**Diapositiva 1: Portada**

**Análisis Exploratorio y Predictivo de la Producción de Caña de Azúcar**

**Análisis Basado en PrediccionAnalisisCaña.ipynb**

**Fecha:** 28 de agosto de 2025

**Diapositiva 2: Agenda**

1. **Introducción y Objetivos**
2. **Metodología y Fuentes de Datos**
3. **Análisis de Calidad de Datos: Valores Faltantes**
4. **Distribución y Outliers: TCH**
5. **Distribución y Outliers: % Sacarosa**
6. **Creación de Categorías de Desempeño**
7. **Análisis de Correlación**
8. **Relación entre Categorías de Desempeño**
9. **Análisis por Variedad de Caña**
10. **Conclusiones y Próximos Pasos**

**Diapositiva 3: Introducción y Objetivos**

**Objetivo Principal:** Realizar un Análisis Exploratorio de Datos (EDA) para comprender los factores que influyen en la producción de caña de azúcar, con un enfoque en dos indicadores clave:

* **TCH (Toneladas de Caña por Hectárea)**
* **%Sac.Caña (Porcentaje de Sacarosa)**

**Objetivos Específicos:**

* Evaluar la calidad y estructura de los datos.
* Identificar valores faltantes y atípicos.
* Visualizar las distribuciones de las variables de interés.
* Crear categorías de desempeño (**Bajo, Medio, Alto**) para facilitar futuros modelos de clasificación.

**Fuentes de Datos:**

* HISTORICO\_SUERTES.xlsx
* BD\_IPSA\_1940.xlsx

**Diapositiva 4: Calidad de Datos - Valores Faltantes**

**Dataset Analizado:** HISTORICO\_SUERTES.xlsx

**Hallazgos Clave:**

* El conjunto de datos presenta una cantidad significativa de **valores faltantes**.
* Variables críticas como **clima y fertilizantes** tienen más del 95% de datos ausentes, lo que las hace poco fiables para el modelado.
* La variable objetivo **%Sac.Caña** también tiene un porcentaje considerable de datos nulos.

**[Gráfico de Mapa de Calor de Valores Faltantes]**

**Explicación del Gráfico:** Este mapa de calor visualiza la ausencia de datos en las columnas más relevantes. Las áreas más claras indican una alta concentración de valores faltantes. Como se observa, variables como Sum Oscilacion Temp Ciclo y Fert.Nitrogen. están completamente vacías.

**Implicación:** Es crucial definir una estrategia para manejar estos datos (imputación o eliminación) antes de entrenar un modelo predictivo.

**Diapositiva 5: Distribución y Outliers - TCH**

**Análisis de Toneladas de Caña por Hectárea (TCH)**

**Distribución:**

* La producción de TCH muestra una distribución aproximadamente normal, con una ligera **asimetría positiva**.
* La mayoría de los lotes se concentran en un rendimiento promedio, pero existen valores excepcionalmente altos.

**Detección de Outliers:**

* Se identificaron **304 valores atípicos (1.45% del total)**.
* El rango normal se estableció entre **46.12 y 212.98 TCH**. Los valores fuera de este rango requieren una investigación más detallada.

**[Gráfico de Boxplot e Histograma para TCH]**

**Explicación del Gráfico:** El histograma (derecha) confirma la distribución de los datos, mientras que el boxplot (izquierda) resalta visualmente los valores atípicos que se encuentran muy por encima del rango intercuartílico.

**Diapositiva 6: Distribución y Outliers - % Sacarosa**

**Análisis del Porcentaje de Sacarosa (%Sac.Caña)**

**Distribución:**

* A diferencia del TCH, el % de Sacarosa presenta una distribución más **simétrica y centrada** alrededor de la media.
* Esto sugiere una menor variabilidad en los niveles de sacarosa en comparación con la producción por hectárea.

**Detección de Outliers:**

* Solo se detectaron **3 valores atípicos (0.01% del total)**.
* La consistencia en esta medición es mucho mayor, lo que aumenta su fiabilidad como variable objetivo.

