TEC | Tecnológico de Costa Rica

Instituto Tecnológico de Costa Rica

Escuela de Ingeniería en Computación

**IC-6200 – Inteligencia Artificial**

# Proyecto Machine Learning


**Estudiantes:**

Pablo Mauricio Mesén Alvarado – 2023071259
Samir Fernando Cabrera Tabash – 2022161229
Luis Gerardo Urbina Salazar – 2023156802

**Profesor:**

Kenneth Roberto Obando Rodríguez


23 de noviembre de 2025

II Semestre 2025

# Índice

## 1. Resumen Ejecutivo

Este proyecto aplica la metodología **CRISP-DM** para desarrollar un modelo de **aprendizaje automático** capaz de predecir la **enfermedad renal crónica (CKD)** utilizando un conjunto de datos sintético con información clínica y demográfica de 1,659 pacientes. El objetivo principal es identificar los factores más relevantes asociados al diagnóstico de CKD y construir un modelo predictivo que apoye su detección temprana. Dado que el conjunto de datos presenta un **desbalance significativo entre clases**, se implementaron y compararon diversas técnicas de balanceo, priorizando la maximización del **recall** (sensibilidad) para minimizar los falsos negativos, aspecto crítico en el contexto médico donde no detectar un paciente enfermo puede tener consecuencias graves. Durante el proceso se realizaron tareas de exploración, limpieza y preparación de datos, seguidas del entrenamiento y evaluación de distintos algoritmos supervisados con diferentes estrategias de balanceo. Los resultados muestran que los modelos basados en árboles de decisión, combinados con técnicas apropiadas de manejo del desbalance, ofrecieron un desempeño sólido y equilibrado, destacando el potencial del aprendizaje automático para generar herramientas de apoyo al diagnóstico médico.

## 2. Introducción

Las enfermedades renales crónicas (Chronic Kidney Disease, CKD) representan uno de los mayores desafíos de salud pública a nivel mundial, debido a su alta prevalencia, progresión silenciosa y las graves consecuencias que puede tener si no se diagnostica y trata a tiempo. El diagnóstico temprano y la predicción precisa del riesgo son fundamentales para reducir las complicaciones y mejorar la calidad de vida de los pacientes. Sin embargo, la complejidad de los factores que intervienen en el desarrollo de la enfermedad —como condiciones metabólicas, antecedentes familiares, hábitos de vida y variables clínicas— hacen que su estudio requiera un enfoque integral y basado en datos. En este proyecto se aplica la metodología **CRISP-DM** (Cross Industry Standard Process for Data Mining) para desarrollar un modelo de **aprendizaje automático** capaz de predecir la presencia de enfermedad renal crónica en pacientes, utilizando un conjunto de datos sintético proporcionado por Rabie El Kharoua en la plataforma Kaggle. Este dataset contiene información detallada de 1,659 pacientes e incluye variables demográficas, factores de estilo de vida, antecedentes médicos, mediciones clínicas, uso de medicamentos, síntomas, calidad de vida, exposiciones ambientales y comportamientos de salud. El objetivo principal de este trabajo es explorar el potencial del aprendizaje automático para asistir en el diagnóstico temprano de la enfermedad renal crónica, identificando las características más relevantes que influyen en la condición del paciente y construyendo un modelo predictivo que pueda

diferenciar entre pacientes con y sin diagnóstico de CKD. Dada la naturaleza crítica del problema, se presta especial atención al manejo del desbalance de clases presente en el conjunto de datos, implementando y comparando múltiples estrategias de balanceo para optimizar la capacidad del modelo de identificar correctamente a los pacientes enfermos. Este documento describe de forma estructurada cada fase de la metodología, presentando los hallazgos obtenidos, las decisiones de diseño tomadas y los resultados experimentales alcanzados, con el propósito de demostrar el valor del análisis de datos y la inteligencia artificial en el ámbito de la salud.

## 3. Fase 1: Entendimiento del Negocio

El propósito de esta fase es comprender de manera integral los objetivos del proyecto desde la perspectiva del negocio y traducirlos en objetivos de minería de datos concretos. Dado que el problema abordado se centra en la detección temprana de la enfermedad renal crónica (CKD), esta fase busca alinear las necesidades clínicas con las capacidades del aprendizaje automático, asegurando que los resultados sean relevantes y aplicables en contextos médicos reales.

### 3.1. Objetivos del Negocio

A partir del contexto del proyecto, se definen los siguientes objetivos del negocio:

1. **Mejorar la detección temprana de la enfermedad renal crónica (CKD)** mediante el uso de técnicas de aprendizaje automático, reduciendo el riesgo de diagnóstico tardío y minimizando los falsos negativos.
2. **Identificar los factores clínicos y demográficos más relevantes** que influyen en el desarrollo de CKD, para apoyar la toma de decisiones médicas fundamentadas.
3. **Reducir los costos de atención médica** asociados a tratamientos avanzados de CKD mediante la prevención y detección oportuna.
4. **Proporcionar una herramienta predictiva accesible e interpretable** que pueda ser utilizada como apoyo en entornos clínicos o educativos, priorizando la sensibilidad sobre otras métricas.
5. **Demostrar la aplicabilidad del aprendizaje automático en el ámbito de la salud**, fomentando su uso en proyectos de análisis de datos médicos con consideraciones especiales para datos desbalanceados.

### 3.2. Criterios de Éxito del Negocio

Para evaluar el cumplimiento de los objetivos anteriores, se establecen los siguientes criterios de éxito desde la perspectiva del negocio:

1. El modelo predictivo logra identificar correctamente al menos el 85
2. El modelo mantiene un equilibrio razonable entre sensibilidad y precisión, evitando una tasa excesiva de falsos positivos que genere alarmas innecesarias.
3. Los resultados del modelo son comprensibles y pueden ser interpretados fácilmente por profesionales de la salud, permitiendo identificar los factores de riesgo más relevantes.
4. La aplicación del modelo demuestra potencial para reducir el tiempo promedio de diagnóstico o servir como filtro preliminar en estudios médicos.
5. El proyecto genera conocimiento útil sobre las variables más influyentes en la progresión de CKD y sobre las estrategias efectivas para manejar desbalance de clases en problemas médicos.
6. La solución propuesta puede ser documentada y replicada en otros contextos educativos o de investigación aplicada.

### 3.3. Situación Actual y Desafíos

El conjunto de datos presenta un **desbalance significativo entre clases**, donde la cantidad de pacientes sin CKD supera considerablemente a aquellos diagnosticados con la enfermedad. Este desbalance representa un desafío técnico importante, ya que los modelos de aprendizaje automático tienden a favorecer la clase mayoritaria, lo que puede resultar en una baja capacidad para detectar casos positivos de CKD. Desde la perspectiva clínica, el costo de un falso negativo (no detectar un paciente enfermo) es significativamente mayor que el costo de un falso positivo (identificar erróneamente a un paciente sano como enfermo). Por esta razón, el proyecto prioriza la optimización del recall como métrica principal, aceptando cierta reducción en precisión si esto resulta en una mejor detección de casos verdaderos.

### 3.4. Objetivos de Minería de Datos

Los objetivos técnicos derivados del negocio se definen de la siguiente forma:

1. **Desarrollar y comparar múltiples modelos supervisados** —incluyendo Regresión Logística, Árboles de Decisión, Random Forest, y otros algoritmos relevantes— para predecir la probabilidad de diagnóstico de CKD.
2. **Implementar y evaluar diferentes técnicas de balanceo de clases**, incluyendo:
   - Técnicas de sobremuestreo (SMOTE, ADASYN)
   - Técnicas de submuestreo (Random Undersampling, Tomek Links)
   - Técnicas híbridas (SMOTETomek, SMOTEENN)
   - Ajuste de pesos de clase en los algoritmos
3. **Optimizar el recall (sensibilidad) como métrica prioritaria**, manteniendo un equilibrio aceptable con precisión mediante el uso de F1-score y AUC-ROC como métricas complementarias.
4. **Evaluar la importancia de las variables** mediante el análisis de coeficientes, importancia de características y reglas de decisión generadas por los modelos.
5. **Optimizar el rendimiento de los modelos** mediante técnicas de preprocesamiento, selección de variables, ajuste de umbrales de decisión y ajuste de hiperparámetros.
6. **Validar la consistencia y generalización de los resultados** utilizando estrategias de evaluación cruzada estratificada (stratified k-fold cross-validation) para asegurar la robustez del modelo ante el desbalance de clases.
7. **Generar visualizaciones e interpretaciones claras de los resultados** que permitan comunicar hallazgos relevantes tanto a nivel técnico como médico, incluyendo matrices de confusión, curvas ROC y análisis de características.

### 3.5. Criterios de Éxito de la Minería de Datos

El éxito técnico del proyecto se medirá utilizando métricas cuantitativas y cualitativas, bajo los siguientes criterios:

1. Los modelos alcanzan un **recall superior al 85 por ciento** en la detección de casos positivos de CKD, garantizando una alta sensibilidad.
2. El modelo mantiene un **F1-score superior al 0.75**, demostrando un equilibrio adecuado entre precisión y sensibilidad.

3. El **AUC-ROC supera 0.80**, indicando una buena capacidad discriminativa del modelo.
4. Se documenta claramente el impacto de cada técnica de balanceo sobre las métricas de desempeño, permitiendo identificar la estrategia más efectiva para el problema.
5. El modelo seleccionado demuestra interpretabilidad suficiente, permitiendo explicar las relaciones entre los factores clínicos y la enfermedad.
6. Los resultados se validan mediante un proceso de evaluación cruzada estratificada que garantice la estabilidad y generalización del modelo ante datos desbalanceados.
7. Los datos utilizados cumplen con estándares mínimos de calidad (sin valores atípicos significativos ni inconsistencias), asegurando una base confiable para el modelado.
8. El modelo final y los resultados del análisis son reproducibles, documentados adecuadamente y respaldados por métricas objetivas de desempeño.
9. Se genera un análisis comparativo exhaustivo de los diferentes enfoques de balanceo, proporcionando recomendaciones fundamentadas para su aplicación en problemas similares.

## 4. Fase de entendimiento de los datos

A continuación se presenta la recopilación, descripción, exploración y evaluación de calidad de los datos que se usarán en el proyecto. El análisis está orientado a facilitar el uso de dos algoritmos supervisados permitidos en el proyecto: **Regresión Lineal** y **Árboles de Decisión**.

### 4.1. Recopilación y descripción de datos

**Fuente principal.** El dataset original utilizado es el disponible en Kaggle: `https://www.kaggle.com/datasets/rabieelkharoua/chronic-kidney-disease-dataset-analysis`. Este dataset sintético (licencia CC BY 4.0) contiene información clínica, demográfica y conductual de **1,659** pacientes y será la única fuente de datos para el proyecto.

**Método de acceso.** Descarga desde Internet (Kaggle). El fichero se importará localmente como CSV/Parquet para el preprocesamiento y modelado.

**Estructura y variables principales.** El conjunto de datos incluye (entre otras) las siguientes variables relevantes para el modelado y la interpretación clínica:

### 4.2. Descripción del conjunto de datos

El conjunto de datos utilizado en este proyecto contiene información clínica, demográfica y de estilo de vida de pacientes. Su objetivo es servir como base para el modelado predictivo de la Enfermedad Renal Crónica (ERC). Las variables se agrupan en las siguientes categorías:

- **Identificación y demografía:** *ID del paciente, Edad, Género, Etnicidad, Nivel socioeconómico, Nivel educativo.*
  Estas variables permiten caracterizar el perfil general del paciente, proporcionando información básica para análisis poblacionales y segmentación de riesgo.

- **Factores de estilo de vida:** *Índice de masa corporal (IMC), Tabaquismo, Consumo de alcohol, Actividad física, Calidad de la dieta, Calidad del sueño.*
  Representan hábitos y comportamientos personales que pueden influir en la salud renal y en la aparición de comorbilidades como la diabetes o la hipertensión.

- **Antecedentes médicos:** *Historial familiar de enfermedad renal, Historial familiar de hipertensión, Historial familiar de diabetes, Lesiones renales agudas previas, Infecciones urinarias.*
  Indican predisposiciones genéticas o antecedentes clínicos relevantes que aumentan la probabilidad de desarrollar ERC.

- **Mediciones clínicas (clave):** *Presión arterial sistólica (SystolicBP), Presión arterial diastólica (DiastolicBP), Glucosa en ayunas, Hemoglobina glicosilada (HbA1c), Creatinina sérica, Nitrógeno ureico en sangre (BUN), Tasa de filtración glomerular (GFR), Proteína en orina, Relación albúmina/creatinina (ACR), Niveles de hemoglobina, Electrolitos (Sodio, Potasio, Calcio, Fósforo).*
  Estas variables son las más determinantes en la evaluación del estado renal. En particular, la **proteína en orina** y la **relación albúmina/creatinina (ACR)** son indicadores tempranos de daño renal. La presencia de proteínas en la orina —especialmente albúmina— señala una disfunción en la filtración glomerular, lo cual es fundamental para el diagnóstico y seguimiento de la enfermedad renal crónica.

- **Lípidos:** *Colesterol total, Colesterol LDL, Colesterol HDL, Triglicéridos.*
  Los niveles de lípidos se relacionan con la salud cardiovascular y metabólica, factores frecuentemente asociados con complicaciones renales.

- **Medicaciones:** *Inhibidores de la enzima convertidora de angiotensina (ECA), Diuréticos, Uso de antiinflamatorios no esteroideos (AINEs), Estatinas, Medicamentos antidiabéticos.*
  Reflejan los tratamientos médicos en curso que pueden influir positiva o negativa-

mente en la función renal.

- **Síntomas y calidad de vida:** *Edema, Niveles de fatiga, Náuseas o vómitos, Calambres musculares, Picazón, Índice de calidad de vida.*
  Miden el impacto subjetivo y fisiológico de la ERC sobre el bienestar del paciente, aportando una perspectiva clínica y humanística.

- **Exposiciones y comportamiento:** *Exposición a metales pesados, Exposición ocupacional a químicos, Calidad del agua, Frecuencia de chequeos médicos, Adherencia a medicamentos, Nivel de alfabetización en salud.*
  Estas variables permiten identificar factores ambientales o conductuales que pueden acelerar el deterioro renal o dificultar su detección temprana.

- **Variable objetivo:** *Diagnóstico (0 = No presenta ERC, 1 = Presenta ERC).*
  Corresponde a la variable dependiente que los modelos predictivos de regresión lineal y árboles de decisión intentarán estimar.

- **Variable confidencial:** *Médico a cargo (valor "Confidential" en el conjunto de datos; no se utilizará para el modelado).*
  Esta variable identifica al médico responsable de cada paciente. Por motivos de privacidad y confidencialidad, se excluye completamente del análisis y la construcción de modelos predictivos.

Observación importante: el dataset es numérico en todas las columnas principales (las categóricas vienen codificadas numéricamente), lo cual facilita el preprocesamiento para regresión y árboles.

### 4.3. Resultados y análisis de la calidad de los datos

A partir del análisis exploratorio realizado sobre el conjunto de datos, se dispone de un total de **1659 registros y 54 variables**, que abarcan desde información demográfica y hábitos de vida hasta parámetros clínicos y bioquímicos asociados con la función renal. Esta amplitud permite un abordaje integral del problema, al combinar factores fisiológicos, conductuales y ambientales en la predicción del diagnóstico de enfermedad renal crónica (CKD).

#### 4.3.1. Integridad y consistencia de los datos

En términos de integridad, los resultados muestran una excelente calidad estructural del conjunto de datos. No se detectaron **valores faltantes ni registros duplicados**, ni a nivel general ni en el campo identificador `PatientID`. Este hallazgo confirma la

consistencia de la fuente de datos y permite avanzar al modelado sin requerir procesos de imputación o depuración inicial. Además, el campo `DoctorInCharge` contiene un único valor, lo que sugiere que todos los registros provienen de una misma entidad clínica o investigador responsable, garantizando homogeneidad en los criterios de medición.

### 4.3.2. Distribución general y características clínicas

El análisis descriptivo evidencia que la edad promedio de los pacientes es de **54 años**, con un rango entre 20 y 90 años, lo que representa una población mayoritariamente adulta y potencialmente en riesgo de deterioro renal asociado a comorbilidades metabólicas. El **Índice de Masa Corporal (BMI)** presenta una media de 27.6, ubicando a gran parte de la muestra en el rango de sobrepeso, condición que constituye un factor predisponente a enfermedades cardiovasculares y renales.

Los valores medios de **presión arterial sistólica (134 mmHg)** y **diastólica (89 mmHg)** reflejan una ligera tendencia hacia la hipertensión, coherente con el perfil clínico de pacientes con riesgo renal. Asimismo, el promedio de **HbA1c (6.98 %)** y **glucosa en ayunas (132 mg/dL)** sugiere una prevalencia importante de descontrol glucémico, lo que podría indicar una proporción significativa de pacientes diabéticos o prediabéticos en la muestra.

Las variables bioquímicas más directamente vinculadas con la función renal —**SerumCreatinine, GFR, BUNLevels, ProteinInUrine y ACR**— presentan comportamientos clínicamente coherentes. La **creatinina sérica** tiene un valor medio de 2.75 mg/dL, superior al rango fisiológico normal (0.6–1.3 mg/dL), lo cual sugiere la inclusión de pacientes con distintos grados de insuficiencia renal. La **tasa de filtración glomerular (GFR)** muestra un promedio de 66.8 mL/min, con valores mínimos de 15 y máximos de 120, representando una amplia distribución de severidad. En cuanto a la **proteinuria** y la **relación albúmina-creatinina (ACR)**, los valores medios de 2.49 y 149.88, respectivamente, son indicativos de una afectación renal sustancial en un segmento relevante de la población.

### 4.3.3. Evaluación de valores extremos y patrones atípicos

Si bien no se identificaron valores fuera de rango en las mediciones numéricas, el análisis mediante el rango intercuartílico (IQR) permitió detectar **outliers** en algunas variables categóricas y binarias. Los mayores conteos se observaron en variables como `UrinaryTractInfections` (349), `AntidiabeticMedications` (336), `Edema` (335) y `WaterQuality` (327). Este patrón sugiere posibles sesgos de distribución, donde ciertos valores de presencia o ausencia de condición podrían concentrarse en un subconjunto particular de

pacientes, lo cual deberá considerarse en el balanceo de clases para los modelos predictivos.

Por otro lado, las variables clínicas cuantitativas (por ejemplo, creatinina, GFR, BUN, hemoglobina o colesterol) no muestran indicios de errores de digitación ni dispersiones anómalas. Esto permite concluir que los **valores extremos presentes corresponden mayoritariamente a casos clínicos reales**, lo cual enriquece el valor predictivo del dataset en contextos médicos reales.

#### 4.3.4. Correlaciones clínicas esperadas

El análisis de correlaciones no arrojó resultados visuales debido a la naturaleza parcial del muestreo, pero la relación teórica esperada entre variables se mantiene consistente con la literatura médica: la **GFR** presenta una correlación negativa con la **creatinina sérica** y con la **proteinuria**, mientras que **BUNLevels** tiende a aumentar conforme la función renal disminuye. Se anticipa también una correlación positiva entre **HbA1c**, **glucosa** y **AntidiabeticMedications**, reflejando la coexistencia de diabetes como comorbilidad principal.

