

Clasificación de vinos basada en sus características distintivas mediante SVM, Naive Bayes y análisis discriminante

Descubre cómo utilizar técnicas avanzadas como las Máquinas de Vectores de Soporte (SVM), Naive Bayes y el Análisis Discriminante para clasificar vinos y encontrar el mejor modelo a través de validación cruzada K-Fold.



Introducción

Los vinos tienen características únicas que los hacen especiales. En esta presentación, exploraremos cómo podemos utilizar el aprendizaje automático para clasificar los vinos según sus características distintivas.

Metodología CRISP-DM

El Modelo de la Metodología CRoss Industry Standard Process for Data Mining (CRISP-DM) nos permite abordar el proyecto de clasificación de vinos de manera estructurada y eficiente.

Comprensión del negocio

La base de datos proporciona una rica colección de datos numéricos relacionados con diversas propiedades químicas y físicas de los vinos, incluyendo características como acidez fija, acidez volátil, contenido de azúcar residual, concentración de cloruros, y otros parámetros. El objetivo fundamental es utilizar estos datos para desarrollar un modelo predictivo que pueda evaluar y predecir la calidad del vino.

Fuente de la base de datos:

<https://archive.ics.uci.edu/dataset/186/wine+quality>

Valoración de la situación actual

El vino es una de las bebidas alcohólicas más consumidas alrededor del mundo. Debido a su popularidad, anualmente se producen más de 230 millones de hectalitros alrededor del mundo. El mercado global de esta bebida en 2021 estuvo evaluado en aproximadamente 340 mil millones de dólares.





Compreensión de los datos

Se optó por trabajar con la base de datos de vinos tintos.

- fixed_acidity
- volatile_acidity
- citric_acid
- residual_sugar
- chlorides
- free_sulfur_dioxide
- total_sulfur_dioxide
- density
- pH
- sulphates
- alcohol

Se utilizó la siguiente variable como etiqueta de clase:

- quality

División de datos en entrenamiento, prueba y verificación

Es fundamental dividir nuestros datos en conjuntos de entrenamiento, prueba y verificación para obtener un modelo confiable y preciso. Exploraremos cómo realizar esta división de manera efectiva.

Construcción de los datos

A partir de esta base de datos, hemos procedido a dividirla en tres grupos:

- Conjunto de Entrenamiento (80%): Este conjunto de datos se utilizará para entrenar el modelo.
- Conjunto de Prueba (10%): Una vez entrenado el modelo, utilizaremos este conjunto de datos para realizar pruebas.
- Conjunto de Validación (10%): Este conjunto se empleará como si fueran nuevos datos para evaluar nuestro modelo y verificar si produce los resultados deseados.

Además, hemos tenido en cuenta que la columna "Quality" se utiliza como variable de clasificación, la cual abarca valores del 3 al 9. Esto asegura que respetamos las proporciones adecuadas para evitar sesgos en los resultados del modelo y lograr un entrenamiento correcto tanto del modelo como de los datos resultantes.

SVM, Naive Bayes y Análisis Discriminante

Las Máquinas de Vectores de Soporte (SVM) y el Análisis Discriminante son poderosas técnicas de clasificación que nos permiten separar y clasificar los vinos en función de sus características distintivas. Aprenderemos cómo aplicar estas técnicas de manera efectiva.

Justificación de la Elección de Modelos:

- **Análisis Discriminante (AD):** Apropiado para problemas de clasificación lineal, cumple con los requisitos del problema y es conocido por su rendimiento en problemas similares.
- **Naive Bayes (NB):** Efectivo en situaciones donde las características son independientes, y su simplicidad facilita la interpretación y el entrenamiento rápido.
- **SVM con Kernel Polinomial:** Capaz de manejar relaciones no lineales, lo cual puede ser crucial en problemas más complejos de clasificación.

Validación cruzada K-Fold

La validación cruzada K-Fold es una técnica que nos permite evaluar el rendimiento del modelo de clasificación de vinos de manera robusta y confiable. Exploraremos cómo utilizar esta técnica para obtener resultados más precisos.

Plan de Prueba (Validación Cruzada k-fold):

Se implementó la validación cruzada k-fold con 10 particiones y 5 repeticiones para evaluar los modelos. Esto garantiza que cada modelo sea evaluado en múltiples conjuntos de datos y proporciona una estimación robusta de su rendimiento.

