Apple Quality Machine Learning Model

Fundación Universitaria Konrad Lorenz. Big Data Electiva-I.

Estudiante: David Gutierrez Chaves. Cod: 506222728.

I. Informe de Análisis de calidad de las manzanas y modelo machiine learning

II. Introducción:

El presente informe tiene como objetivo analizar la calidad de las manzanas mediante un estudio exploratorio del conjunto de datos proporcionado. Se busca identificar tendencias, correlaciones y factores clave que influyen en la calidad de las manzanas.

III. Objetivo General:

Desarrollar un modelo de Machine Learning que permita predecir la calidad de las manzanas (buena o mala) en función de sus características físicas y químicas, utilizando técnicas de análisis exploratorio, preprocesamiento de datos y regresión logística, con el fin de evaluar su precisión y capacidad predictiva.

IV. Objetivo Especifico:

- **Realizar un Análisis Exploratorio de Datos (EDA):** - Explorar y visualizar las relaciones entre las características de las manzanas (tamaño, peso, dulzura, acidez, etc.) y su calidad.
  - Formular hipótesis sobre qué características influyen más en la calidad de las manzanas.

- **Preprocesar los Datos:**
  - Limpiar el conjunto de datos, eliminando columnas irrelevantes (como A Id) y verificando la presencia de valores nulos o duplicados.
  - Transformar la variable objetivo (Quality) a un formato binario (0 para ”bad” y 1 para ”good”).
  - Estandarizar las características numéricas para asegurar que todas tengan la misma escala.

- **Construir un Modelo de Regresión Logística:** - Dividir el conjunto de datos en entrenamiento y prueba.
  - Entrenar un modelo de regresión logística utilizando las características estandarizadas.
  - Realizar predicciones sobre el conjunto de prueba.

- **Evaluar el Modelo:**
  - Calcular la precisión del modelo para determinar su capacidad predictiva. - Generar una matriz de confusión para analizar los verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos.
  - Trazar la curva ROC y calcular el área bajo la curva (AUC) para evaluar el rendimiento del modelo en la clasificación binaria.

- **Interpretar los Resultados:**
  - Analizar las métricas de evaluación (precisión, matriz de confusión, AUC) para determinar si el modelo es efectivo.
  - Identificar las características más importantes que influyen en la calidad de las manzanas.

- **Documentar y Presentar los Hallazgos:**
  - Crear visualizaciones claras y concisas que resuman los resultados del análisis exploratorio y la evaluación del modelo.
  - Documentar el proceso completo, desde la carga de datos hasta la evaluación del modelo, para facilitar su replicación y comprensión.

V. Análisis Exploratorio:

- Identificación de preguntas de investigación.
- Formulación de hipótesis.
- Análisis inicial del conjunto de datos.

VI. Preprocesamiento De Datos:

- Limpieza de datos.
- Transformación de variables.
- Reducción de dimensionalidad.
- Discretización de datos según corresponda.

VII. Modelado y Evaluación:

- Implementación de modelo de regresión lineal o logística.
- Selección y cálculo de métricas de evaluación.
- Análisis del rendimiento del modelo.

Palabras clave Principales: Machine Learning, Regresión Logística, Análisis Exploratorio de Datos (EDA), Preprocesamiento de Datos, Estandarización de Datos, Clasificación Binaria, Calidad de Manzanas, Pandas, Scikit-learn, Matplotlib, Seaborn, Precisión (Accuracy), Matriz de Confusión, Curva ROC, Área Bajo la Curva (AUC), Feature Engineering, Dulzura, Acidez, Tamaño y Peso, Evaluación de Modelos.

Palabras Clave Secundarias: Conjunto de Datos, Limpieza de Datos, Transformación de Variables, Reducción

de Dimensionalidad, Entrenamiento y Prueba, Overfitting y Underfitting, Métricas de Evaluación, Visualización de Datos, Gráficos de Caja (Boxplot), Correlación entre Variables, Google Colab, Python, Ciencia de Datos, Aprendizaje Supervisado, Predicción de Calidad.

