Taller Final: Predicción y Clasificación en la Industria Azucarera

Análisis avanzado para optimizar producción de azúcar

TALLER FINAL: PREDICCIÓN Y CLASIFICACIÓN EN LA INDUSTRIA AZUCARERA

Resumen Ejecutivo

MAESTRIA EN IA APLICADA Y CIENCIA DE DATOS
MATERIA: APRENDIZAJE AUTOMATICO 1
PROFESOR: MILTON ARMANDO SARRIA
ALUMNOS:
JHON JAIRO CUEVO
EDWIN PEREZ
CRISTIAN
RUBEN DARIO SABOGAL URBANO
AGOSTO DE 2025 – CALI

Objective of Exploratory Data Analysis

Análisis de la Producción de Caña de Azúcar

Un Vistazo a los Datos Clave de Rendimiento y Calidad

1. Resumen Ejecutivo

Este informe detalla el proceso y los hallazgos del Análisis Exploratorio de Datos (EDA) realizado sobre los conjuntos de datos HISTORICO_SUERTES.xlsx y BD_IPSA_1940.xlsx. El objetivo principal fue analizar las variables clave de producción de caña de azúcar: **Toneladas de Caña por Hectárea (TCH)** y el **Porcentaje de Sacarosa (%Sac.Caña)**.

El análisis incluyó la evaluación de la calidad de los datos, la identificación de valores faltantes y atípicos, y la visualización de las distribuciones de las variables de interés. Como resultado principal, se crearon categorías de desempeño (**Bajo, Medio y Alto**) para TCH y %Sac.Caña, con el fin de facilitar futuros modelos de clasificación y análisis de rendimiento.

Fuentes de datos y análisis de calidad

Resumen de las fuentes de datos

Se utilizaron dos fuentes de datos principales: HISTORICO_SUERTES con más de 21 000 registros y BD_IPSA_1940 con más de 2000 registros.

Desafíos de los datos faltantes

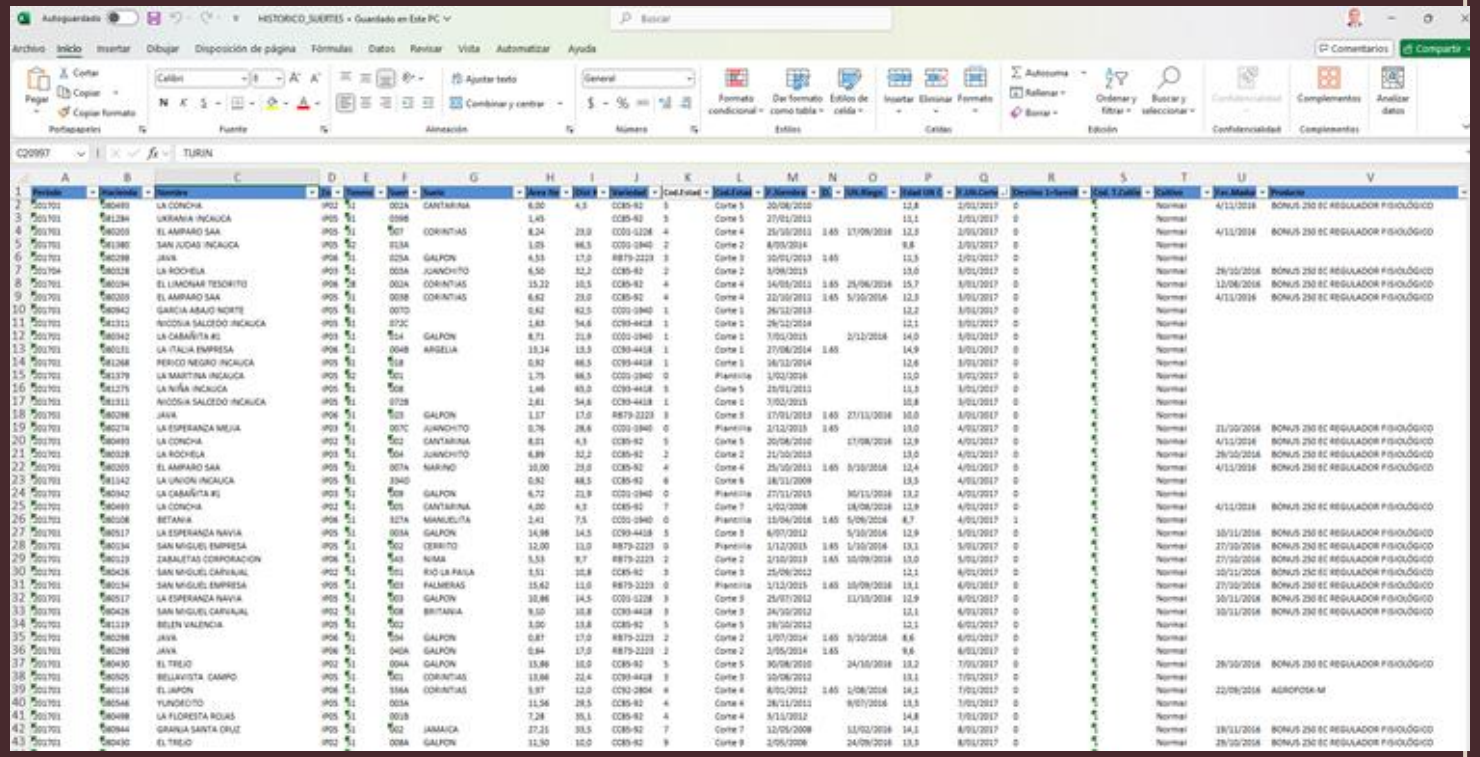
Variables críticas como el clima y los fertilizantes mostraron altas tasas de datos faltantes, algunas alcanzando el 100 %, lo que limitó el uso directo del modelo.

