UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola d'Enginyeria de Barcelona Est

Predicción de enfermedad renal - análisis, modelado y evaluación

Miembros:

Rocío Ávalos Morillas y Ainhoa Fraile Pulido

Profesores:

Magda Ruiz, Luis Eduardo Mujica y Santiago Alferez

Asignatura:

Aprendizaje Bioestadístico

Escola d'Enginyeria de Barcelona Est
Universitat Politècnica de Catalunya

20/04/2025

1. OBJETIVO

El objetivo principal de este proyecto es adquirir experiencia práctica en el desarrollo de un flujo completo de trabajo en ciencia de datos. En particular, nos centramos en la aplicación de modelos de regresión para predecir la variable objetivo 'bu' (urea en sangre), así como en la implementación y comparación de tres clasificadores: LDA, QDA y SVM con un kernel sigmoidal.

2. PRINCIPALES HALLAZGOS

2.1. Análisis exploratorio de datos (EDA) y limpieza de datos

2.1.1 Corrección de valores corrompidos

Durante el análisis inicial, se detectó que algunas variables numéricas estaban almacenadas como tipo object debido a la presencia de caracteres no numéricos. Este fue el caso de las variables 'pc', 'wc' y 'rc', que fueron convertidas correctamente a tipo numérico para su análisis posterior. Esta corrección permitió su inclusión en procesos de imputación, transformación y modelado.

2.1.2. Revisión de valores únicos

Se identificaron inconsistencias en variables categóricas, originadas por errores de codificación, como espacios en blanco y tabulaciones: la variable 'classification' contenía valores como 'ckd\t', en la variable 'cad' se hallaron registros como '\tno', y en la variable 'dm' se observaron múltiples variantes de la misma categoría: '\tyes', ' yes', '\tno'.

Estas inconsistencias fueron limpiadas para garantizar la correcta agrupación de categorías y evitar problemas durante la codificación o el entrenamiento de modelos (Figura 1).

2.1.3. Análisis de valores faltantes

Se identificó la presencia de valores ausentes en múltiples columnas (Figura 2). Para profundizar en su análisis, hemos utilizado un dendrograma que permitió visualizar patrones de correlación entre los datos faltantes (Figura 3).

Variables relacionadas con análisis de laboratorio, como 'rbc', 'pc', 'pcc', 'ba', 'sod', 'pot', 'hemo', 'pcv', 'wc' y 'rc', mostraron una alta proporción de valores nulos. Esta ausencia puede atribuirse tanto a la falta de realización de ciertas pruebas en pacientes con enfermedad renal crónica como a decisiones clínicas individuales.

También se detectaron patrones por fila: algunos pacientes presentaban múltiples resultados faltantes simultáneamente, lo que sugiere posibles omisiones sistemáticas en la recolección de datos clínicos.

2.1.4. Imputación de datos

Para llevar a cabo una imputación adecuada, las variables se dividieron en numéricas y categóricas. Además, se aplicó validación cruzada para comparar el rendimiento de cada método de imputación.

- Variables numéricas: se evaluaron varios métodos, incluyendo imputación por media, mediana, KNN con diferentes valores de k (buscando optimizar su elección), y MICE. La selección del método óptimo se realizó basándonos en múltiples métricas: rendimiento predictivo (RMSE, R^2), pruebas estadísticas (KS, t-test, Shapiro-Wilk), entre otras.
- Variables categóricas: se imputaron utilizando moda, KNN Imputer e Iterative Imputer. Posteriormente, se seleccionó el método más apropiado considerando métricas como la proporción de la categoría modal, la entropía, y otros indicadores de estabilidad y consistencia.

2.1.5. Análisis de la variable objetivo

Dado que la variable objetivo del estudio es 'bu', inicialmente se analizó sin aplicar transformaciones.

Para ello, se utilizaron histogramas, diagramas de caja y estadísticas descriptivas (media, mediana, desviación estándar, asimetría y curtosis) para evaluar su comportamiento.

Este análisis permitió identificar aspectos clave sobre la naturaleza de la variable, como su grado de asimetría, la presencia de valores extremos y el comportamiento general de su distribución (Figura 4).

2.1.6. Outliers

Con el fin de minimizar el impacto de los valores atípicos en el modelo (Figura 5), se abordaron desde dos enfoques:

- Clínica: se definieron rangos fisiológicos plausibles, basados en literatura médica, para detectar posibles errores de registro o mediciones anómalas.
- Estadística: se analizaron distribuciones, diagramas de caja (boxplots), gráficos Q-Q y medidas de asimetría.

Se aplicaron diversas transformaciones (logarítmica, raíz cuadrada, winsorización y Yeo-Johnson) con el objetivo de mejorar la simetría sin comprometer la interpretación clínica de las variables.

Finalmente, aunque Yeo-Johnson fue la más eficaz para corregir la asimetría en la mayoría de casos, se descartó por generar valores negativos, lo cual no resulta coherente en el contexto clínico (Figura 6).

Respecto a la variable 'bu', se seleccionó la transformación 'sqrt_bu', ya que redujo la asimetría de forma efectiva sin comprometer su interpretación médica (Figura 7).

2.2. Modelo de regresión

En esta etapa, al incorporar variables categóricas, se procedió primero a su codificación numérica mediante *one-hot encoding*. Una vez completado, se analizó la relación entre las variables independientes y la variable 'sqrt_bu'. Para ello se emplearon:

- Un heatmap de correlación para identificar asociaciones lineales relevantes (Figura 8).
- Un gráfico de barras que muestra la magnitud y dirección de cada correlación (Figura 9).

Para completar este proceso de preparación, nos aseguramos de que todas las variables del conjunto de datos estuvieran representadas en un formato numérico compatible. Por eso, los valores de texto y booleanos fueron mapeados, asignando False = 0 y True = 1.