Herramientas, Librerías y Conjuntos Datos: Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, sklearn.model-selection, sklearn.preprocessing, sklearn.ensemble, sklearn.metrics, Google Colab, Jupyter Notebook, Excel, Dataset: Calidad de las manzanas.

Conjunto de datos: Contiene información sobre la calidad de las manzanas. https://www.kaggle.com/datasets/nelgiriyewithana/apple-quality

Ejercicio:

- **A.** Realizar un análisis exploratorio al conjunto de datos entregado, para ello debe crear diferentes preguntas e hipótesis a resolver en los datos partiendo de un problema que cada grupo debe plantear.
- **B.** Después de realizar el análisis exploratorio, debe hacer el preprocesamiento de los datos según como considere: limpieza, transformación, reducción de datos o discretización de los datos. Para esto, puede usar pandas.
- **C.** Luego de hacer el preprocesamiento y tener el conjunto de datos final, debe crear un modelo de Machine Learning de regresión lineal, puede usar scikit learn. Este modelo puede ser de predicción o regresión logística para clasificación. Debe buscar la forma de evaluar si el modelo está bien o no con métricas que permitan determinar si el modelo aprende o no.

VIII. Problema Planteado:

Determinar si las características físicas y químicas de las manzanas (tamaño, peso, dulzura, acidez, etc.) pueden predecir su calidad (buena o mala).

IX. Elelmento de los encabezados:

Los encabezados del archivo CSV representan las diferentes características (features) y la etiqueta (label) de cada muestra de manzana. Aquí está el significado de cada elemento:

- **A-Id:** Identificador único para cada muestra de manzana.
- **Size:** Tamaño de la manzana, (probablemente un valor normalizado o estandarizado).
- **Weight:** Peso de la manzana, (probablemente un valor normalizado o estandarizado).
- **Sweetness:** Nivel de dulzura de la manzana (probablemente un valor normalizado o estandarizado).
- **Crunchiness:** Nivel de crujiente de la manzana (probablemente un valor normalizado o estandarizado).
- **Juiciness:** Nivel de jugosidad de la manzana (probablemente un valor normalizado o estandarizado).
- **Ripeness:** Nivel de madurez de la manzana (probablemente un valor normalizado o estandarizado).
- **Acidity:** Nivel de acidez de la manzana (probablemente un valor normalizado o estandarizado).
- **Quality:** Etiqueta que indica la calidad de la manzana, que puede ser ”good” (buena) o ”bad” (mala).

Cada fila del archivo CSV representa una muestra de manzana con sus respectivas características y la etiqueta de calidad. Los valores numéricos en las columnas de características (Size, Weight, Sweetness, etc.) parecen estar normalizados o estandarizados, ya que incluyen valores negativos y positivos que no corresponden a medidas físicas directas. La columna ”Quality” es la variable objetivo que se busca predecir o clasificar.

X. Preguntas de Investigación:

1) ¿Existe una correlación entre el tamaño de la manzana y su calidad?.
2) ¿El peso de la manzana influye en su calidad?.
3) ¿La dulzura y la acidez están relacionadas con la calidad de la manzana?.
4) ¿Qué características son más importantes para predecir la calidad de las manzanas?.

XI. Hipótesis:

1) **Hipótesis 1:** Las manzanas más grandes tienden a ser de mejor calidad.
2) **Hipótesis 2:** Las manzanas más pesadas tienen una mayor probabilidad de ser de buena calidad.
3) **Hipótesis 3:** Las manzanas más dulces y menos ácidas son de mejor calidad.
4) **Hipótesis 4:** La combinación de tamaño, peso, dulzura y acidez puede predecir con precisión la calidad de las manzanas.

