Taller Final: Predicción y Clasificación en la Industria Azucarera

Análisis avanzado para optimizar producción de azúcar

Resumen Ejecutivo

Objective of Exploratory Data Analysis

Analysis Purpose

The EDA focused on key sugarcane production variables, including Tons of Cane per Hectare and Sucrose Percentage.

Data Quality Assessment

Data quality was reviewed by identifying missing and outlier values to ensure accurate analysis results.

Performance Categories

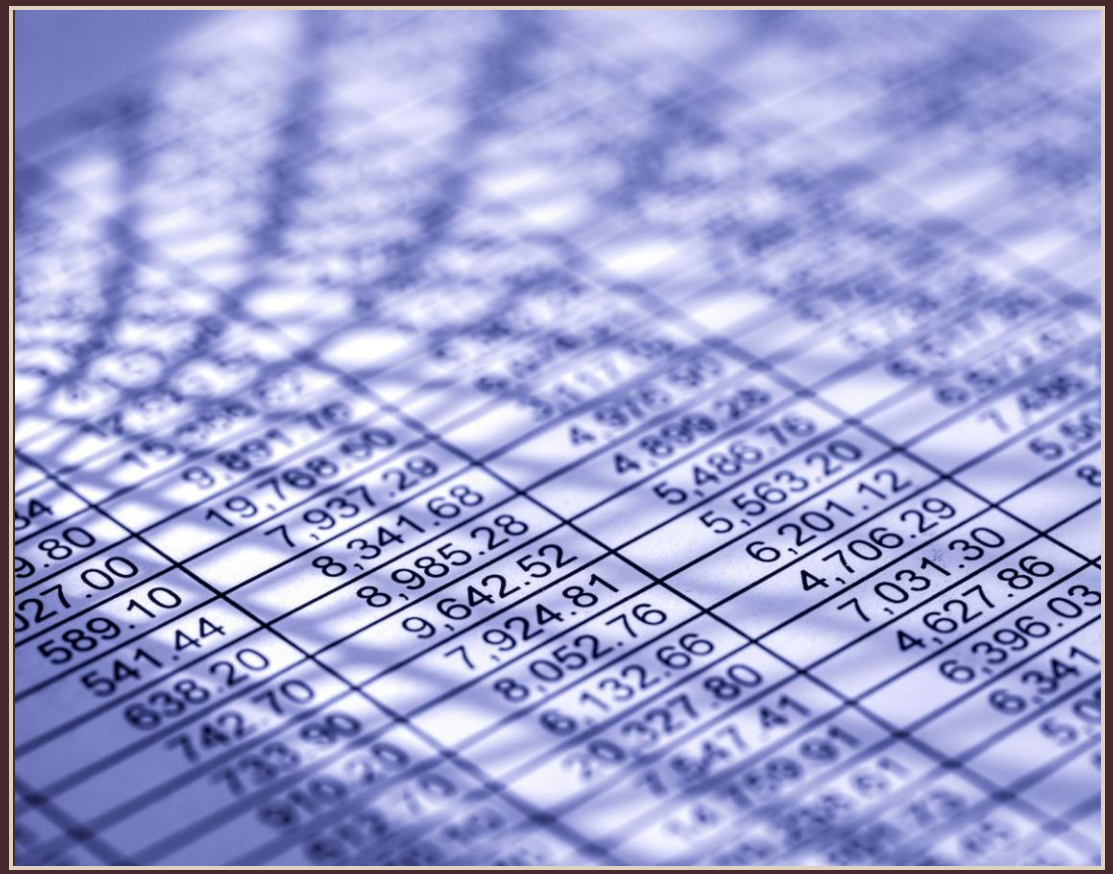
Creation of Low, Medium, and High performance categories for key variables supports future classification modeling.

Foundation for Predictive Models

The exploratory analysis establishes a solid base for developing predictive models in the sugar industry.

Methodology and Data Processing

Fuentes de datos y análisis de calidad



Data Sources Overview

Two main data sources were used: HISTORICO_SUERTES with over 21,000 records and BD_IPSA_1940 with over 2,000 records.

Missing Data Challenges

Critical variables like climate and fertilizers showed high missing data rates, some reaching 100%, limiting direct model use.

