuente: elaboración propia con datos de Dua, D. and Graff, C. (2019)

Tabla 3: Cantidad de registros agregados con respecto a la variable “quality” y su porcentaje con respecto al total

Calidad	**Incidencias**	**Porcentaje (%)**
3	10	0.65
4	53	3.31
5	681	42.58
6	638	39.89
7	199	12.44
8	18	1.12

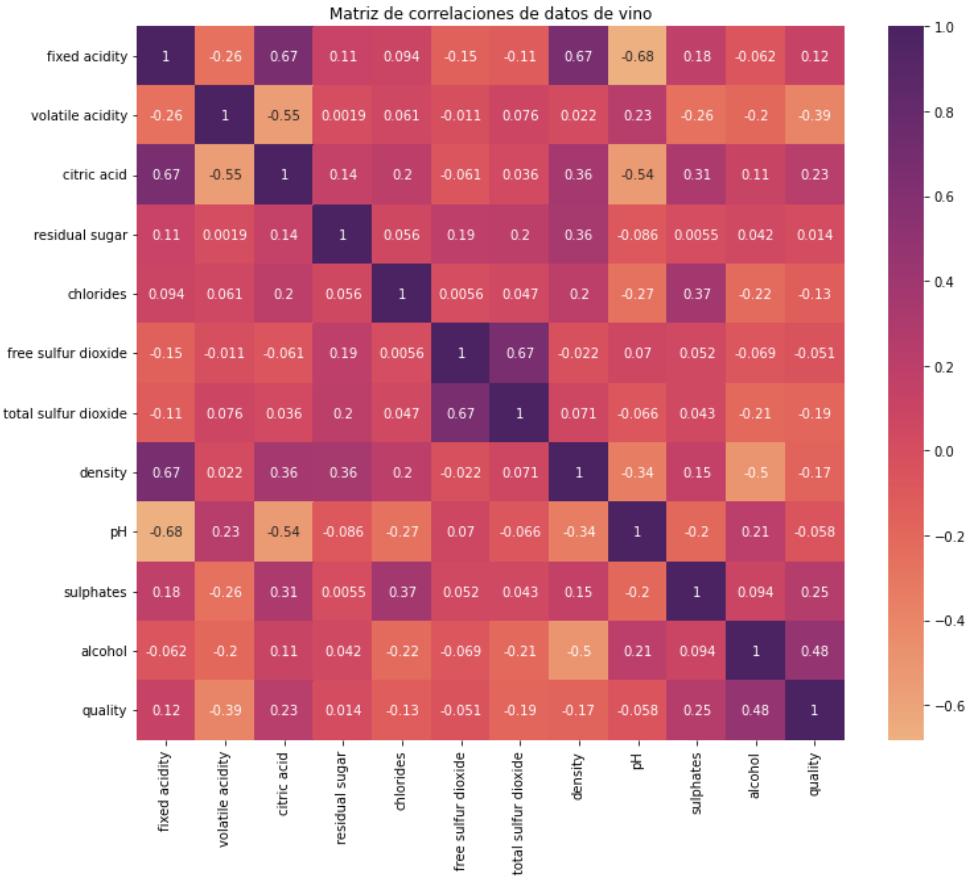
Fuente: elaboración propia con datos de Dua, D. and Graff, C. (2019)

Gráfico 1: Diagramas de caja por columna de los datos.



Fuente: elaboración propia con datos de Dua, D. and Graff, C. (2019)

Gráfico 2: Matriz de correlaciones de los datos.



Fuente: elaboración propia con datos de Dua, D. and Graff, C. (2019)

3. Modelado

3.1. Escoger técnica de modelado

Debido a que el problema es de naturaleza clasificatoria, se consideraron para entrenamiento los modelos Random Forest, SVC, SVC lineal y Naive Bayes con el objetivo de discernir cuál era el que mejor se ajustaba a la problemática.

3.2. Generar plan de prueba

Por medio del software MLflow se entrenaron exhaustivamente los modelos seleccionados considerando segmentaciones del conjunto de datos en proporciones de 80-20 para los segmentos de entrenamiento y pruebas, respectivamente. Las medidas de desempeño para el estudio fueron *Accuracy* (exactitud), *Precisison* (precisión) y *F1-Score* (Valor-F).

