UNIVERSIDAD
Panamericana®

ANÁLISIS AUTOMATIZADO DE CALIDAD DE VINOS BLANCOS

Herramienta de Análisis de Datos con Inteligencia Artificial

Materia:
Minería de Datos

Equipo:
TEAM SOPO

Integrantes:
Yahwthani Morales Gómez
Gabriel Zaid Gutiérrez González
Gabriel Torres Zacarias
Sebastián Avilez Hernández

Profesores:
Ronaldo Jesús Ruíz Álvarez
Oswaldo Rojas Bravo

Fecha:
5 de diciembre de 2025

Resumen

Analizamos una base de datos para generar información usando machine learning, y generación de contenido con API’s de Claude.

Analizamos más de 4500 tipos de vinos blancos donde los datos que nos importa es la Quality que se mide de 3-9

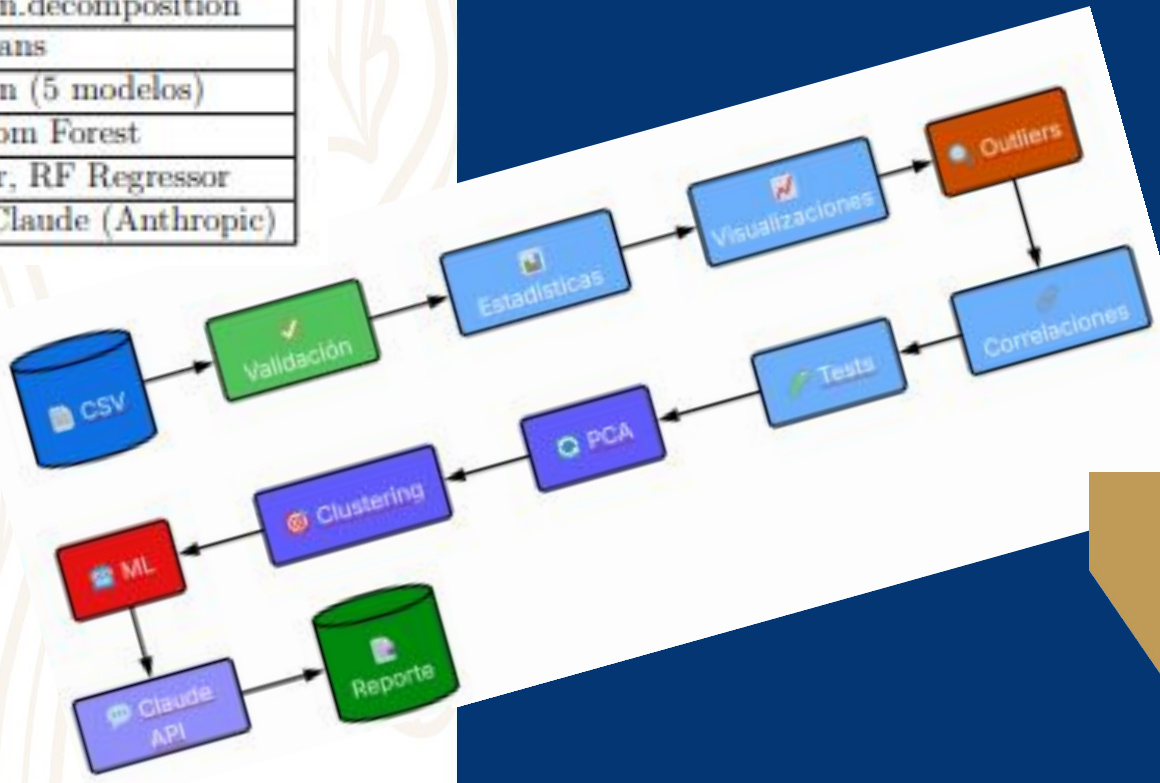
Objetivo del Proyecto

Desarrollar una herramienta para:

- **Analizar estadistica**
- **Visualizar datos**
- **Generar comentarios por IA**

Metodología

Fase	Descripción	Herramientas
1	Carga y validación de datos	pandas
2	Análisis de valores faltantes	pandas, seaborn
3	Estadísticas descriptivas	pandas, numpy
4	Tests de normalidad	scipy.stats
5	Visualizaciones	matplotlib, seaborn
6	Detección de outliers	IQR, Z-score
7	Análisis de correlaciones	Pearson, Spearman
8	Tests estadísticos	ANOVA, Kruskal-Wallis
9	PCA	sklearn.decomposition
10	Clustering	K-Means
11	Machine Learning	sklearn (5 modelos)
12	Feature Importance	Random Forest
13	Regresión	Linear, RF Regressor
14	Generación de insights	API Claude (Anthropic)



Estrategia con la API

Para mayor eficacia le dimos al prompt elemento importante para que nos diera mejor las cosas:

- Rol definido: Se establece que Claude actúa como experto en análisis de datos
- Estructura clara: Se especifican las 8 secciones requeridas
- Datos concretos: Se incluye el resumen completo del análisis
- Formato de salida: Se solicita respuesta en español y formato profesional