Outlier Detection Using IQR

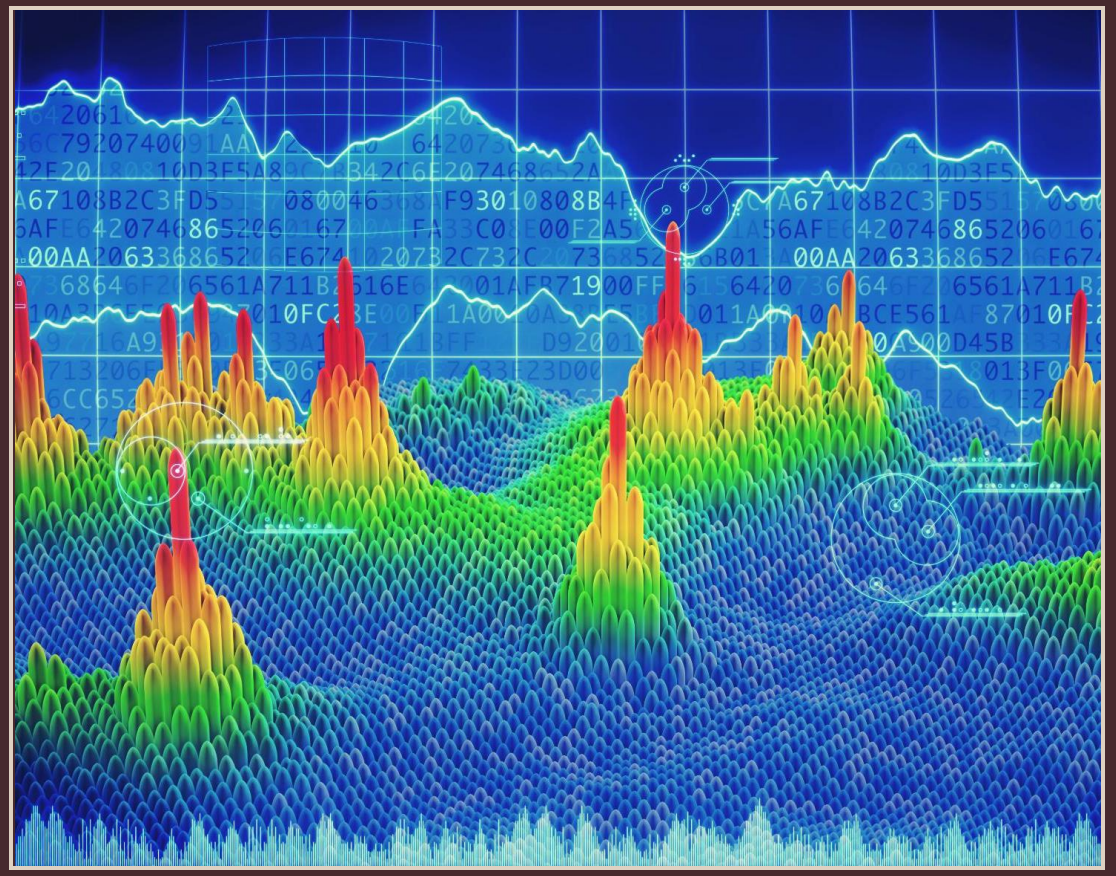
The Interquartile Range method identified 304 outliers in TCH and only 3 in %Sac.Caña, showing varying data consistency.

Performance Categories Creation

Performance categories for TCH and %Sac.Caña were defined using 33rd and 66th percentiles for balanced data distribution.

Resultados y Análisis Visual

Distribución y segmentación de variables clave



Distribution of TCH Variable

TCH distribution is nearly normal with slight positive skew, showing some exceptionally high production batches.

Distribution of %Sac.Caña Variable

%Sac.Caña distribution is symmetric and concentrated around the mean, indicating low variability.

Segmentation Effectiveness

Segmentation into Low, Medium, and High categories is balanced for TCH and skewed for %Sac.Caña, aiding classification.

Relationship and Variety Impact

Heatmap shows no strong correlation between TCH and %Sac.Caña; variety influences both variables' performance.

Conclusions and Recommendations

Key Findings and Suggestions for Modeling

Data Quality Challenges

The dataset has many missing values in climatic and fertilization variables, affecting data quality for modeling.

Outlier Investigation

TCH variable contains numerous outliers requiring validation to distinguish errors from valid data points.

Categorical Performance Groups

Performance categorization by percentiles generated balanced groups useful as classification model labels.

Modeling Recommendations

Use strategies for missing data, expert validation of outliers, and multivariate models like logistic regression.