El primer modelo desarrollado se basó exclusivamente en variables numéricas. Se realizó un análisis detallado de sus estadísticas globales, coeficientes y significancia, junto con un diagnóstico de residuos para comprobar los supuestos clásicos de la regresión lineal.

Ante indicios de multicolinealidad, se construyó un segundo modelo que integraba variables numéricas y categóricas. Se identificaron variables altamente correlacionadas, y se utilizó el índice de inflación de la varianza (VIF) para evaluar redundancias (Figura 10).

Con el objetivo de mejorar la capacidad predictiva y la robustez, se aplicaron técnicas de regularización como Ridge, Lasso y ElasticNet, además de enfoques avanzados como regresión polinómica y XGBoost.

Todos los modelos fueron comparados frente a la regresión lineal tradicional, utilizando métricas cuantitativas y visualizaciones (Figura 11). XGBoost demostró ser el más preciso, aunque los modelos lineales regularizados destacaron por su simplicidad e interpretabilidad.

2.3. Modelo de clasificación

Para mejorar la precisión del modelo y reducir la complejidad del conjunto de datos, se realizó una selección de características mediante dos enfoques complementarios: ANOVA F-test y Random Forest. Ambos métodos permitieron identificar las 20 variables más relevantes, combinando criterios estadísticos y de aprendizaje automático.

Un hallazgo clave fue la coincidencia de varias variables seleccionadas por ambos métodos, lo que evidencia su alto poder discriminativo y valor predictivo, añadiendo valor real al modelo.

A continuación, se evaluaron tres modelos de clasificación sobre un conjunto de prueba de forma (100, 20), utilizando métricas estándar como exactitud, precisión, recall, F1-score, AUC-ROC y la matriz de confusión (Figura 12 y 13). Los resultados fueron los siguientes:

- SVM (con función sigmoide) mostró el mejor rendimiento general, alcanzando una precisión perfecta y un F1-score sobresaliente en ambas clases.
- QDA también obtuvo buenos resultados, particularmente en la clasificación de la clase 0.

- LDA, si bien presentó una precisión global más baja, demostró buen desempeño en la clasificación de la clase 1.

2.4. Modelo de clasificación por Entrenamiento Conjunto

Con el objetivo de mejorar aún más el rendimiento del sistema de clasificación, se optó por implementar un modelo basado en entrenamiento conjunto, específicamente mediante la técnica de Stacking Ensemble. Esta técnica combina las predicciones de múltiples modelos base a través de un meta-modelo que aprende a optimizar la combinación de sus salidas.

Se utilizó como modelo comparativo el SVM, ya que había demostrado el mejor rendimiento individual. Además, para ambos modelos se calcularon las métricas y visualizaciones ya mencionadas (Figura 14 y 15).

3. RENDIMIENTO DE LOS MODELOS

En la etapa final del análisis, ambos modelos mostraron un rendimiento muy similar. Sin embargo, es relevante destacar ciertos aspectos que permiten diferenciar sutilmente su desempeño.

El modelo SVM cometió un único error de clasificación: un falso negativo, al identificar incorrectamente a un paciente enfermo como sano, lo que redujo su sensibilidad al 98%. Este tipo de error puede ser crítico en contextos clínicos, donde pasar por alto un diagnóstico puede afectar gravemente la salud del paciente. A pesar de ello, el modelo mantuvo una precisión y especificidad del 100%, sin falsos positivos.

En contraste, el modelo Stacking Ensemble logró un rendimiento perfecto, clasificando correctamente todas las instancias del conjunto de prueba. Este resultado sugiere una mayor capacidad para captar patrones complejos, lo que refuerza su robustez como herramienta de apoyo diagnóstico.

4. LIMITACIONES DE LOS MODELOS Y POSIBLES MEJORAS FUTURAS

A pesar del excelente rendimiento de los modelos SVM y Stacking Ensemble, existen algunas limitaciones. El conjunto de prueba fue relativamente reducido y equilibrado artificialmente mediante SMOTE, lo que podría comprometer la capacidad de generalización de los modelos a datos reales no balanceados.

El preprocesamiento del conjunto de datos fue riguroso (imputación, detección de errores, tratamiento de atípicos, etc.), y la selección de variables se apoyó en métodos sólidos como ANOVA F-test y Random Forest. Sin embargo, podrían explorarse enfoques multivariados más complejos para capturar interacciones entre variables.

Para futuras mejoras, podríamos:

- Incluir variables temporales, genéticas o contextuales.
- Aplicar modelos más interpretables y avanzados, como SHAP o redes neuronales.
- Explorar nuevas técnicas de selección y reducción de características.

5. CONCLUSIONES FINALES Y POTENCIALES IMPLICACIONES

Con la realización del proyecto se logró desarrollar modelos predictivos altamente precisos para la detección de enfermedad renal crónica, destacando el Stacking Ensemble, que clasificó correctamente todas las instancias del conjunto de prueba. El modelo SVM también mostró un rendimiento excelente, con solo un falso negativo.

Estos resultados validan tanto la calidad del preprocesamiento como la efectividad de las técnicas aplicadas. Además, demuestran el potencial de estas herramientas para apoyar la toma de decisiones clínicas, especialmente en enfermedades crónicas como la ERC, donde una detección temprana puede mejorar significativamente el pronóstico del paciente y la eficiencia del sistema sanitario.

Este trabajo subraya el valor de integrar conocimiento médico con analítica avanzada. Con la incorporación de nuevas fuentes de datos y modelos más robustos e interpretables, será posible avanzar hacia sistemas predictivos más precisos y aplicables en entornos clínicos reales.