3.3. Construcción del modelo

Se realizó una primera iteración realizando una segmentación de los datos 80 - 20, entrenando los modelos considerando los datos como fueron obtenidos de la fuente original. Con el objetivo de discernir el mejor modelo clasificatorio para la problemática, dados los modelos previamente entrenados, se realizó un cálculo de las métricas de rendimiento accuracy, precission y F1-score sobre los resultados de clasificación del modelo haciendo uso de los datos del mismo entrenamiento, obteniendo la información mostrada en la Tabla 4. Se hace notar que, con excepción del Random Forest, los métodos SVC, SVC lineal y Naive Bayes presentaron un desempeño pobre con métricas menores a 0.6 en todos los casos. En contraste, el Random Forest entrenado obtuvo en las tres métricas el valor 1, realizando una clasificación sin errores.

Se conjeturó que los resultados bajos en el desempeño de los modelos fue resultado del imbalance de datos detectado en la Tabla 3, por lo que una técnica de sobre muestreo fue aplicado obteniendo el balance final presentado en la Tabla 5, donde se tuvieron un total de 3306 registros en el cual cada clase se componía de 551 registros, contrastado con los 1599 registros originales.

Se efectúo una nueva segmentación 80-20 con los datos balanceados, y a partir de los datos de entrenamiento, se realizó el cálculo de las métricas de

rendimiento (ver Tabla 6), en el cual se hace notar que el imbalance de datos no presentó ser un factor determinante para el desempeño de los modelos, donde de nueva cuenta el modelo con los mejores resultados fue el Random Forest con valor de 1 en todas sus métricas. Debido a lo anterior, se seleccionó el Random Forest para la fase de pruebas.

Tabla 4: Valores de métricas de evaluación de los modelos SVC lineal, Naive Bayes, SVC y Random Forest haciendo uso de la porción de datos utilizados para el entrenamiento.

Method name	Training accuracy (exactitud de entrenamiento)	Training precision (precisión de entrenamiento)	Training F1 Score (Valor-F de entrenamiento)
Linear SVC	0.557	0.544	0.543
Gaussian NB	0.573	0.579	0.573
SVC	0.513	0.577	0.464
Random Forest	1	1	1

Fuente: elaboración propia con datos de Dua, D. and Graff, C. (2019)

Tabla 5: Cantidad de registros obtenidos por la técnica de sobre muestreo agregados con respecto a la variable “quality” y su porcentaje con respecto al total.

Calidad	Incidencias	Porcentaje (%)
3	551	16.6%
4	551	16.6%
5	551	16.6%
6	551	16.6%
7	551	16.6%
8	551	16.6%

Fuente: elaboración propia con datos de Dua, D. and Graff, C. (2019)

Tabla 6: Valores de métricas de evaluación de los modelos SVC lineal, Naive Bayes, SVC y Random Forest haciendo uso de la porción de datos utilizados para el entrenamiento con la técnica de sobre muestreo.

Method name	Training accuracy (exactitud de entrenamiento)	Training precision (precisión de entrenamiento)	Training F1 Score (Valor-F de entrenamiento)
Linear SVC	0.384	0.359	0.333
Gaussian NB	0.528	0.501	0.498
SVC	0.412	0.422	0.388
Random Forest	1	1	1

Fuente: elaboración propia con datos de Dua, D. and Graff, C. (2019)

Por medio de una búsqueda exhaustiva y aleatorizada de hiper parámetros se determinó que los parámetros que presentaron mejor desempeño tanto con el segmento de datos de la fuente original designado para pruebas como el segmento de datos de prueba de la muestra con técnica de sobre muestreo fueron los mostrados en la Tabla 6. Los resultados clasificatorios de los conjuntos de prueba originales y remuestreados no presentaron diferencias significativas en las métricas de desempeño (ver Tabla 7) en donde la matriz de confusión de ambas pruebas resultaron idénticas en todas sus entradas (ver Gráfico 3 y 4). Por lo anterior se determinó que el sesgo de la muestra original no afectó el entrenamiento del modelo, por lo que el uso de los datos sobre muestreados fueron descartados para el entrenamiento del modelo final.