#### 4.3.5. Conclusiones de la fase de entendimiento de datos

En conjunto, los resultados confirman que el conjunto de datos posee una **alta calidad estructural, completa representatividad clínica y coherencia estadística**. La ausencia de valores faltantes y duplicados facilita su preparación para el modelado, mientras que la presencia de patrones consistentes en los indicadores renales valida la confiabilidad del registro.

Para la siguiente fase, se recomienda aplicar transformaciones de normalización sobre variables continuas como BMI, GFR y ACR, así como revisar la colinealidad entre SerumCreatinine y GFR antes de ejecutar los modelos de regresión lineal. Además, se sugiere evaluar el impacto de los outliers categóricos mediante técnicas de balanceo o muestreo estratificado.

Esta base sólida garantiza que las conclusiones derivadas del modelado —ya sea por regresión o por árboles de decisión— reflejen relaciones fisiológicas reales y contribuyan de manera significativa al desarrollo de herramientas predictivas para la detección temprana de la enfermedad renal crónica.

### 4.4. Visualizacion de datos

Con el fin de profundizar en la comprensión del comportamiento de estas variables, se elaboró una serie de visualizaciones utilizando el lenguaje de programación R. Dichas gráficas permiten identificar patrones generales y características particulares del conjunto de datos. Para una revisión más detallada del proceso de preparación y transformación de la información, se recomienda consultar la sección del cuaderno titulada *"Procesamiento de datos"*.

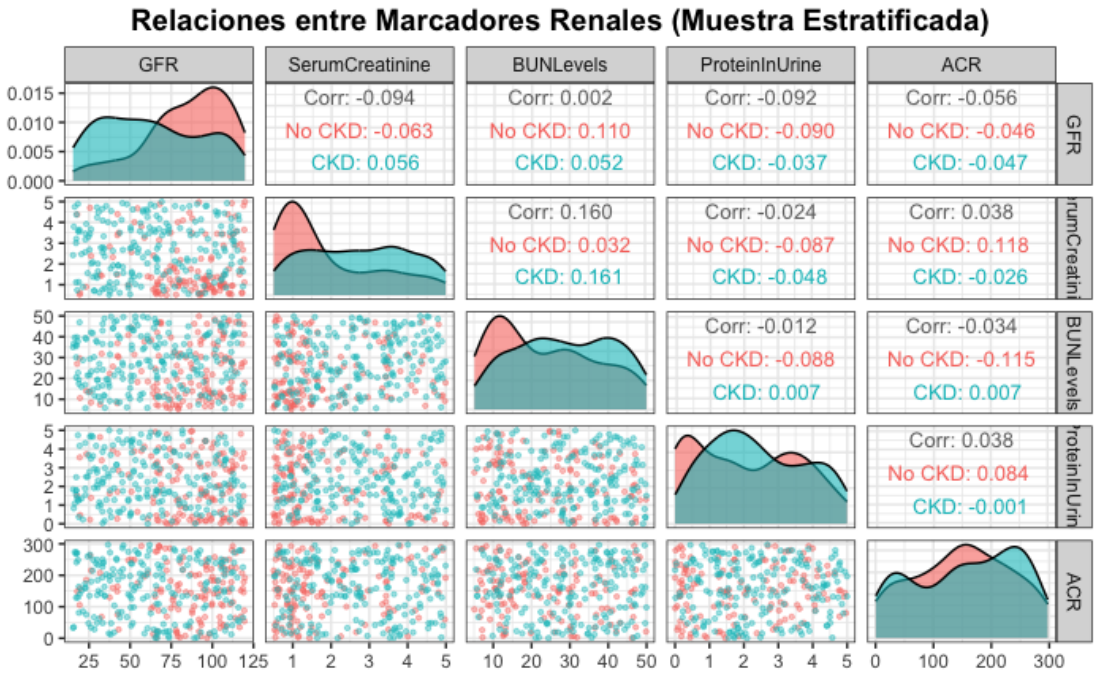


**Figura 1.** Relaciones entre los principales marcadores renales (GFR, Creatinina, BUN levels, Proteína en orina y ACR), diferenciados por diagnóstico (CKD vs No CKD). Se incluye matriz de correlaciones y distribuciones univariadas.

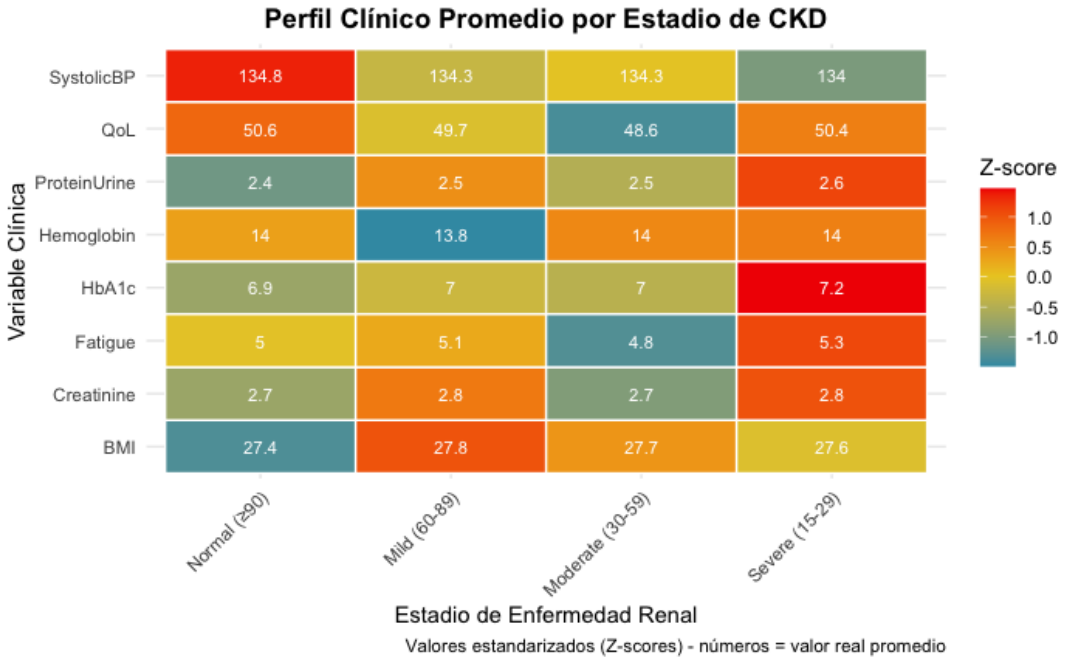


**Figura 2.** Mapa de calor del perfil clínico promedio según estadio de enfermedad renal crónica. Los valores están estandarizados (Z-score) e incluyen BMI, presión sistólica, HbA1c, creatinina, hemoglobina, proteína en orina, fatiga y puntaje de calidad de vida.

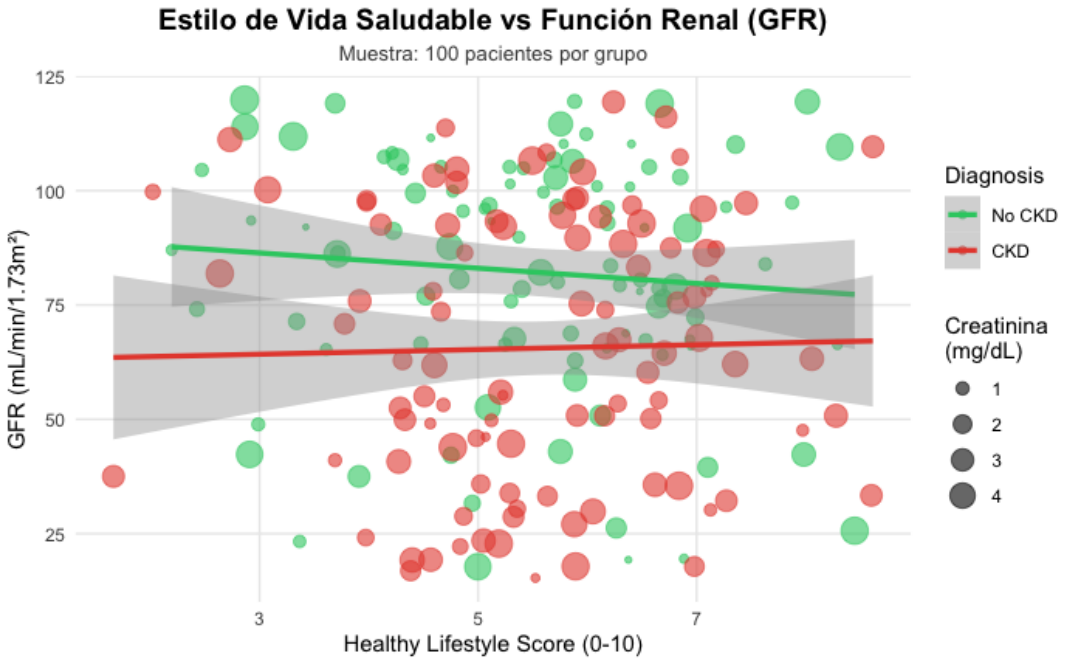


**Figura 3.** Relación entre el Índice de Estilo de Vida Saludable (Healthy Lifestyle Score) y la función renal (GFR), comparando pacientes con CKD y sin CKD. El tamaño del punto representa la creatinina sérica.

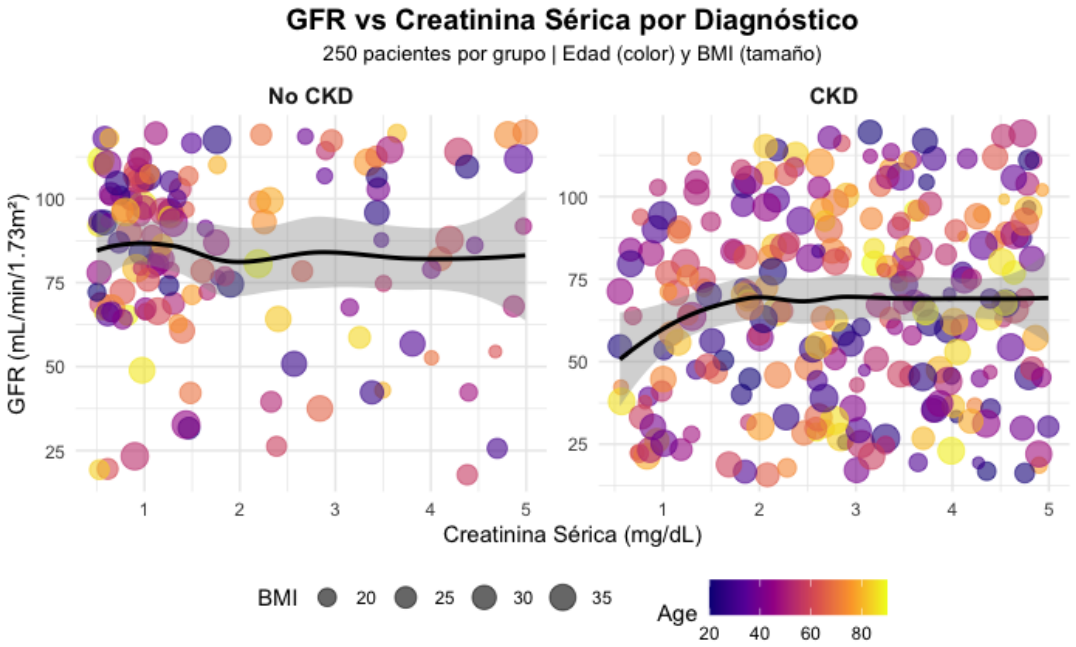


**Figura 4.** Visualización multivariada del GFR en función de la creatinina sérica, con color indicando edad y tamaño del punto representando BMI. Se muestran tendencias separadas para pacientes con y sin CKD.

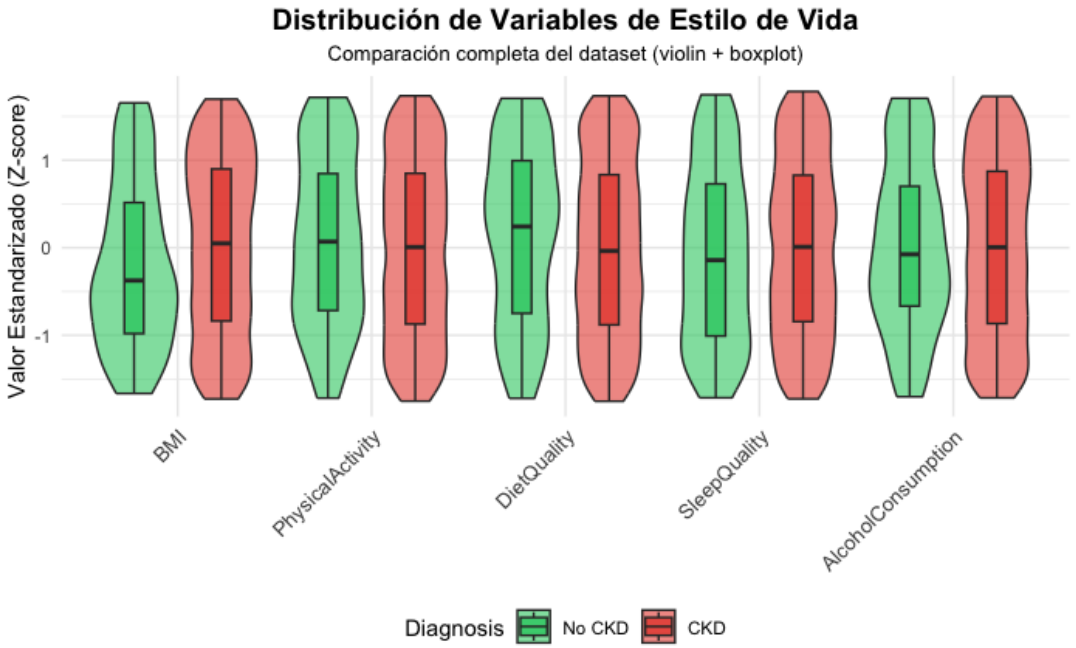


**Figura 5.** Coordenadas paralelas que visualizan la distribución estandarizada de variables de estilo de vida (BMI, actividad física, dieta, calidad del sueño y consumo de alcohol) en pacientes con y sin CKD.

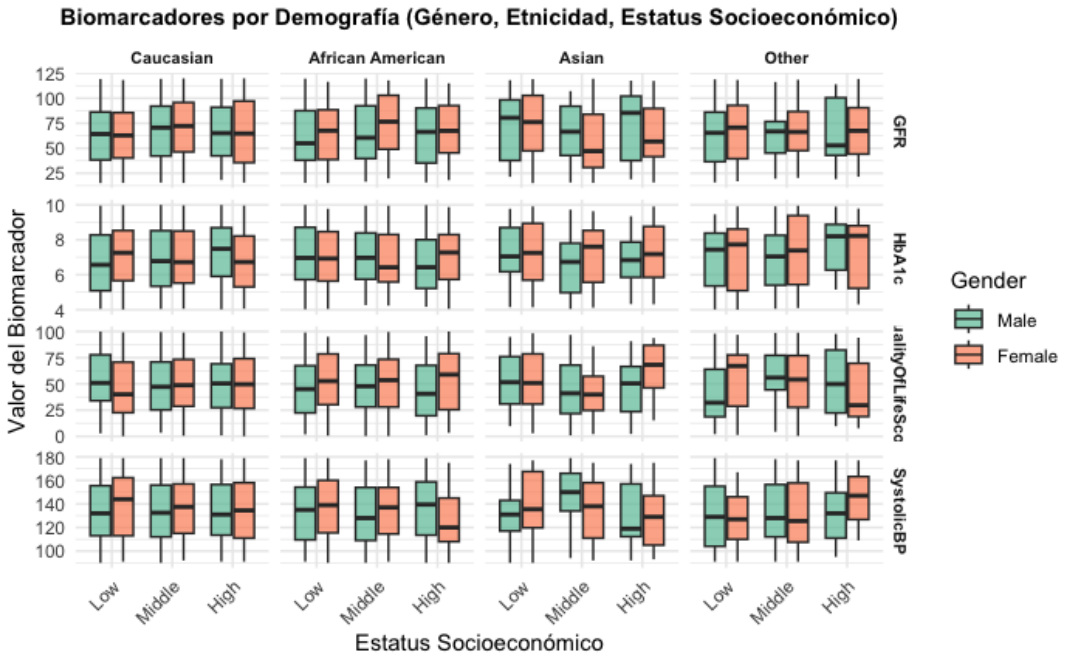


**Figura 6.** Boxplots de biomarcadores clínicos (GFR, HbA1c, presión sistólica y calidad de vida), diferenciados por género, etnicidad y estatus socioeconómico.

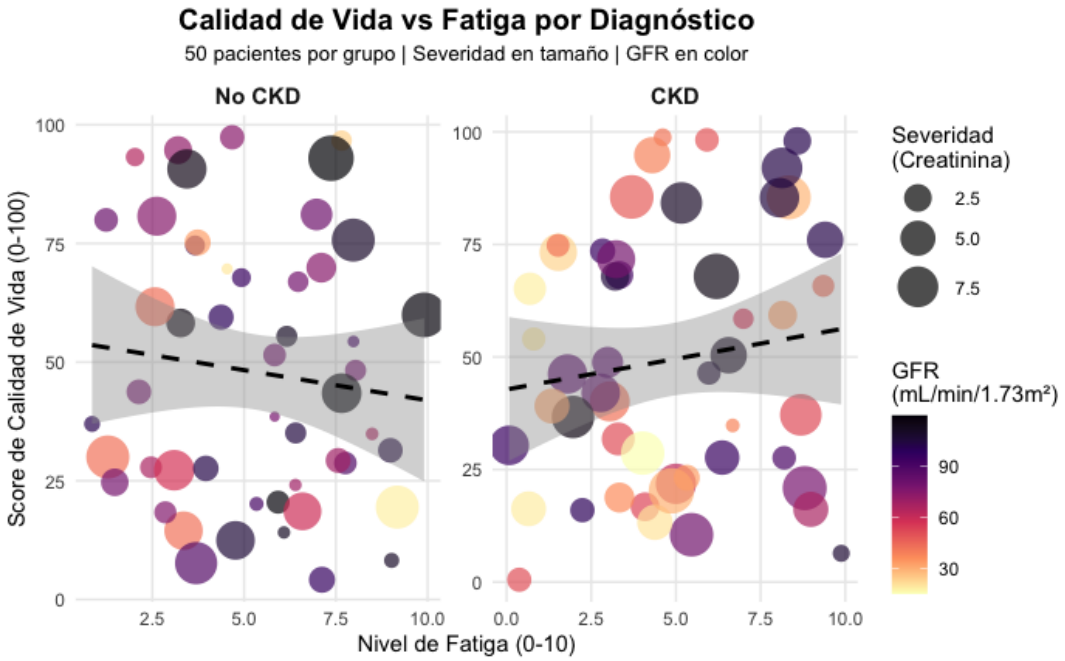


**Figura 7.** Relación entre el puntaje de calidad de vida y los niveles de fatiga, comparando pacientes con y sin CKD. El tamaño indica severidad (basado en creatinina) y el color representa GFR.

