Análisis Predictivo y Clasificación de Rendimiento y Calidad en Cultivos de Caña de Azúcar: Un Estudio de Caso en el Ingenio La Providencia

Eduardo José Avendaño Caicedo[1], Sebastián Dow Valenzuela[1], and Supervisado por: PhD. Milton Sarria-Paja[2]

[1]*Maestría en Ciencia de Datos, Universidad Icesi, Cali, Colombia*

Marzo de 2024 (Revisado: 31 de marzo de 2025)

Resumen

Resumen: La optimización de la producción agrícola mediante ciencia de datos es crucial para la industria azucarera. Este estudio presenta un análisis predictivo aplicado a datos operativos del Ingenio La Providencia (Valle del Cauca, Colombia). Se utilizaron dos conjuntos de datos: uno histórico (`HISTORICO_SUERTES.xlsx`, N=21027) para desarrollar modelos de regresión lineal múltiple destinados a predecir el rendimiento (TCH: Toneladas de Caña por Hectárea) y la calidad (%Sac.Caña: Porcentaje de Sacarosa); y otro con detalles agronómicos (`BD_IPSA_1940.xlsx`) para construir modelos de clasificación (k-NN, Regresión Logística L1/L2) que categorizan el desempeño (TCH y %Sac.Caña) en niveles Alto, Medio y Bajo definidos por percentiles. El flujo de trabajo incluyó preprocesamiento exhaustivo, análisis exploratorio de datos (univariado y multivariado), entrenamiento de modelos, diagnóstico de supuestos de regresión y evaluación rigurosa mediante validación cruzada utilizando métricas estándar (R^2, RMSE, MAE, Accuracy, Matriz de Confusión, Kappa). Los modelos de regresión mostraron una capacidad explicativa moderada (R^2 promedio para TCH ≈ 0.21), con evidencia de violación de supuestos clave. Los modelos de clasificación alcanzaron una exactitud promedio cercana al 45 %. Se presentan visualizaciones clave como distribuciones, correlaciones, diagnósticos de residuos y matrices de confusión. Estos resultados establecen una línea base cuantitativa y destacan áreas para futuras investigaciones.

Palabras Clave: Caña de azúcar, regresión lineal, clasificación, aprendizaje automático, TCH, sacarosa, agricultura de precisión, Ingenio La Providencia.

1 Introducción

La industria azucarera enfrenta el desafío constante de optimizar la producción y la calidad de la caña de azúcar. La aplicación de técnicas de ciencia de datos ofrece oportunidades para mejorar la toma de decisiones mediante el análisis predictivo de indicadores clave. Este trabajo se centra en el Ingenio La Providencia, un actor relevante en el sector azucarero del Valle del Cauca, Colombia.

El objetivo principal es desarrollar y evaluar modelos de aprendizaje automático para predecir y clasificar el rendimiento y la calidad de la caña de azúcar utilizando datos operativos reales del ingenio. Específicamente, se abordan dos problemas complementarios:

1. **Predicción (Regresión):** Estimar los valores de Toneladas de Caña por Hectárea (TCH), un indicador primario de rendimiento, y el Porcentaje de Sacarosa en Caña (%Sac.Caña), un indicador clave de calidad.
2. **Clasificación:** Categorizar los lotes de cultivo ('suertes') en niveles de desempeño (Alto, Medio, Bajo) tanto para TCH como para %Sac.Caña, basados en umbrales definidos por percentiles.

Se utilizaron dos conjuntos de datos distintos proporcionados por el ingenio: `HISTORICO_SUERTES.xlsx` y `BD_IPSA_1940.xlsx`. Este estudio busca proporcionar una metodología reproducible y resultados cuantitativos, incluyendo visualizaciones relevantes, que sirvan como base para futuras optimizaciones.

2 Materiales y Métodos

2.1 Conjuntos de Datos

Se emplearon dos archivos de datos en formato Excel:

- `HISTORICO_SUERTES.xlsx`: Contiene 21027 registros y 85 variables iniciales. Incluye identificadores, características del cultivo, datos de manejo, variables climáticas agregadas y las variables objetivo TCH y %Sac.Caña.

- `BD_IPSA_1940.xlsx`: Dataset con información agronómica detallada para 1940 observaciones, utilizado para la clasificación. Incluye variables como tipo de corte, variedad, edad, número de cortes, 'sacarosa', etc.

