# Clasificación de Diabetes Tipo 2 - Pima Indians

Carlos Castillo, Juan Saldaña, Ángela Torres

Universidad de los Andes

February 21, 2025

# Motivación del Proyecto

- La diabetes tipo 2 afecta a millones de personas en todo el mundo.
- La detección temprana puede prevenir complicaciones graves.
- Los modelos predictivos mejoran la precisión del diagnóstico.

# Objetivo del Proyecto

- Desarrollar un modelo predictivo para clasificar la presencia de diabetes tipo 2.
- Minimizar falsos negativos para mejorar la detección de casos positivos.
- Evaluar el impacto de técnicas de balanceo de clases como SMOTE.

# Descripción del Dataset

- Dataset: Pima Indians Diabetes Database (Kaggle).
- Tamaño: 768 observaciones, 8 variables predictoras y 1 variable objetivo.
- Variables predictoras: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age.
- Variable objetivo: Outcome (0: No Diabetes, 1: Diabetes).

# Distribución de Clases

- Clase 0 (No Diabetes): 65%
- Clase 1 (Diabetes): 35%
- Desequilibrio de clases que puede afectar el rendimiento del modelo.

| Value | Count | Frequency (%) |
|---|---|---|
| 0 | 500 | 65.1% |
| 1 | 268 | 34.9% |

Figure: Distribución de Clases en el Dataset

# Distribución de Clases train vs test

| Conjunto | Clase 0 (%) | Clase 1 (%) |
|---|---|---|
| Train | 65.15 | 34.85 |
| Test | 64.94 | 35.06 |