Descripción

Variable	Descripción	Media	Std
fixed acidity	Acidez fija (g/L)	6.85	0.84
volatile acidity	Acidez volátil (g/L)	0.28	0.10
citric acid	Ácido cítrico (g/L)	0.33	0.12
residual sugar	Azúcar residual (g/L)	6.39	5.07
chlorides	Cloruros (g/L)	0.046	0.022
free sulfur dioxide	SO2 libre (mg/L)	35.31	17.01
total sulfur dioxide	SO2 total (mg/L)	138.36	42.50
density	Densidad (g/cm^3)	0.994	0.003
pH	pH	3.19	0.15
sulphates	Sulfatos (g/L)	0.49	0.11
alcohol	Alcohol (% vol)	10.51	1.23
quality	Calidad (0-10)	5.88	0.89

a		Porcentaje	Acumulado
3	20	0.41 %	0.41 %
4	163	3.33 %	3.74 %
5	1,457	29.75 %	33.49 %
6	2,198	44.88 %	78.37 %
7	880	17.97 %	96.34 %
8	175	3.57 %	99.91 %
9	5	0.10 %	100.00 %

Correlaciones

#	Variable 1	Variable 2	Correlación
1	residual sugar	density	**0.839**
2	density	alcohol	**-0.780**
3	free sulfur dioxide	total sulfur dioxide	0.616
4	total sulfur dioxide	density	0.530
5	residual sugar	alcohol	-0.451

Variable	Pearson	Spearman
alcohol	**0.436**	**0.440**
density	-0.307	-0.348
chlorides	-0.210	-0.315
volatile acidity	-0.195	-0.197
total sulfur dioxide	-0.175	-0.197

Resultados de Machine Learning

- **Random Forest: mejor modelo con 67.55% accuracy.**
- **Comparación con Decision Tree, KNN, Logistic Regression, Gradient Boosting.**

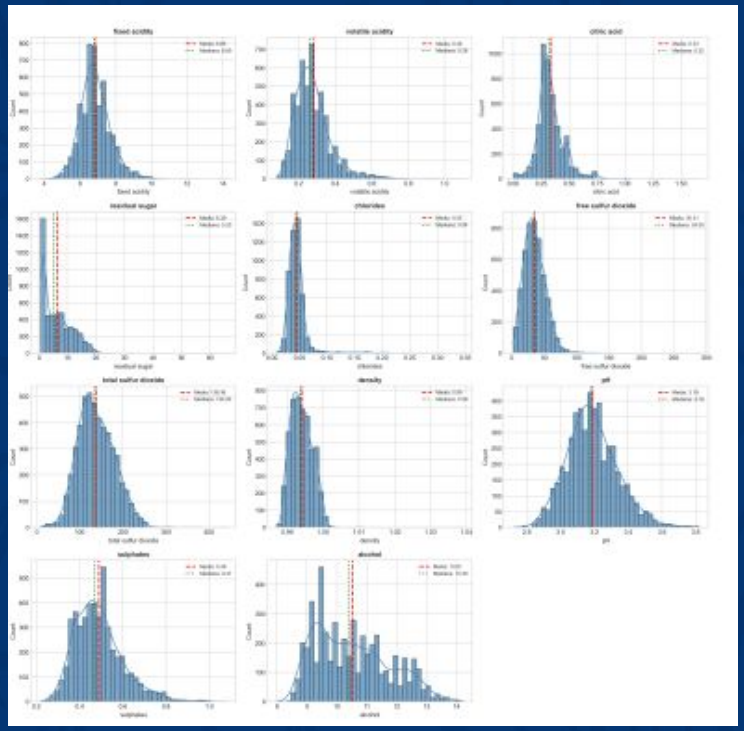
Feature Importance

Alcohol es la variable más importante. Le siguen density, volatile acidity y sulfuros.

Ranking	Variable	Importancia
1	**alcohol**	**11.65 %**
2	density	10.30 %
3	volatile acidity	10.05 %
4	free sulfur dioxide	9.22 %
5	total sulfur dioxide	9.08 %
6	pH	8.66 %
7	chlorides	8.64 %
8	residual sugar	8.63 %
9	citric acid	8.18 %
10	sulphates	7.83 %
11	fixed acidity	7.76 %

Visualizaciones Generadas

Con todo los datos, se generaron 16 graficas de las que incluyen Heatmaps, PCA, clustering, boxplots, histogramas y matriz de confusión.



Insights de Claude

- Supremacía del alcohol como predictor
- Relación inversa densidad-alcohol
- Significancia estadística universal
- Ausencia de normalidad
- Estructura de correlación compleja

Recomendaciones de Preprocesamiento

- Isolation Forest para outliers.
- Transformación log para chlorides.
- SMOTE para balanceo.
- Ratios derivados como alcohol/azúcar.
- Validación cruzada estratificada.

Limitaciones

- Restricciones del dataset
- Desbalance de clases
- Subjetividad en evaluación
- Limitaciones del modelo

Conclusiones

El análisis confirma que el contenido alcohólico es el predictor dominante de calidad en vinos blancos. Los modelos de Machine Learning alcanzan rendimiento moderado (67.55 % accuracy), limitados principalmente por el desbalance severo de clases.

Gracias

UNIVERSIDAD
Panamericana®