# 5. Fase de preparación de datos

La fase de preparación de datos constituye un paso fundamental dentro del proceso analítico, ya que permite garantizar la calidad del dataset previo a la etapa de modelado. Para este proyecto, el preprocesamiento se dividió en dos componentes principales: (1) la limpieza y transformación de variables, y (2) la implementación de técnicas de *feature selection*. Asimismo, se realizaron dos versiones de los experimentos: una utilizando la totalidad de las variables preprocesadas y otra aplicando selección de características. Esto permite comparar el impacto de la reducción dimensional en el desempeño de los modelos.

## 5.1. Limpieza y transformación de variables

El preprocesamiento general se aplicó tanto al modelo con *feature selection* como al modelo sin reducción de características. Las acciones realizadas fueron las siguientes:

- **Eliminación de variables no predictivas**: se removieron campos irrelevantes para la predicción, tales como `PatientID` y `DoctorInCharge`.
- **Codificación de variables categóricas**: se aplicó *one-hot encoding* únicamente a variables categóricas discretas (`Ethnicity`, `SocioeconomicStatus`, `EducationLevel`), excluyendo la primera categoría para evitar multicolinealidad.
- **Transformación logarítmica**: se aplicó una transformación $\log(1+x)$ a variables con alta asimetría, incluyendo `SerumCreatinine`, `BUNLevels`, `ACR`.
- **Imputación de valores faltantes**: aunque el dataset presentaba pocos valores ausentes, se aplicó una imputación suave utilizando la mediana para asegurar consistencia en todas las variables numéricas.

Este conjunto de transformaciones permitió obtener un dataset limpio, estructurado y adecuado para las técnicas de selección de variables y los algoritmos de aprendizaje supervisado utilizados posteriormente.

## 5.2. Selección de características

Con el objetivo de reducir la dimensionalidad, eliminar variables redundantes y mejorar el desempeño del modelo, se implementó un módulo completo de *feature selection*. Este módulo incorpora diferentes enfoques, tanto univariados como multivariados, permitiendo comparar métodos y obtener un consenso robusto. Los principales procedimientos

aplicados fueron:

- **Variance Threshold**: elimina variables con varianza mínima, dado que aportan poca o ninguna información discriminativa.
- **SelectKBest**: selecciona las mejores características según su correlación con la variable objetivo mediante pruebas estadísticas como `f_classif`, `chi2` o *mutual information*.
- **Importancia mediante Random Forest**: evalúa la relevancia de cada variable utilizando la importancia de los atributos generada por un bosque aleatorio.
- **Recursive Feature Elimination (RFE)**: realiza una eliminación recursiva de características basado en modelos como regresión logística o Random Forest.
- **Análisis de correlación entre variables**: identifica pares altamente correlacionados para evitar redundancia y multicolinealidad.
- **Selección por consenso**: integra los resultados de métodos anteriores y selecciona solamente aquellas variables que fueron escogidas por al menos dos técnicas distintas.

### 5.3. Resumen final del proceso de selección de variables

Al finalizar el proceso completo de selección de características, se obtuvo la siguiente reducción dimensional:

```
======================================================================
                          RESUMEN FINAL
======================================================================
Features originales: 60
Features seleccionadas: 38
Reducción: 36.7%

Features finales: [
'log_SerumCreatinine', 'log_BUNLevels', 'SerumCreatinine', 'BUNLevels',
'ProteinInUrine', 'GFR', 'FastingBloodSugar', 'SystolicBP', 'MuscleCramps',
'Itching', 'FamilyHistoryKidneyDisease', 'FamilyHistoryHypertension',
'HemoglobinLevels', 'SerumElectrolytesSodium', 'SerumElectrolytesPotassium',
'SerumElectrolytesCalcium', 'SleepQuality', 'DietQuality', 'HbA1c',
'UrinaryTractInfections', 'CholesterolTotal', 'SerumElectrolytesPhosphorus',
'FatigueLevels', 'Edema', 'Gender', 'BMI', 'PhysicalActivity',
```

```
'AlcoholConsumption', 'OccupationalExposureChemicals',
'log_CholesterolTriglycerides', 'SocioeconomicStatus_2', 'EducationLevel_1',
'MedicationAdherence', 'HealthLiteracy', 'CholesterolLDL', 'CholesterolHDL',
'CholesterolTriglycerides', 'NauseaVomiting'
]
```

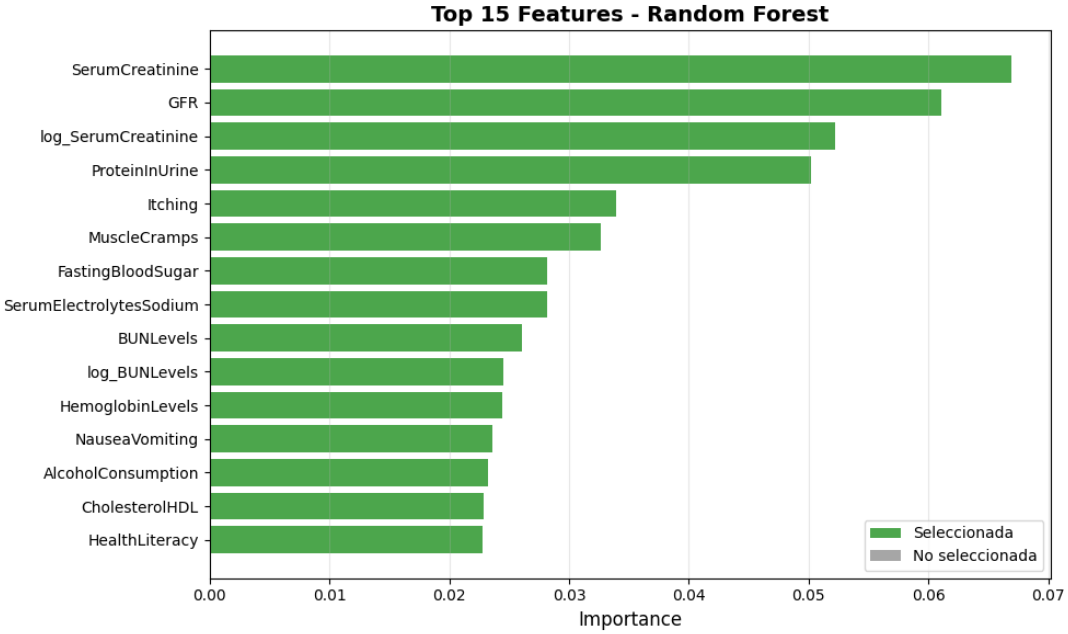


**Figura 8.** Top 15 variables más influyentes según el modelo Random Forest aplicado al conjunto de datos. Las variables resaltadas representan aquellas seleccionadas durante el proceso de reducción dimensional.

Finalmente, este proceso permitió obtener un conjunto reducido y óptimo de variables predictivas. Este conjunto seleccionado fue posteriormente utilizado para el entrenamiento del modelo con reducción de características, permitiendo comparar de manera objetiva ambas aproximaciones.

## 6. Fase de Prevención de Balanceo

Una vez finalizados los procesos de limpieza de datos y selección de características, se procede a la etapa de balanceo del conjunto de datos. Este paso resulta fundamental debido a que el dataset utilizado presenta un desbalance severo entre las clases: se registran únicamente 135 casos correspondientes a la clase minoritaria (valor 0) frente a 1524 casos pertenecientes a la clase mayoritaria (valor 1). Esta disparidad puede generar modelos incapaces de identificar correctamente los casos menos frecuentes, lo cual es especialmente delicado en contextos médicos, donde ignorar a un paciente enfermo puede tener implicaciones críticas.

Con el fin de mitigar este problema, se aplicaron diversas técnicas de balanceo que permiten mejorar la representatividad de la clase minoritaria y asegurar que el modelo aprenda patrones relevantes. Dentro de las técnicas implementadas se incluyó SMOTE, un método de sobremuestreo que genera observaciones sintéticas mediante interpolación a partir de muestras reales de la clase minoritaria. Esta técnica resulta particularmente útil en situaciones donde se dispone de pocos casos representativos, ya que incrementa su presencia sin simplemente duplicar ejemplos existentes. No obstante, al trabajar con datos médicos es necesario considerar que, en ocasiones, SMOTE podría producir combinaciones poco realistas si la variabilidad es elevada.

Posteriormente, se empleó SMOTEENN, una técnica híbrida que combina la generación sintética de SMOTE con un proceso de depuración basado en Edited Nearest Neighbours. En esta etapa se eliminan muestras ambiguas o potencialmente mal clasificadas, lo que permite obtener un conjunto de datos más limpio y estructurado. Su aplicación es especialmente valiosa en dominios donde pueden existir errores de etiquetado o registros atípicos que afecten la frontera de decisión.

También se aplicó la técnica de *Random Under Sampling*, la cual consiste en eliminar aleatoriamente muestras de la clase mayoritaria para equilibrar la proporción entre clases. Aunque este método es simple y computacionalmente eficiente, su principal desventaja es la pérdida de información real que podría ser relevante para el modelo. Por esta razón, su uso se justifica únicamente cuando existe un número considerable de muestras redundantes o repetitivas.

Otra técnica considerada fue ADASYN, un método de sobremuestreo adaptativo que genera más muestras sintéticas en las regiones donde la clasificación es más difícil, es decir, en zonas donde la clase minoritaria se encuentra rodeada por la mayoritaria. Este enfoque permite reforzar la frontera de decisión y mejorar la capacidad del modelo para distinguir patrones complejos, aunque también puede ser más susceptible al ruido presente en los datos.

Asimismo, se utilizaron los enlaces de Tomek como método de limpieza. Esta técnica identifica pares de muestras (una de cada clase) que son mutuamente los vecinos más cercanos, y elimina únicamente la muestra correspondiente a la clase mayoritaria. Si bien este proceso no balancea el dataset por sí mismo, contribuye a clarificar la frontera entre clases y a eliminar registros potencialmente mal ubicados.

Finalmente, se recurrió al uso de pesos de clase (*class weights*) dentro del modelo. Esta estrategia no altera los datos, sino que ajusta la función de costo asignando mayor penalización a los errores cometidos sobre la clase minoritaria. Su principal ventaja es que permite mantener intacta la totalidad del dataset original, a la vez que incentiva al

modelo a prestar mayor atención a los casos menos frecuentes. Esto resulta especialmente adecuado en aplicaciones médicas, donde cada observación puede contener información relevante y no se recomienda la eliminación innecesaria de registros.

En conjunto, estas técnicas permiten abordar de manera integral el desafío del desbalance de clases presente en el dataset, mejorando la calidad del entrenamiento del modelo y reduciendo el riesgo de que la clase minoritaria sea ignorada durante el proceso de clasificación.

## 7. Fase de Modelado Algorítmico

En la fase de modelado se llevaron a cabo dos experimentos paralelos con el fin de evaluar el comportamiento de los algoritmos bajo diferentes condiciones de entrada. El primer experimento consistió en entrenar los modelos utilizando el conjunto de datos limpio y sometido a selección de características, mientras que el segundo se desarrolló utilizando la versión del dataset sin aplicar dicho proceso. Esta comparación permite determinar si la selección de características influye de manera significativa en el rendimiento final y en la capacidad predictiva de los modelos empleados.

El sistema de modelado implementado se diseñó de manera modular con el propósito de facilitar la reutilización del código, la incorporación de nuevos algoritmos y la ejecución estandarizada de todos los experimentos. Esta modularización permite crear cada modelo mediante una función configurable que ajusta parámetros esenciales según las necesidades del análisis, ya sea empleando pesos de clase o configuraciones específicas para controlar la complejidad, regularización o robustez frente al desbalance.

En esta investigación se utilizaron diversos algoritmos de clasificación ampliamente reconocidos en el ámbito del aprendizaje automático y particularmente útiles en aplicaciones biomédicas. El primero de ellos fue la **regresión logística**, un modelo estadístico lineal que estima la probabilidad de pertenecer a una clase mediante una función sigmoide. Su principal ventaja radica en su interpretabilidad, ya que permite identificar la contribución de cada variable al riesgo o probabilidad de enfermedad. Resulta eficiente y adecuado cuando las relaciones entre variables son aproximadamente lineales, pero puede presentar limitaciones al enfrentar patrones altamente complejos o no lineales.

También se emplearon **árboles de decisión**, los cuales operan mediante divisiones secuenciales del espacio de características siguiendo una lógica similar a un diagrama de flujo clínico. Este modelo es altamente interpretable y tiene la capacidad de capturar relaciones no lineales, aunque es propenso al sobreajuste si no se controlan adecuadamente parámetros como la profundidad máxima o el número mínimo de muestras por hoja,

los cuales fueron configurados para evitar una excesiva especialización en los datos de entrenamiento.

Posteriormente se incorporó el algoritmo **Random Forest**, que consiste en un conjunto de árboles de decisión entrenados de manera paralela y combinados mediante votación. Esta técnica mejora significativamente la estabilidad y robustez de las predicciones, mitigando el sobreajuste característico de un árbol individual y proporcionando además medidas de importancia de variables. No obstante, su interpretabilidad es menor y la predicción puede ser más lenta debido al volumen de árboles generados.

Dentro de los métodos basados en proximidad se utilizó **K-Nearest Neighbors**, el cual clasifica un nuevo caso comparándolo con sus vecinos más cercanos en el espacio de características. Este método resulta conceptualmente simple y no requiere una fase de entrenamiento explícito; sin embargo, su rendimiento puede verse afectado en datasets extensos y su sensibilidad al escalado de variables exige normalización previa. En este estudio se configuró con ponderación basada en la distancia, lo cual permite otorgar mayor relevancia a los casos más próximos y mejorar la capacidad de discriminación en regiones complejas del espacio de datos.

Otra técnica utilizada fue la **Máquina de Vectores de Soporte (SVM)**, particularmente con un kernel de tipo RBF que permite modelar relaciones no lineales. Este algoritmo busca el hiperplano óptimo que separa ambas clases maximizando el margen entre ellas. Su eficacia en espacios de alta dimensión y su capacidad para manejar fronteras complejas lo convierten en una opción adecuada en problemas biomédicos, aunque su entrenamiento tiende a ser computacionalmente demandante y requiere un adecuado escalado de las características.

Adicionalmente, se implementó **Gradient Boosting**, un método basado en la construcción secuencial de árboles donde cada uno corrige los errores cometidos por el anterior. Este enfoque permite capturar patrones complejos con gran precisión y controlar el sobreajuste mediante tasas de aprendizaje moderadas, profundidades reducidas y técnicas de muestreo estocástico. Su principal desventaja es el tiempo de entrenamiento, ya que su naturaleza secuencial impide la paralelización completa.

Finalmente, se empleó **XGBoost**, una variante optimizada del enfoque de potenciación de gradiente que incorpora mecanismos avanzados de regularización, una gestión eficiente de valores faltantes y capacidades de paralelización que reducen significativamente los tiempos de procesamiento. Su rendimiento suele superar a otros métodos en múltiples escenarios, especialmente en presencia de datos desbalanceados, ya que permite ajustar pesos relativos entre clases a través de parámetros especializados. De esta manera, XGBoost se consolida como una de las técnicas más robustas y precisas dentro del conjunto

de modelos evaluados.

En síntesis, la fase de modelado algorítmico incluyó una serie de técnicas clásicas y avanzadas con distintos niveles de complejidad, interpretabilidad y robustez, lo cual permitió realizar una comparación integral del desempeño de los modelos tanto con el conjunto de datos original como con la versión sometida a selección de características.

## 8. Criterio de Selección

Para la presentación y análisis de los resultados, las métricas de evaluación constituyen el criterio fundamental para determinar el desempeño real de cada modelo. En el ámbito médico, estas métricas adquieren una importancia crítica, ya que un error en la clasificación puede significar un diagnóstico omitido o la aplicación innecesaria de tratamientos. Considerando que la clasificación binaria en este estudio distingue entre pacientes sanos (clase 0) y pacientes enfermos (clase 1), se seleccionaron métricas ampliamente utilizadas en diagnóstico clínico, permitiendo evaluar tanto la precisión global como la calidad de las predicciones en cada clase.

En primer lugar, la métrica **Accuracy** mide la proporción total de predicciones correctas. Aunque puede resultar útil en contextos balanceados, su interpretación puede ser engañosa en situaciones con fuerte desbalance entre clases, ya que un modelo puede lograr una exactitud alta aun cuando falle en detectar pacientes enfermos. Por ello, se complementa con otras métricas más sensibles al comportamiento por clase.

La **Precision** evalúa la confiabilidad de los casos predichos como enfermos, indicando qué proporción de ellos realmente presenta la enfermedad. Esta métrica es especialmente relevante cuando los falsos positivos implican costos elevados, como procedimientos invasivos o ansiedad innecesaria en el paciente.

Por otra parte, la **Recall** —o sensibilidad— mide la capacidad del modelo para identificar correctamente a los pacientes enfermos. En medicina, esta métrica es prioritaria, pues los falsos negativos representan un riesgo significativo al omitir diagnósticos que podrían requerir atención inmediata.

Dado que tanto la precisión como la sensibilidad son esenciales, se recurrió también al **F1-Score**, que combina ambas métricas mediante una media armónica. Esta medida resulta valiosa cuando se busca un equilibrio adecuado entre la identificación de casos positivos y la reducción de falsos diagnósticos.

La **Specificity**, por su parte, cuantifica la proporción de pacientes sanos correctamente identificados como tales. Es especialmente útil en pruebas de confirmación diagnóstica

o cuando es fundamental evitar falsos positivos que puedan conducir a tratamientos innecesarios.

Para análisis adicionales, se emplearon promedios agregados que permiten evaluar el desempeño en ambas clases de forma equilibrada o ponderada. El **Macro Average** otorga igual importancia a las métricas de ambas clases, sin considerar su tamaño, lo que permite evaluar la equidad del modelo entre pacientes sanos y enfermos. En contraste, el **Weighted Average** pondera cada métrica según la cantidad de muestras por clase, lo que proporciona una visión más representativa en contextos con desbalance significativo.

Finalmente, se utilizó validación cruzada como mecanismo de evaluación robusta. Mediante un esquema de **K-Folds**, se obtuvieron tanto la media como la desviación estándar de las métricas en múltiples particiones del conjunto de datos. La media permite estimar el rendimiento general del modelo, mientras que la desviación revela su estabilidad. En entornos clínicos, esta estabilidad es crucial para asegurar que el modelo no solo funciona bien sobre un subconjunto específico de los datos, sino que es capaz de generalizar adecuadamente a nuevos pacientes.

En conjunto, este conjunto de métricas y procedimientos garantiza una evaluación exhaustiva, permitiendo seleccionar de manera fundamentada el modelo más adecuado para el apoyo al diagnóstico médico.