6. ANEXO

6.1. Documento de análisis y código

El documento IPython Notebook (.ipynb) que contiene el código de análisis, gráficos y resultados se encuentra en el siguiente enlace de GitHub:

6.2. Figuras y tablas

	Variable	Valores únicos	Distribución de valores
0	rbc	nan, normal, abnormal	normal: 201 (50.25%) abnormal: 47 (11.75%)
1	pc	normal, abnormal, nan	normal: 259 (64.75%) abnormal: 76 (19.00%)
2	pcc	notpresent, present, nan	notpresent: 354 (88.50%) present: 42 (10.50%)
3	ba	notpresent, present, nan	notpresent: 374 (93.50%) present: 22 (5.50%)
4	htn	yes, no, nan	no: 251 (62.75%) yes: 147 (36.75%)
5	dm	yes, no, nan	no: 261 (65.25%) yes: 137 (34.25%)
6	cad	no, yes, nan	no: 364 (91.00%) yes: 34 (8.50%)
7	appet	good, poor, nan	good: 317 (79.25%) poor: 82 (20.50%)
8	pe	no, yes, nan	no: 323 (80.75%) yes: 76 (19.00%)
9	ane	no, yes, nan	no: 339 (84.75%) yes: 60 (15.00%)
10	classification	ckd, notckd	ckd: 250 (62.50%) notckd: 150 (37.50%)

Figura 1. Tabla de variables categóricas y sus valores únicos y distribución antes de la imputación

