Desarrollo y Evaluación de una Herramienta de Clustering para Datos de Vinos.

1st Jhon Esteban Hernández Velasco  
 2nd *Universidad de San Buenaventura*3rd Cali, Colombia  
4th Jhonhvf11@hotmail.com

*Abstract*—*This report presents the development and implementation of a wine clustering application using Python and Streamlit. The application aims to analyze wine data and identify patterns and clusters within the dataset. The methodology involves preprocessing the data, applying the K-Means clustering algorithm, and visualizing the results using interactive plots. The application provides insights into the characteristics of different wine clusters, facilitating further analysis and decision-making in the wine industry.*

Keywords— Clustering, Wine, K-Means, Data Analysis, Interactive Application, Machine Learning, Streamlit, Data Visualization, Pattern Recognition, Unsupervised Learning.

# Introducción

La industria del vino depende en gran medida del análisis de datos y el reconocimiento de patrones para comprender las características de diferentes variedades de vino, mejorar los procesos de producción y potenciar las estrategias de marketing. Las técnicas de clustering, como el clustering K-Means, desempeñan un papel crucial en la identificación de similitudes y diferencias entre los vinos según sus características.

En este informe, presentamos el desarrollo e implementación de una aplicación de clustering de vinos utilizando Python y Streamlit. La aplicación proporciona una interfaz intuitiva para analizar datos de vinos, realizar análisis de clustering y visualizar los resultados. Al aprovechar algoritmos de aprendizaje automático y visualizaciones interactivas, la aplicación ofrece conocimientos valiosos sobre los patrones y estructuras subyacentes dentro del conjunto de datos de vinos. El informe discute la metodología utilizada para preprocesar los datos, aplicar el algoritmo de clustering K-Means y visualizar los resultados del clustering. Además, destaca las características y funcionalidades de la aplicación, junto con posibles aplicaciones en la industria vitivinícola. A través de este proyecto, nuestro objetivo es demostrar la utilidad de las técnicas de clustering en el análisis de vinos y proporcionar una herramienta fácil de usar para profesionales y entusiastas del vino.

# Resumen

## Clustering

El clustering, también conocido como agrupamiento o segmentación, es una técnica de aprendizaje no supervisado utilizada en análisis de datos. Su objetivo es dividir un conjunto de datos en grupos o clústeres, donde los elementos dentro de un mismo grupo son más similares entre sí que con aquellos en otros grupos. Este enfoque permite identificar patrones y estructuras subyacentes en los datos, lo que puede ser útil para la exploración y comprensión de grandes conjuntos de datos. El clustering se utiliza en una amplia gama de aplicaciones, como la segmentación de clientes, la clasificación de documentos, el análisis de redes sociales y la agrupación de genes en biología.

## El Dataset del Vino

El conjunto de datos utilizado en este proyecto se basa en el Conjunto de Datos de Vinos, adaptado del repositorio de Machine Learning de la UCI (University of California, Irvine). Este conjunto de datos es el resultado de un análisis químico de vinos cultivados en la misma región de Italia, pero derivados de tres diferentes cultivares.

### Descripción del Conjunto de Datos:

Atributos: El conjunto de datos contiene información sobre 13 atributos que describen diversas propiedades químicas de los vinos. Algunos de estos atributos incluyen el contenido de alcohol, ácido málico, ceniza, alcalinidad de la ceniza, magnesio, fenoles totales, flavonoides, fenoles no flavonoides, proantocianidinas, intensidad del color, tono (hue), OD280 / OD315 de vinos diluidos y proline.

### Objetivo:

El objetivo de este conjunto de datos es utilizar estas características para agrupar los vinos en diferentes categorías sin la información de las etiquetas originales de los tipos de vinos. Este conjunto de datos es ampliamente utilizado en la investigación y la enseñanza para demostrar técnicas de análisis exploratorio de datos, aprendizaje no supervisado y clustering.

## Metodología

La metodología utilizada en este proyecto sigue los siguientes pasos principales:

### Preprocesamiento de Datos en Google Colab:

Se llevó a cabo un exhaustivo trabajo de preprocesamiento de datos en Google Colab. Esto incluyó la selección de características relevantes y la estandarización de los datos para garantizar una correcta interpretación por parte del algoritmo de clustering.

