

**TALLER FINAL: PREDICCIÓN Y CLASIFICACIÓN EN LA INDUSTRIA AZUCARERA**

**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**

**Análisis Exploratorio de Datos para la Producción de Caña de Azúcar**

Fecha: 28 de agosto de 2025

Basado en: PrediccionAnalisisCaña.ipynb

**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.

**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.
* **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.

**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 %

**3. Resultados y Análisis Visual**

**3.1. Distribución de Variables Clave**

* **Distribución de TCH**: El histograma mostró una distribución aproximadamente normal, aunque con una ligera asimetría positiva (cola a la derecha), indicando que si bien la mayoría de los lotes tienen un rendimiento promedio, existen algunos con producciones excepcionalmente altas.
* **Distribución de %Sac.Caña**: Esta variable presentó una distribución más simétrica y concentrada alrededor de la media, lo que sugiere una menor variabilidad en los niveles de sacarosa en comparación con el TCH.

**3.2. Análisis de Categorías de Desempeño**

La segmentación de los datos en los niveles Bajo, Medio y Alto arrojó la siguiente distribución para el dataset HISTORICO\_SUERTES:

* **Distribución de TCH**: Cada categoría (Bajo, Medio, Alto) contiene aproximadamente un tercio de los registros (alrededor del 33% cada una), lo que confirma que la segmentación por percentiles fue efectiva para crear grupos balanceados.
* **Distribución de %Sac.Caña**: La distribución también fue equilibrada, con la categoría "Bajo" conteniendo un 54%, "Medio" un 12%, y "Alto" un 34% de los registros.

**3.3. Relación entre Categorías y Análisis por Variedad**

* **Relación TCH vs. %Sac.Caña**: El análisis cruzado de las categorías (heatmap) no mostró una correlación visualmente fuerte, lo que sugiere que un alto rendimiento en TCH no garantiza necesariamente un alto porcentaje de sacarosa, y viceversa.
* **Análisis por Variedad**: Se observó que ciertas variedades de caña tienden a tener un mejor desempeño tanto en TCH como en %Sac.Caña, lo que indica que la **variedad es un factor relevante** para predecir el rendimiento.

**4. Conclusiones y Recomendaciones**

**Conclusiones**

1. **Calidad de Datos**: El conjunto de datos HISTORICO\_SUERTES presenta desafíos importantes debido a la alta cantidad de valores faltantes, especialmente en variables climáticas y de fertilización, lo que limita su uso directo en modelos predictivos.
2. **Valores Atípicos**: La variable TCH contiene un número considerable de outliers que deben ser investigados para determinar si son errores o datos válidos de cosechas excepcionales.
3. **Categorización Exitosa**: La creación de categorías de desempeño (Bajo, Medio, Alto) mediante percentiles resultó en grupos bien balanceados, proporcionando una base sólida para futuros modelos de clasificación.
4. **Factores Relevantes**: El análisis preliminar sugiere que la **variedad de la caña** es un factor influyente en el rendimiento. La relación entre TCH y %Sac.Caña no es lineal, lo que justifica un análisis multivariado más profundo.

**Recomendaciones**

1. **Estrategia de Datos Faltantes**: Antes de construir un modelo, es crucial definir una estrategia para manejar los datos faltantes, ya sea mediante la eliminación de registros incompletos o aplicando técnicas de imputación estadística.
2. **Investigación de Outliers**: Se recomienda consultar con expertos en agronomía para validar los valores atípicos de TCH y decidir si deben ser excluidos o ajustados.
3. **Modelado de Clasificación**: Utilizar las nuevas variables categóricas (Categoria\_TCH y Categoria\_Sacarosa) como etiquetas objetivo para entrenar modelos de clasificación (ej. Regresión Logística, Árboles de Decisión) que puedan identificar los factores que determinan un rendimiento alto, medio o bajo.
4. **Análisis de Variables Adicionales**: Profundizar en el análisis de otras variables con datos más completos (como Edad.Precosecha, Lluvias Ciclo, Tenencia) para descubrir su impacto en las categorías de desempeño.