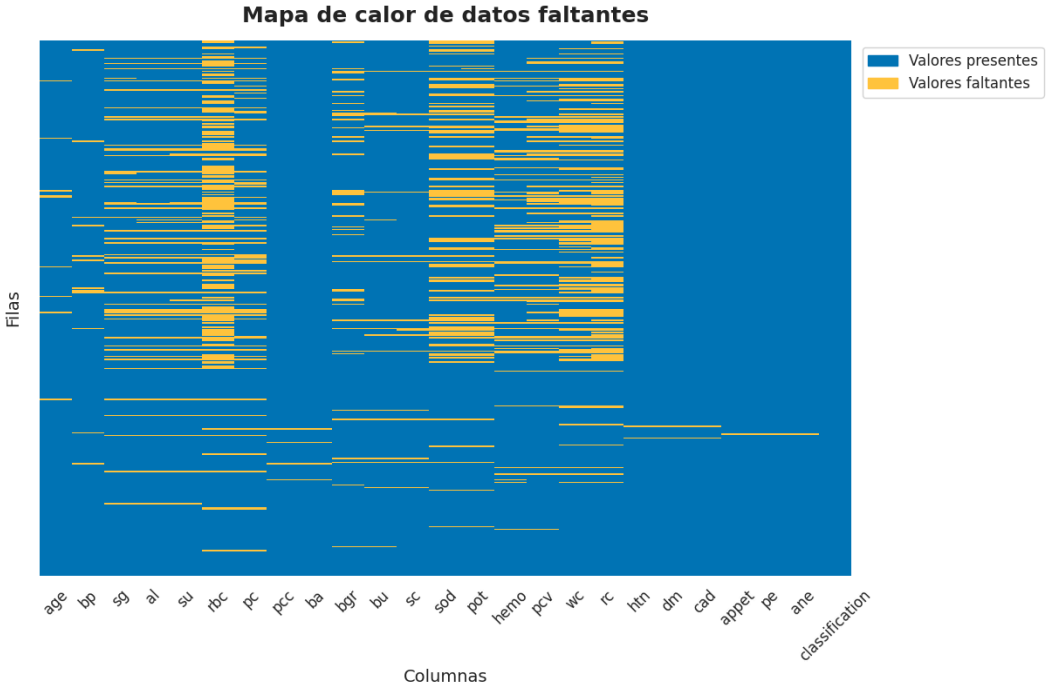

Figura 2. Mapa de calor de datos faltantes en el dataset

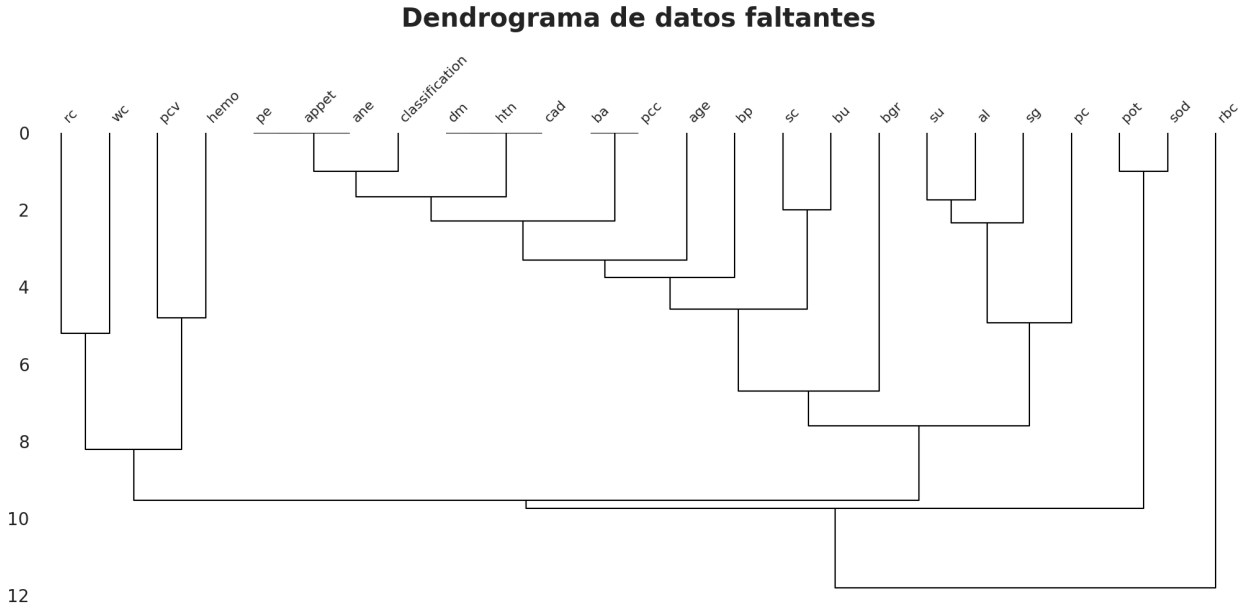

Figura 3. Dendrograma de datos faltantes en el dataset

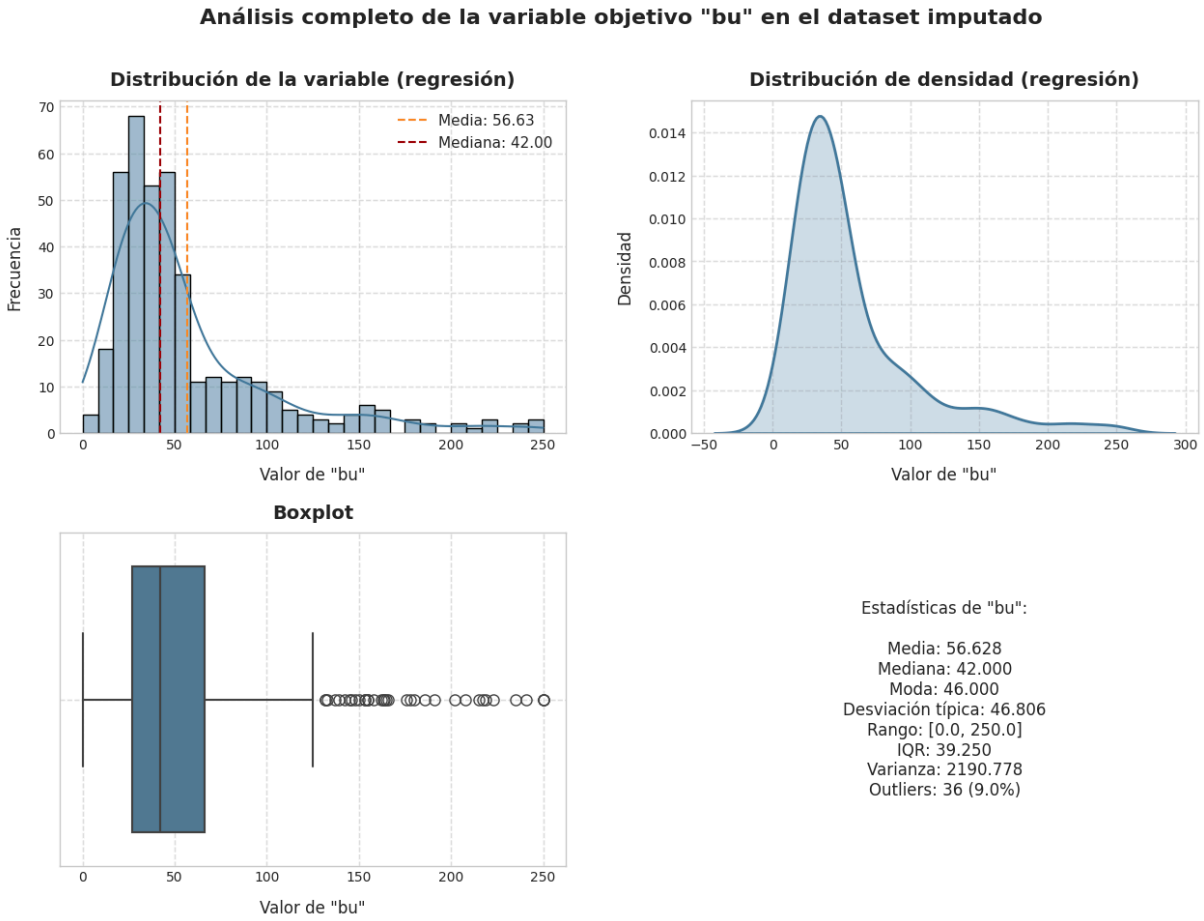

Figura 4. Análisis estadístico de urea en sangre ('bu') en el dataset imputado

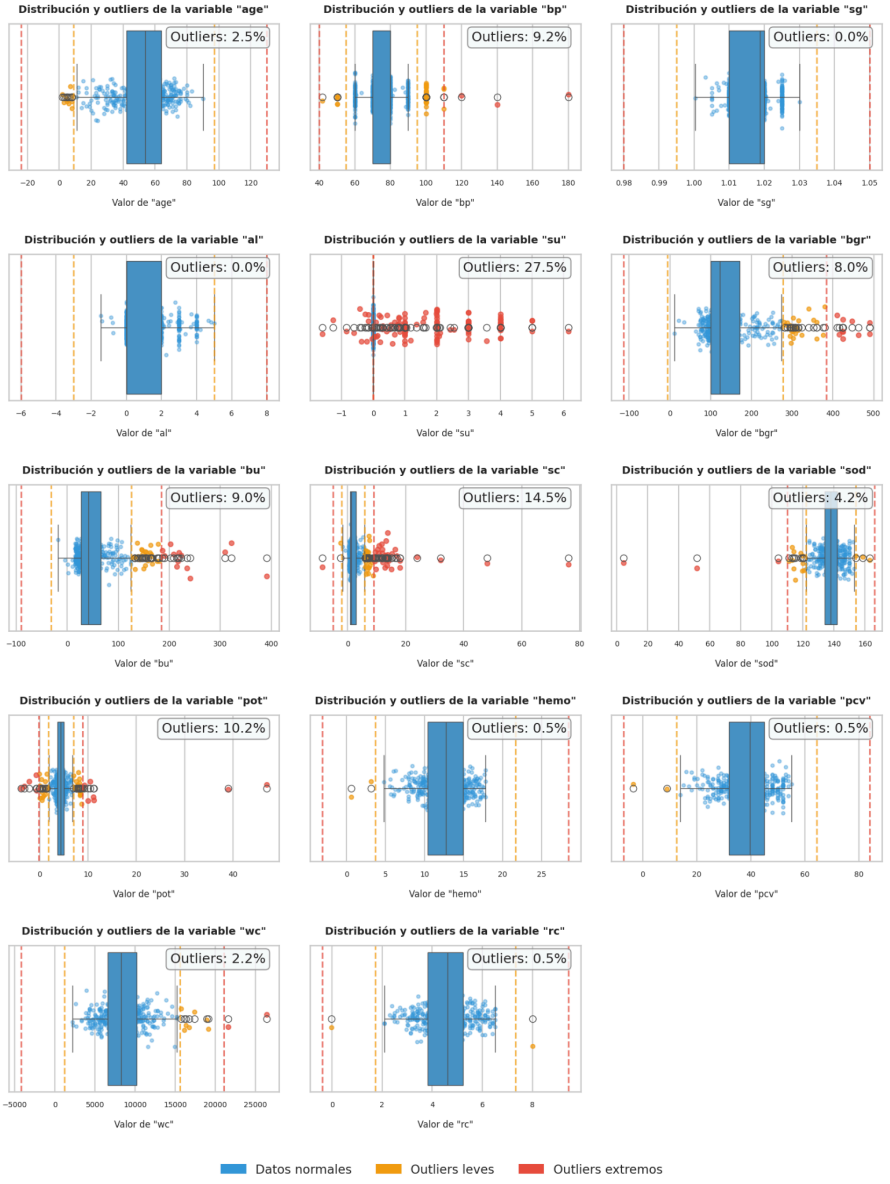

Figura 5. Outliers presentes en las variables numéricas

	Variable	Media	Mediana	Desv. Típica	CV (%)	Asimetría	P-valor Shapiro	Normal
0	Urea en sangre	51.343	42.000	32.570	63.435	0.964	0.000	False
1	Niveles de sodio	137.454	138.000	7.980	5.806	-0.958	0.000	False