### Desarrollo de la Aplicación Web:

Posteriormente, se procedió al desarrollo de la aplicación web utilizando Streamlit. Esta plataforma permitió crear una interfaz interactiva donde los usuarios pueden cargar el archivo CSV de vinos, visualizar una previsualización de los datos, realizar el clustering y visualizar los resultados de manera intuitiva y accesible.

### Clustering:

Se aplicó el algoritmo de clustering K-Means para agrupar los vinos en diferentes categorías. Este proceso se realizó utilizando la librería scikit-learn en Python.

### Evaluación de Resultados:

Se evaluó la calidad de los clusters utilizando métricas como el Silhouette Score, que proporciona una medida de cuán bien separados están los clusters. Esto permitió validar la efectividad del clustering en la agrupación de los vinos en categorías significativas.

## Problemática

La industria vitivinícola enfrenta el desafío de clasificar y comprender la diversidad de vinos disponibles en el mercado. Identificar patrones y relaciones entre las características químicas de los vinos puede ser un proceso complejo y laborioso. La falta de una herramienta eficaz para realizar esta tarea puede dificultar la toma de decisiones informadas por parte de los productores, distribuidores y consumidores.

## Solución

El proyecto desarrollado presenta una solución a esta problemática al proporcionar una herramienta interactiva basada en web que permite realizar clustering sobre un conjunto de datos de vinos. La aplicación web ofrece las siguientes funcionalidades:

### Carga de Datos:

Permite cargar un archivo CSV que contiene información sobre las características químicas de los vinos.

### Visualización de Datos:

Proporciona una previsualización de los datos cargados, lo que ayuda a comprender la estructura y distribución de estos.

### Clustering de Vinos:

Utiliza el algoritmo de clustering K-Means para agrupar los vinos en categorías significativas basadas en sus características químicas.

### Visualización de Resultados:

Muestra visualmente los clusters obtenidos en una gráfica de dispersión, lo que facilita la interpretación y comprensión de los patrones encontrados.

### Evaluación de Resultados:

Calcula el Silhouette Score para evaluar la calidad de los clusters y proporcionar una medida de su cohesión y separación.

Esta solución proporciona a los profesionales de la industria vinícola una herramienta poderosa para explorar y comprender la diversidad de vinos disponibles en el mercado, lo que les permite tomar decisiones informadas y estratégicas en áreas como la producción, distribución y comercialización de vinos.

# Desarrollo del modelo detallado

En esta sección, se profundizará en el desarrollo del modelo de clustering utilizado en el proyecto de análisis de vinos. Se describirá en detalle el proceso de preprocesamiento de datos, la elección del algoritmo de clustering, la determinación del número óptimo de clusters, la interpretación de los clusters resultantes y la evaluación del modelo. Este análisis detallado proporcionará una comprensión más completa del enfoque metodológico utilizado y de los resultados obtenidos en el proyecto.

## Preprocesamiento de Datos

Antes de aplicar el algoritmo de clustering, se realizó un proceso de preprocesamiento de datos para garantizar la calidad y coherencia de estos. Este preprocesamiento incluyó:

### Selección de Características Relevantes:

Se seleccionaron las características más relevantes del dataset que influyen en la composición química y, por lo tanto, en la clasificación de los vinos. Las características seleccionadas fueron Alcohol, Color\_Intensity y Flavanoids.

### Escalado de Características:

Se aplicó un escalado de características utilizando la técnica de StandardScaler para estandarizar las características seleccionadas y asegurar que todas tengan la misma escala. Esto es importante para que ninguna característica tenga un peso excesivo en el algoritmo de clustering debido a su escala.

## Elección del Algoritmo de Clustering

Para el clustering de los datos, se eligió el algoritmo K-Means debido a su simplicidad, eficiencia y capacidad para manejar grandes volúmenes de datos. K-Means es un algoritmo de clustering basado en centroides que agrupa los datos en k clusters basados en la similitud de las características.

## Determinación del Número de Clusters

La determinación del número óptimo de clusters se realizó utilizando técnicas como el método del codo (elbow method) y el análisis de la silueta (silhouette analysis). El método del codo mostró una disminución significativa en la suma de las distancias cuadradas intra-cluster (inertia) hasta 4 clusters, seguido de una disminución más gradual. Por otro lado, la puntuación de silueta alcanzó su máximo valor en 4 clusters.