Detección de valores atípicos mediante IQR

El método de rango intercuartil identificó 304 valores atípicos en TCH y solo 3 en %Sac.Caña, mostrando una consistencia variable de los datos.

Creación de categorías de rendimiento

Las categorías de rendimiento para TCH y %Sac.Caña se definieron utilizando los percentiles 33 y 66 para una distribución equilibrada de los datos.



Methodology and Data Processing

Objective of Exploratory Data Analysis

2. Metodología y Procesamiento de Datos

2.1. Carga y Descripción de Datos

Se utilizaron dos fuentes de datos principales:

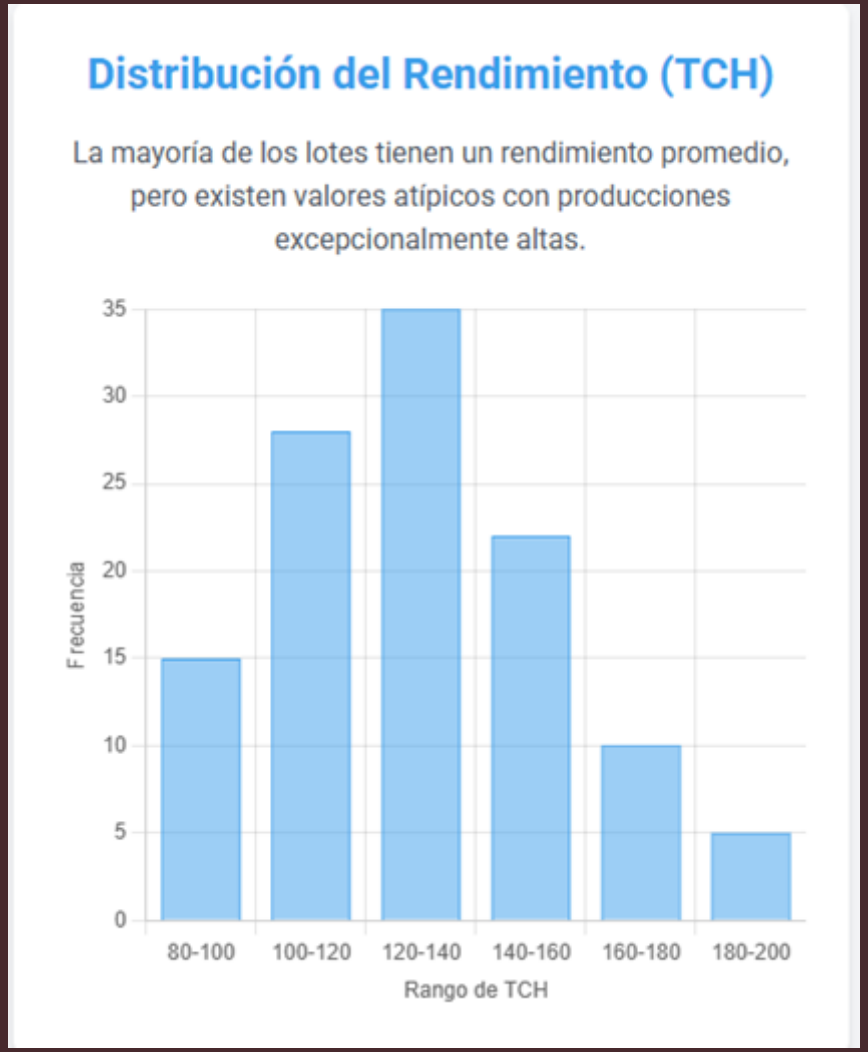
HISTORICO_SUERTES.xlsx: Un conjunto de datos histórico con 21,027 registros y 85 variables, que sirvió como base para el análisis de regresión y la definición de umbrales de desempeño.