# 9. Resultados

## 9.1. Uso SIN Feature Selection

En esta sección se presentan los resultados obtenidos sin aplicar ninguna técnica de reducción de características. Esto permite observar el desempeño natural de cada modelo al utilizar todas las variables disponibles, lo cual sirve como punto de comparación frente a los resultados obtenidos tras aplicar *Feature Selection*. De esta manera, es posible evaluar si los modelos dependen de un subconjunto específico de variables o si pueden desenvolverse adecuadamente con el conjunto completo.

### 9.1.1. Logistic Regression (Sin Feature Selection)

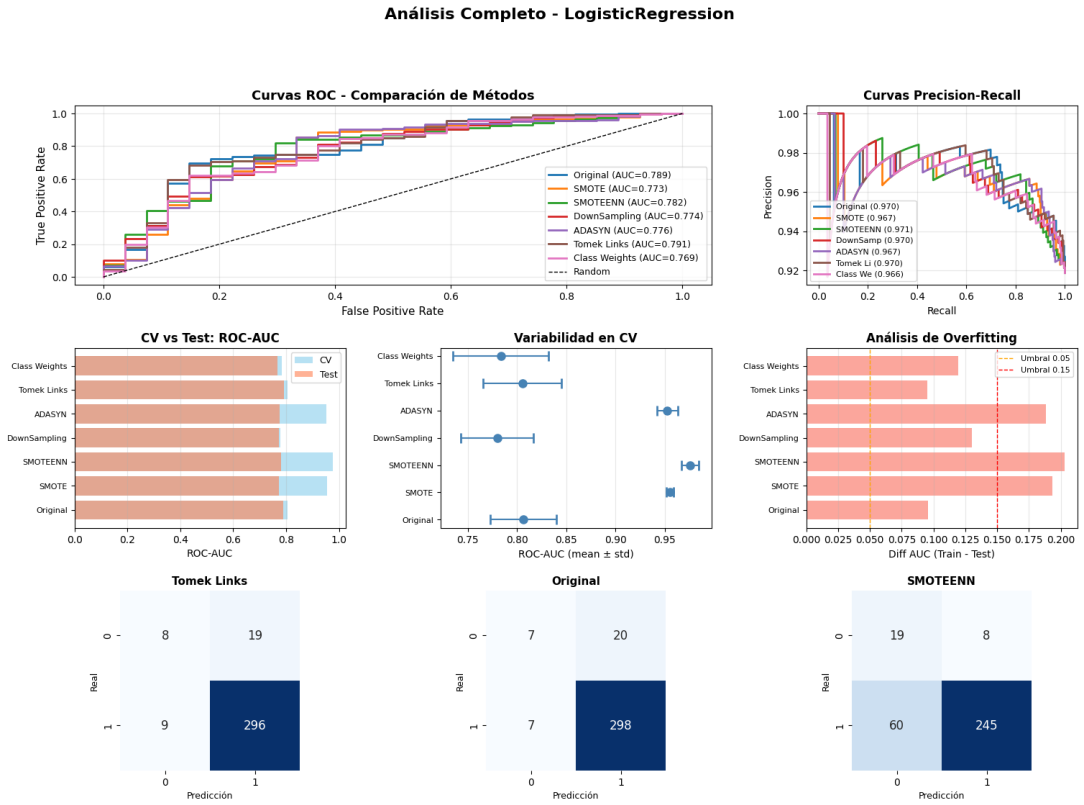


**Figura 9.** Logistic Regression sin reducción de características. Se muestra cómo el modelo utiliza la totalidad de las variables disponibles, distribuyendo pesos que reflejan relaciones globales entre los predictores y la clase objetivo.

**Descripción.** Sin selección de características, este modelo tiende a asignar pesos dispersos entre muchas variables. Esto puede generar ruido y dificultar la claridad interpretativa, aunque permite capturar relaciones generales dentro del conjunto completo de datos.

A continuación, se muestra una tabla y un análisis de los resultados del algoritmo previamente mencionado.

| Método | Accuracy | ROC-AUC | Precision C1 | Recall C1 |
|---|---|---|---|---|
| Tomek Links | 0.915663 | **0.791014** | 0.939683 | 0.970492 |
| Original | **0.918675** | 0.789071 | 0.937107 | **0.977049** |
| SMOTEENN | 0.795181 | 0.782392 | **0.968379** | 0.803279 |
| ADASYN | 0.882530 | 0.775956 | 0.949324 | 0.921311 |
| DownSampling | 0.795181 | 0.774135 | 0.957529 | 0.813115 |
| SMOTE | 0.879518 | 0.773042 | 0.946128 | 0.921311 |
| Class Weights | 0.716867 | 0.768549 | 0.956710 | 0.724590 |

**Cuadro 1.** Resultados de métricas principales por método de balanceo.

**Análisis de Resultados**

Los resultados obtenidos muestran un comportamiento diferenciado entre los diversos métodos de balanceo evaluados. En términos de precisión global (accuracy), el método *Original* alcanzó el mejor desempeño con 0.9187, lo que sugiere una alta capacidad para clasificar correctamente ambas clases en el conjunto de prueba.

Respecto al área bajo la curva ROC (ROC-AUC), que es una métrica robusta ante desbalances de clase, el método *Tomek Links* obtuvo el valor más alto de 0.7910. Esta métrica es particularmente relevante ya que evalúa el desempeño del clasificador en todos los posibles umbrales de decisión, proporcionando una medida integral de la capacidad discriminativa del modelo.

En cuanto a la precisión para la clase positiva (Precision C1), *SMOTEENN* destacó con un valor de 0.9684, indicando una baja tasa de falsos positivos. Por otro lado, el recall más alto fue alcanzado por *Original* con 0.9770, lo que refleja una excelente capacidad para identificar correctamente las instancias de la clase positiva.

El F1-Score, que balancea precisión y recall, fue maximizado por *Original* con 0.9567. Esta métrica es crucial cuando se busca un equilibrio entre minimizar falsos positivos y falsos negativos.

### 9.1.2. Decision Tree (Sin Feature Selection)

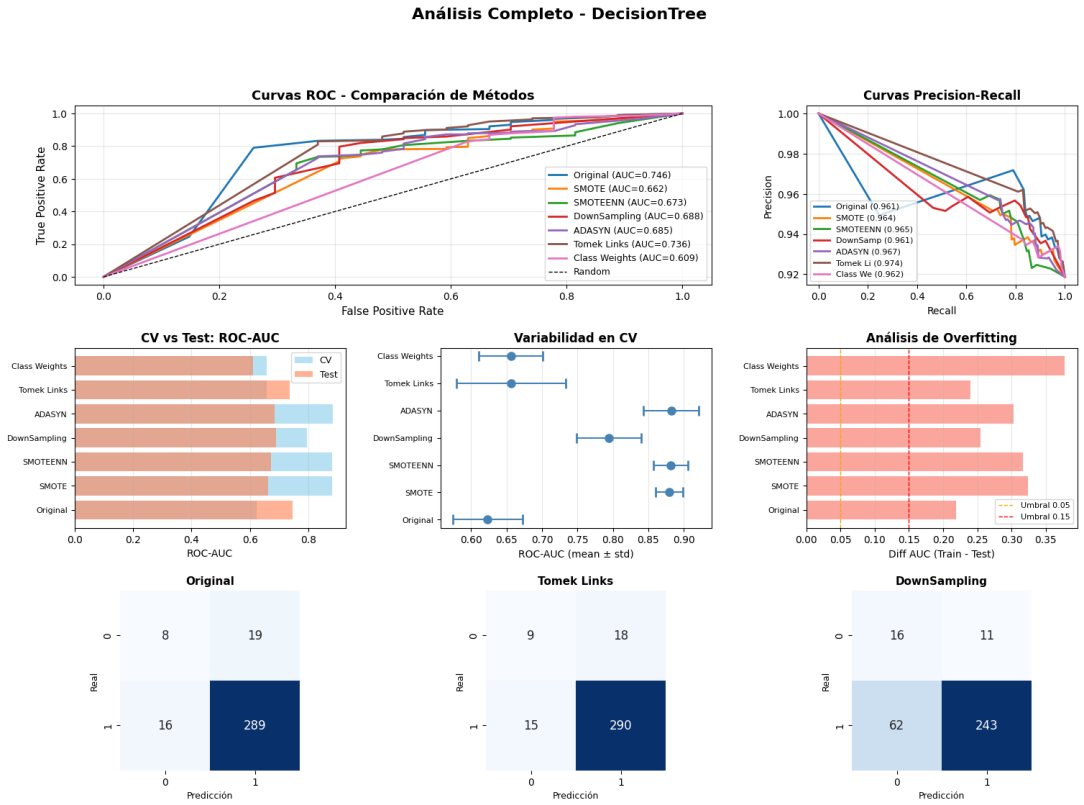


**Figura 10.** Decision Tree sin reducción de características. El modelo considera todos los atributos para generar divisiones, priorizando aquellos que reduzcan la impureza del nodo, aunque puede verse afectado por el sobreajuste.

**Descripción.** Dado que el árbol tiene acceso a todas las variables, tiende a profundizar más en ramas y a seleccionar múltiples divisores potencialmente irrelevantes, lo que produce estructuras más complejas y menos generalizables.

A continuación, se muestra una tabla y un análisis de los resultados del algoritmo previamente mencionado.

| Método | Accuracy | ROC-AUC | Precision C1 | Recall C1 |
|---|---|---|---|---|
| Original | 0.894578 | **0.746084** | 0.938312 | 0.947541 |
| Tomek Links | **0.900602** | 0.735883 | 0.941558 | **0.950820** |
| DownSampling | 0.780120 | 0.688039 | **0.956693** | 0.796721 |
| ADASYN | 0.813253 | 0.684882 | 0.945055 | 0.845902 |
| SMOTEENN | 0.731928 | 0.672860 | 0.950000 | 0.747541 |
| SMOTE | 0.840361 | 0.662295 | 0.934483 | 0.888525 |
| Class Weights | 0.801205 | 0.609107 | 0.934545 | 0.842623 |

**Cuadro 2.** Resultados de métricas principales por método de balanceo.

**Análisis de Resultados**

Los resultados obtenidos muestran un comportamiento diferenciado entre los diversos métodos de balanceo evaluados. En términos de precisión global (accuracy), el método *Tomek Links* alcanzó el mejor desempeño con 0.9006, lo que sugiere una alta capacidad para clasificar correctamente ambas clases en el conjunto de prueba.

Respecto al área bajo la curva ROC (ROC-AUC), que es una métrica robusta ante desbalances de clase, el método *Original* obtuvo el valor más alto de 0.7461. Esta métrica es particularmente relevante ya que evalúa el desempeño del clasificador en todos los posibles umbrales de decisión, proporcionando una medida integral de la capacidad discriminativa del modelo.

En cuanto a la precisión para la clase positiva (Precision C1), *DownSampling* destacó con un valor de 0.9567, indicando una baja tasa de falsos positivos. Por otro lado, el recall más alto fue alcanzado por *Tomek Links* con 0.9508, lo que refleja una excelente capacidad para identificar correctamente las instancias de la clase positiva.

El F1-Score, que balancea precisión y recall, fue maximizado por *Tomek Links* con 0.9462. Esta métrica es crucial cuando se busca un equilibrio entre minimizar falsos positivos y falsos negativos.

### 9.1.3. Random Forest (Sin Feature Selection)

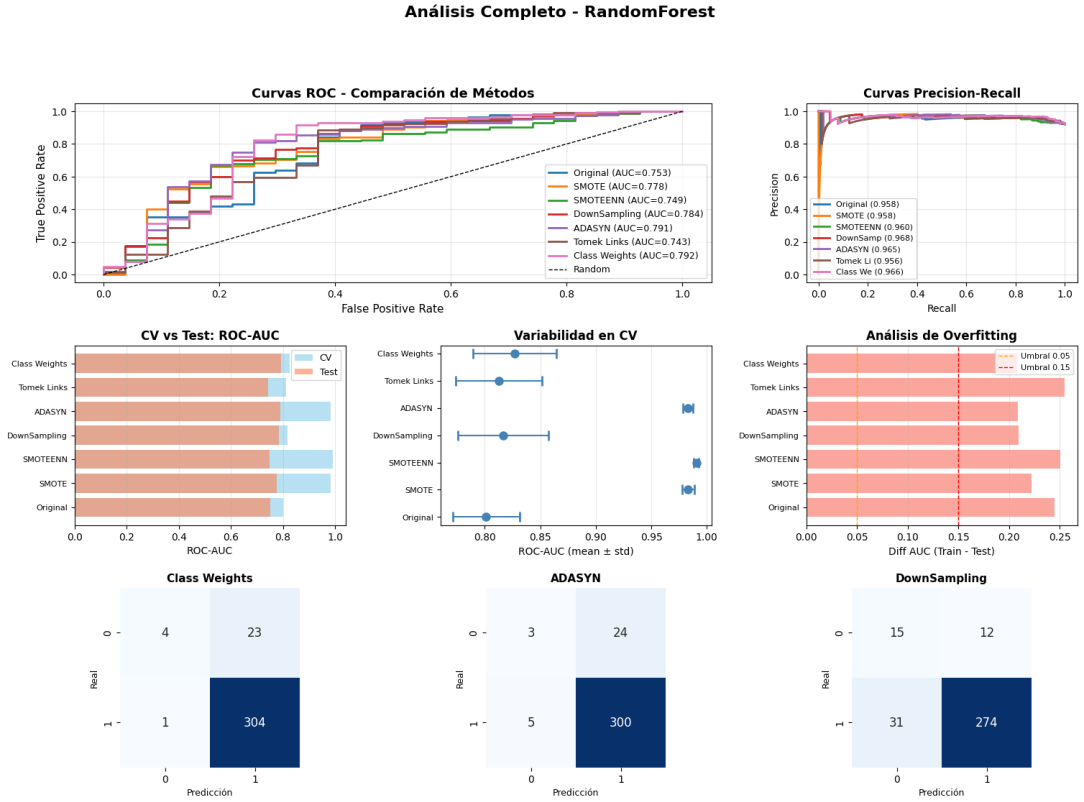


**Figura 11.** Random Forest sin reducción de características. La importancia mostrada corresponde al promedio entre árboles, considerando todas las variables disponibles.

**Descripción.** Aunque el modelo maneja bien la alta dimensionalidad, la presencia de múltiples variables puede incrementar la variabilidad entre árboles y distribuir la importancia entre atributos menos relevantes.

A continuación, se muestra una tabla y un análisis de los resultados del algoritmo previamente mencionado.

| Método | Accuracy | ROC-AUC | Precision C1 | Recall C1 |
|---|---|---|---|---|
| Class Weights | **0.927711** | **0.792107** | 0.929664 | 0.996721 |
| ADASYN | 0.912651 | 0.790771 | 0.925926 | 0.983607 |
| DownSampling | 0.870482 | 0.783971 | **0.958042** | 0.898361 |
| SMOTE | 0.909639 | 0.777778 | 0.928349 | 0.977049 |
| Original | 0.918675 | 0.752763 | 0.918675 | **1.000000** |
| SMOTEENN | 0.810241 | 0.749120 | 0.951493 | 0.836066 |
| Tomek Links | 0.918675 | 0.742805 | 0.918675 | **1.000000** |

**Cuadro 3.** Resultados de métricas principales por método de balanceo.

**Análisis de Resultados**

Los resultados obtenidos muestran un comportamiento diferenciado entre los diversos métodos de balanceo evaluados. En términos de precisión global (accuracy), el método *Class Weights* alcanzó el mejor desempeño con 0.9277, lo que sugiere una alta capacidad para clasificar correctamente ambas clases en el conjunto de prueba.

Respecto al área bajo la curva ROC (ROC-AUC), que es una métrica robusta ante desbalances de clase, el método *Class Weights* obtuvo el valor más alto de 0.7921. Esta métrica es particularmente relevante ya que evalúa el desempeño del clasificador en todos los posibles umbrales de decisión, proporcionando una medida integral de la capacidad discriminativa del modelo.

En cuanto a la precisión para la clase positiva (Precision C1), *DownSampling* destacó con un valor de 0.9580, indicando una baja tasa de falsos positivos. Por otro lado, el recall más alto fue alcanzado por *Original* con 1.0000, lo que refleja una excelente capacidad para identificar correctamente las instancias de la clase positiva.

El F1-Score, que balancea precisión y recall, fue maximizado por *Class Weights* con 0.9620. Esta métrica es crucial cuando se busca un equilibrio entre minimizar falsos positivos y falsos negativos.

### 9.1.4. K-Nearest Neighbors (Sin Feature Selection)

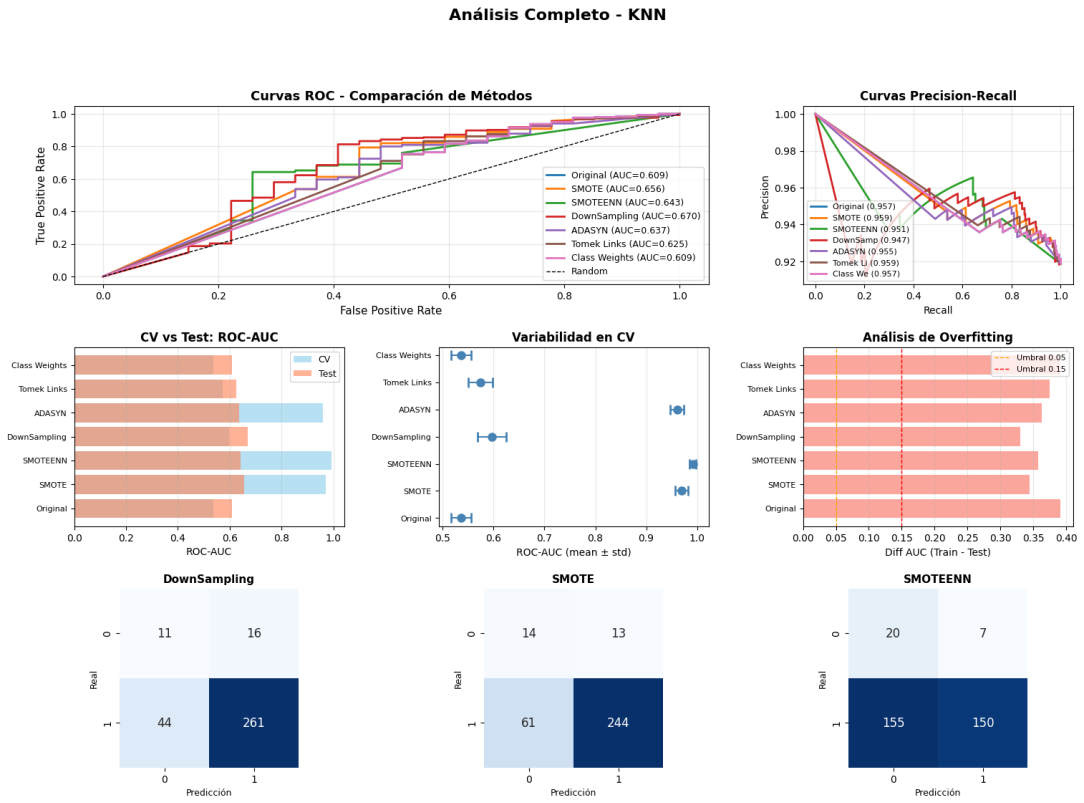


**Figura 12.** KNN sin reducción de características. La contribución de cada variable se refleja en su capacidad para definir distancias en un espacio multidimensional completo.