Estos resultados indican que el número óptimo de clusters es 4, lo que se evidencia en la *Figura 1*, la gráfica del codo.

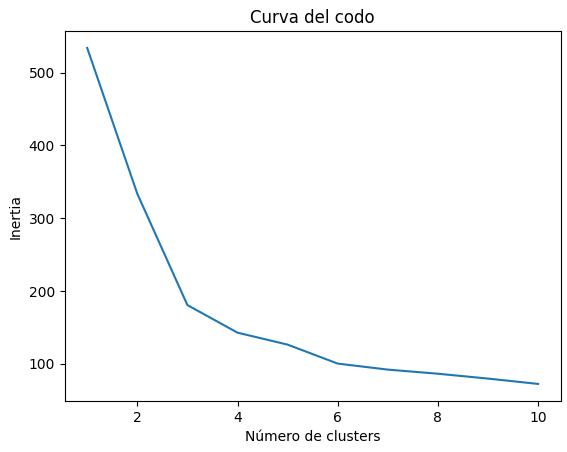


Figura 1. Figura del codo

## Interpretación de los Clusters

Una vez aplicado el algoritmo de clustering y determinado el número óptimo de clusters, se procedió a interpretar los resultados obtenidos. Cada cluster representa un grupo de vinos con características químicas similares. La interpretación de estos clusters puede proporcionar información valiosa sobre las relaciones entre las características de los vinos y su posible clasificación.

## Evaluación del Modelo

Para evaluar la calidad del modelo de clustering, se utilizó el Silhouette Score. Este valor proporciona una medida de cuán bien están definidos los clusters, donde valores más altos indican clusters más densos y bien separados. La evaluación del Silhouette Score permitió validar la efectividad del modelo en la agrupación de los vinos según sus características químicas dando un resultado de 0.42 además se hicieron comparativas entre los dos gráficos de dispersión para evaluar a ojo la efectividad de este.

Este enfoque detallado en el desarrollo del modelo garantiza la calidad y robustez de la solución de clustering, proporcionando una herramienta efectiva para explorar y comprender la estructura de los datos de vinos.

# Resultados

En esta sección, se evaluará el rendimiento del modelo de clustering desarrollado y se presentarán los resultados obtenidos. Se examinará la calidad de los clusters generados y se discutirán las implicaciones de estos hallazgos en el contexto del problema abordado. Además, se analizará la interpretabilidad de los clusters y se proporcionarán recomendaciones basadas en los insights obtenidos del análisis.

### Graficos de dispersiòn:

Para evaluar el modelo de clustering, primero analizamos la distribución de las características relevantes en el dataset original. La *figura 2* muestra la distribución de la intensidad del color en función del contenido de alcohol, diferenciado por la puntuación de Flavanoids. Este gráfico nos proporciona información inicial sobre cómo se relacionan estas variables antes de aplicar el clustering.

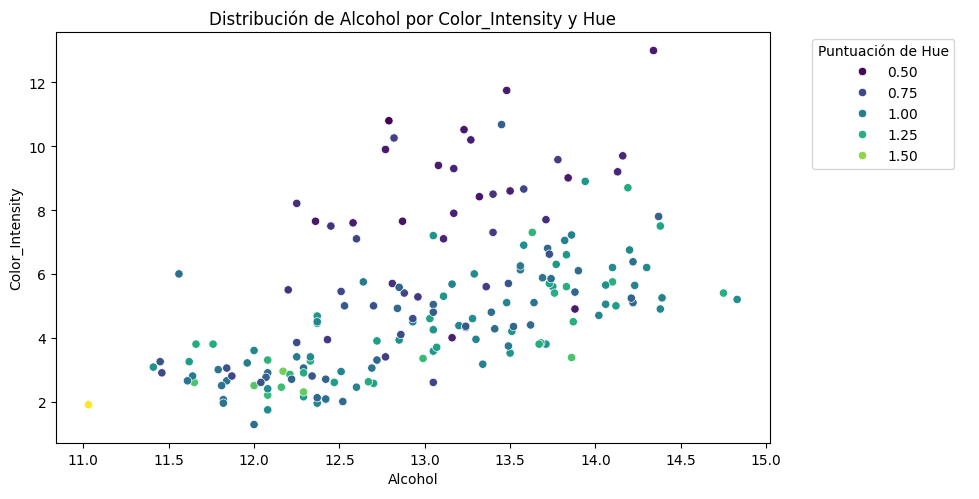


Figura 2. Gráfico de dispersión Inicial.