Table: Distribución de clases en Train y Test

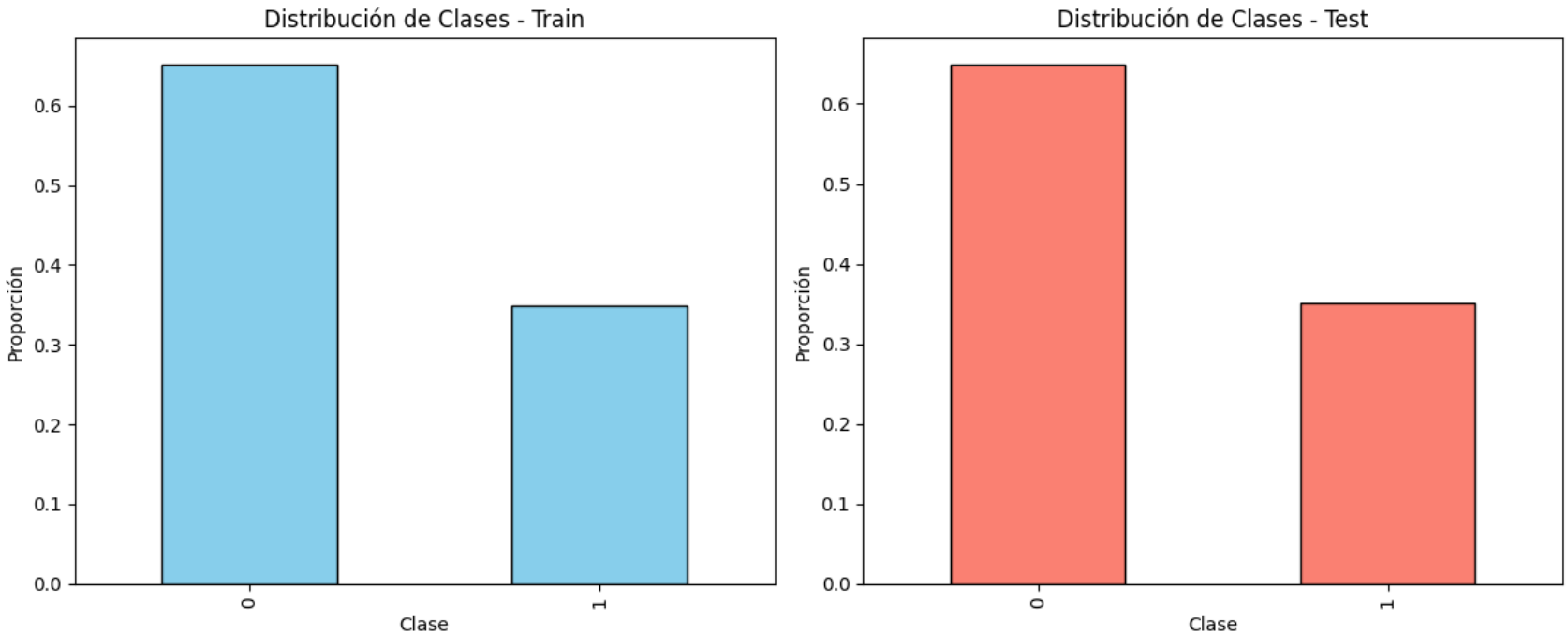



Figure: Distribución de Clases

# Distribución de Variables

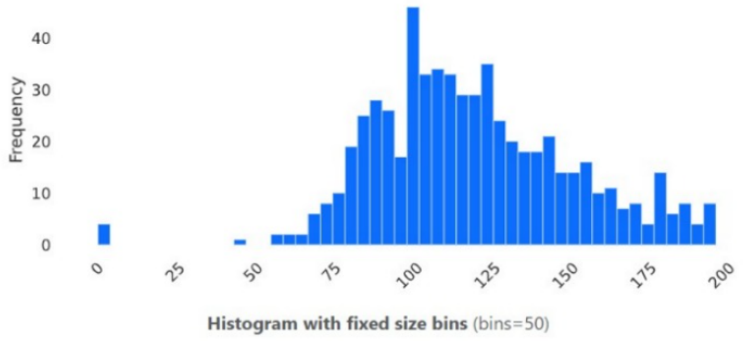
**Glucose train vs test**

Quantile statistics

| | |
|---|---|
| Minimum | 0 |
| 5-th percentile | 78 |
| Q1 | 99 |
| median | 116 |
| Q3 | 140 |
| 95-th percentile | 181 |
| Maximum | 198 |
| Range | 198 |
| Interquartile range (IQR) | 41 |

Descriptive statistics

| | |
|---|---|
| Standard deviation | 32.01518 |
| Coefficient of variation (CV) | 0.26644239 |
| Kurtosis | 0.59176733 |
| Mean | 120.15798 |
| Median Absolute Deviation (MAD) | 20 |
| Skewness | 0.17465286 |
| Sum | 73777 |
| Variance | 1024.9717 |
| Monotonicity | Not monotonic |




Quantile statistics

| | |
|---|---|
| Minimum | 0 |
| 5-th percentile | 83.65 |
| Q1 | 102 |
| median | 120.5 |
| Q3 | 140.75 |
| 95-th percentile | 185.05 |
| Maximum | 199 |
| Range | 199 |
| Interquartile range (IQR) | 38.75 |

Descriptive statistics

| | |
|---|---|
| Standard deviation | 31.735798 |
| Coefficient of variation (CV) | 0.25628279 |
| Kurtosis | 0.92638189 |
| Mean | 123.83117 |
| Median Absolute Deviation (MAD) | 19 |
| Skewness | 0.18019364 |
| Sum | 19070 |
| Variance | 1007.1609 |
| Monotonicity | Not monotonic |

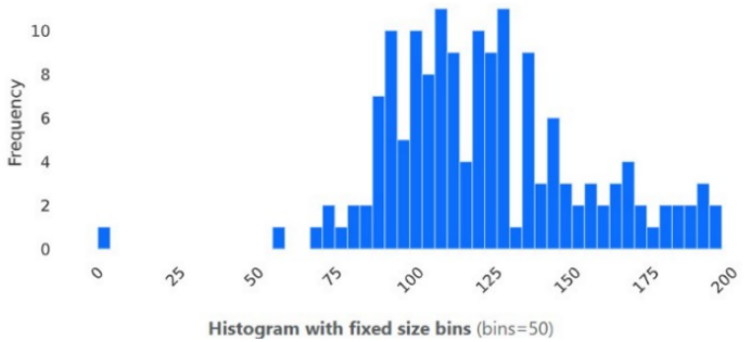


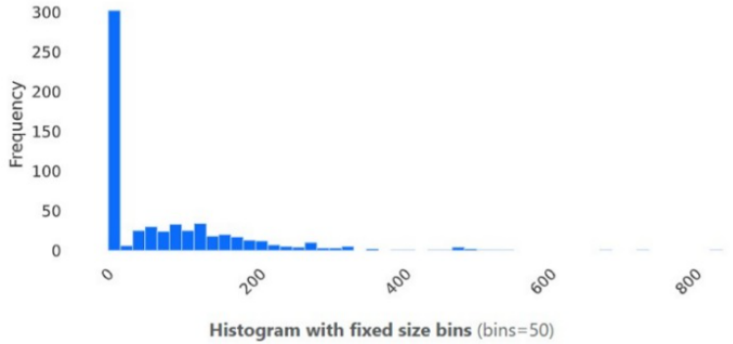

Figure: Distribución de Variables Predictoras

# Distribución de Variables (sesgo)

**Insulin train vs test**

| Quantile statistics | |
|---|---|
| Minimum | 0 |
| 5-th percentile | 0 |
| Q1 | 0 |
| median | 28.5 |
| Q3 | 125.75 |
| 95-th percentile | 284.35 |
| Maximum | 846 |
| Range | 846 |
| Interquartile range (IQR) | 125.75 |