Tabla 6: Hiper parámetros del Random Forest.

Hiper Parámetro	**Valor**
Número de árboles	1000
Número mínimo de muestras necesarias para dividir un nodo	2
Número mínimo de muestras necesarias para ser un nodo	1
Profundidad máxima del árbol	50

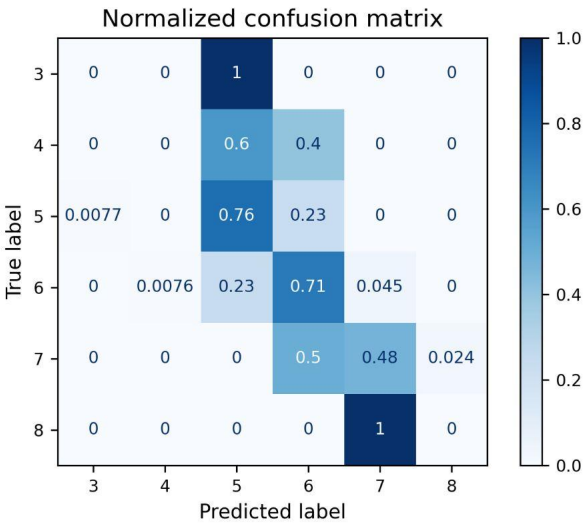
Fuente: elaboración propia con datos de Dua, D. and Graff, C. (2019)

Tabla 7: Valores de métricas de desempeño del Random Forest utilizando diferentes muestras como conjunto de prueba.

	Testing accuracy (exactitud de prueba)	**Testing precision (precisión de prueba)**	**Testing F1 Score (Valor-F de prueba)**
Muestra original	0.666	0.635	0.648
Muestra sobre muestreada	0.666	0.638	0.649

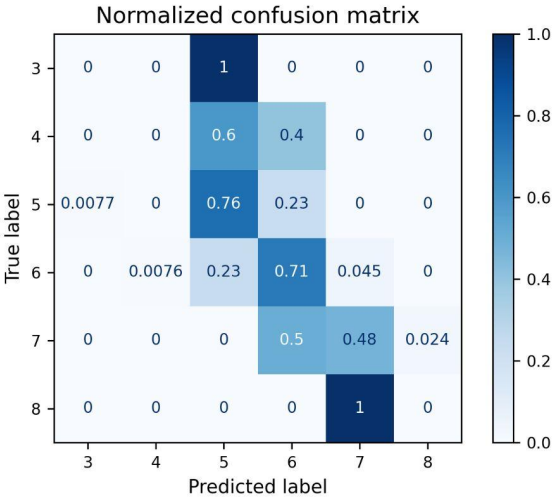
Fuente: elaboración propia con datos de Dua, D. and Graff, C. (2019)

Gráfico 3: Matriz de confusión del Random Forest de los datos de prueba originales.



Fuente: elaboración propia con datos de Dua, D. and Graff, C. (2019)

Gráfico 4: Matriz de confusión del Random Forest de los datos de prueba con técnica de sobre muestreo.



Fuente: elaboración propia con datos de Dua, D. and Graff, C. (2019)

4. Evaluación

4.1. Evaluación de resultados

Dado que el modelo final tuvo una precisión, exactitud y valor-F menores a 0.7, se sugiere realizar una implantación de prueba y analizar la evolución del modelo en nuevas muestras.

4.2. Revisar procesos

Se conjetura que en el análisis exploratorio de datos se puede expandir si se estudian las correlaciones multivariadas detectadas por el VIF, lo cual podría llevar a un aumento del entendimiento de las cualidades del vino si se realiza una interpretación de estás relaciones.

A su vez, la posibilidad de entrenamiento de un modelo clasificatorio a través de características derivadas por medio de un método de agregación de las variables multicolineales detectadas por el VIF, o por medio de un análisis de componentes principales, podrían llevar a nuevos resultados en las evaluaciones de los modelos entrenados.

5. Implantación

5.1. Planear implementación

La implantación del modelo entrenado se compone principalmente de dos secciones: programación y diseño del tablero de uso del modelo, y puesta a disposición de uso (host) para el usuario final. Se hace uso de la programación del tablero a través del framework de alto nivel disponible denominado *Streamlit*, siendo utilizado el sistema de host de la nube comunitaria de Streamlit.