2	Creatinina sérica	2.869	1.300	3.682	128.342	2.618	0.000	False
3	Niveles de potasio	4.434	4.300	1.711	38.597	1.521	0.000	False
4	Recuento de glóbulos blancos	8486.790	8259.878	2836.917	33.427	0.833	0.000	False
5	Glucosa en sangre aleatoria	148.926	123.000	75.503	50.698	1.596	0.000	False
6	Tensión arterial	76.395	80.000	13.674	17.899	1.528	0.000	False
7	Niveles de azucar en sangre	0.504	0.000	1.115	221.351	2.394	0.000	False

Figura 6. Estadísticas para algunas variables numéricas

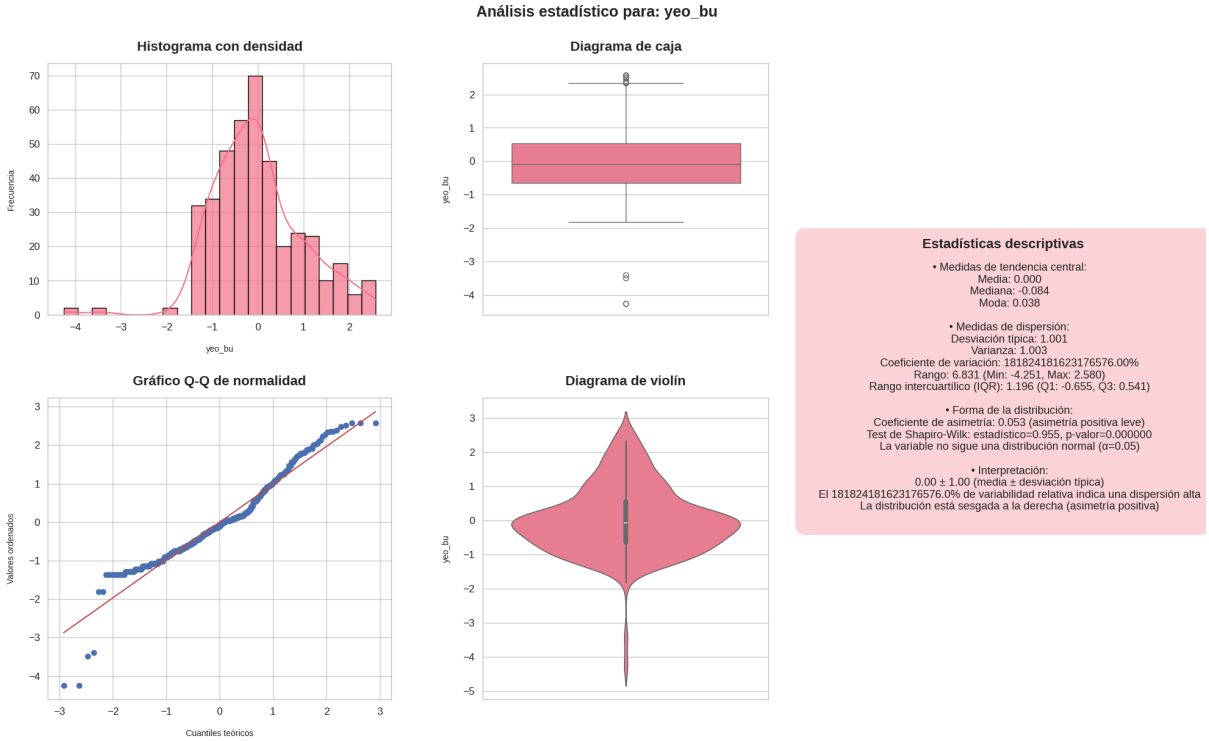

Figura 6. Análisis estadístico para 'yeo_bu'