| Descriptive statistics | |
|---|---|
| Standard deviation | 113.51686 |
| Coefficient of variation (CV) | 1.4461947 |
| Kurtosis | 8.148442 |
| Mean | 78.493485 |
| Median Absolute Deviation (MAD) | 28.5 |
| Skewness | 2.3568949 |
| Sum | 48195 |
| Variance | 12886.077 |
| Monotonicity | Not monotonic |




| Quantile statistics | |
|---|---|
| Minimum | 0 |
| 5-th percentile | 0 |
| Q1 | 0 |
| median | 35 |
| Q3 | 135 |
| 95-th percentile | 337.45 |
| Maximum | 600 |
| Range | 600 |
| Interquartile range (IQR) | 135 |

| Descriptive statistics | |
|---|---|
| Standard deviation | 122.13453 |
| Coefficient of variation (CV) | 1.436767 |
| Kurtosis | 4.4733269 |
| Mean | 85.006494 |
| Median Absolute Deviation (MAD) | 35 |
| Skewness | 1.9966482 |
| Sum | 13091 |
| Variance | 14916.843 |
| Monotonicity | Not monotonic |

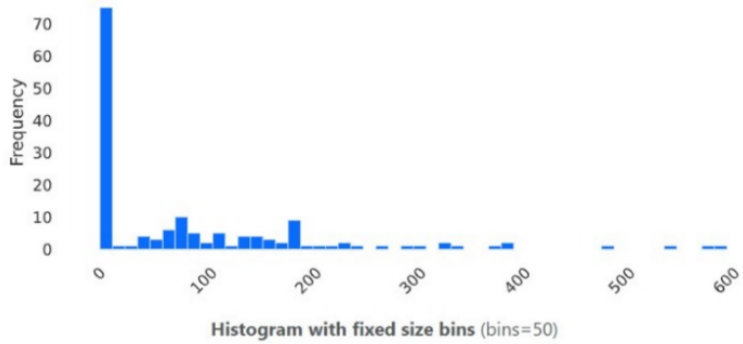



Figure: Distribución de Variables Predictoras

# Análisis Exploratorio de Datos

- Detección de valores faltantes en variables fisiológicamente improbables:
  - BloodPressure, SkinThickness, Insulin, BMI.
- Imputación de valores faltantes utilizando la media.

# Distribución de Variables

- Distribución antes y después de la imputación.
- Evaluación de normalidad en las variables.
- Identificación de posibles valores atípicos.

# Estandarización de Variables

- Se aplicó StandardScaler() para centrar las variables en media 0 y desviación estándar 1.
- Estandarización necesaria para modelos sensibles a escalas como SVM y KNN.
- Aceleración de convergencia en modelos como XGBoost.

$$\text{Standardization:}$$
$$z = \frac{x - \mu}{\sigma}$$
$$\text{with mean:}$$
$$\mu = \frac{1}{N} \sum_{i=1}^{N} (x_i)$$
$$\text{and standard deviation:}$$
$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2}$$

Figure: StandardScaler()

# Justificación de la Estandarización

- Evitar que variables con escalas grandes dominen el modelo.
- Mejorar la estabilidad numérica en redes neuronales y algoritmos basados en gradiente.
- Normalización de datos para mejorar el rendimiento de KNN.

# Análisis de Correlación

- Evaluación de la correlación entre variables predictoras.
- Identificación de multicolinealidad.