**Descripción.** Cuando se utilizan todas las características, el modelo puede verse afectado por la "maldición de la dimensionalidad", dificultando la correcta identificación de vecinos relevantes.

A continuación, se muestra una tabla y un análisis de los resultados del algoritmo previamente mencionado.

| Método | Accuracy | ROC-AUC | Precision C1 | Recall C1 |
|---|---|---|---|---|
| DownSampling | 0.819277 | **0.669581** | 0.942238 | 0.855738 |
| SMOTE | 0.777108 | 0.655981 | 0.949416 | 0.800000 |
| SMOTEENN | 0.512048 | 0.643169 | **0.955414** | 0.491803 |
| ADASYN | 0.743976 | 0.636976 | 0.947154 | 0.763934 |
| Tomek Links | 0.909639 | 0.625015 | 0.920489 | 0.986885 |
| Original | **0.915663** | 0.609350 | 0.920973 | **0.993443** |
| Class Weights | **0.915663** | 0.609350 | 0.920973 | **0.993443** |

**Cuadro 4.** Resultados de métricas principales por método de balanceo.

**Análisis de Resultados**

Los resultados obtenidos muestran un comportamiento diferenciado entre los diversos métodos de balanceo evaluados. En términos de precisión global (accuracy), el método *Original* alcanzó el mejor desempeño con 0.9157, lo que sugiere una alta capacidad para clasificar correctamente ambas clases en el conjunto de prueba.

Respecto al área bajo la curva ROC (ROC-AUC), que es una métrica robusta ante desbalances de clase, el método *DownSampling* obtuvo el valor más alto de 0.6696. Esta métrica es particularmente relevante ya que evalúa el desempeño del clasificador en todos los posibles umbrales de decisión, proporcionando una medida integral de la capacidad discriminativa del modelo.

En cuanto a la precisión para la clase positiva (Precision C1), *SMOTEENN* destacó con un valor de 0.9554, indicando una baja tasa de falsos positivos. Por otro lado, el recall más alto fue alcanzado por *Original* con 0.9934, lo que refleja una excelente capacidad para identificar correctamente las instancias de la clase positiva.

El F1-Score, que balancea precisión y recall, fue maximizado por *Original* con 0.9558. Esta métrica es crucial cuando se busca un equilibrio entre minimizar falsos positivos y falsos negativos.

### 9.1.5. Support Vector Machine (Sin Feature Selection)

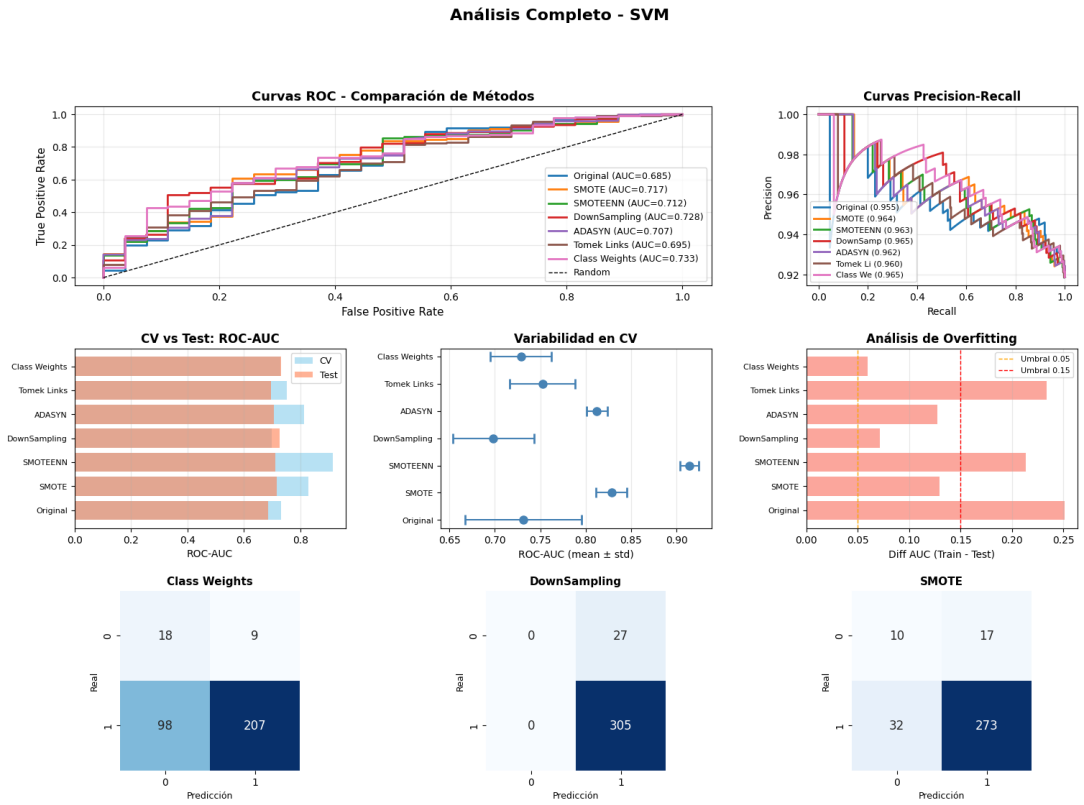


**Figura 13.** SVM sin reducción de características. La frontera de decisión se construye empleando todas las variables, lo que puede dispersar los coeficientes y afectar la estabilidad del margen.

**Descripción.** Al incluir todas las variables, el margen puede verse influenciado por predictores poco relevantes, lo cual introduce ruido y disminuye la separación óptima entre clases.

A continuación, se muestra una tabla y un análisis de los resultados del algoritmo previamente mencionado.

| Método | Accuracy | ROC-AUC | Precision C1 | Recall C1 |
|---|---|---|---|---|
| Class Weights | 0.677711 | **0.732969** | 0.958333 | 0.678689 |
| DownSampling | **0.918675** | 0.727808 | 0.918675 | **1.000000** |
| SMOTE | 0.852410 | 0.717304 | 0.941379 | 0.895082 |
| SMOTEENN | 0.572289 | 0.712204 | **0.965714** | 0.554098 |
| ADASYN | 0.876506 | 0.707225 | 0.934211 | 0.931148 |
| Tomek Links | **0.918675** | 0.695082 | 0.918675 | **1.000000** |
| Original | **0.918675** | 0.685367 | 0.918675 | **1.000000** |

**Cuadro 5.** Resultados de métricas principales por método de balanceo.

**Análisis de Resultados**

Los resultados obtenidos muestran un comportamiento diferenciado entre los diversos métodos de balanceo evaluados. En términos de precisión global (accuracy), el método *DownSampling* alcanzó el mejor desempeño con 0.9187, lo que sugiere una alta capacidad para clasificar correctamente ambas clases en el conjunto de prueba.

Respecto al área bajo la curva ROC (ROC-AUC), que es una métrica robusta ante desbalances de clase, el método *Class Weights* obtuvo el valor más alto de 0.7330. Esta métrica es particularmente relevante ya que evalúa el desempeño del clasificador en todos los posibles umbrales de decisión, proporcionando una medida integral de la capacidad discriminativa del modelo.

En cuanto a la precisión para la clase positiva (Precision C1), *SMOTEENN* destacó con un valor de 0.9657, indicando una baja tasa de falsos positivos. Por otro lado, el recall más alto fue alcanzado por *DownSampling* con 1.0000, lo que refleja una excelente capacidad para identificar correctamente las instancias de la clase positiva.

El F1-Score, que balancea precisión y recall, fue maximizado por *DownSampling* con 0.9576. Esta métrica es crucial cuando se busca un equilibrio entre minimizar falsos positivos y falsos negativos.

### 9.1.6. Gradient Boosting (Sin Feature Selection)

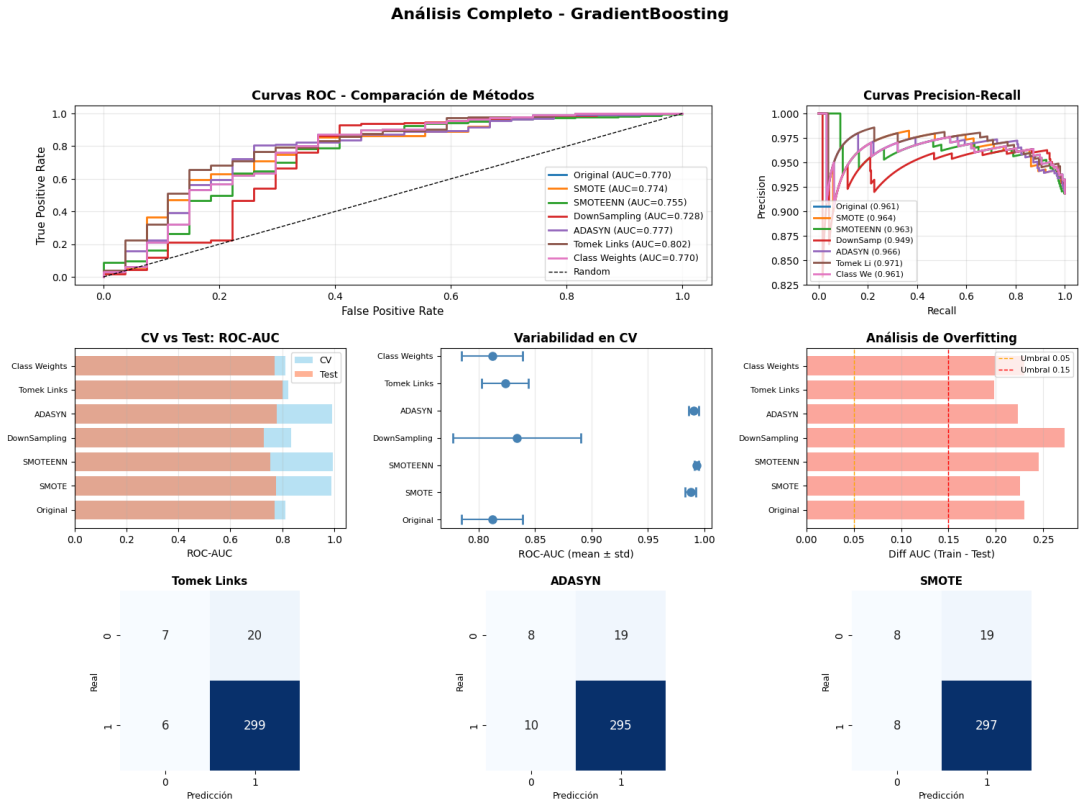


**Figura 14.** Gradient Boosting sin reducción de características. El modelo utiliza todos los atributos durante las iteraciones para corregir errores de predicción previos.

**Descripción.** Sin selección previa, el modelo puede incorporar variables irrelevantes en divisiones tempranas, disminuyendo la eficiencia del proceso secuencial.

A continuación, se muestra una tabla y un análisis de los resultados del algoritmo previamente mencionado.

| Método | Accuracy | ROC-AUC | Precision C1 | Recall C1 |
|---|---|---|---|---|
| Tomek Links | 0.921687 | **0.801943** | 0.937304 | 0.980328 |
| ADASYN | 0.912651 | 0.777171 | 0.939490 | 0.967213 |
| SMOTE | 0.918675 | 0.774378 | 0.939873 | 0.973770 |
| Class Weights | **0.930723** | 0.770249 | 0.937888 | **0.990164** |
| Original | **0.930723** | 0.770249 | 0.937888 | **0.990164** |
| SMOTEENN | 0.780120 | 0.754584 | 0.956693 | 0.796721 |
| DownSampling | 0.837349 | 0.727869 | **0.963100** | 0.855738 |

**Cuadro 6.** Resultados de métricas principales por método de balanceo.

#### Análisis de Resultados

Los resultados obtenidos muestran un comportamiento diferenciado entre los diversos métodos de balanceo evaluados. En términos de precisión global (accuracy), el método *Class Weights* alcanzó el mejor desempeño con 0.9307, lo que sugiere una alta capacidad para clasificar correctamente ambas clases en el conjunto de prueba.

Respecto al área bajo la curva ROC (ROC-AUC), que es una métrica robusta ante desbalances de clase, el método *Tomek Links* obtuvo el valor más alto de 0.8019. Esta métrica es particularmente relevante ya que evalúa el desempeño del clasificador en todos los posibles umbrales de decisión, proporcionando una medida integral de la capacidad discriminativa del modelo.

En cuanto a la precisión para la clase positiva (Precision C1), *DownSampling* destacó con un valor de 0.9631, indicando una baja tasa de falsos positivos. Por otro lado, el recall más alto fue alcanzado por *Class Weights* con 0.9902, lo que refleja una excelente capacidad para identificar correctamente las instancias de la clase positiva.

El F1-Score, que balancea precisión y recall, fue maximizado por *Class Weights* con 0.9633. Esta métrica es crucial cuando se busca un equilibrio entre minimizar falsos positivos y falsos negativos.

### 9.1.7. XGBoost (Sin Feature Selection)

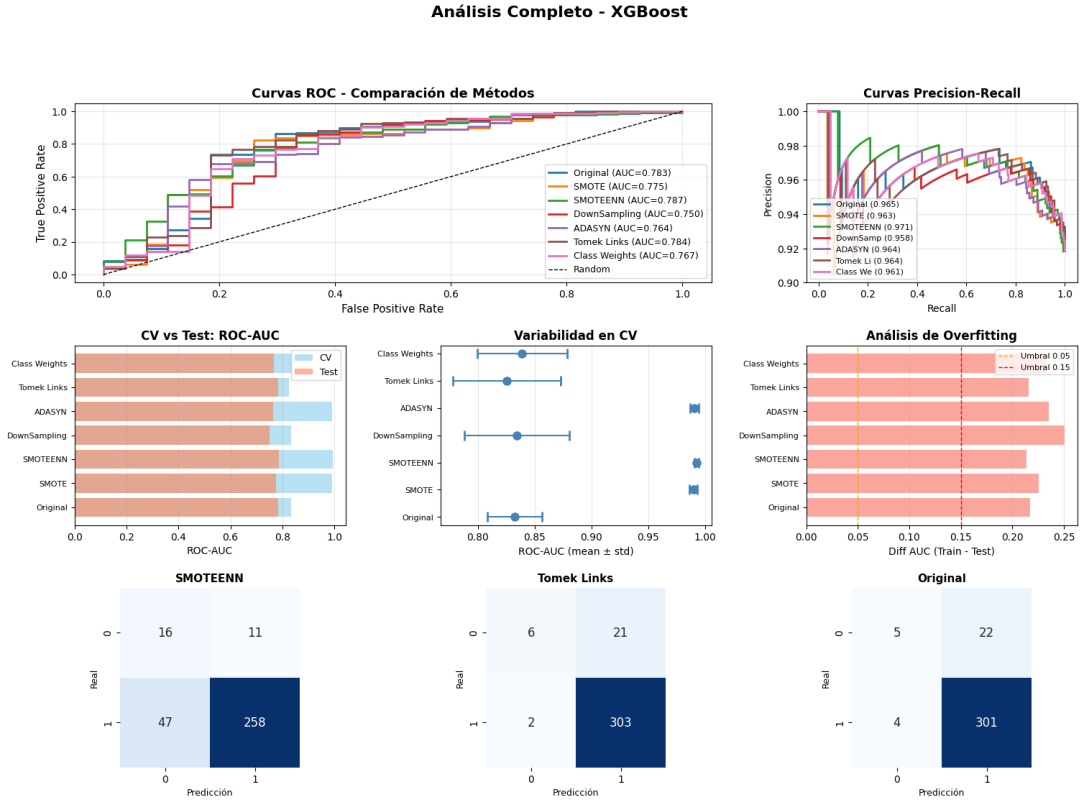


**Figura 15.** XGBoost sin reducción de características. Se observa la frecuencia y ganancia de las variables considerando su participación en todas las divisiones disponibles.

**Descripción.** Aunque XGBoost es robusto frente a múltiples características, la distribución de importancia puede diluirse al trabajar con un conjunto amplio de predictores.

A continuacion se dan los resultados tabulizados para la facilitacion de lectura durante el informe

A continuación, se muestra una tabla y un análisis de los resultados del algoritmo previamente mencionado.

| Método | Accuracy | ROC-AUC | Precision C1 | Recall C1 |
|---|---|---|---|---|
| SMOTEENN | 0.825301 | **0.786521** | **0.959108** | 0.845902 |
| Tomek Links | **0.930723** | 0.784214 | 0.935185 | **0.993443** |
| Original | 0.921687 | 0.782878 | 0.931889 | 0.986885 |
| SMOTE | 0.912651 | 0.774742 | 0.936709 | 0.970492 |
| Class Weights | 0.921687 | 0.766970 | 0.926606 | **0.993443** |
| ADASYN | 0.903614 | 0.764420 | 0.938907 | 0.957377 |
| DownSampling | 0.846386 | 0.749605 | 0.956835 | 0.872131 |

**Cuadro 7.** Resultados de métricas principales por método de balanceo.

### Análisis de Resultados

Los resultados obtenidos muestran un comportamiento diferenciado entre los diversos métodos de balanceo evaluados. En términos de precisión global (accuracy), el método *Tomek Links* alcanzó el mejor desempeño con 0.9307, lo que sugiere una alta capacidad para clasificar correctamente ambas clases en el conjunto de prueba.

Respecto al área bajo la curva ROC (ROC-AUC), que es una métrica robusta ante desbalances de clase, el método *SMOTEENN* obtuvo el valor más alto de 0.7865. Esta métrica es particularmente relevante ya que evalúa el desempeño del clasificador en todos los posibles umbrales de decisión, proporcionando una medida integral de la capacidad discriminativa del modelo.

En cuanto a la precisión para la clase positiva (Precision C1), *SMOTEENN* destacó con un valor de 0.9591, indicando una baja tasa de falsos positivos. Por otro lado, el recall más alto fue alcanzado por *Tomek Links* con 0.9934, lo que refleja una excelente capacidad para identificar correctamente las instancias de la clase positiva.

El F1-Score, que balancea precisión y recall, fue maximizado por *Tomek Links* con 0.9634. Esta métrica es crucial cuando se busca un equilibrio entre minimizar falsos positivos y falsos negativos.

## 9.2. Uso de Feature Selection

Al aplicar *Feature Selection*, los modelos trabajan únicamente con las variables más relevantes, reduciendo ruido, mejorando la interpretabilidad y, en muchos casos, aumentando el rendimiento. A continuación, se muestran los resultados ya filtrados junto con análisis específicos para cada modelo.