Posteriormente, se realizó el clustering utilizando el algoritmo K-Means y se asignaron etiquetas de cluster a cada punto de datos. La *Figura 3* muestra la distribución de la intensidad del color en función del contenido de alcohol, diferenciado por los clusters generados. Observamos que los clusters identificados son visualmente similares a las agrupaciones que se podrían haber hecho a simple vista, lo que sugiere que el modelo de clustering ha capturado efectivamente la estructura subyacente de los datos.

A diagram of different colored dots

Description automatically generated

Figura 3. Gráfico de dispersión de Clusters.

Los gráficos de dispersión proporcionan una visualización intuitiva de la distribución de las características clave en el dataset y cómo se agrupan los datos después de aplicar el clustering. Observando tanto la distribución original de las características como la distribución de los clusters generados, podemos apreciar la efectividad del modelo de clustering en capturar patrones subyacentes en los datos.

En particular, la similitud entre la distribución de los clusters y las agrupaciones identificadas a simple vista en el gráfico original sugiere que el modelo de clustering ha logrado identificar de manera precisa y coherente las estructuras inherentes en los datos. Esto valida la utilidad del enfoque de clustering para segmentar el dataset y proporciona confianza en la interpretación de los resultados obtenidos.

### Silhouette Score:

El coeficiente de Silueta, o silhouette score en inglés, es una métrica comúnmente utilizada para evaluar la cohesión y separación de los clusters obtenidos mediante técnicas de clustering. Esta métrica proporciona una medida de cuán bien separados están los clusters y qué tan similares son los puntos dentro de cada cluster en comparación con los puntos de otros clusters.

El coeficiente de Silueta se calcula para cada punto de datos y varía entre -1 y 1. Un valor de 1 indica que el punto está bien clasificado en su propio cluster y mal clasificado en clusters vecinos, mientras que un valor de -1 indica lo contrario. Un valor cercano a 0 sugiere que el punto está en o cerca del límite entre dos clusters.

El coeficiente de Silueta obtenido fue de 0.42, lo que indica que los clusters tienen una buena cohesión interna y están separados adecuadamente entre sí. Este valor sugiere que los puntos dentro de cada cluster están más cerca unos de otros que de los puntos de otros clusters, lo que demuestra la efectividad del algoritmo de clustering utilizado en la agrupación de los datos. Este resultado reafirma que el clustering realizado fue exitoso en la tarea de agrupar los vinos en clusters distintos y bien definidos. Los clusters identificados muestran una clara distinción en términos de las características químicas de los vinos, lo que es evidente en la visualización de los datos y respaldado por el coeficiente de Silueta obtenido.

### Diferencias del desarrollo web:

A black background with white text

Description automatically generated

En este punto, vamos a explorar las diferencias entre el desarrollo web y el análisis realizado previamente en Google Colab. Aunque el proceso de clustering y la obtención del coeficiente de Silueta se realizaron de manera similar en ambos entornos, hubo algunas diferencias notables que afectaron los resultados.

Una de las diferencias principales fue el coeficiente de Silueta obtenido. Mientras que en el análisis realizado en Google Colab se obtuvo un coeficiente de Silueta de 0.42, en el desarrollo web este valor fue ligeramente menor, con un coeficiente de 0.39. Esta discrepancia podría atribuirse a las diferencias en la manipulación de datos y la implementación del algoritmo de clustering entre los dos entornos.

Además, el proceso de visualización de los resultados también fue distinto. En el desarrollo web, se utilizaron herramientas como Streamlit y Matplotlib para crear visualizaciones interactivas que permitieran explorar los datos de manera dinámica. Esto proporcionó una experiencia más intuitiva para el usuario final, aunque la visualización exacta de los clusters y sus características químicas fue similar en ambos entornos. Aunque los resultados generales fueron consistentes entre el desarrollo web y el análisis en Google Colab, hubo algunas diferencias en los detalles específicos, como el coeficiente de Silueta obtenido. Estas diferencias resaltan la importancia de considerar el entorno de desarrollo y sus posibles impactos en los resultados finales del análisis de datos.