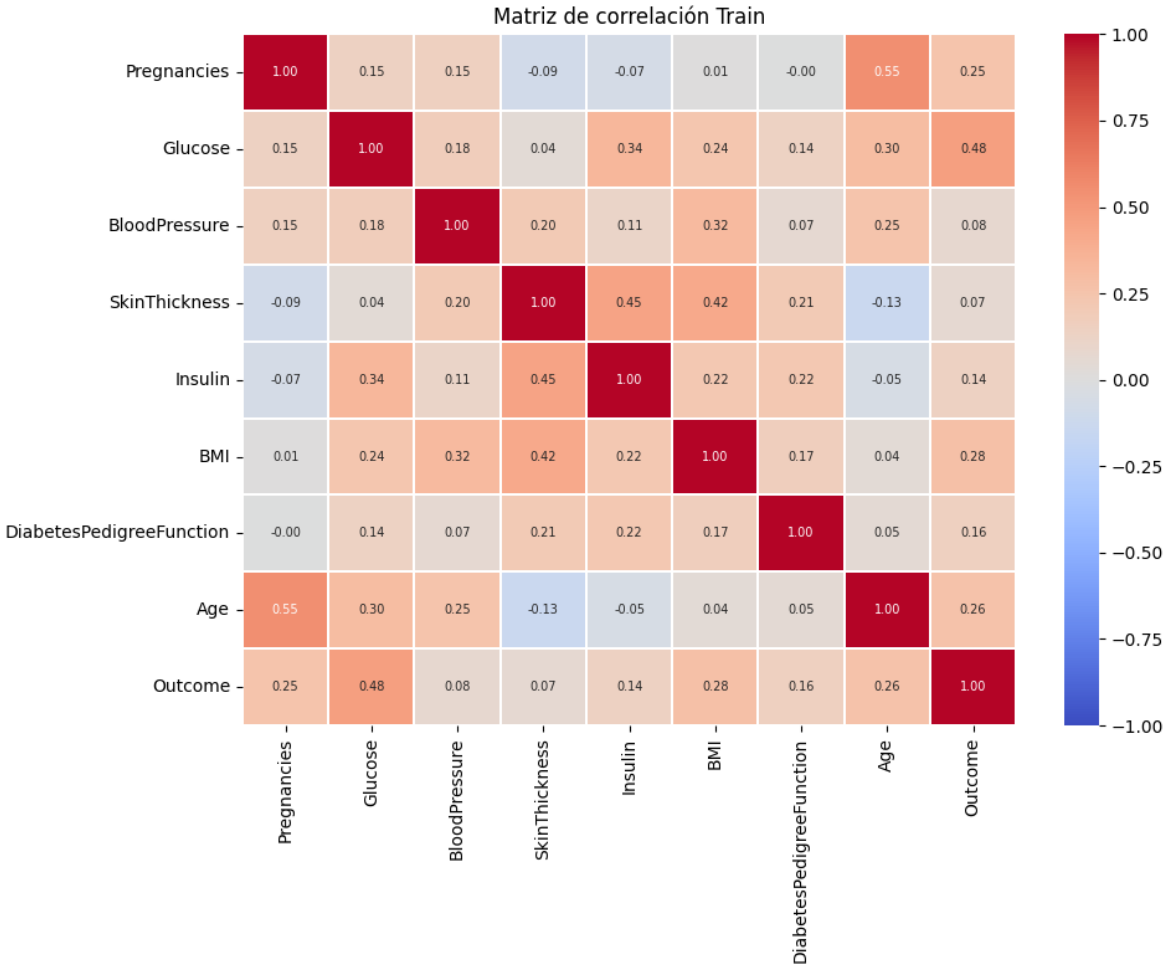



Figure: Matriz de Correlación de Variables

# Conclusiones del Preprocesamiento

- Imputación de valores faltantes mejora la calidad de los datos.
- La estandarización es crucial para modelos basados en distancia.
- La correlación alta sugiere posible multicolinealidad.

# Modelos de Clasificación

- **Regresión Logística (Ridge y Lasso):** Interpretabilidad y eficiencia.
- **SVM:** Alta capacidad de generalización.
- **Random Forest:** Robustez frente a ruido y sobreajuste.
- **XGBoost:** Alta precisión y capacidad de ajuste fino.
- **KNN:** Intuitivo y eficaz en datos balanceados.

# Justificación de Modelos

- **Regresión Logística:** Línea base interpretativa y eficiente.
- **SVM:** Funciona bien con datos no lineales.
- **Random Forest:** Evita sobreajuste con ensamblado de árboles.
- **XGBoost:** Optimizaciones en memoria y velocidad de entrenamiento.
- **KNN:** Captura relaciones locales en datos balanceados.

# Tuning de Hiperparámetros

- GridSearchCV con validación cruzada de 3 folds.
- Métrica de evaluación: Recall Weighted para optimizar la detección de casos positivos.
- Hiperparámetros ajustados:
  - Regresión Logística: C (regularización), max_iter (iteraciones).
  - SVM: C (regularización), kernel (lineal y RBF).
  - Random Forest y XGBoost: n_estimators (número de árboles), max_depth (profundidad).

# Métricas de Evaluación

- **Accuracy:** Proporción de predicciones correctas.
- **Precision:** Exactitud de las predicciones positivas.
- **Recall:** Capacidad de detectar casos positivos.
- **F1-Score:** Promedio armónico de Precision y Recall.
- **AUC-ROC:** Área bajo la curva ROC, mide capacidad de clasificación.

# Modelos de Clasificación

| Escenario | Métrica recomendada |
| --- | --- |
| **Datos balanceados** | `accuracy` |
| **Datos desbalanceados y queremos minimizar falsos negativos** (evitar no detectar diabéticos) | `recall_weighted` |
| **Queremos un equilibrio entre** `precision` **y** `recall` | `f1_weighted` |
| **Queremos evaluar la discriminación del modelo** | `roc_auc` |

Figure: Métrica

# Justificación de Recall Weighted

- **Problema:** El dataset está desbalanceado (35% Diabetes, 65% No Diabetes).
- **Solución:** Recall Weighted da mayor peso a la clase minoritaria.
- **Beneficios:**
  - Reduce los falsos negativos.
  - Mejora la sensibilidad en la detección de diabetes.
  - Mayor utilidad en contextos médicos.