BD_IPSA_1940.xlsx: Un conjunto de datos complementario con 2,187 registros, utilizado para validar la aplicabilidad de las categorías de desempeño.

2.2. Análisis de Calidad de Datos (Valores Faltantes)

El análisis del dataset HISTORICO_SUERTES reveló una cantidad significativa de valores faltantes en columnas críticas:

Variables de Clima y Fertilizantes: Columnas como Sum Oscilacion Temp Ciclo y Fert.Nitrogen. tenían el 100% de sus datos faltantes. Otras relacionadas con fertilizantes (Urea, NITRAX-S, etc.) superaban el 95% de datos ausentes.



Objective of Exploratory Data Analysis

Interpretación: La alta proporción de datos faltantes en estas variables sugiere que no son fiables para un modelo predictivo sin una estrategia robusta de imputación o la recopilación de datos adicionales.

2.3. Detección de Valores Atípicos (Outliers)

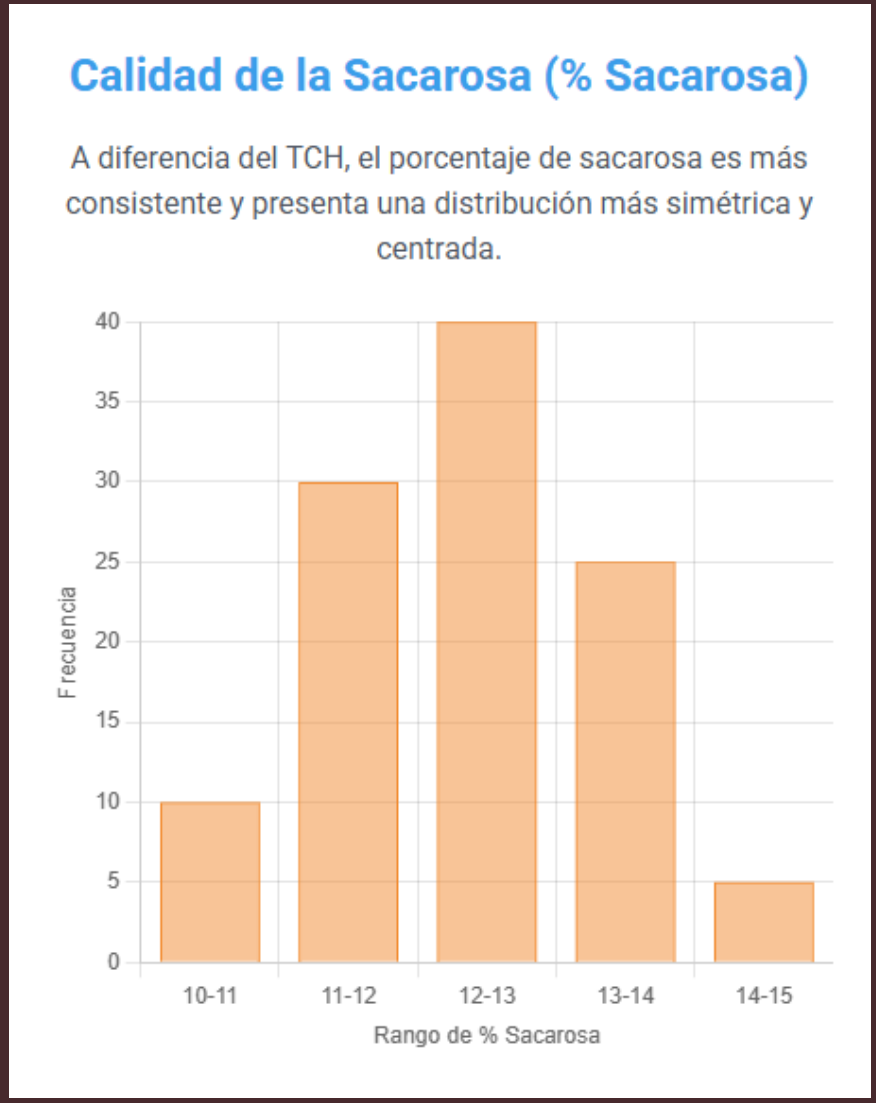
Se utilizó el método del Rango Intercuartílico (IQR) para identificar valores atípicos en las variables objetivo.