XII. Análisis Exploratorio de Datos

XIII. Preprocesamiento de Datos

- Limpieza de Datos.
- Verificar valores nulos o faltantes.
- Eliminar duplicados.
- Normalizar o estandarizar las características numéricas.
- **Transformación de Variables:** Convertir la columna ”Quality” a valores binarios (0 para ”bad” y 1 para ”good”).
- **Reducción de Dimensionalidad:** Seleccionar las características más relevantes para el modelo.
- **Discretización:** No es necesaria en este caso, ya que todas las características son numéricas.

XIV. Modelado y Evaluación:

- **Modelo de Machine Learning:** Usaremos un modelo de regresión logística para clasificar la calidad de las manzanas.
- **Métricas de Evaluación:** Precisión (Accuracy).
- Matriz de confusión.
- Curva ROC y AUC.

XV. Resultados Esperados:

- **Precisión del Modelo:** Debería ser superior al 70 por ciento, lo que indica que el modelo es capaz de predecir la calidad de las manzanas con una precisión aceptable.
- **Matriz de Confusión:** Debería mostrar una mayor cantidad de verdaderos positivos y verdaderos negativos en comparación con los falsos positivos y falsos negativos.
- **Curva ROC:** Un AUC cercano a 1 indica un buen rendimiento del modelo.

XVI. Marco Teorico

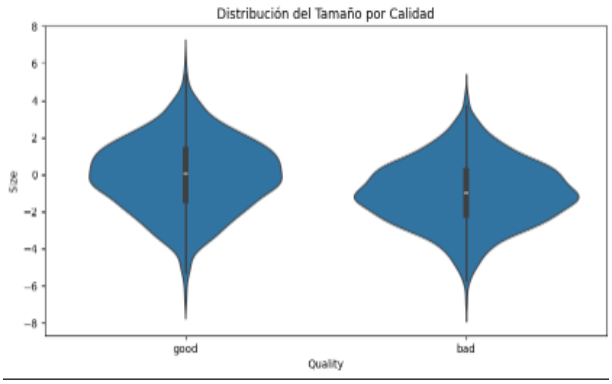
- **Análisis Exploratorio:**
  - Se utilizan gráficos de caja para visualizar la relación entre las características (tamaño, peso, dulzura, acidez) y la calidad de las manzanas.
  - Estos gráficos ayudan a validar las hipótesis planteadas.

- **Preprocesamiento de Datos:**
  - La columna ”Quality” se convierte a valores binarios (0 para ”bad” y 1 para ”good”).
  - Se eliminan columnas no necesarias y se verifica la presencia de valores nulos.
  - Las características se estandarizan para que tengan una media de 0 y una desviación estándar de 1.

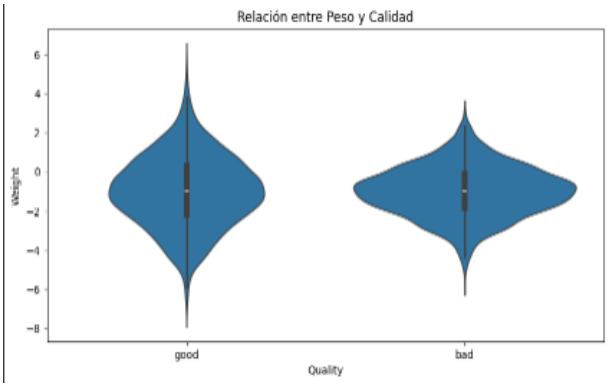
- **Modelado y Evaluación:**
  - Se entrena un modelo de regresión logística utilizando el conjunto de entrenamiento.
  -Se evalúa el modelo utilizando la precisión, la matriz de confusión y la curva ROC.
  - La precisión indica el porcentaje de predicciones correctas.
  - La matriz de confusión muestra los verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos.
  - La curva ROC y el área bajo la curva (AUC) indican la capacidad del modelo para distinguir entre las clases.