**[Gráfico de Boxplot e Histograma para % Sacarosa]**

**Explicación del Gráfico:** El histograma y el boxplot para la sacarosa muestran una concentración de datos bien definida y una cantidad mínima de valores atípicos, lo que indica una alta calidad en esta medición.

**Diapositiva 7: Creación de Categorías de Desempeño**

Para facilitar el análisis de clasificación, se crearon tres niveles de desempeño (**Bajo, Medio, Alto**) utilizando percentiles para asegurar una distribución equilibrada.

**Umbrales Definidos:**

* **TCH:**
  + Bajo: < 115.99 TCH
  + Medio: 115.99 - 142.27 TCH
  + Alto: > 142.27 TCH
* **% Sacarosa:**
  + Bajo: < 11.89 %
  + Medio: 11.89 % - 12.82 %
  + Alto: > 12.82 %

**[Gráfico de Barras de Distribución de Categorías]**

**Explicación del Gráfico:** Los gráficos de barras muestran que los registros se distribuyen de manera bastante uniforme entre las tres categorías para ambas variables, validando el método de segmentación por percentiles.

**Diapositiva 8: Análisis de Correlación**

Se exploró la relación entre las variables numéricas y las nuevas categorías de desempeño.

**Hallazgos:**

* **Correlación Numérica:** El mapa de calor de correlación entre las variables numéricas originales no mostró una relación lineal fuerte entre TCH, % Sacarosa y otros predictores como las lluvias.
* **Relación entre Categorías:** El análisis cruzado de las categorías de desempeño tampoco reveló una dependencia clara. Un TCH "Alto" no siempre corresponde a una Sacarosa "Alta".

**[Gráfico de Matriz de Correlación y Heatmap de Categorías Cruzadas]**

**Explicación de los Gráficos:** A la izquierda, la matriz de correlación muestra coeficientes bajos. A la derecha, el heatmap de categorías cruzadas indica que los registros se distribuyen a través de todas las combinaciones de niveles de TCH y Sacarosa, sin un patrón dominante.

**Diapositiva 9: Análisis por Variedad de Caña**

La variedad de la caña se identificó como un factor potencialmente influyente en el rendimiento.

**Observaciones:**

* **Impacto en TCH y Sacarosa:** Al analizar la distribución de ambas métricas por variedad, se observó que ciertas variedades consistentemente muestran un mejor o peor desempeño.
* **Relevancia para el Modelo:** Este hallazgo confirma que la **variedad es una variable predictora importante** que debe ser incluida en cualquier modelo de clasificación o regresión.

**[Gráfico de Boxplots de TCH y % Sacarosa por Variedad]**

**Explicación del Gráfico:** Estos boxplots comparan el rendimiento de diferentes variedades de caña. Se puede ver claramente cómo algunas variedades tienen medianas de TCH y % Sacarosa más altas y rangos de variabilidad diferentes, destacando su importancia en la predicción del rendimiento.

**Diapositiva 10: Conclusiones y Próximos Pasos**

**Conclusiones Clave:**

1. **Calidad de Datos:** Existen desafíos con los datos faltantes que deben ser abordados.
2. **Outliers en TCH:** La producción por hectárea (TCH) tiene valores atípicos que necesitan ser validados.
3. **Categorización Exitosa:** Se crearon categorías de desempeño balanceadas (Bajo, Medio, Alto) que son ideales para modelos de clasificación.
4. **Factores Relevantes:** La **variedad de la caña** es un predictor clave del rendimiento. No existe una relación lineal simple entre TCH y % Sacarosa.

**Recomendaciones y Próximos Pasos:**

1. **Imputar o Eliminar Datos Faltantes:** Aplicar una estrategia para manejar los valores nulos.
2. **Validar Outliers:** Consultar con expertos para confirmar la validez de los datos atípicos.
3. **Desarrollar Modelos de Clasificación:** Utilizar las nuevas categorías como variable objetivo para predecir el desempeño.
4. **Ingeniería de Características:** Explorar otras variables con datos completos para mejorar la precisión del modelo.