2.2 Preprocesamiento de Datos

Se aplicó un protocolo de preprocesamiento: manejo de valores nulos (imputación/eliminación), detección de outliers (análisis estadístico, sin eliminación explícita), codificación de categóricas (One-Hot Encoding) y escalado de numéricas ('StandardScaler').

2.3 Análisis Exploratorio de Datos (EDA)

Se realizó un EDA para comprender las características de los datos. Se examinaron las distribuciones univariadas y las relaciones multivariadas.

2.4 Modelado de Regresión

Se construyó un modelo de Regresión Lineal Múltiple ('LinearRegression' en Scikit-learn) dentro de un 'Pipeline' para predecir TCH y %Sac.Caña. Se diagnosticaron los supuestos de multicolinealidad (VIF), normalidad de residuos (Shapiro-Wilk) y homocedasticidad (Breusch-Pagan).

2.5 Modelado de Clasificación

Se generaron etiquetas ('Bajo', 'Medio', 'Alto') para TCH y 'sacarosa' usando terciles. Se entrenaron modelos k-NN ($k = 5$), Regresión Logística L1 y L2 usando 'Pipeline'.

2.6 Evaluación de Modelos

Se utilizó Validación Cruzada K-Fold ($k = 5$) para regresión (métricas: R^2, RMSE, MAE) y Estratificada ($k = 5$) para clasificación (métrica: Accuracy). Se realizó evaluación detallada en el conjunto de prueba para clasificación (Matriz de Confusión, Reporte de Clasificación, Kappa).

3 Resultados

3.1 Análisis Exploratorio

Las estadísticas descriptivas para las variables objetivo TCH y %Sac.Caña en el dataset histórico se resumen en la Tabla 1. TCH presenta una media de 129.6 ton/ha con una desviación estándar considerable (29.5 ton/ha), mientras que %Sac.Caña tiene una media de 12.3 % con menor dispersión (std dev 1.15 %).

Cuadro 1: Estadísticas descriptivas para TCH y %Sac.Caña.

Estadística	TCH (ton/ha)	%Sac.Caña (%)
count	21 027,00	20 578,00
mean	129,61	12,32
std	29,48	1,15
min	1,57	1,86
25 %	110,48	11,62
50 %	129,48	12,37
75 %	148,38	13,08
max	401,05	17,63

Las distribuciones de estas variables se visualizan en las Figuras 1 y 2. Ambas se asemejan a distribuciones normales, aunque TCH muestra una cola derecha más pronunciada (posibles outliers o asimetría).

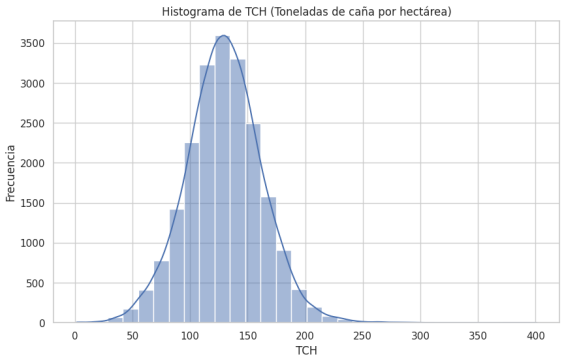


Figura 1: Distribución de TCH (Toneladas de caña por hectárea).

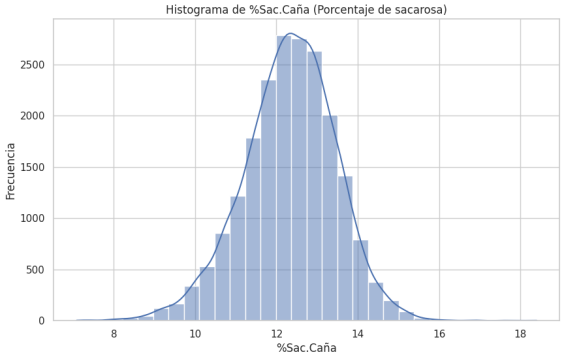


Figura 2: Distribución de %Sac.Caña (Porcentaje de sacarosa).

La matriz de correlación de Pearson entre las variables numéricas (Figura 3) revela relaciones lineales. Por ejemplo, la correlación entre TCH y Edad del último corte es positiva (0.30), mientras que entre TCH y %Sac.Caña es débilmente negativa (-0.17).