# Objetivo

- En este caso, es más importante reducir los falsos negativos (FN), porque no detectar un paciente diabético a tiempo puede ser peligroso.

| Error | Consecuencia | Impacto en Diagnóstico de Diabetes |
|---|---|---|
| **Falso Positivo (FP)** → Se predice "diabetes", pero en realidad **no tiene** | Puede causar exámenes adicionales o tratamientos innecesarios. | **Menos grave**, ya que solo generará más pruebas médicas. |
| **Falso Negativo (FN)** → Se predice "sin diabetes", pero en realidad **sí tiene** | No recibe tratamiento a tiempo, la enfermedad progresa sin control. | **Muy grave**, porque podría poner en riesgo la salud del paciente. |

Figure: Objetivo

# Objetivo

- En un caso médico, como la detección de diabetes, es crítico minimizar los falsos negativos (FN), ya que un diagnóstico incorrecto puede llevar a que un paciente con diabetes no reciba tratamiento oportuno, aumentando el riesgo de complicaciones graves. La métrica recallweighted es la más adecuada porque mide la capacidad del modelo para identificar correctamente todos los casos positivos, asegurando que la mayoría de los pacientes diabéticos sean detectados. Aunque los,falsos positivos (FP) pueden generar exámenes adicionales, es un error menos peligroso que un FN. Por esta razón, en problemas de salud pública y medicina, optimizar recallweighted es clave para priorizar la seguridad del paciente.