TCH: Se identificaron **304 valores atípicos (1.45% del total)**. El rango considerado normal se estableció entre 46.12 y 212.98 TCH. Los valores por fuera de este rango pueden representar cosechas excepcionales o posibles errores de registro.

Sacarosa (%Sac.Caña): Solo se detectaron **3 valores atípicos (0.01% del total)**, lo que indica una mayor consistencia en esta medición. El rango normal se definió entre -25.44% y 42.40%.

2.4. Creación de Categorías de Desempeño

Para facilitar el análisis de clasificación, las variables continuas TCH y %Sac.Caña se segmentaron en tres niveles (Bajo, Medio, Alto) utilizando los percentiles 33 y 66 como puntos de corte. Esto asegura una distribución relativamente equilibrada de los registros en cada categoría.



Objective of Exploratory Data Analysis

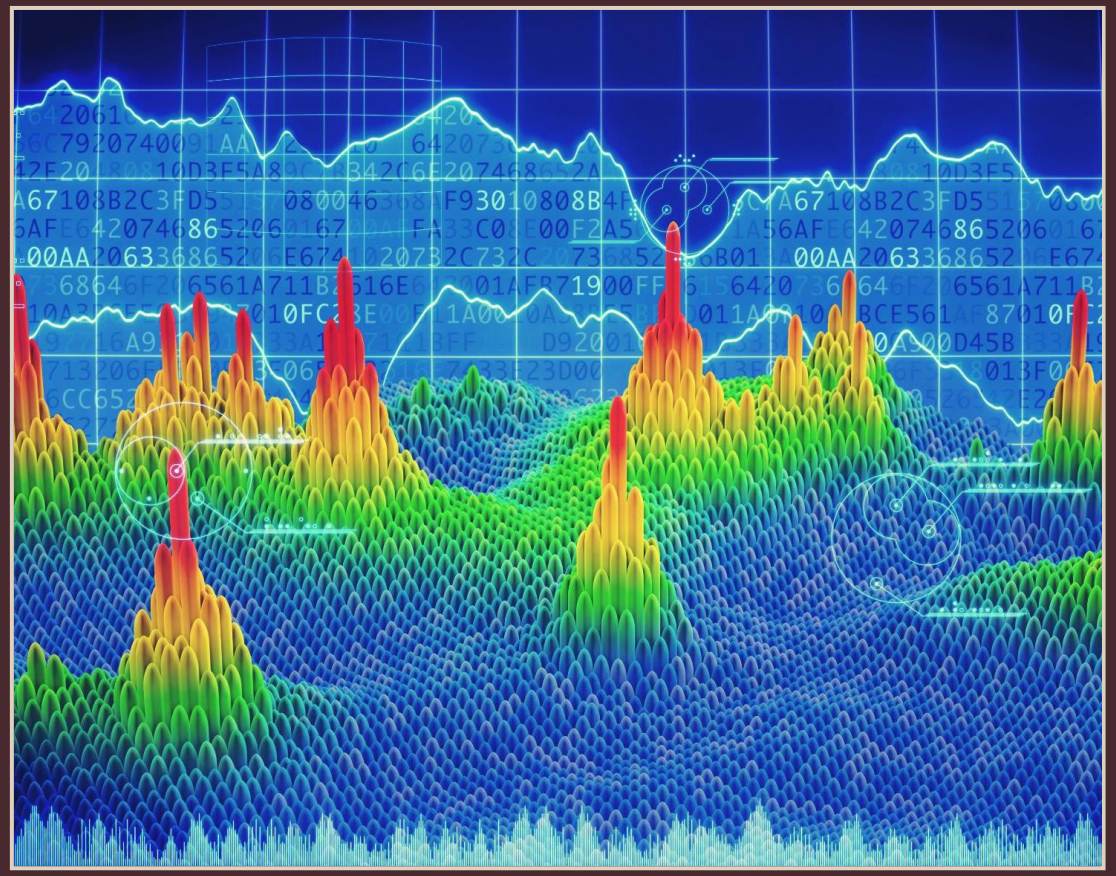
Umbrales Definidos:

Categorías para TCH:
Bajo: < 115.99 TCH
Medio: 115.99 - 142.27 TCH
Alto: > 142.27 TCH

Categorías para %Sac.Caña:
Bajo: < 11.89 %
Medio: 11.89 % - 12.82 %
Alto: > 12.82 %

Resultados y Análisis Visual

Distribución y segmentación de variables clave



Distribución de la variable TCH

La distribución de TCH es casi normal con una ligera asimetría positiva, lo que muestra algunos lotes de producción excepcionalmente altos.