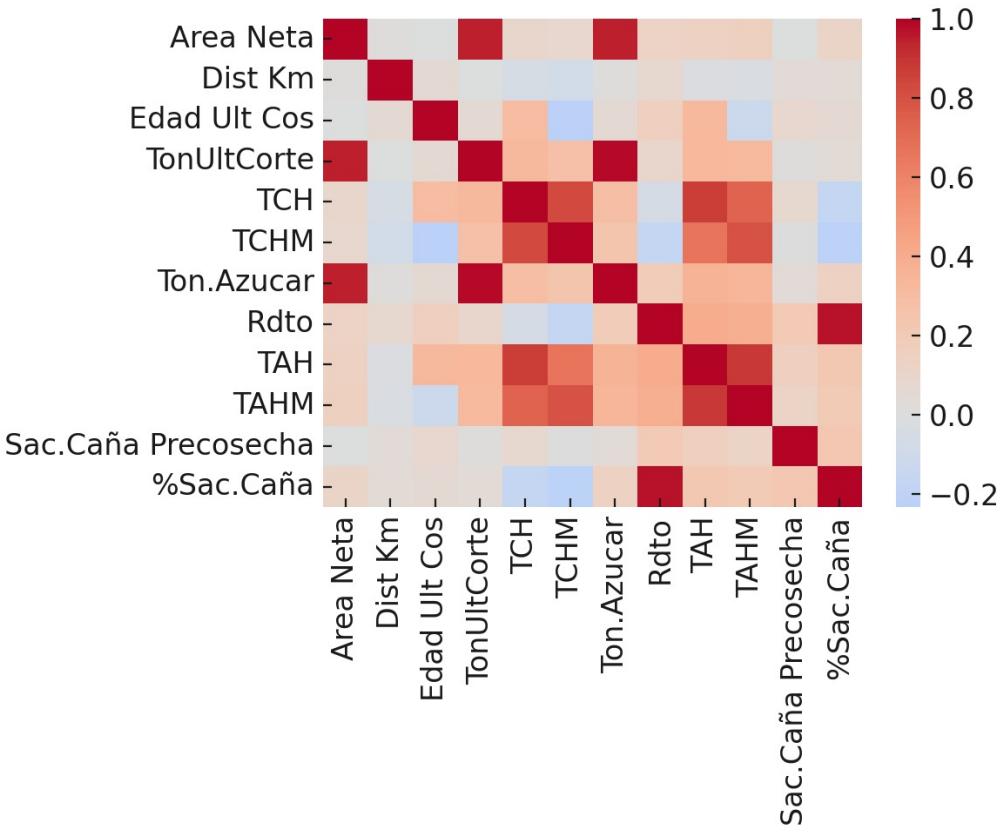


Figura 3: Matriz de Correlación de Pearson entre variables numéricas seleccionadas.

3.2 Resultados del Modelo de Regresión

La validación cruzada del modelo lineal para TCH arrojó $R^2 = 0{,}209$, RMSE= 28,89 ton/ha, y MAE= 22,18 ton/ha.

El diagnóstico de supuestos mostró problemas:

- **Multicolinealidad:** La Figura 4 muestra los VIF para las variables predictoras. Varias dummies de 'Variedad' y 'Zona' superan el umbral de 10, indicando multicolinealidad.
- **Normalidad Residuos:** Test Shapiro-Wilk rechazó normalidad ($p \ll 0{,}001$).
- **Homocedasticidad:** Test Breusch-Pagan rechazó homocedasticidad ($p \ll 0{,}001$). El gráfico de residuos vs. predichos (Figura 5) muestra un patrón cónico.

Variable	VIF ▼
cat__Variedad_CC01-1940	42.588397735198335
cat__Zona_IP05	37.79834237202292
cat__Zona_IP03	25.804110661977447
cat__Variedad_CC85-92	19.67282007873445
cat__Zona_IP06	19.24875380361869
cat__Variedad_CC05-430	11.16303433193149
cat__Zona_IP01	11.013347549424632
cat__Zona_IP02	8.460226770889964

Figura 4: Factor de Inflación de Varianza (VIF) para predictores seleccionados (imagen). Varias variables superaron el umbral VIF=10.

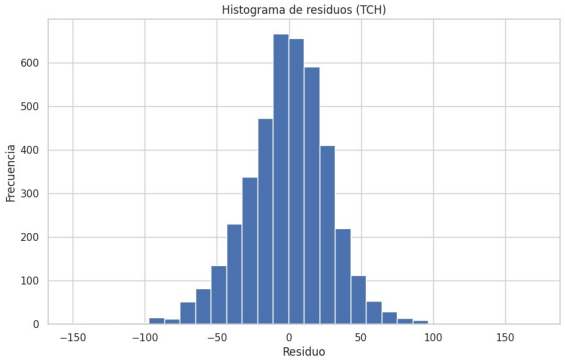


Figura 5: Gráfico de Residuos vs. Valores Predichos (TCH), mostrando heterocedasticidad.