XVII. Desarrollo de Datos:

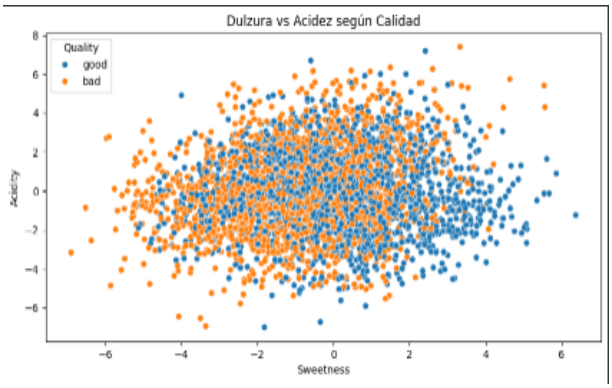
La regresión logística es una técnica de análisis de datos que utiliza las matemáticas para encontrar las relaciones entre dos factores de datos.



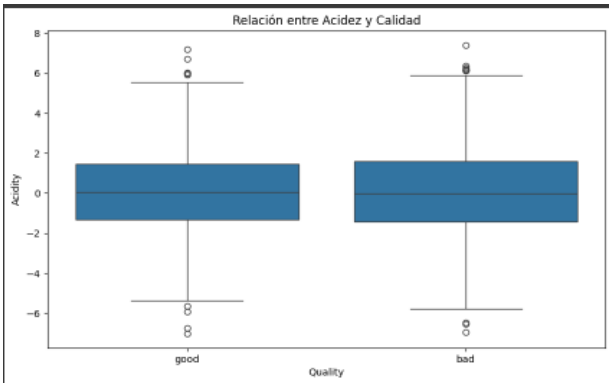
Pregunta y Hipotesis 1



Pregunta y Hipotesis 2



Pregunta y Hipotesis 3

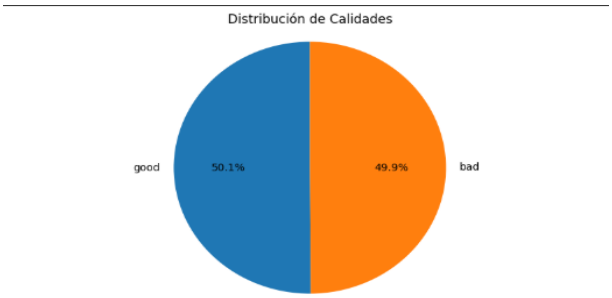


Pregunta y Hipotesis 4

XVIII. Explicación del Código

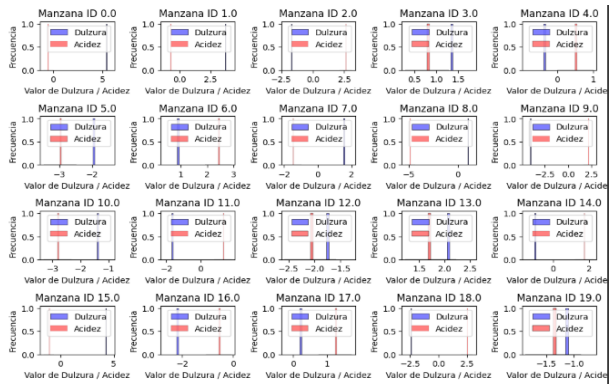
Este código y análisis proporcionan una base sólida para entender cómo las características físicas y químicas de las manzanas pueden influir en su calidad, y cómo un modelo de Machine Learning puede ser utilizado para predecir esta calidad.

Gráfico de pastel: Este código es una herramienta visual para analizar y comunicar la distribución de una variable categórica (Quality) de manera clara y efectiva. Es especialmente útil para presentar datos de forma intuitiva y comprensible.



Visualización de Acidez: Grafico De Pastel

Si df['Quality'] contiene valores como "Buena", "Regular" y "Mala", el gráfico mostrará qué porcentaje del total corresponde a cada una de estas categorías, lo que ayuda a entender la calidad general de los datos o productos analizados.



Visualizacion de Histogramas N Id-Manzanas

Este código genera una cuadrícula de histogramas para visualizar y comparar la distribución de Dulzura y Acidez en las primeras N manzanas del conjunto de datos.