5.2. Producir el informe final

Del presente estudio, se consideran los siguientes puntos como los más relevantes.

- El 41% de las columnas presentaban niveles altos de multicolinealidad.

- Los datos presentan un imbalance importante con respecto a la variable objetivo, donde 85% de los datos recabados estaban etiquetados con calidad 5 y 6, no obstante, este imbalance no parece afectar el proceso de entrenamiento del modelo.
- Se lograron métricas del modelo final menores a 0.7, por lo que se recomienda realizar pruebas sobre resultados obtenidos por el modelo una vez ya implantado.

5.3. Revisar el proyecto

En el presente trabajo se reconoce como área de oportunidad la incorporación de técnicas de tratamiento para datos atípicos, ya que datos de esta naturaleza fueron detectados en la muestra.

Por otro lado, es posible realizar mayores pruebas en una segunda versión de este análisis en el área de creación de variables agregadas (sumar variables con alta multicolinealidad) o de creación sintética (por medio de ACP/PCA) para la evaluación de modelos clasificatorios alternativos que podrían llevar a una mayor generalidad en el proceso de modelado.

Referencias

I. A. L. Waterhouse, G. L. Sacks, and D W. Jeffery. (2016) Understanding Wine Chemistry, John Wiley & Sons

II. Coli, M. S., Rangel, A. G. P., Souza, E. S., Oliveira, M. F., & Chiaradia, A. C. N. (2015, marzo). Chloride concentration in red wines: influence of terroir and grape type. *Food Science and Technology (Campinas)*, *35*(1), 95–99. https://doi.org/10.1590/1678-457x.6493

III. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science

IV. OECD. (2004). *The OECD Glossary of Statistical Terms*. Constado en Septiembre 28, 2022, de https://stats.oecd.org/glossary/

V. Howard, C. (2022, Abril 27). *What is Residual Sugar in Wine?* Whicher Ridge. Consulta en Septiembre 27, 2022, de https://whicherridge.com.au/blog/what-is-residual-sugar-in-wine/

VI. Iowa State University. (2018, Febrero 27). *Total Sulfur Dioxide – Why it Matters, Too!* Iowa State University Extension and Outreach. Retrieved Septiembre 27, 2022, de https://www.extension.iastate.edu/wine/total-sulfur-dioxide-why-it-matters-too

VII. Mazzeo, J. (2021, Noviembre 9). *What Does 'Volatile Acidity' Mean in Wine?* Wine Enthusiast. COnsultado en Septiembre 27, 2022, de https://www.winemag.com/2021/11/09/volatile-acidity-wine/

VIII. Mooney, A. (2019, Junio 17). *Why is testing for citric acid important in winemaking?* Randox Food. Consutlado en September 27, 2022, de https://www.randoxfood.com/why-is-testing-for-citric-acid-important-in-winemaking/

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

X. Pedregosa et al (2011) Scikit-learn: Machine Learning in Python [https://scikit-learn.org/stable/index.html], ., JMLR 12, pp. 2825-2830.

XI. *Use and Measurement of Sulfur Dioxide in Wine | Page 1 of 1*. (n.d.). Consultado en Septiembre 27, 2022, de https://www.piwine.com/use-and-measurement-of-sulfur-dioxide-in-wine.html

XII. What do "pH" and "TA" numbers mean to a wine? (2009, abril 15). Wine Spectator. Consultado en septiembre 28, 2022, de https://www.winespectator.com/articles/what-do-ph-and-ta-numbers-mean-to-a-wine-5035

XIII. What to Know About Sulfites in Wine. (2021, abril). WebMD. Retrieved September 28, 2022, tomado de https://www.webmd.com/diet/what-to-know-sulfites-in-wine.

XIV. Wikipedia contributors. (2022a, julio7). *Acids in wine*. Wikipedia. Consuta en Septiembre 27, 2022, de https://en.wikipedia.org/wiki/Acids_in_wine#Citric

XV. —— (2022b, Marzo 24). *Vinho verde*. Wikipedia, La Enciclopedia Libre. Consulta en septiembre29, 2022, de https://es.wikipedia.org/wiki/Vinho_verde