3.3 Resultados de los Modelos de Clasificación

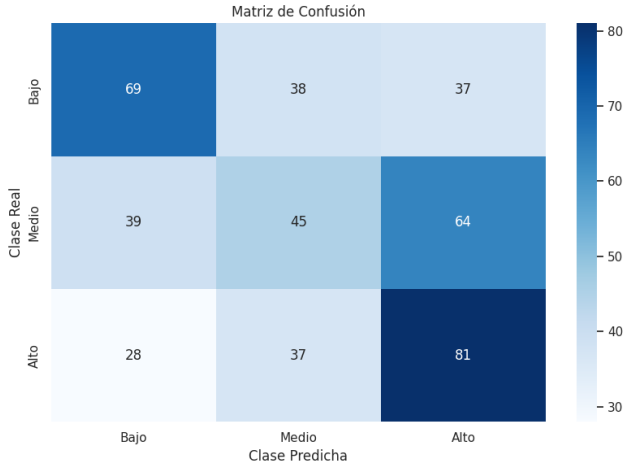
La validación cruzada estratificada ($k = 5$) para clasificar TCH (umbrales $\approx$ 132 y 153 ton/ha) dio exactitudes promedio de: KNN (0.446), Regresión Logística L2 (0.447) y L1 (0.447).

El modelo de **Regresión Logística L1** realizó selección automática de características, reduciendo el número de predictores activos de 39 (en L2) a **17 coeficientes no nulos**, sin pérdida de exactitud en validación cruzada.

La evaluación detallada del modelo KNN en el conjunto de prueba se muestra en la Figura 6. La exactitud

fue del 45 % y Kappa=0.17. La matriz de confusión revela errores distribuidos, particularmente entre clases adyacentes. Las métricas de Precisión y Recall por clase fueron modestas.

predicción moderada del TCH ($R^2 \approx 0.21$) con regresión lineal, aunque con violación de supuestos. La clasificación del desempeño en categorías Alto/Medio/Bajo resultó desafiante (Accuracy≈45 %, Kappa≈0.17). Las



Métricas de Evaluación

Reporte de Clasificación:

	precision	recall	f1-score	support
Alto	0.45	0.55	0.49	146
Bajo	0.51	0.48	0.49	144
Medio	0.38	0.30	0.34	148
accuracy			0.45	438
macro avg	0.44	0.45	0.44	438
weighted avg	0.44	0.45	0.44	438

Kappa de Cohen: 0.17

Figura 6: Evaluación del modelo KNN en el conjunto de prueba: (Izquierda) Matriz de Confusión. (Derecha) Reporte de Clasificación y Kappa de Cohen.

Resultados similares se anticipan para la clasificación de 'sacarosa'.

4 Discusión

Los resultados sugieren que los modelos lineales simples capturan solo una parte de la variabilidad del rendimiento ($R^2 \approx 0{,}21$) y sus supuestos son violados, lo cual es común con datos agrícolas complejos. La multicolinealidad identificada podría requerir estrategias de selección de variables más robustas o el uso de modelos regularizados (como se hizo en clasificación). La falta de normalidad y homocedasticidad podría abordarse con transformaciones o modelos más avanzados (GLMs, GAMLSS, etc.).

La baja exactitud en clasificación (≈45 %) y el bajo Kappa (0.17) indican que, con las variables y la definición de clases actual, es difícil separar los lotes en categorías de desempeño distintas. Esto puede deberse a fronteras de clase difusas, alta variabilidad intra-clase o la influencia de factores no medidos. La regularización L1 fue efectiva en reducir la dimensionalidad sin perder rendimiento predictivo en este caso.

Las limitaciones incluyen la calidad y granularidad de los datos, la definición de clases basada en percentiles y el alcance de los modelos explorados. Sin embargo, las visualizaciones y métricas proporcionan información valiosa. Por ejemplo, la matriz de confusión (Fig. 6) detalla dónde falla el modelo clasificador.

5 Conclusiones

Este estudio aplicó modelos de regresión y clasificación a datos del Ingenio La Providencia. Se logró una visualizaciones y diagnósticos aportaron información sobre las limitaciones y características de los datos y modelos. Futuras líneas incluyen explorar variables adicionales, modelos no lineales y definiciones de clases alternativas.

A Librerías de Software Utilizadas

Python 3.x, Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, Statsmodels, SciPy.