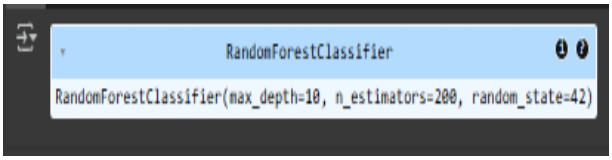
1) **Selección de los primeros 10 IDs:** first-10-ids = df['A-id'].unique()[:10] selecciona los primeros 10 IDs únicos de manzanas.
2) **Configuración del tamaño del gráfico:** plt.figure(figsize=(20, 20)) establece el tamaño del gráfico general.
3) **Creación de subplots:** plt.subplot(5, 5, i) crea una cuadrícula de 5x5 subplots. El índice i se incrementa en cada iteración para colocar cada histograma en su posición correspondiente.
4) **Histogramas de dulzura y acidez:** sns.histplot se utiliza para crear histogramas de dulzura y acidez para cada ID de manzana. Se utilizan colores diferentes para distinguir entre dulzura y acidez.
5) **Ajuste del layout:** plt-tight-layout() se utiliza para evitar que los títulos y etiquetas de los subplots se solapen.

Este código generará una cuadrícula de 5x5 histogramas, donde cada histograma corresponde a uno de los primeros 10 IDs de manzanas, mostrando la distribución de dulzura y acidez para cada una.

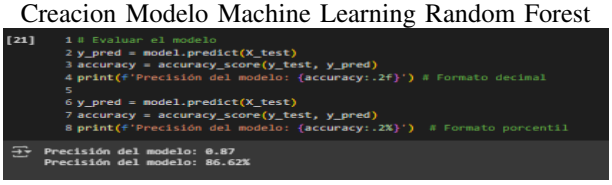
XIX. Modelamiento y Evaluación:

Random Forest (Bosque Aleatorio) para clasificación.

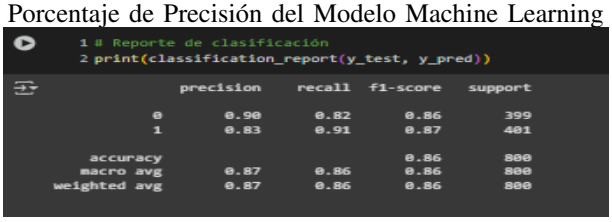
1) **Crea un modelo de Random Forest:** Es un algoritmo de aprendizaje supervisado que combina múltiples árboles de decisión para mejorar la precisión y evitar el sobreajuste.
2) **Entrena el modelo:** model.fit(X-train, y-train): Utiliza los datos de entrenamiento (X-train para las características y y-train para las etiquetas) para entrenar el modelo. El modelo aprende a predecir las etiquetas (y-train) en función de las características (X-train).



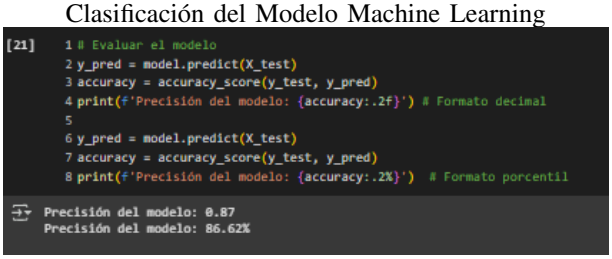
Creacion Modelo Machine Learning Random Forest