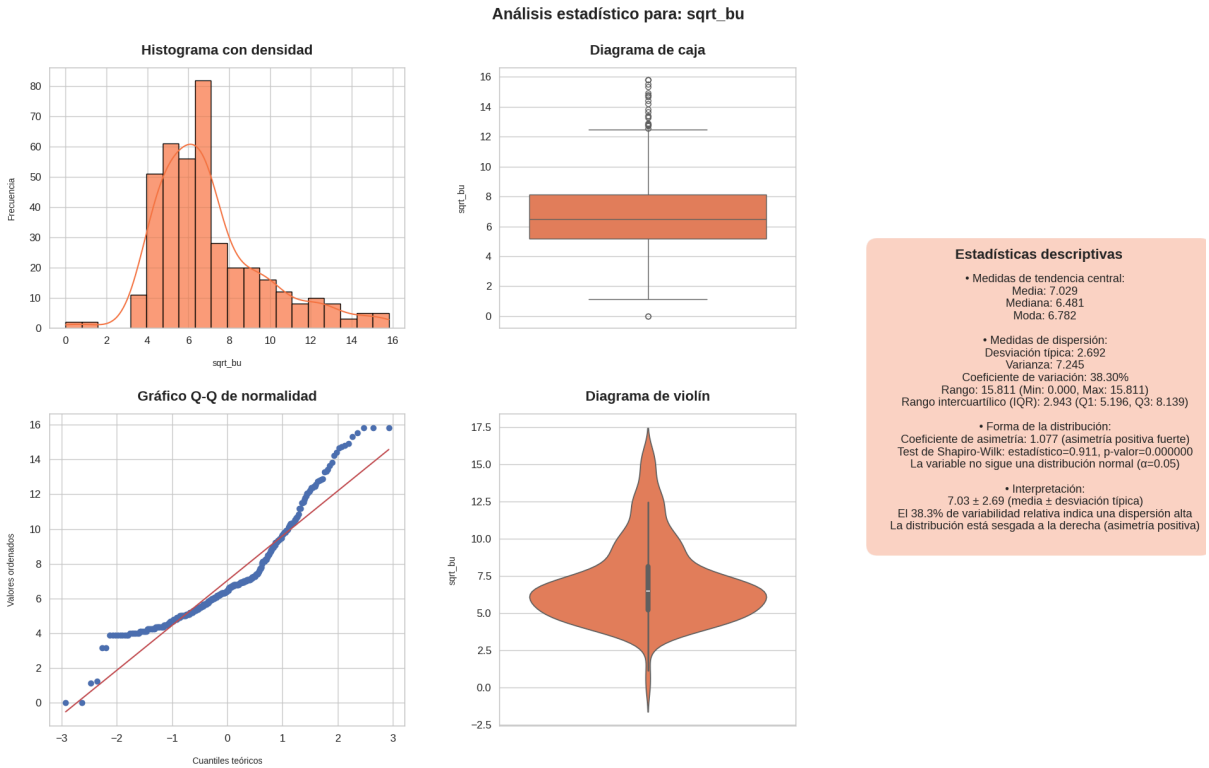

Figura 7. Análisis estadístico para 'sqrt_bu'

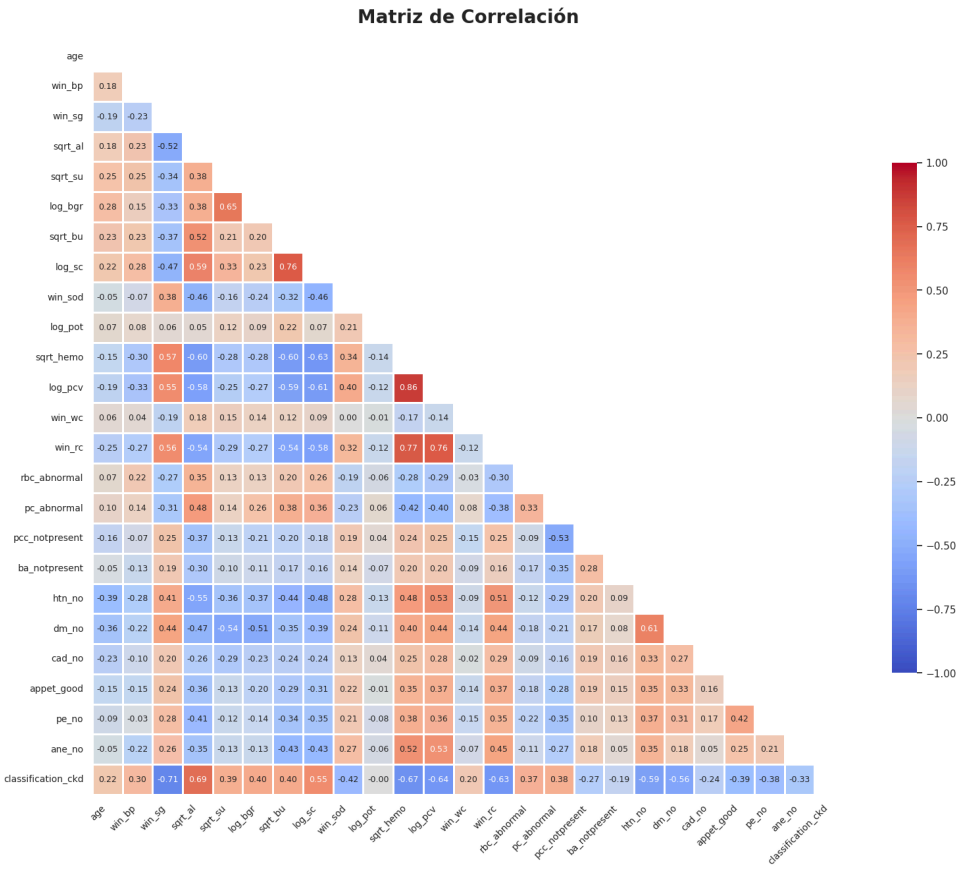

Figura 8. Matriz de correlación después de imputación y transformación de variables