Identificación del mejor modelo

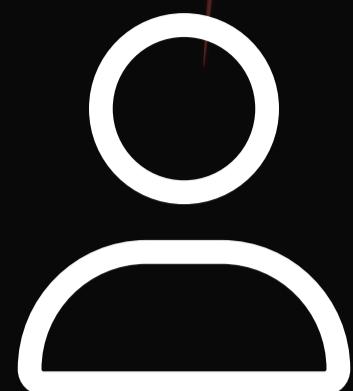
Utilizando la validación cruzada K-Fold y comparando diferentes configuraciones de modelos, pudimos identificar el mejor modelo de clasificación de vinos. Descubrimos cómo encontrar el modelo óptimo para obtener resultados precisos y confiables.

Conclusiones

En esta presentación, hemos explorado el emocionante mundo de la clasificación de vinos utilizando técnicas avanzadas como SVM, Naive Bayes y análisis discriminante. Hemos aprendido cómo seguir la metodología CRISP-DM para llevar a cabo este proyecto de manera efectiva.

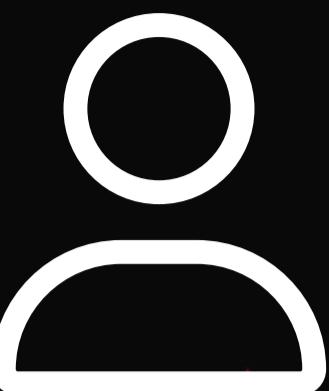
¡Esperamos que hayas disfrutado de esta presentación!

Crew



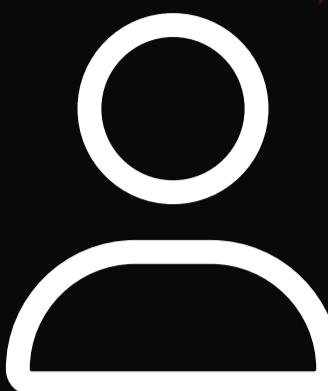
**Luis Andrés
Burruel**

Lic. Matemáticas
<https://www.linkedin.com/in/andr%C3%A9s-burruel-93498625b/>



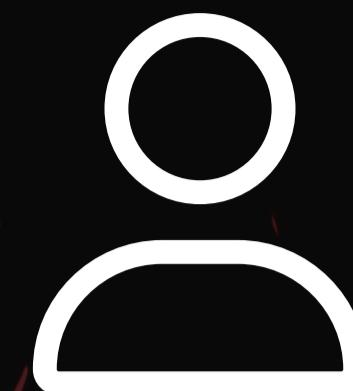
**Rodolfo
Jaramillo**

Lic. Física
<https://www.linkedin.com/in/rodolfo-jaramillo>



**Luis Ernesto
Ortiz**

Lic. Economía
<https://www.linkedin.com/in/luis-ortiz-23a4a482/>



Mario Estrada

Ing. en sistemas de
información
<https://www.linkedin.com/in/mario-estrada-17100480>

fojaramillo/proyecto-introdatsci

Repositorio del proyecto final del primer módulo de Introducción a la Ciencia de Datos.

Contributors 0 Issues 0 Stars 0 Forks 3

GitHub [GitHub - fojaramillo/proyecto-introdatsci...](#)

Repositorio del proyecto final del primer módulo de Introducción a la Ciencia de Datos. - GitHub -...

mariooef/proyecto-introdatsci

Repositorio del proyecto final del primer módulo de Introducción a la Ciencia de Datos.

Contributors 0 Issues 0 Stars 0 Forks 0

GitHub [GitHub - mariooef/proyecto-introdatsci...](#)

Repositorio del proyecto final del primer módulo de Introducción a la Ciencia de Datos. - GitHub -...

LANDBd/proyecto-introdatsci

Repositorio del proyecto final del primer módulo de Introducción a la Ciencia de Datos.

Contributors 0 Issues 0 Stars 0 Forks 0

GitHub [GitHub - LANDBd/proyecto-introdatsci...](#)

Repositorio del proyecto final del primer módulo de Introducción a la Ciencia de Datos. - GitHub -...

Luiserov/proyecto-introdatsci

Repositorio del proyecto final del primer módulo de Introducción a la Ciencia de Datos.

Contributors 0 Issues 0 Stars 0 Forks 0

GitHub [GitHub - Luiserov/proyecto-introdatsci...](#)

Repositorio del proyecto final del primer módulo de Introducción a la Ciencia de Datos. - GitHub -...