# Resultados, Matriz de Confusión

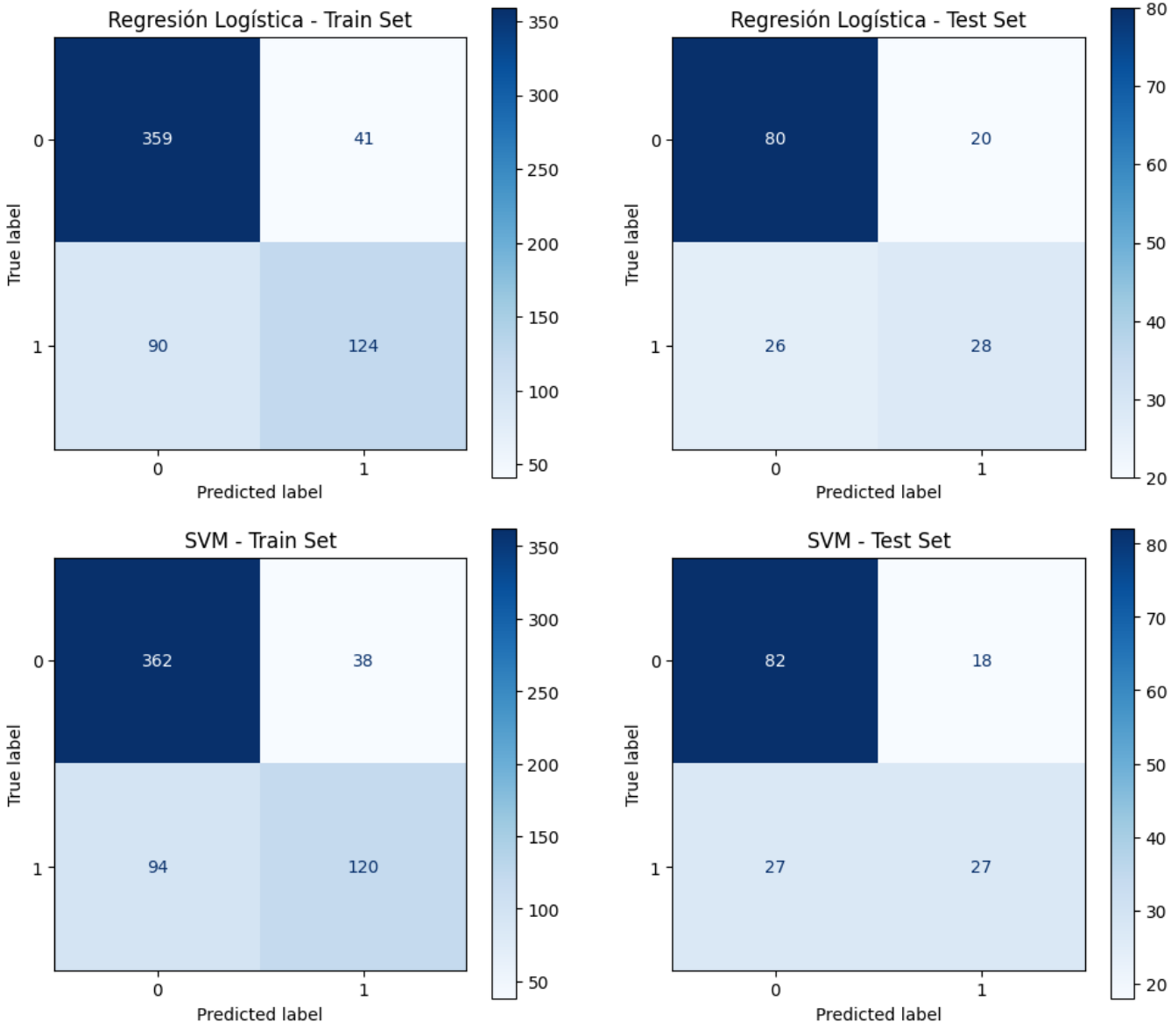



Figure: Matrices de Confusión para Modelos 1, 2 y 3

# Resultados, Matriz de Confusión

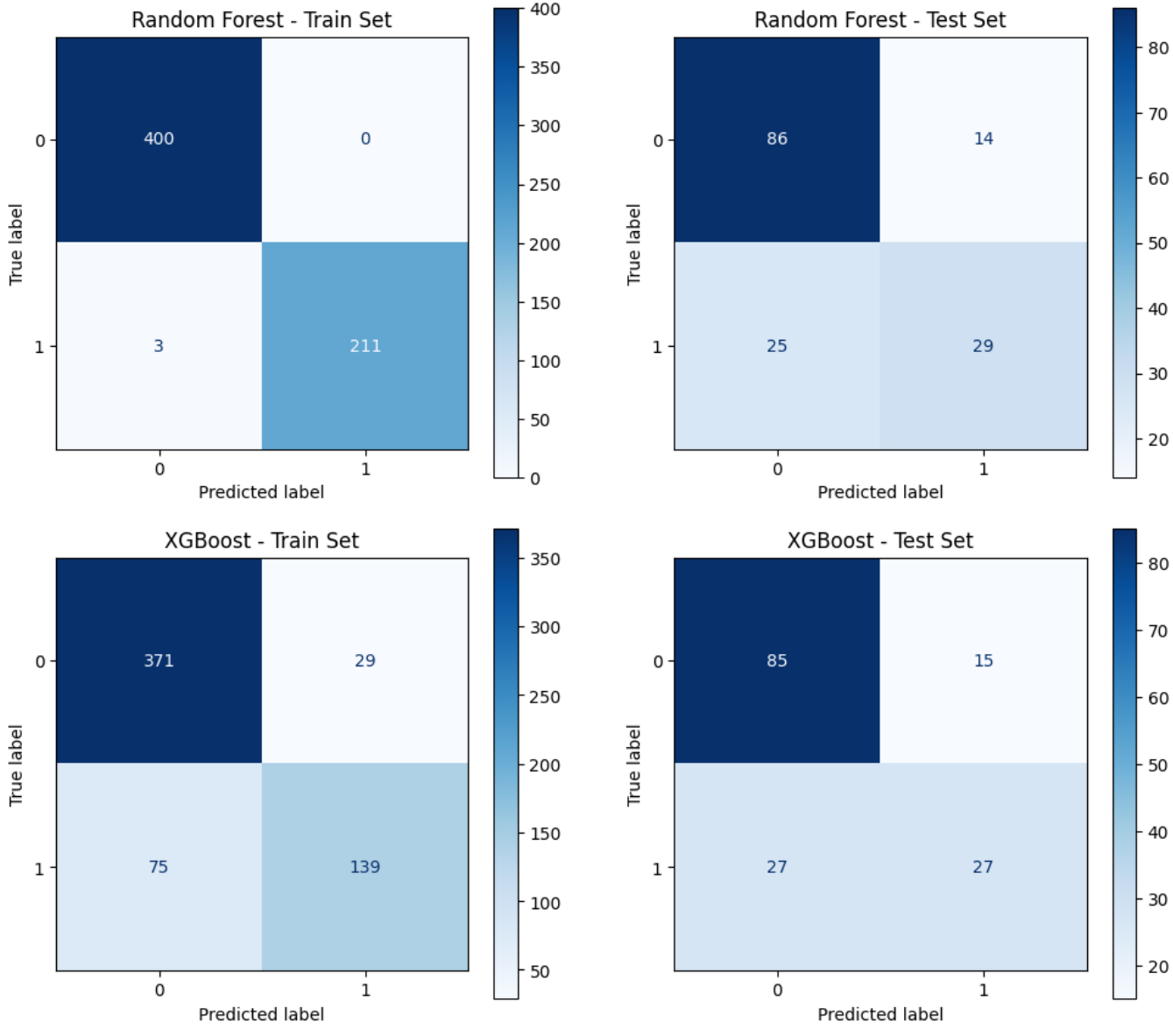



Figure: Matrices de Confusión para Modelos 4 y 5

# Resultados, Matriz de Confusión

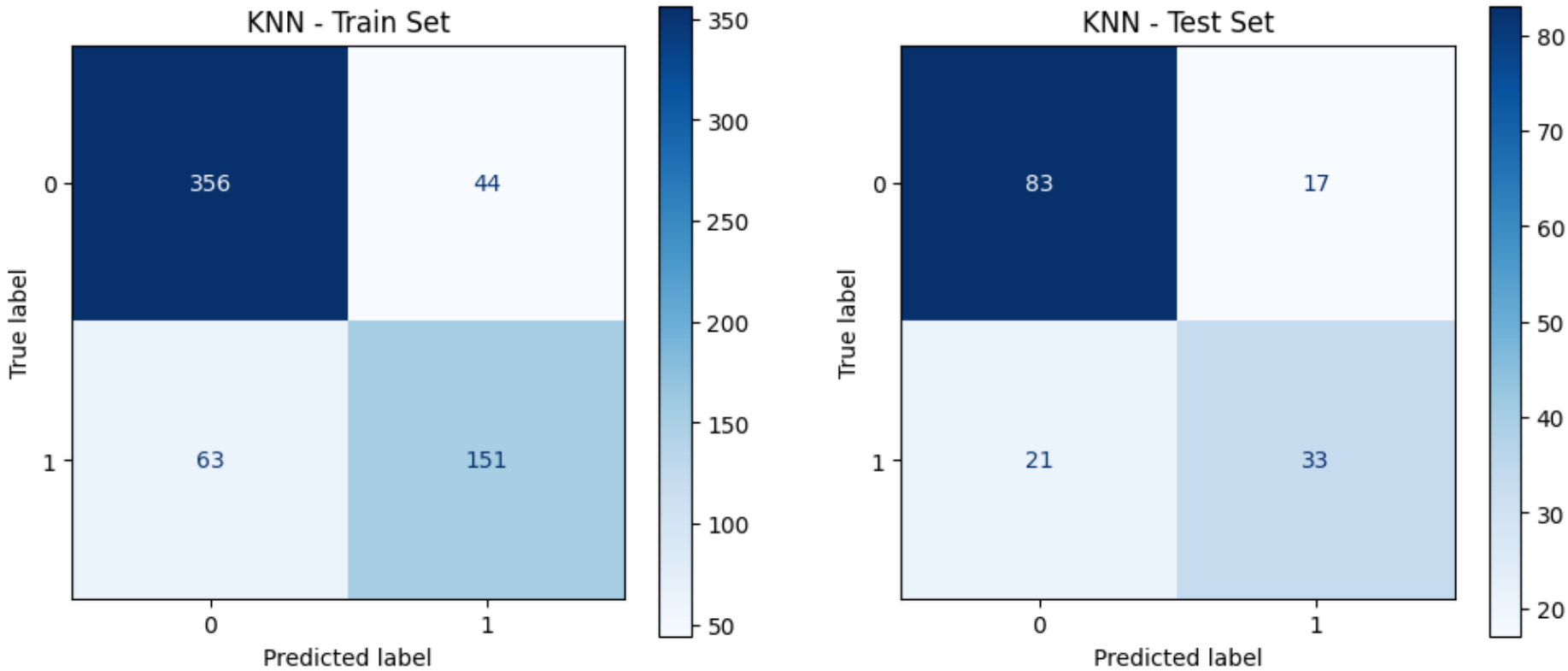



Figure: Matrices de Confusión para Modelos 4 y 5

# Resultados

- KNN mostró el mejor desempeño en Recall Weighted antes de SMOTE.
- Random Forest tuvo un buen balance entre Precision y Recall.
- SVM y Regresión Logística mostraron un desempeño similar con un ligero sesgo hacia la clase mayoritaria.

| Modelo | Recall Weighted | Accuracy | Precision | Recall | F1-Score | Best Params |
|---|---|---|---|---|---|---|
| KNN | 0.7532 | 0.7532 | 0.7497 | 0.7532 | 0.7509 | {'n_neighbors': 7, 'weights': 'uniform'} |
| Random Forest | 0.7468 | 0.7468 | 0.7396 | 0.7468 | 0.7390 | {'max_depth': 10, 'n_estimators': 100} |
| XGBoost | 0.7273 | 0.7273 | 0.7182 | 0.7273 | 0.7179 | {'learning_rate': 0.01, 'max_depth': 3, 'n_estimators': 200} |
| SVM | 0.7078 | 0.7078 | 0.6989 | 0.7078 | 0.7008 | {'C': 10, 'kernel': 'linear'} |
| Regresión Logística | 0.7013 | 0.7013 | 0.6946 | 0.7013 | 0.6969 | {'C': 1, 'max_iter': 200} |