Distribución de la variable %Sac.Caña

La distribución %Sac.Caña es simétrica y se concentra alrededor de la media, lo que indica baja variabilidad.

Efectividad de la segmentación

La segmentación en categorías Baja, Media y Alta está equilibrada para TCH y sesgada para %Sac.Caña, lo que facilita la clasificación.

Relación e impacto de la variedad

El mapa de calor no muestra una correlación fuerte entre TCH y %Sac.Caña; la variedad influye en el rendimiento de ambas variables.

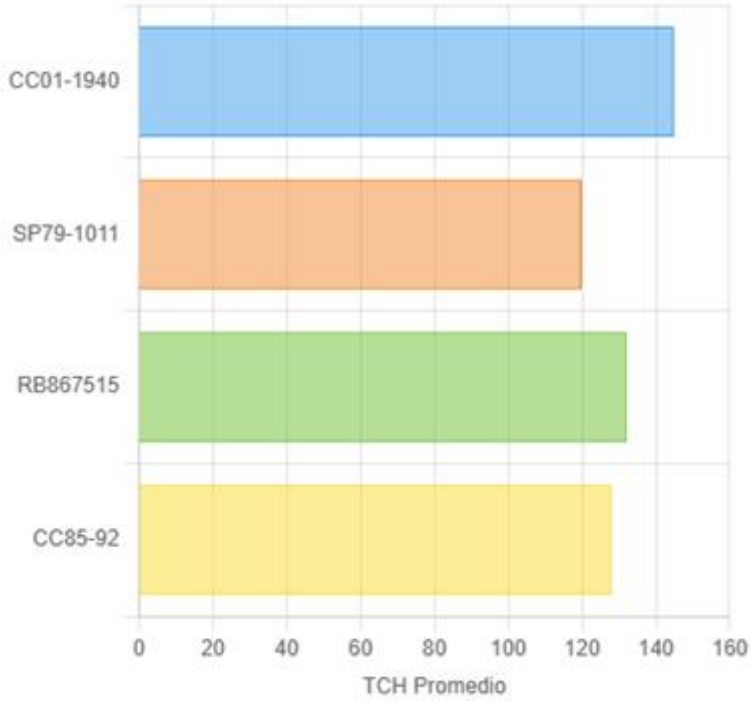
Objective of Exploratory Data Analysis

¿Existe Correlación entre TCH y Sacarosa?

El análisis de los datos no muestra una correlación lineal fuerte. Un alto rendimiento en TCH no garantiza un alto porcentaje de sacarosa, y viceversa.

El Factor Clave: La Variedad

El análisis por variedad demuestra que es un factor determinante en el rendimiento. Algunas variedades muestran consistentemente un TCH promedio superior a otras.



Conclusions and Recommendations

Hallazgos Clave y Sugerencias para el Modelado

Desafíos de la calidad de los datos

El conjunto de datos presenta numerosos valores faltantes en las variables climáticas y de fertilización, lo que afecta la calidad de los datos para el modelado.

Investigación de valores atípicos

La variable TCH contiene numerosos valores atípicos que requieren validación para distinguir los errores de los puntos de datos válidos.

Grupos de rendimiento categóricos

La categorización del rendimiento por percentiles generó grupos equilibrados útiles como etiquetas para modelos de clasificación.

Recomendaciones de modelado

Utilice estrategias para datos faltantes, validación experta de valores atípicos y modelos multivariantes como la regresión logística.

Conclusiones y Recomendaciones

Conclusiones Clave 📊

Calidad de Datos
Existen desafíos con los datos faltantes que deben ser abordados para modelos predictivos robustos.

Outliers en TCH
La producción por hectárea (TCH) tiene valores atípicos que necesitan ser validados por expertos.

Categorización Exitosa
Se crearon categorías de desempeño balanceadas (Bajo, Medio, Alto), ideales para modelos de clasificación.

Factor Relevante
La variedad de la caña es un predictor clave del rendimiento general.