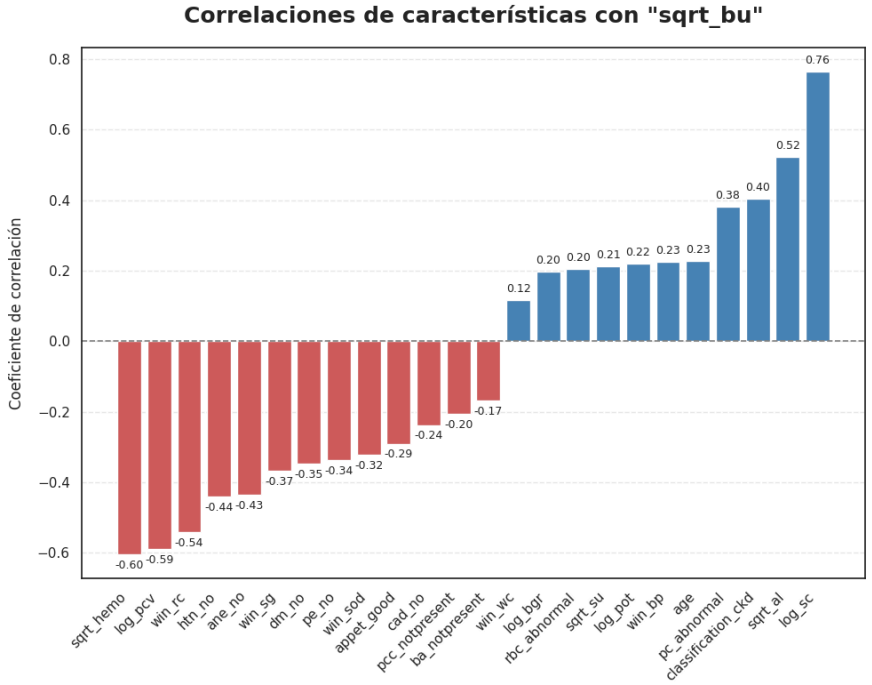

Figura 9. Diagrama de barras de correlaciones de las variables (numéricas y categóricas) con el target 'sqrt_bu'

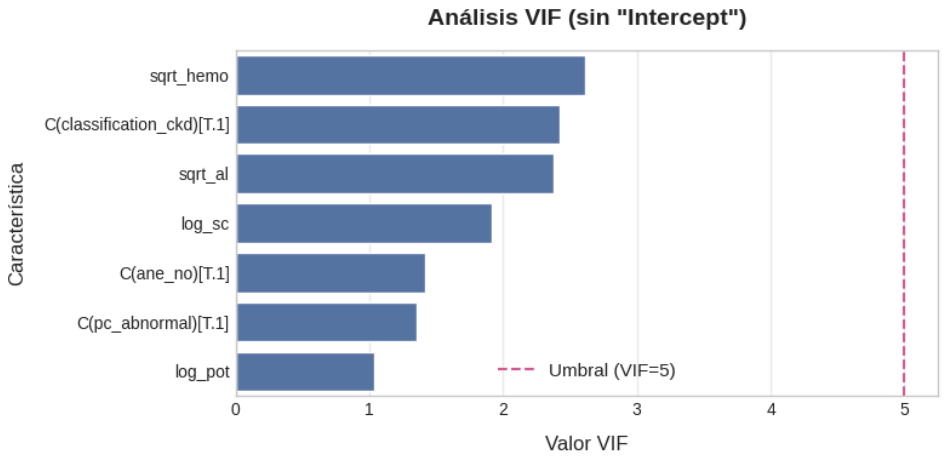

Figura 10. Resultados obtenidos del análisis del Factor de Inflación de la Varianza (VIF)

Modelo	MSE	RMSE	MAE	R^2	AIC	Precisión Cuartiles
Linear	3.5806	1.8922	1.3732	0.6063	1182.72	0.5750
Ridge	3.5551	1.8855	1.3705	0.6091	—	—
Lasso	3.5797	1.8920	1.3727	0.6064	1182.72	—
ElasticNet	3.5569	1.8860	1.3697	0.6089	—	—
XGBoost	**3.2358**	**1.7988**	**1.3314**	**0.6442**	—	**0.5375**