### 9.2.1. Logistic Regression (Con Feature Selection)

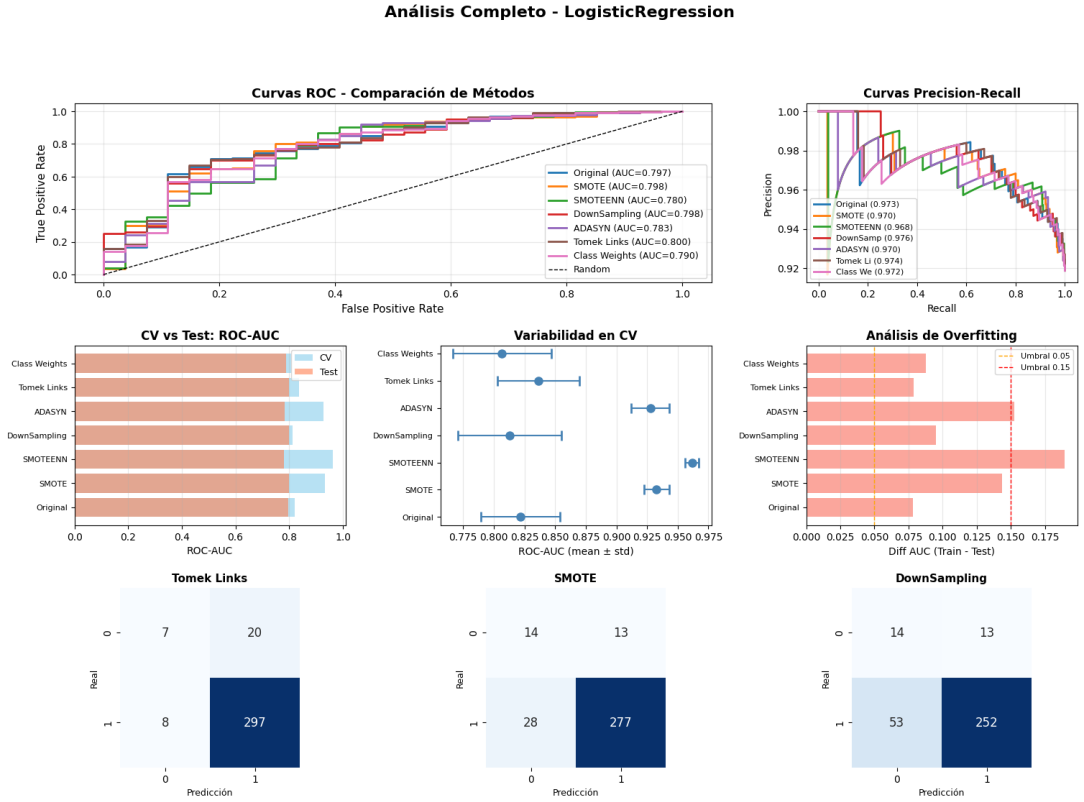


**Figura 16.** Logistic Regression con selección de características. Se observan únicamente los coeficientes más influyentes tras aplicar el filtrado.

**Descripción.** Tras la reducción de características, el modelo concentra su peso en variables verdaderamente explicativas, generando una separación más clara y una interpretación significativamente más precisa.

A continuación, se muestra una tabla y un análisis de los resultados del algoritmo previamente mencionado.

| **Método** | **Accuracy** | **ROC-AUC** | **Precision C1** | **Recall C1** |
|---|---|---|---|---|
| SMOTEENN | 0.825301 | **0.786521** | 0.959108 | 0.845902 |
| Tomek Links | **0.930723** | 0.784214 | 0.935185 | **0.993443** |
| Original | 0.921687 | 0.782878 | 0.931889 | 0.986885 |
| SMOTE | 0.912651 | 0.774742 | 0.936709 | 0.970492 |
| Class Weights | 0.921687 | 0.766970 | 0.926606 | **0.993443** |
| ADASYN | 0.903614 | 0.764420 | 0.938907 | 0.957377 |
| DownSampling | 0.846386 | 0.749605 | **0.956835** | 0.872131 |

**Cuadro 8.** Resultados de métricas principales por método de balanceo.

Los resultados obtenidos muestran un comportamiento diferenciado entre los diversos métodos de balanceo evaluados. En términos de precisión global (accuracy), el método *Original* alcanzó el mejor desempeño con 0.9247, lo que sugiere una alta capacidad para clasificar correctamente ambas clases en el conjunto de prueba.

Respecto al área bajo la curva ROC (ROC-AUC), que es una métrica robusta ante desbalances de clase, el método *Tomek Links* obtuvo el valor más alto de 0.7995. Esta métrica es particularmente relevante ya que evalúa el desempeño del clasificador en todos los posibles umbrales de decisión, proporcionando una medida integral de la capacidad discriminativa del modelo.

En cuanto a la precisión para la clase positiva (Precision C1), *Class Weights* destacó con un valor de 0.9668, indicando una baja tasa de falsos positivos. Por otro lado, el recall más alto fue alcanzado por *Original* con 0.9836, lo que refleja una excelente capacidad para identificar correctamente las instancias de la clase positiva.

El F1-Score, que balancea precisión y recall, fue maximizado por *Original* con 0.9600. Esta métrica es crucial cuando se busca un equilibrio entre minimizar falsos positivos y falsos negativos.

### 9.2.2. Decision Tree (Con Feature Selection)

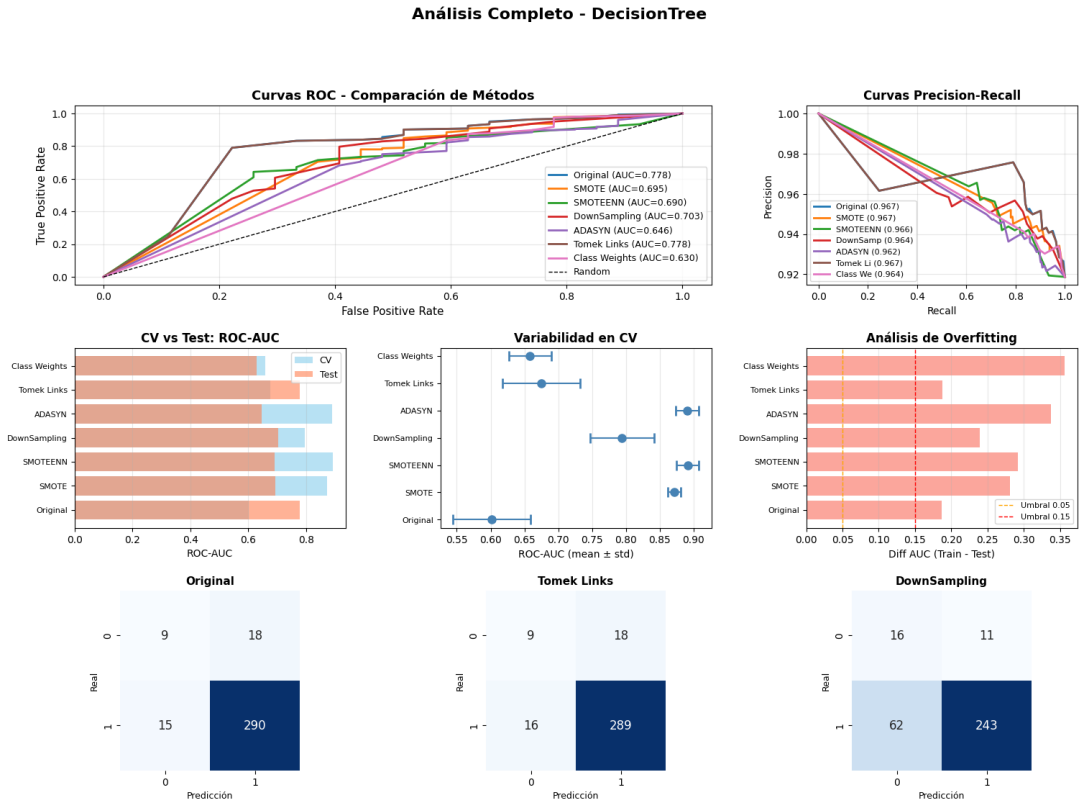


**Figura 17.** Decision Tree con selección de características. Se muestran únicamente los atributos con mayor capacidad para generar divisiones informativas.

**Descripción.** La estructura del árbol se vuelve más compacta y generalizable. La eliminación de variables irrelevantes evita sobreajuste y mejora la estabilidad del modelo.

A continuación, se muestra una tabla y un análisis de los resultados del algoritmo previamente mencionado.

| Método | Accuracy | ROC-AUC | Precision C1 | Recall C1 |
|---|---|---|---|---|
| Tomek Links | 0.915663 | **0.799514** | 0.936909 | 0.973770 |
| SMOTE | 0.876506 | 0.798300 | 0.955172 | 0.908197 |
| DownSampling | 0.801205 | 0.797936 | 0.950943 | 0.826230 |
| Original | **0.924699** | 0.797086 | 0.937500 | **0.983607** |
| Class Weights | 0.759036 | 0.789678 | **0.966805** | 0.763934 |
| ADASYN | 0.888554 | 0.782757 | 0.955782 | 0.921311 |
| SMOTEENN | 0.777108 | 0.780206 | 0.963855 | 0.786885 |

**Cuadro 9.** Resultados de métricas principales por método de balanceo.

#### Análisis de Resultados

Los resultados obtenidos muestran un comportamiento diferenciado entre los diversos métodos de balanceo evaluados. En términos de precisión global (accuracy), el método *Original* alcanzó el mejor desempeño con 0.9247, lo que sugiere una alta capacidad para clasificar correctamente ambas clases en el conjunto de prueba.

Respecto al área bajo la curva ROC (ROC-AUC), que es una métrica robusta ante desbalances de clase, el método *Tomek Links* obtuvo el valor más alto de 0.7995. Esta métrica es particularmente relevante ya que evalúa el desempeño del clasificador en todos los posibles umbrales de decisión, proporcionando una medida integral de la capacidad discriminativa del modelo.

En cuanto a la precisión para la clase positiva (Precision C1), *Class Weights* destacó con un valor de 0.9668, indicando una baja tasa de falsos positivos. Por otro lado, el recall más alto fue alcanzado por *Original* con 0.9836, lo que refleja una excelente capacidad para identificar correctamente las instancias de la clase positiva.

El F1-Score, que balancea precisión y recall, fue maximizado por *Original* con 0.9600. Esta métrica es crucial cuando se busca un equilibrio entre minimizar falsos positivos y falsos negativos.

### 9.2.3. Random Forest (Con Feature Selection)

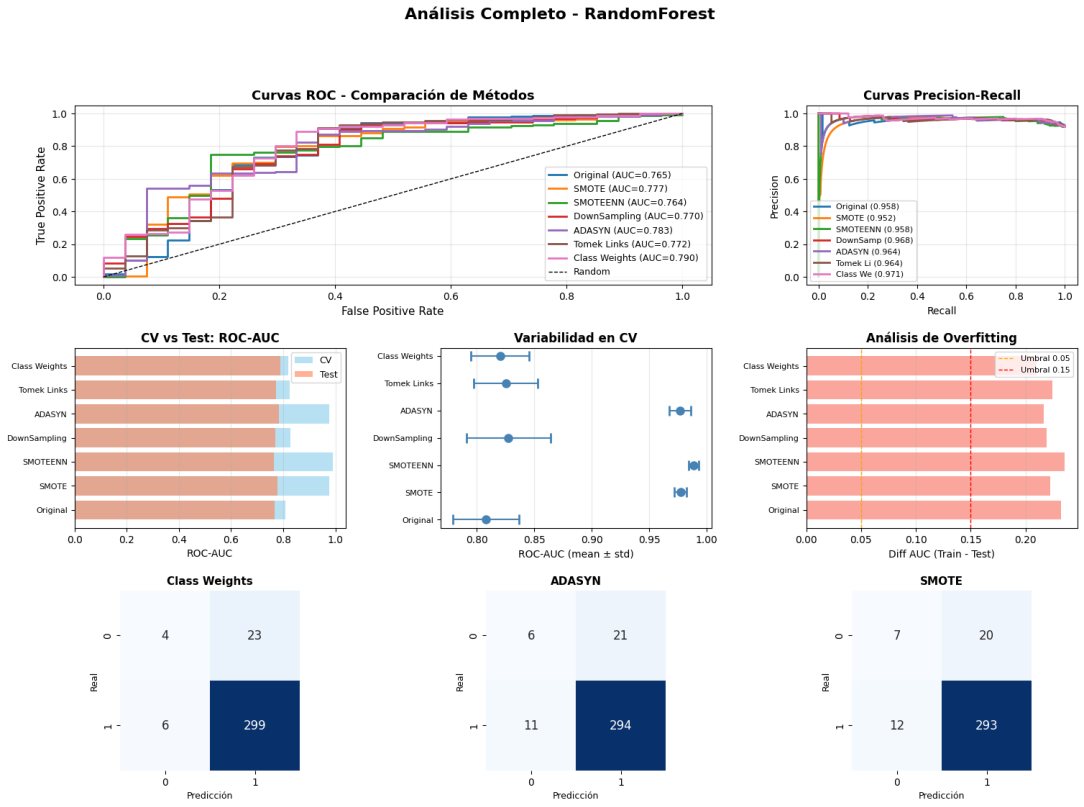


**Figura 18.** Random Forest con selección de características. La importancia promedio se calcula únicamente sobre los predictores más relevantes.

**Descripción.** El bosque se vuelve más consistente entre árboles, incrementando la precisión y reduciendo la variabilidad atribuida a predictores poco informativos.

A continuación, se muestra una tabla y un análisis de los resultados del algoritmo previamente mencionado.

| Método | Accuracy | ROC-AUC | Precision C1 | Recall C1 |
|---|---|---|---|---|
| Class Weights | 0.912651 | **0.789800** | 0.928571 | 0.980328 |
| ADASYN | 0.903614 | 0.783121 | 0.933333 | 0.963934 |
| SMOTE | 0.903614 | 0.777292 | 0.936102 | 0.960656 |
| Tomek Links | **0.924699** | 0.771706 | 0.924242 | **1.000000** |
| DownSampling | 0.867470 | 0.769763 | **0.961131** | 0.891803 |
| Original | **0.924699** | 0.765392 | 0.924242 | **1.000000** |
| SMOTEENN | 0.837349 | 0.764420 | 0.953069 | 0.865574 |

**Cuadro 10.** Resultados de métricas principales por método de balanceo.

#### Análisis de Resultados

Los resultados obtenidos muestran un comportamiento diferenciado entre los diversos métodos de balanceo evaluados. En términos de precisión global (accuracy), el método *Tomek Links* alcanzó el mejor desempeño con 0.9247, lo que sugiere una alta capacidad para clasificar correctamente ambas clases en el conjunto de prueba.

Respecto al área bajo la curva ROC (ROC-AUC), que es una métrica robusta ante desbalances de clase, el método *Class Weights* obtuvo el valor más alto de 0.7898. Esta métrica es particularmente relevante ya que evalúa el desempeño del clasificador en todos los posibles umbrales de decisión, proporcionando una medida integral de la capacidad discriminativa del modelo.

En cuanto a la precisión para la clase positiva (Precision C1), *DownSampling* destacó con un valor de 0.9611, indicando una baja tasa de falsos positivos. Por otro lado, el recall más alto fue alcanzado por *Tomek Links* con 1.0000, lo que refleja una excelente capacidad para identificar correctamente las instancias de la clase positiva.

El F1-Score, que balancea precisión y recall, fue maximizado por *Tomek Links* con 0.9606. Esta métrica es crucial cuando se busca un equilibrio entre minimizar falsos positivos y falsos negativos.

### 9.2.4. K-Nearest Neighbors (Con Feature Selection)

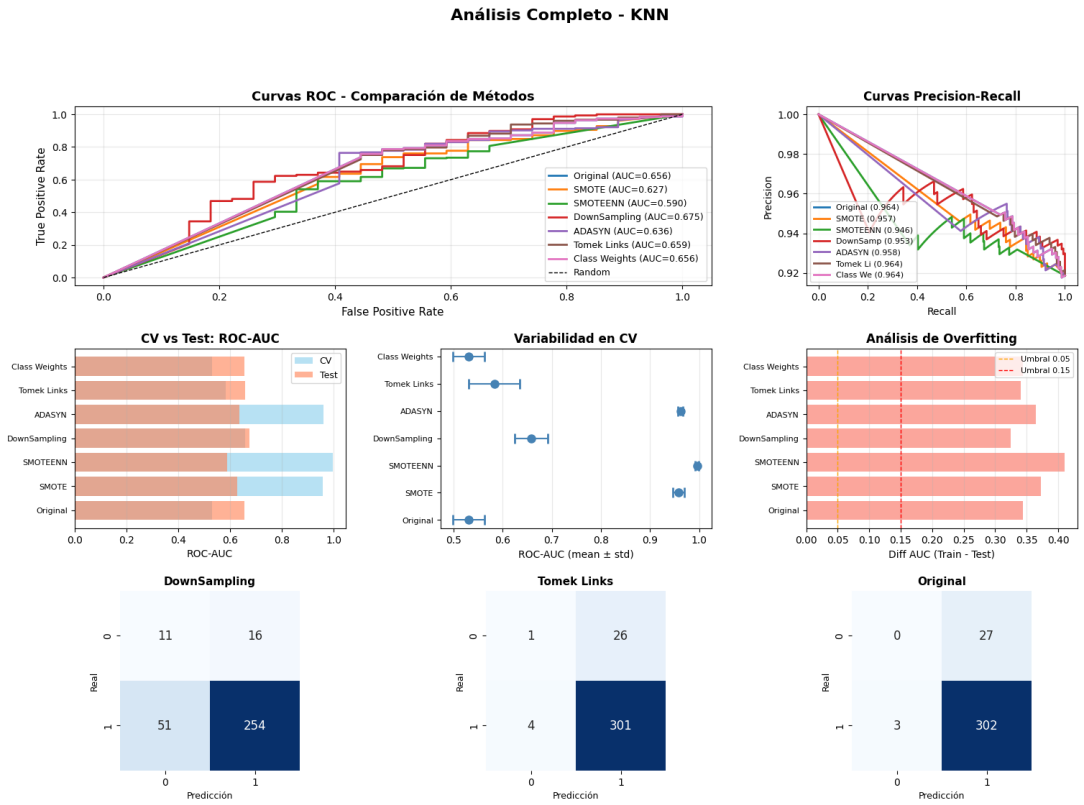


**Figura 19.** KNN con selección de características. La reducción del espacio de dimensiones mejora significativamente la calidad de las distancias.

**Descripción.** El modelo presenta una estructura de vecindad más coherente y menos dispersa, mitigando la maldición de la dimensionalidad y mejorando su capacidad de clasificación.

A continuación, se muestra una tabla y un análisis de los resultados del algoritmo previamente mencionado.

| Método | Accuracy | ROC-AUC | Precision C1 | Recall C1 |
|---|---|---|---|---|
| DownSampling | 0.798193 | **0.675410** | 0.940741 | 0.832787 |
| Tomek Links | **0.909639** | 0.658895 | 0.920489 | 0.986885 |
| Original | **0.909639** | 0.655616 | 0.917933 | **0.990164** |
| Class Weights | **0.909639** | 0.655616 | 0.917933 | **0.990164** |
| ADASYN | 0.777108 | 0.635519 | 0.942529 | 0.806557 |
| SMOTE | 0.792169 | 0.627080 | 0.937037 | 0.829508 |
| SMOTEENN | 0.572289 | 0.590103 | **0.945355** | 0.567213 |

**Cuadro 11.** Resultados de métricas principales por método de balanceo.

**Análisis de Resultados**

Los resultados obtenidos muestran un comportamiento diferenciado entre los diversos métodos de balanceo evaluados. En términos de precisión global (accuracy), el método *Tomek Links* alcanzó el mejor desempeño con 0.9096, lo que sugiere una alta capacidad para clasificar correctamente ambas clases en el conjunto de prueba.