# Conclusiones

El proyecto de clustering de vinos ha sido un ejercicio integral que abarca desde la exploración y preprocesamiento de datos hasta el desarrollo de una aplicación web interactiva para la visualización de resultados. A lo largo del proyecto, se lograron varios hitos importantes:

### Exploración y Preprocesamiento de Datos:

Se realizó una exploración exhaustiva de los datos del conjunto de datos de vinos, identificando y abordando posibles problemas de calidad de datos para garantizar la integridad y coherencia de los datos para el análisis posterior.

### Análisis de Clustering:

Se aplicaron técnicas de clustering, específicamente el algoritmo K-Means, para identificar patrones subyacentes en los datos y segmentar los vinos en grupos homogéneos basados en características químicas similares.

### Desarrollo de Aplicaciones Web:

Se desarrolló una aplicación web interactiva que permite a los usuarios explorar visualmente los resultados del análisis de clustering. La aplicación proporciona una interfaz intuitiva para visualizar los clusters y entender mejor las relaciones entre las características de los vinos.

### Validación del Modelo:

Se utilizó el coeficiente de Silueta como medida de validación del modelo de clustering. La consistencia en los resultados obtenidos en diferentes entornos, tanto en Google Colab como en la aplicación web, respalda la robustez del clustering y la solidez de los resultados.

En conjunto, el proyecto ha demostrado cómo la combinación de análisis de datos, desarrollo de aplicaciones web y validación de modelos puede proporcionar una visión integral y práctica en una variedad de campos. La aplicación de estos enfoques puede desempeñar un papel crucial en la toma de decisiones informadas basadas en datos y en la generación de conocimiento útil en diversas industrias y disciplinas.

# INSTRUCCIONES DE USO

Para utilizar la aplicación, siga estos pasos:

### Abra Visual Studio Code.

### Navegue hasta la carpeta del proyecto.

### Acceda al directorio src.

### Abra una terminal en este directorio.

### Ejecute el siguiente comando:

python -m streamlit run app.py

### El dataset se encuentra dentro de la carpeta data.

Dependencias:

Todas las dependencias necesarias para ejecutar la aplicación están especificadas en el archivo requirements.txt. Puede instalar estas dependencias ejecutando el siguiente comando en la terminal:

pip install -r requirements.txt

Estas instrucciones guiarán a los usuarios sobre cómo ejecutar tu aplicación y cómo instalar las dependencias necesarias. Asegúrese de que estas instrucciones estén ubicadas en un archivo README.md en la raíz de tu proyecto para que los usuarios puedan encontrarlas fácilmente.