Figura 11. Comparación estadística de los modelos de regresión

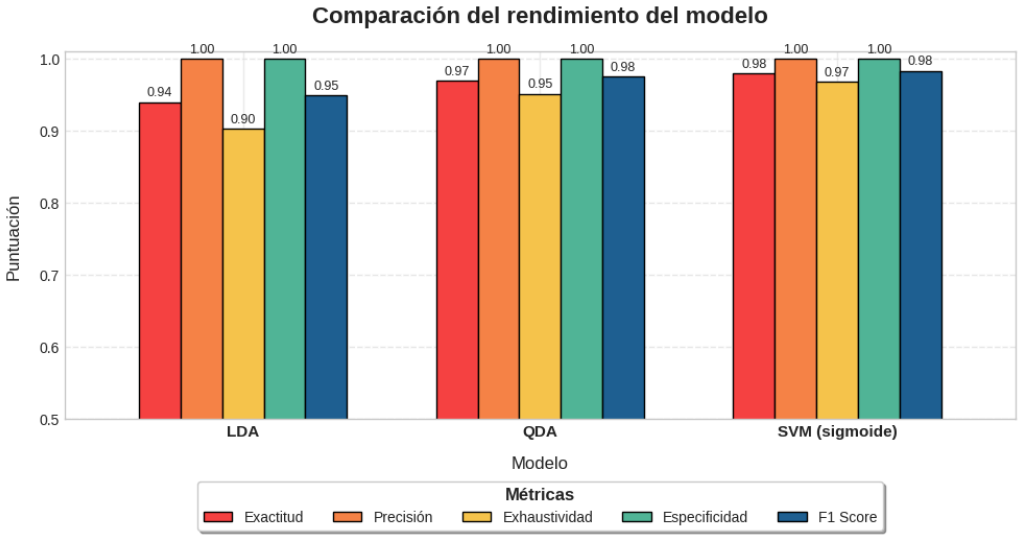

Figura 12. Resultados del rendimiento de los modelos de clasificación

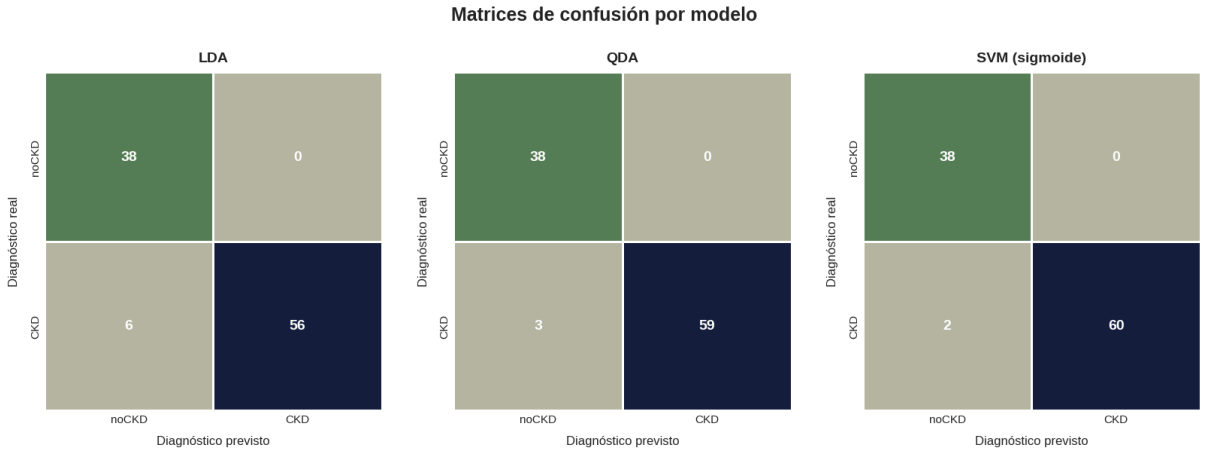

Figura 13. Comparación de las matrices de confusión de los modelos de clasificación

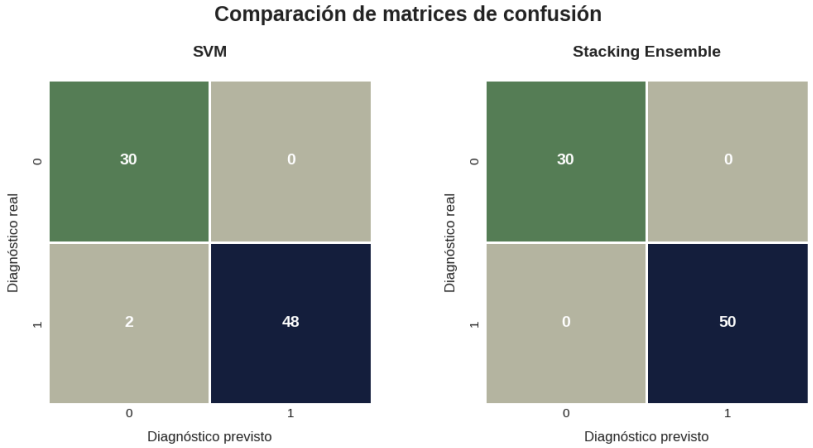

Figura 14. Comparación de las matrices de confusión entre el mejor modelo de clasificación (SVM sigmoidal) y el modelo de clasificación ensemble por stacking

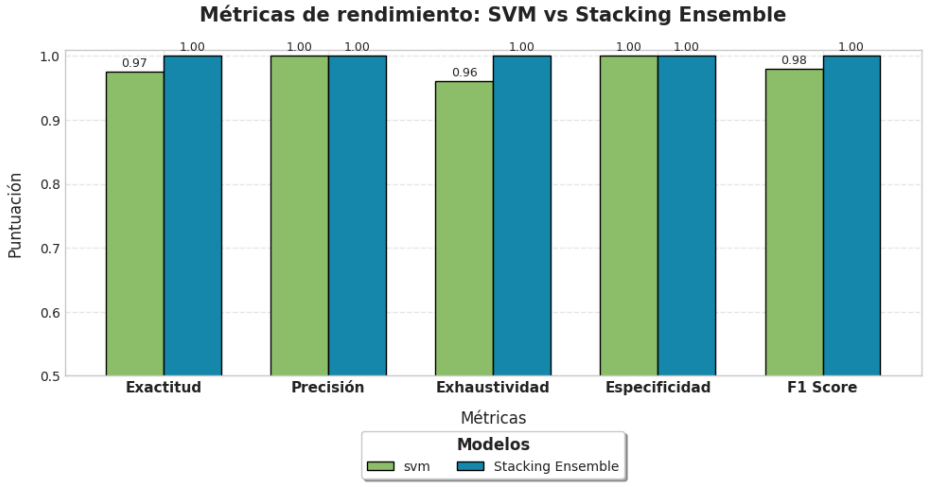

Figura 15. Comparación de las métricas de rendimiento entre el mejor modelo de clasificación (SVM sigmoidal) y el modelo de clasificación ensemble por stacking