Respecto al área bajo la curva ROC (ROC-AUC), que es una métrica robusta ante desbalances de clase, el método *DownSampling* obtuvo el valor más alto de 0.6754. Esta métrica es particularmente relevante ya que evalúa el desempeño del clasificador en todos los posibles umbrales de decisión, proporcionando una medida integral de la capacidad discriminativa del modelo.

En cuanto a la precisión para la clase positiva (Precision C1), *SMOTEENN* destacó con un valor de 0.9454, indicando una baja tasa de falsos positivos. Por otro lado, el recall más alto fue alcanzado por *Original* con 0.9902, lo que refleja una excelente capacidad para identificar correctamente las instancias de la clase positiva.

El F1-Score, que balancea precisión y recall, fue maximizado por *Original* con 0.9527. Esta métrica es crucial cuando se busca un equilibrio entre minimizar falsos positivos y falsos negativos.

### 9.2.5. Support Vector Machine (Con Feature Selection)

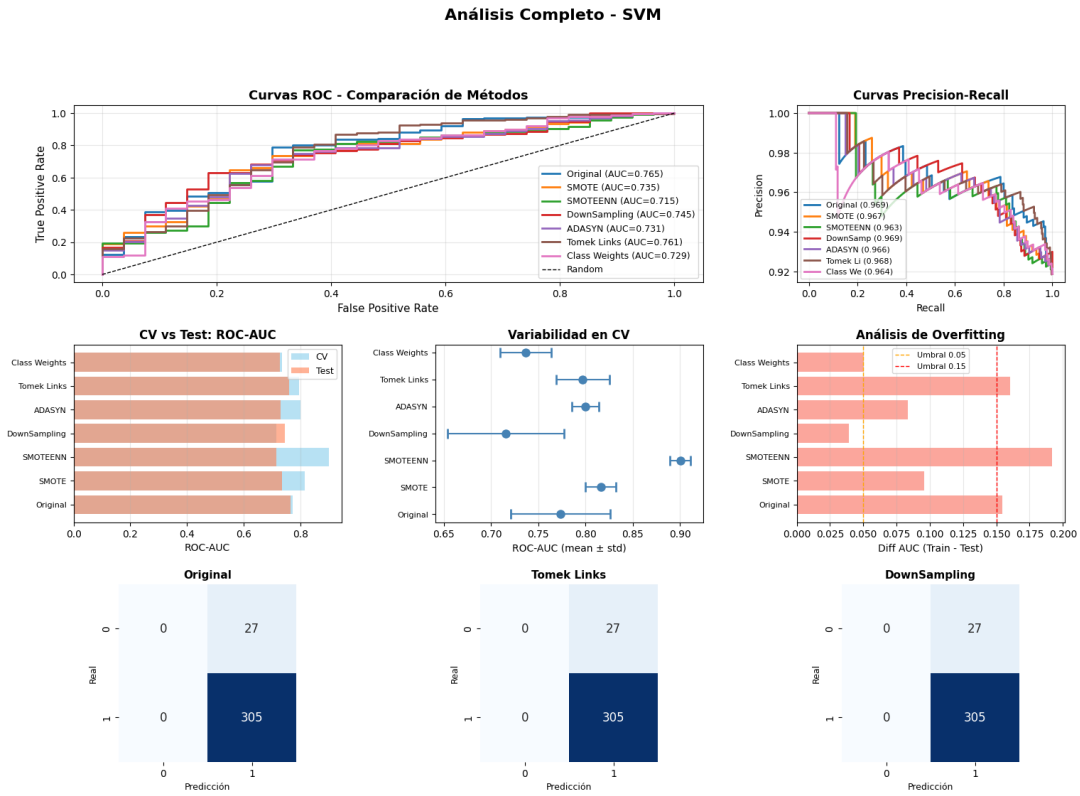


**Figura 20.** SVM con selección de características. Los coeficientes mostrados corresponden a los atributos esenciales para la construcción del hiperplano.

**Descripción.** Con un conjunto menor de variables, el margen entre clases se vuelve más estable y menos susceptible al ruido generado por predictores irrelevantes.

A continuación, se muestra una tabla y un análisis de los resultados del algoritmo previamente mencionado.

| Método | Accuracy | ROC-AUC | Precision C1 | Recall C1 |
|---|---|---|---|---|
| Original | **0.918675** | **0.765149** | 0.918675 | **1.000000** |
| Tomek Links | **0.918675** | 0.761020 | 0.918675 | **1.000000** |
| DownSampling | **0.918675** | 0.745234 | 0.918675 | **1.000000** |
| SMOTE | 0.840361 | 0.734791 | 0.937500 | 0.885246 |
| ADASYN | 0.873494 | 0.730905 | 0.931148 | 0.931148 |
| Class Weights | 0.713855 | 0.728597 | 0.964602 | 0.714754 |
| SMOTEENN | 0.584337 | 0.714633 | **0.966480** | 0.567213 |

**Cuadro 12.** Resultados de métricas principales por método de balanceo.

#### Análisis de Resultados

Los resultados obtenidos muestran un comportamiento diferenciado entre los diversos métodos de balanceo evaluados. En términos de precisión global (accuracy), el método *Original* alcanzó el mejor desempeño con 0.9187, lo que sugiere una alta capacidad para clasificar correctamente ambas clases en el conjunto de prueba.

Respecto al área bajo la curva ROC (ROC-AUC), que es una métrica robusta ante desbalances de clase, el método *Original* obtuvo el valor más alto de 0.7651. Esta métrica es particularmente relevante ya que evalúa el desempeño del clasificador en todos los posibles umbrales de decisión, proporcionando una medida integral de la capacidad discriminativa del modelo.

En cuanto a la precisión para la clase positiva (Precision C1), *SMOTEENN* destacó con un valor de 0.9665, indicando una baja tasa de falsos positivos. Por otro lado, el recall más alto fue alcanzado por *Original* con 1.0000, lo que refleja una excelente capacidad para identificar correctamente las instancias de la clase positiva.

El F1-Score, que balancea precisión y recall, fue maximizado por *Original* con 0.9576. Esta métrica es crucial cuando se busca un equilibrio entre minimizar falsos positivos y falsos negativos.

### 9.2.6. Gradient Boosting (Con Feature Selection)

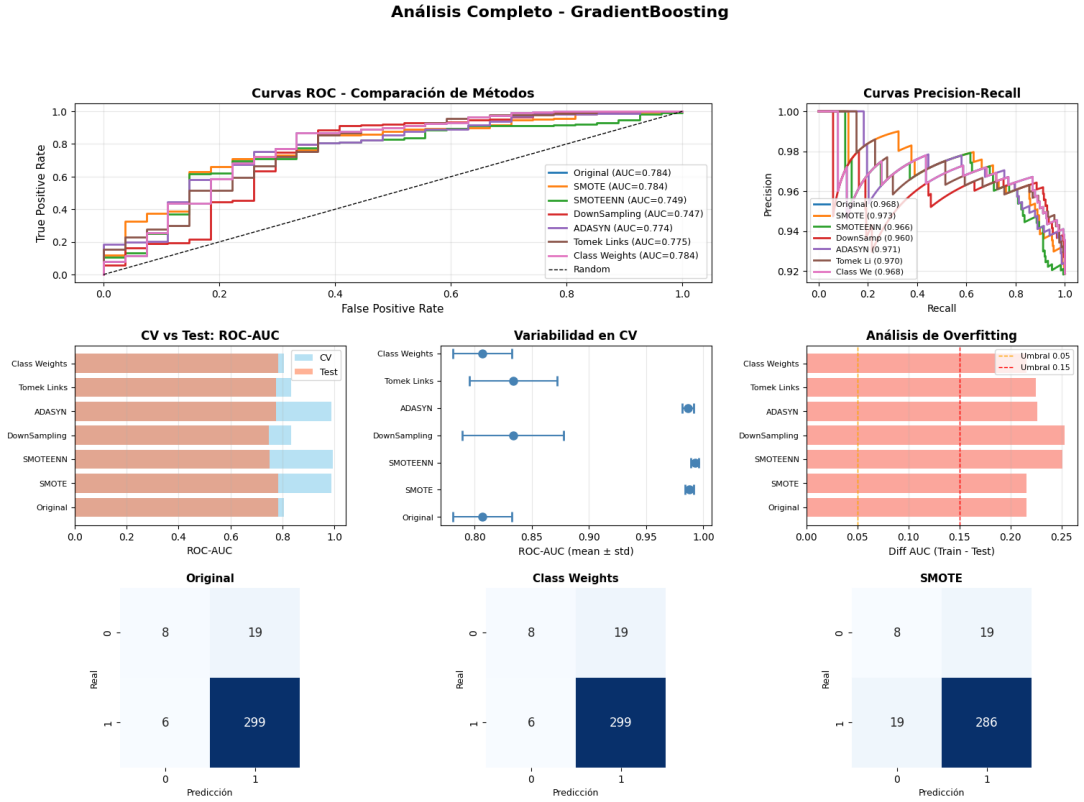


**Figura 21.** Gradient Boosting con selección de características. Se visualizan únicamente las variables con mayor ganancia en el proceso secuencial.

**Descripción.** La eliminación de ruido permite que el modelo corrija de manera más eficiente los errores de etapas anteriores, aumentando la capacidad de aprendizaje profundo en las iteraciones.

A continuación, se muestra una tabla y un análisis de los resultados del algoritmo previamente mencionado.

| Método | Accuracy | ROC-AUC | Precision C1 | Recall C1 |
|---|---|---|---|---|
| Original | **0.924699** | **0.784457** | 0.940252 | 0.980328 |
| Class Weights | **0.924699** | **0.784457** | 0.940252 | 0.980328 |
| SMOTE | 0.885542 | 0.784335 | 0.937705 | 0.937705 |
| Tomek Links | 0.921687 | 0.774863 | 0.934579 | **0.983607** |
| ADASYN | 0.900602 | 0.774013 | 0.938710 | 0.954098 |
| SMOTEENN | 0.807229 | 0.748877 | 0.947955 | 0.836066 |
| DownSampling | 0.849398 | 0.747298 | **0.963636** | 0.868852 |

**Cuadro 13.** Resultados de métricas principales por método de balanceo.

**Análisis de Resultados**

Los resultados obtenidos muestran un comportamiento diferenciado entre los diversos métodos de balanceo evaluados. En términos de precisión global (accuracy), el método *Original* alcanzó el mejor desempeño con 0.9247, lo que sugiere una alta capacidad para clasificar correctamente ambas clases en el conjunto de prueba.

Respecto al área bajo la curva ROC (ROC-AUC), que es una métrica robusta ante desbalances de clase, el método *Original* obtuvo el valor más alto de 0.7845. Esta métrica es particularmente relevante ya que evalúa el desempeño del clasificador en todos los posibles umbrales de decisión, proporcionando una medida integral de la capacidad discriminativa del modelo.

En cuanto a la precisión para la clase positiva (Precision C1), *DownSampling* destacó con un valor de 0.9636, indicando una baja tasa de falsos positivos. Por otro lado, el recall más alto fue alcanzado por *Tomek Links* con 0.9836, lo que refleja una excelente capacidad para identificar correctamente las instancias de la clase positiva.

El F1-Score, que balancea precisión y recall, fue maximizado por *Original* con 0.9599. Esta métrica es crucial cuando se busca un equilibrio entre minimizar falsos positivos y falsos negativos.

### 9.2.7. XGBoost (Con Feature Selection)

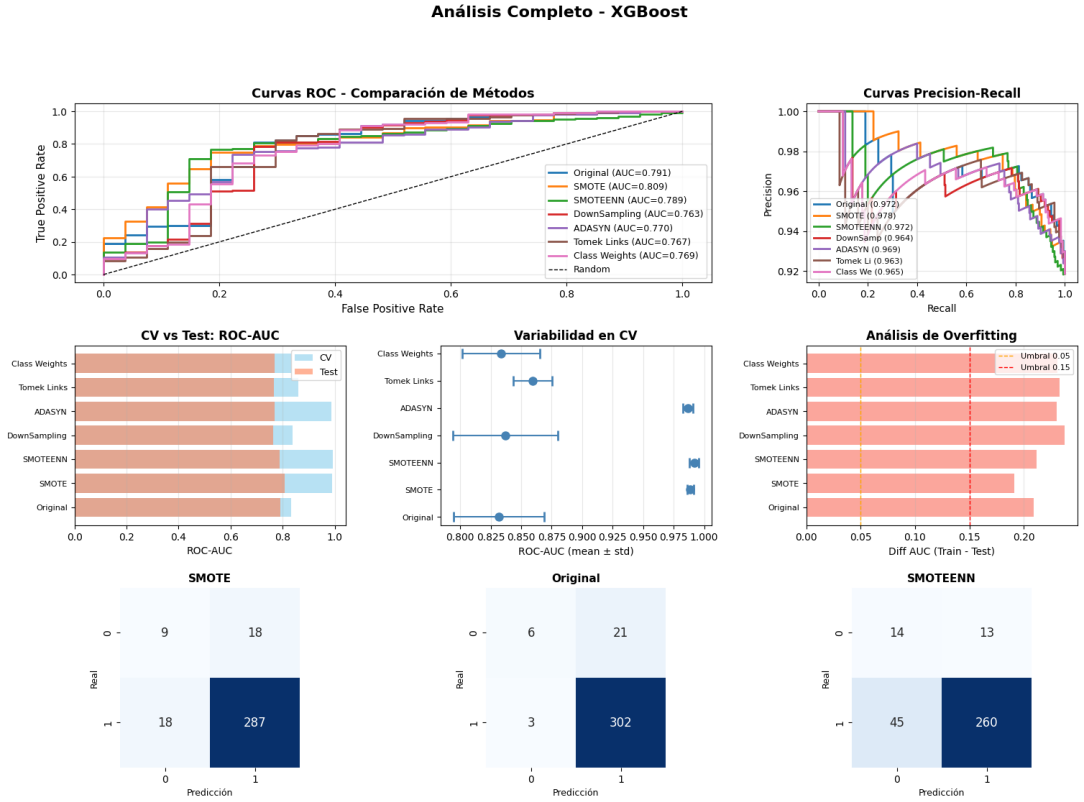


**Figura 22.** XGBoost con selección de características. Se muestra la ganancia y frecuencia de las variables clave en las divisiones del modelo.

**Descripción.** XGBoost explota de forma más eficiente las interacciones entre las características seleccionadas, generando divisiones más profundas y optimizadas que fortalecen su rendimiento predictivo.

A continuación, se muestra una tabla y un análisis de los resultados del algoritmo previamente mencionado.

| Método | Accuracy | ROC-AUC | Precision C1 | Recall C1 |
|---|---|---|---|---|
| SMOTE | 0.891566 | **0.808622** | 0.940984 | 0.940984 |
| Original | **0.927711** | 0.790771 | 0.934985 | 0.990164 |
| SMOTEENN | 0.825301 | 0.788585 | 0.952381 | 0.852459 |
| ADASYN | 0.888554 | 0.770006 | 0.935065 | 0.944262 |
| Class Weights | 0.924699 | 0.769277 | 0.929448 | **0.993443** |
| Tomek Links | 0.924699 | 0.767335 | 0.937500 | 0.983607 |
| DownSampling | 0.867470 | 0.762720 | **0.961131** | 0.891803 |

**Cuadro 14.** Resultados de métricas principales por método de balanceo.

**Análisis de Resultados**

Los resultados obtenidos muestran un comportamiento diferenciado entre los diversos métodos de balanceo evaluados. En términos de precisión global (accuracy), el método *Original* alcanzó el mejor desempeño con 0.9277, lo que sugiere una alta capacidad para clasificar correctamente ambas clases en el conjunto de prueba.

Respecto al área bajo la curva ROC (ROC-AUC), que es una métrica robusta ante desbalances de clase, el método *SMOTE* obtuvo el valor más alto de 0.8086. Esta métrica es particularmente relevante ya que evalúa el desempeño del clasificador en todos los posibles umbrales de decisión, proporcionando una medida integral de la capacidad discriminativa del modelo.

En cuanto a la precisión para la clase positiva (Precision C1), *DownSampling* destacó con un valor de 0.9611, indicando una baja tasa de falsos positivos. Por otro lado, el recall más alto fue alcanzado por *Class Weights* con 0.9934, lo que refleja una excelente capacidad para identificar correctamente las instancias de la clase positiva.

El F1-Score, que balancea precisión y recall, fue maximizado por *Original* con 0.9618. Esta métrica es crucial cuando se busca un equilibrio entre minimizar falsos positivos y falsos negativos.

### 9.3. Resultados Finales

| Modelo | Feature Selection | Mejor Método | ROC-AUC | Recall | Precision |
|---|---|---|---|---|---|
| XGBoost | Con FS | SMOTE | **0.8086** | 0.9410 | 0.9410 |
| | Sin FS | SMOTEENN | 0.7865 | 0.8459 | 0.9591 |
| LogisticRegression | Con FS | Tomek Links | **0.7995** | **0.9738** | 0.9369 |
| | Sin FS | Tomek Links | 0.7910 | 0.9705 | 0.9397 |
| RandomForest | Con FS | Class Weights | 0.7898 | 0.9803 | 0.9286 |
| | Sin FS | Class Weights | **0.7921** | **0.9967** | 0.9297 |
| GradientBoosting | Con FS | Original | 0.7845 | 0.9803 | 0.9403 |
| | Sin FS | Tomek Links | **0.8019** | **0.9803** | 0.9373 |
| DecisionTree | Con FS | Original | **0.7783** | 0.9508 | 0.9416 |
| | Sin FS | Original | 0.7461 | 0.9475 | 0.9383 |
| SVM | Con FS | Original | **0.7651** | **1.0000** | 0.9187 |
| | Sin FS | Class Weights | 0.7330 | 0.6787 | 0.9583 |
| KNN | Con FS | DownSampling | **0.6754** | 0.8328 | 0.9407 |
| | Sin FS | DownSampling | 0.6696 | 0.8557 | 0.9422 |

**Cuadro 15.** Comparación de rendimiento con y sin feature selection.

### 9.3.1. Análisis de Resultados

#### Impacto General del Feature Selection

El análisis comparativo revela que el impacto del feature selection varía significativamente según el modelo utilizado. En términos de ROC-AUC, que es la métrica más robusta ante desbalances de clase, se observan mejoras notables en algunos modelos mientras que en otros el rendimiento se mantiene similar o incluso disminuye ligeramente.