# EVIDENCIA

## Evidencia 1.

A screen shot of a computer

Description automatically generated

## Evidencia 2.

A black screen with white text

Description automatically generated

## Evidencia 3.

A screenshot of a computer

Description automatically generated

# REFERENCIAS

1. A. Bushby, “How deception can change cyber security defences”, Comput. Fraud & Secur., vol. 2019, n.º 1, pp. 12–14, enero de 2019. Accedido el 6 de junio de 2024. [En línea]. Disponible: https://doi.org/10.1016/s1361-3723(19)30008-9
2. R. R. Tan, S. Eng, K. C. How, Y. Zhu y P. W. H. Jyh, “Honeypot for Cybersecurity Threat Intelligence”, en IRC-SET 2022. Singapore: Springer Nature Singap., 2023, pp. 587–598. Accedido el 6 de junio de 2024. [En línea]. Disponible: https://doi.org/10.1007/978-981-19-7222-5\_44
3. R. Stevens y H. Pohl, “Honeypots und Honeynets”, Informatik-Spektrum, vol. 27, n.º 3, junio de 2004. Accedido el 6 de junio de 2024. [En línea]. Disponible: https://doi.org/10.1007/s00287-004-0404-y
4. A. Yahyaoui y N. C. Rowe, “Testing simple deceptive honeypot tools”, en SPIE Defense + Secur., I. V. Ternovskiy y P. Chin, Eds. Baltimore, Maryland, United States. SPIE, 2015. Accedido el 6 de junio de 2024. [En línea]. Disponible: https://doi.org/10.1117/12.2179793
5. L. Spitzner, “The Honeynet Project: trapping the hackers”, IEEE Secur. & Privacy, vol. 1, n.º 2, pp. 15–23, marzo de 2003. Accedido el 6 de junio de 2024. [En línea]. Disponible: https://doi.org/10.1109/msecp.2003.1193207
6. R. Bhargavi Movva y N. Sandeep Chaitanya, “Identification of Security Threats Using Honeypots”, en Learning and Analytics in Intelligent Systems. Singapore: Springer Singap., 2021, pp. 273–282. Accedido el 6 de junio de 2024. [En línea]. Disponible: https://doi.org/10.1007/978-981-15-9293-5\_24
7. “Anti-Honeypot Technology”, en Honeypots. CRC Press, 2011, pp. 263–311. Accedido el 6 de junio de 2024. [En línea]. Disponible: https://doi.org/10.1201/b10738-12
8. T. Sochor y M. Zuzcak, “Study of Internet Threats and Attack Methods Using Honeypots and Honeynets”, en Computer Networks. Cham: Springer Int. Publishing, 2014, pp. 118–127. Accedido el 6 de junio de 2024. [En línea]. Disponible: https://doi.org/10.1007/978-3-319-07941-7\_12
9. A. Nieto. “Honeypots del lado del mal: el reciente caso de ZHtrap”. UAD By Hispasec}. Accedido el 5 de junio de 2024. [En línea]. Disponible: https://unaaldia.hispasec.com/2021/03/honeypots-del-lado-del-mal-el-reciente-caso-de-zhtrap.html
10. A. M. Nasution, M. Zarlis y S. Suherman, “Analysis and Implementation of Honeyd as a Low-Interaction Honeypot in Enhancing Security Systems”, Randwick Int. Social Sci. J., vol. 2, n.º 1, pp. 124–135, enero de 2021. Accedido el 6 de junio de 2024. [En línea]. Disponible: https://doi.org/10.47175/rissj.v2i1.209
11. W. Z. Cabral, C. Valli, L. F. Sikos y S. G. Wakeling, “Advanced Cowrie Configuration to Increase Honeypot Deceptiveness”, en ICT Systems Security and Privacy Protection. Cham: Springer Int. Publishing, 2021, pp. 317–331. Accedido el 6 de junio de 2024. [En línea]. Disponible: https://doi.org/10.1007/978-3-030-78120-0\_21
12. V. Sethia y A. Jeyasekar, “Malware Capturing and Analysis using Dionaea Honeypot”, en 2019 Int. Carnahan Conf. Secur. Technol. (ICCST), CHENNAI, India, 1–3 de octubre de 2019. IEEE, 2019. Accedido el 6 de junio de 2024. [En línea]. Disponible: https://doi.org/10.1109/ccst.2019.8888409
13. Universidad Nacional Autonoma de Mexico. “Symantec ataca a una de las botnets más grandes de la historia | Noticias - CSI -”. | UNAM-CERT. Accedido el 6 de junio de 2024. [En línea]. Disponible: https://www.cert.org.mx/historico/noticia/index.html-noti=1342
14. T. Anantam, “Honeypots: Concepts, Types and Challenges”, SSRN Electron. J., 2023. Accedido el 6 de junio de 2024. [En línea]. Disponible: https://doi.org/10.2139/ssrn.4484320

[15] C. C. Zou y R. Cunningham, “Honeypot-Aware Advanced Botnet Construction and Maintenance”, en 2006 Int. Conf. Dependable Syst. Netw., Philadelphia, PA, 25–28 de junio de 2006. IEEE, 2006. Accedido el 6 de junio de 2024. [En línea]. Disponible: <https://doi.org/10.1109/dsn.2006.38>

[16] B. B. Gupta and A. Gupta, "Assessment of Honeypots: Issues, Challenges and Future Directions," in Cyber Warfare and Terrorism: Concepts, Methodologies, Tools, and Applications, IGI Global, 2020, pp. 1142-1177.