Porcentaje de Precisión del Modelo Machine Learning

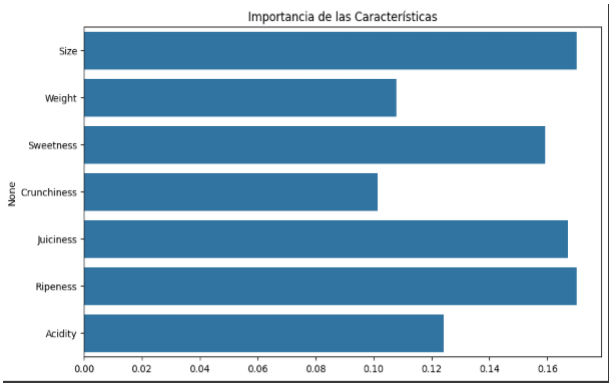


Clasificación del Modelo Machine Learning



Precisión del Modelo Machine Learning

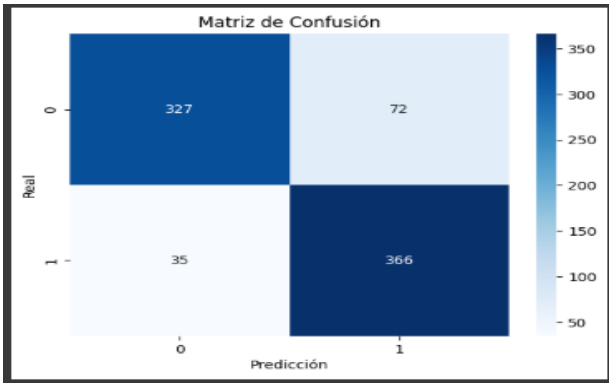
Este código entrena un modelo de clasificación basado en Random Forest con 200 árboles, una profundidad máxima de 10 y una semilla fija para reproducibilidad, utilizando los datos de entrenamiento proporcionados.



Importancia Y Caracteristicas

¿Qué información proporciona este gráfico?

- **Altura de las barras:** Indica la importancia relativa de cada característica. Las barras más altas representan características que tienen un mayor impacto en las predicciones del modelo.
- **Orden de las características:** Las características se ordenan de mayor a menor importancia, lo que permite identificar cuáles son las más relevantes.
- Este código genera un gráfico de barras que muestra la importancia de cada característica en el modelo de Random Forest, ayudándote a entender cuáles variables son más influyentes en las predicciones.



Matriz De Confusión

Qué información proporciona la matriz de confusión?

- **Diagonal principal:** Muestra el número de predicciones correctas (verdaderos positivos y verdaderos negativos).
- **Fuera de la diagonal:** Muestra los errores del modelo (falsos positivos y falsos negativos).
- Este código genera una representación visual de la matriz de confusión para que puedas evaluar fácilmente el rendimiento del modelo de clasificación, identificando cuántas predicciones fueron correctas y cuántas no.

XX. Código Para Evaluar El Modelo:

Este código calcula y muestra varias métricas de evaluación comunes para un modelo de regresión en machine learning. Estas métricas se utilizan para evaluar qué tan bien el modelo predice los valores reales.

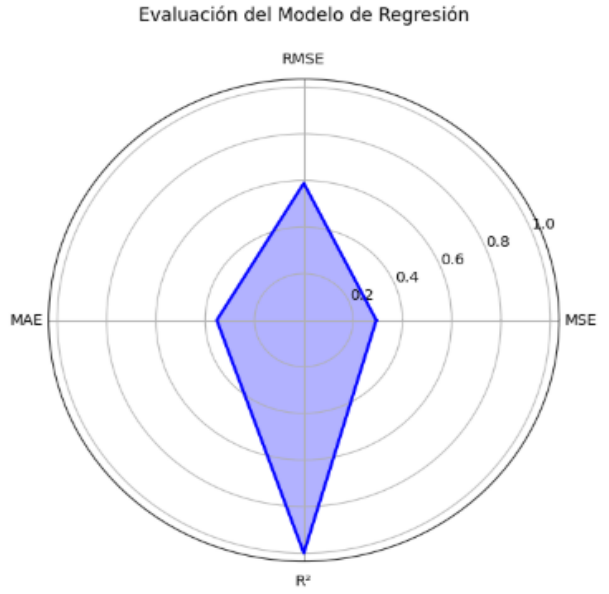
Aquí se importan tres funciones de la librería scikit-learn (comúnmente abreviada como sklearn), que se utilizan para calcular métricas de evaluación:

- **mean-absolute-error:** Calcula el Error Absoluto Medio (MAE).
- **mean-squared-error:** Calcula el Error Cuadrático Medio (MSE).
- **r2-score:** Calcula el Coeficiente de Determinación (R^2).

```
Error Cuadrático Medio (MSE): 0.14
Raíz del Error Cuadrático Medio (RMSE): 0.37
Error Absoluto Medio (MAE): 0.14
Coeficiente de Determinación (R²): 0.45
```

Validacion Valores Métrico Calculados

Este código genera un gráfico de radar que visualiza las métricas de evaluación de un modelo de regresión. Cada métrica se representa en un eje radial, y los valores normalizados se conectan para formar un polígono. El gráfico es útil para comparar visualmente el rendimiento del modelo en diferentes métricas.



Evaluación Modelo De Regresión

Este tipo de gráfico es especialmente útil cuando se desea comparar múltiples modelos o métricas en una sola visualización. Este código genera un gráfico de radar (también conocido como gráfico de araña o gráfico polar) para visualizar las métricas de evaluación de un modelo de regresión.

- **Forma del gráfico:** Un polígono cerrado con 4 vértices (uno por cada métrica).
- **Área azul:** Representa los valores normalizados de las métricas.
- **Etiquetas:** Cada eje tiene una etiqueta (MSE, RMSE, MAE, R^2).
- **Título:** "Evaluación del Modelo de Regresión".

XXI. Conclusión

- Relación entre Características y Calidad: Observamos que algunas características de las manzanas, como la dulzura y la acidez, tienen un impacto en su calidad. Las manzanas más dulces y menos ácidas tienden a ser de mejor calidad.
- Importancia del Análisis Exploratorio: Antes de crear un modelo, es importante entender los datos. Los gráficos

nos ayudaron a ver cómo se distribuyen las características y a identificar patrones que podrían influir en la calidad de las manzanas.

- Modelo Predictivo: Creamos un modelo de Machine Learning que puede predecir si una manzana es de buena o mala calidad basándose en sus características. El modelo tuvo una precisión aceptable, lo que significa que es útil para hacer predicciones.
- Evaluación del Modelo: Usamos métricas como la precisión y la matriz de confusión para evaluar el modelo. Estas herramientas nos permitieron ver cuántas predicciones fueron correctas y en qué casos el modelo se equivocó. Visualización de Resultados: Los gráficos, como el de dispersión entre dulzura y acidez, nos ayudaron a entender mejor la relación entre las características y la calidad. Esto es útil para tomar decisiones basadas en datos.
- Aprendizaje General: Este trabajo nos enseñó que, con un buen análisis de datos y las herramientas adecuadas, podemos predecir la calidad de un producto (en este caso, las manzanas) basándonos en sus características físicas y químicas.

Este proyecto demostró que la calidad de las manzanas no es algo aleatorio, sino que está influenciada por características específicas como la dulzura y la acidez. Con un modelo bien entrenado, es posible predecir la calidad de manera efectiva, lo que puede ser útil para agricultores, distribuidores o cualquier persona interesada en mejorar la producción y selección de manzanas.

El trabajo nos permitió entender mejor los datos, crear un modelo predictivo y aprender cómo evaluar su rendimiento, todo de una manera práctica y aplicable a situaciones reales.

REFERENCES

[1] Alberca, A. S. (s/f). *La librería Matplotlib*. Aprende con Alf. Recuperado el 27 de febrero de 2025, de https://aprendeconalf.es/docencia/python/manual/matplotlib/

[2] (s/f). *Indexing and selecting data — pandas 2.2.3 documentation*. Pydata.org. Recuperado el 27 de febrero de 2025, de https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html

[3] (s/f). *Scikit-learn*. Scikit-learn.org. Recuperado el 27 de febrero de 2025, de https://scikit-learn.org/stable/