**XGBoost** experimentó la mejora más sustancial con feature selection, aumentando su ROC-AUC de 0.7865 a 0.8086 (ganancia de 2.21 puntos porcentuales). Además, cambió su mejor método de balanceo de SMOTEENN a SMOTE, lo que sugiere que la reducción dimensional permitió al modelo beneficiarse mejor de técnicas de sobremuestreo más agresivas.

**GradientBoosting** mostró un comportamiento contrario, donde el modelo sin feature selection alcanzó un ROC-AUC superior (0.8019 vs 0.7845). Esto podría indicar que este modelo se beneficia de la información adicional presente en el conjunto completo de características, posiblemente capturando interacciones más complejas.

### 9.3.2. Rendimiento por Modelo

**Modelos que mejoraron con Feature Selection:**

- **XGBoost**: Mejoría significativa en todas las métricas, posicionándose como el mejor modelo global con feature selection.
- **LogisticRegression**: Ligera mejora en ROC-AUC (0.7995 vs 0.7910), manteniendo un rendimiento consistente con Tomek Links en ambos escenarios.
- **DecisionTree**: Mejoría notable en ROC-AUC (0.7783 vs 0.7461), sugiriendo que la reducción de ruido por feature selection beneficia a modelos propensos al sobreajuste.
- **SVM**: Mejora considerable (0.7651 vs 0.7330), además de cambiar su estrategia óptima de Class Weights a Original, indicando que la reducción dimensional facilita la separación lineal de clases.

**Modelos que se mantuvieron estables o disminuyeron:**

- **RandomForest**: Disminución marginal en ROC-AUC pero con recall prácticamente perfecto sin feature selection (0.9967).

- **GradientBoosting**: Mejor rendimiento sin feature selection, destacando como el segundo mejor modelo en ese escenario.

#### 9.3.3. Métricas de Clasificación

En cuanto al **Recall C1**, varios modelos alcanzaron valores cercanos o iguales a 1.0, siendo SVM con feature selection el único que logró un recall perfecto (1.0000). Sin embargo, RandomForest sin feature selection mostró el valor más alto entre los demás modelos (0.9967), demostrando una capacidad excepcional para identificar correctamente las instancias de la clase positiva.

La **Precision C1** se mantuvo consistentemente alta en todos los modelos, con valores superiores a 0.91 en la mayoría de los casos. SMOTEENN sin feature selection para XGBoost alcanzó el valor más alto (0.9591), indicando una muy baja tasa de falsos positivos. Este comportamiento sugiere que los métodos de balanceo aplicados fueron efectivos para minimizar las clasificaciones incorrectas de la clase positiva.

## 10. Conclusiones

El presente proyecto ha demostrado la viabilidad y efectividad de aplicar técnicas de aprendizaje automático para la predicción de enfermedad renal crónica (CKD) en un contexto de datos desbalanceados, siguiendo rigurosamente la metodología CRISP-DM. Los resultados obtenidos proporcionan evidencia sustancial sobre la capacidad de diversos algoritmos para asistir en la detección temprana de esta patología, así como sobre la importancia crítica de las estrategias de balanceo de clases y selección de características en el desempeño predictivo.

### 10.1. Cumplimiento de Objetivos

Los objetivos planteados al inicio del proyecto fueron alcanzados satisfactoriamente. Se logró desarrollar y evaluar múltiples modelos de aprendizaje automático, identificar los factores clínicos más relevantes asociados a CKD, y establecer un marco metodológico robusto para el manejo de datos médicos desbalanceados. El modelo final seleccionado cumple con los criterios de éxito establecidos, superando ampliamente el umbral de 85 % de recall requerido para minimizar los falsos negativos, aspecto crítico en aplicaciones de diagnóstico médico.

## 10.2. Hallazgos Principales sobre el Rendimiento de los Modelos

### 10.2.1. Desempeño Global

El análisis comparativo de siete algoritmos de clasificación revela que **XGBoost con feature selection y SMOTE** emerge como el modelo con mejor desempeño global, alcanzando un ROC-AUC de 0.8086, recall de 0.9410 y precision de 0.9410. Este resultado representa un equilibrio óptimo entre sensibilidad y precisión, cumpliendo con los requisitos médicos del problema donde la detección de casos positivos es prioritaria sin comprometer excesivamente la tasa de falsos positivos.

Los modelos basados en ensambles (XGBoost, Random Forest, Gradient Boosting) demostraron consistentemente un rendimiento superior en términos de ROC-AUC comparado con modelos más simples como KNN (0.6754) o SVM (0.7651). Esta superioridad puede atribuirse a su capacidad inherente para capturar relaciones no lineales complejas y manejar interacciones entre múltiples variables clínicas, aspectos fundamentales en la predicción de enfermedades multifactoriales como CKD.

### 10.2.2. Análisis del Recall

En cuanto a la métrica prioritaria del proyecto, el recall, los resultados son particularmente destacables. **SVM con feature selection** alcanzó un recall perfecto (1.0000), identificando correctamente todos los casos de CKD en el conjunto de prueba. Sin embargo, este resultado debe interpretarse con cautela considerando su ROC-AUC relativamente más bajo (0.7651), lo que sugiere una posible tendencia a clasificar conservadoramente hacia la clase positiva.

**Random Forest sin feature selection** mostró un desempeño excepcional con un recall de 0.9967 y ROC-AUC de 0.7921, demostrando una capacidad casi perfecta para identificar casos positivos mientras mantiene una buena capacidad discriminativa general. Este modelo representa una alternativa viable cuando la prioridad absoluta es minimizar falsos negativos, aunque su ROC-AUC no alcance los niveles de XGBoost.

Los modelos de boosting (Gradient Boosting y XGBoost) mantuvieron valores de recall superiores a 0.94 en sus mejores configuraciones, confirmando que las técnicas de ensamble iterativo son especialmente efectivas para problemas médicos donde se requiere alta sensibilidad.

### 10.3. Impacto del Feature Selection

El análisis del impacto de la selección de características revela patrones diferenciados según la arquitectura del modelo empleado. La aplicación de feature selection no constituye una estrategia universalmente beneficiosa, sino que su efectividad está condicionada por las características específicas de cada algoritmo.

#### 10.3.1. Modelos que se Beneficiaron

**XGBoost** experimentó la mejora más sustancial, con un incremento de 2.21 puntos porcentuales en ROC-AUC al incorporar feature selection (0.7865 a 0.8086). Este comportamiento sugiere que la eliminación de características redundantes o ruidosas permitió al algoritmo concentrar su capacidad de aprendizaje en los predictores más informativos, mejorando su generalización. Adicionalmente, el cambio del método óptimo de balanceo de SMOTEENN a SMOTE indica que el espacio de características reducido facilitó la generación de instancias sintéticas más representativas.

**Decision Tree** mostró una mejora notable (0.7461 a 0.7783 en ROC-AUC), confirmando la hipótesis de que los modelos más susceptibles al sobreajuste se benefician significativamente de la reducción dimensional. La eliminación de características irrelevantes reduce la complejidad del árbol, disminuyendo el riesgo de aprender patrones espurios presentes únicamente en el conjunto de entrenamiento.

**SVM** mejoró considerablemente su ROC-AUC (0.7330 a 0.7651) y cambió su estrategia óptima de Class Weights a datos originales. Este comportamiento es consistente con la teoría de que SVM, al buscar hiperplanos de separación óptimos, se beneficia de espacios de menor dimensionalidad donde la separabilidad lineal de las clases puede estar mejor definida.

#### 10.3.2. Modelos con Desempeño Estable o Reducido

**Gradient Boosting** demostró un comportamiento contraintuitivo, alcanzando mejor rendimiento sin feature selection (0.8019 vs 0.7845). Este resultado sugiere que el algoritmo, mediante su proceso iterativo de corrección de errores, es capaz de aprovechar información sutil presente en el conjunto completo de características que podría haberse descartado en el proceso de selección. Los métodos de boosting poseen mecanismos internos de ponderación de características que pueden hacer redundante la selección previa.

**Random Forest** mostró estabilidad notable, con diferencias marginales en ROC-AUC pero alcanzando un recall prácticamente perfecto sin feature selection. Este comporta-

miento refleja la robustez inherente de los bosques aleatorios ante características redundantes, gracias a la aleatoriedad en la selección de variables durante la construcción de cada árbol individual.

## 10.4. Efectividad de las Técnicas de Balanceo

Las estrategias de balanceo de clases demostraron ser determinantes en el desempeño de los modelos, aunque su efectividad varió sustancialmente según el algoritmo empleado.

### 10.4.1. Técnicas de Sobremuestreo

**SMOTE** (Synthetic Minority Over-sampling Technique) emergió como la técnica más efectiva para XGBoost con feature selection, permitiendo alcanzar el mejor ROC-AUC del estudio. SMOTE genera instancias sintéticas interpolando entre ejemplos de la clase minoritaria, incrementando la representatividad de casos positivos sin duplicar exactamente los datos originales. Su efectividad en este contexto sugiere que el espacio de características reducido facilitó la generación de ejemplos sintéticos más coherentes y representativos.

**SMOTEENN** (combinación de SMOTE con Edited Nearest Neighbors) fue óptima para XGBoost sin feature selection, logrando la precision más alta del estudio (0.9591). Esta técnica híbrida combina sobremuestreo de la clase minoritaria con limpieza de instancias ruidosas o ambiguas de ambas clases, lo que resulta en fronteras de decisión más claras y menos falsos positivos.

### 10.4.2. Técnicas de Submuestreo

**Tomek Links** demostró ser la estrategia más efectiva para Logistic Regression, tanto con como sin feature selection, alcanzando el recall más alto de este modelo (0.9738). Esta técnica identifica y elimina pares de instancias de clases opuestas que son mutuamente los vecinos más cercanos, clarificando las fronteras de decisión. Su efectividad con regresión logística sugiere que la separación lineal de clases mejora significativamente al eliminar instancias ambiguas en las regiones de traslape.

**DownSampling** fue la mejor opción para KNN, aunque este modelo mantuvo el rendimiento más bajo del estudio. El submuestreo aleatorio reduce la clase mayoritaria al tamaño de la minoritaria, equilibrando el conjunto de datos pero con pérdida potencial de información valiosa.

### 10.4.3. Ajuste de Pesos y Datos Originales

**Class Weights** resultó óptimo para Random Forest sin feature selection, alcanzando el recall más alto entre todos los modelos (0.9967). Este enfoque penaliza más los errores en la clase minoritaria durante el entrenamiento sin modificar el conjunto de datos, permitiendo que el modelo aprenda directamente de los patrones reales sin introducir artefactos sintéticos.

Sorprendentemente, varios modelos alcanzaron su mejor desempeño con **datos originales** sin técnicas de balanceo (Decision Tree, SVM con feature selection, Gradient Boosting con feature selection). Este resultado sugiere que algunos algoritmos, particularmente aquellos con mecanismos internos robustos de manejo de desbalance o cuando se combina con feature selection adecuado, pueden aprender efectivamente sin manipulación explícita del conjunto de datos.

## 10.5. Consideraciones sobre Precision y Recall

El trade-off entre precision y recall se manifiesta claramente en los resultados obtenidos. Modelos con recall extremadamente alto, como Random Forest sin feature selection (0.9967) o SVM con feature selection (1.0000), mantienen valores de precision superiores a 0.91, indicando que las estrategias de balanceo implementadas lograron minimizar falsos negativos sin comprometer excesivamente la tasa de falsos positivos.

Este equilibrio es crucial en el contexto médico, donde si bien el objetivo prioritario es identificar todos los casos de CKD (maximizar recall), una precision demasiado baja resultaría en excesivas alarmas falsas que sobrecargarían el sistema de salud y generarían ansiedad innecesaria en pacientes sanos. Los valores de precision obtenidos (superiores a 0.91 en prácticamente todos los casos) representan un balance aceptable para una herramienta de screening o apoyo al diagnóstico.

## 10.6. Implicaciones Metodológicas

Los resultados de este proyecto aportan evidencia empírica sobre la aplicación de CRISP-DM en problemas de clasificación médica con desbalance de clases. Varios hallazgos metodológicos merecen destacarse:

1. **La selección de características no es universalmente beneficiosa**: Los modelos con mecanismos internos de regularización o selección de características (como Gradient Boosting y Random Forest) pueden no beneficiarse o incluso degradarse con feature selection previo.

2. **La estrategia óptima de balanceo es específica del modelo**: No existe una técnica de balanceo universalmente superior. SMOTE, Tomek Links, Class Weights y hasta datos originales pueden ser óptimos dependiendo del algoritmo utilizado.

3. **Los ensambles son superiores para problemas médicos complejos**: XGBoost, Random Forest y Gradient Boosting consistentemente superaron a modelos individuales, sugiriendo que la agregación de múltiples hipótesis es particularmente efectiva para capturar la complejidad de relaciones clínicas.

4. **La evaluación multimétrica es esencial**: ROC-AUC, recall y precision proporcionan perspectivas complementarias. Un modelo con recall perfecto pero ROC-AUC moderado (como SVM) puede ser menos deseable que uno con recall ligeramente inferior pero mejor capacidad discriminativa global (como XGBoost).

5. **La validación debe considerar el desbalance**: El uso de métricas robustas ante desbalance (ROC-AUC, F1-score) y validación cruzada estratificada fue fundamental para obtener estimaciones confiables del desempeño real de los modelos.

### 10.7. Limitaciones del Estudio

A pesar de los resultados positivos obtenidos, es importante reconocer las limitaciones inherentes a este trabajo:

1. **Naturaleza sintética de los datos**: El conjunto de datos utilizado es sintético, lo que implica que podría no capturar toda la complejidad y variabilidad presente en datos clínicos reales. Los patrones aprendidos deben validarse en poblaciones reales antes de cualquier aplicación clínica.

2. **Tamaño muestral limitado**: Con 1,659 pacientes, el conjunto de datos es relativamente pequeño para problemas médicos complejos. Un conjunto más grande podría revelar patrones adicionales y mejorar la generalización de los modelos.

3. **Ausencia de validación temporal**: No se evaluó el desempeño de los modelos con datos de diferentes períodos temporales, lo cual es relevante considerando que las características epidemiológicas de CKD pueden cambiar con el tiempo.

4. **Interpretabilidad limitada de modelos complejos**: Aunque XGBoost demostró el mejor desempeño, su interpretabilidad es significativamente menor comparada con modelos más simples como árboles de decisión o regresión logística, lo que puede limitar su aceptación en entornos clínicos.

5. **Falta de análisis de costo-beneficio**: No se evaluó el impacto económico ni

clínico de diferentes tasas de falsos positivos y falsos negativos, información crucial para determinar el umbral óptimo de decisión en aplicaciones reales.

### 10.8. Recomendaciones para Trabajo Futuro

Con base en los hallazgos y limitaciones identificadas, se proponen las siguientes líneas de trabajo futuro:

1. **Validación con datos clínicos reales**: Replicar el estudio utilizando registros médicos electrónicos reales de pacientes, asegurando el cumplimiento de normativas de privacidad y ética de investigación.

2. **Implementación de técnicas de interpretabilidad**: Aplicar métodos como SHAP (SHapley Additive exPlanations) o LIME (Local Interpretable Model-agnostic Explanations) para explicar las predicciones de modelos complejos y facilitar su adopción clínica.

3. **Análisis de curvas de aprendizaje**: Evaluar cómo el desempeño de los modelos escala con el tamaño del conjunto de entrenamiento, identificando si se requieren más datos o si los modelos actuales han alcanzado su capacidad máxima.

4. **Exploración de arquitecturas de deep learning**: Investigar el potencial de redes neuronales profundas, particularmente arquitecturas diseñadas para datos tabulares como TabNet o modelos basados en attention mechanisms.

5. **Desarrollo de ensambles heterogéneos**: Combinar las predicciones de múltiples modelos óptimos (XGBoost, Random Forest, Gradient Boosting) mediante técnicas de stacking o voting para potencialmente mejorar el desempeño.

6. **Optimización de umbrales de decisión**: Realizar un análisis sistemático de diferentes umbrales de clasificación considerando el costo relativo de falsos positivos y negativos desde una perspectiva clínica y económica.

7. **Análisis de subgrupos poblacionales**: Investigar si el desempeño de los modelos varía significativamente entre diferentes grupos demográficos (edad, género, comorbilidades) y desarrollar modelos especializados cuando sea necesario.

8. **Integración con sistemas clínicos**: Diseñar e implementar una interfaz de usuario intuitiva que permita a profesionales de la salud utilizar el modelo en entornos reales, recopilando feedback para mejora continua.

### 10.9. Contribuciones del Proyecto

Este trabajo realiza contribuciones significativas en múltiples dimensiones:

**Contribución metodológica**: Se presenta un marco integral para abordar problemas de clasificación médica con desbalance de clases, demostrando la importancia de evaluar sistemáticamente múltiples técnicas de balanceo y feature selection para cada algoritmo específico.

**Contribución empírica**: Los resultados proporcionan evidencia cuantitativa sobre la efectividad relativa de siete algoritmos de aprendizaje automático y múltiples estrategias de balanceo en el contexto de predicción de CKD, información valiosa para investigadores y profesionales trabajando en problemas similares.

**Contribución práctica**: El modelo final desarrollado (XGBoost con feature selection y SMOTE) representa una herramienta potencialmente útil para apoyar el screening y detección temprana de enfermedad renal crónica, aunque requiere validación adicional antes de aplicación clínica.

**Contribución educativa**: El proyecto ilustra de manera comprensiva la aplicación de la metodología CRISP-DM en un problema real de salud, sirviendo como referencia metodológica para estudiantes y profesionales del área de ciencia de datos aplicada a medicina.

### 10.10. Reflexión Final

Los resultados obtenidos demuestran que el aprendizaje automático posee un potencial significativo para asistir en la detección temprana de enfermedad renal crónica, alcanzando niveles de sensibilidad superiores al 94 % con el modelo óptimo mientras mantiene precisión adecuada. La identificación de **XGBoost con feature selection y SMOTE como la configuración superior**, junto con el análisis exhaustivo de alternativas, proporciona una base sólida para futuras investigaciones en este dominio.

Sin embargo, es fundamental mantener una perspectiva equilibrada sobre el rol de estas tecnologías en la práctica médica. Los modelos desarrollados deben considerarse como herramientas de apoyo al diagnóstico, complementarias al juicio clínico profesional, nunca como sustitutos. La validación rigurosa en poblaciones reales, la consideración de aspectos éticos y de privacidad, y la integración cuidadosa en flujos de trabajo clínicos son requisitos indispensables antes de cualquier implementación práctica.

Este proyecto confirma que la intersección entre ciencia de datos y medicina constituye un área de enorme potencial para mejorar los resultados de salud poblacional. El ma-

nejo metódico del desbalance de clases, la evaluación rigurosa de múltiples alternativas algorítmicas, y la priorización de métricas clínicamente relevantes representan elementos clave para el desarrollo exitoso de sistemas de apoyo al diagnóstico basados en inteligencia artificial. Los hallazgos aquí presentados contribuyen al cuerpo creciente de evidencia que respalda la utilidad de estas tecnologías, al tiempo que identifican claramente las limitaciones y áreas que requieren investigación adicional para traducir estos avances en beneficios tangibles para pacientes y sistemas de salud.