Figure: Resultados antes de aplicar SMOTE

# Resultados

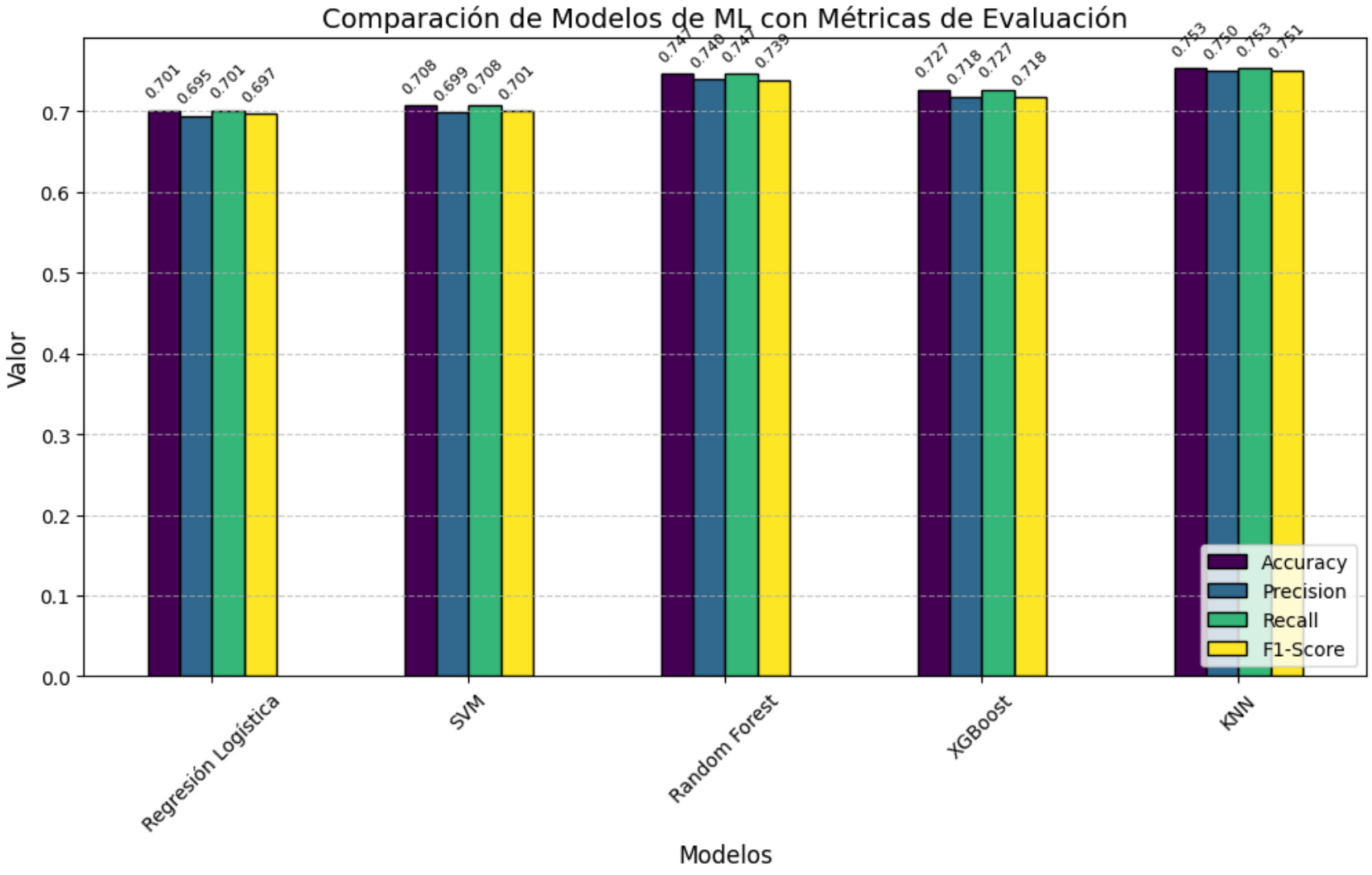



Figure: Resultados gráfico de barras

# Resultados

- Para evaluar el desempeño de los modelos de clasificación en la detección de diabetes, implementamos la Curva ROC (Receiver Operating Characteristic) y calculamos el AUC (Área Bajo la Curva). Esta métrica nos permite analizar la capacidad de los modelos para distinguir entre pacientes con y sin diabetes, considerando el equilibrio entre la tasa de verdaderos positivos (sensibilidad) y la tasa de falsos positivos en distintos umbrales de decisión. Dado que en problemas médicos como este es crucial minimizar los falsos negativos, la Curva ROC nos ayuda a comparar qué modelo ofrece la mejor discriminación entre clases y cuál podría requerir ajustes adicionales en su umbral de decisión.

# Resultados

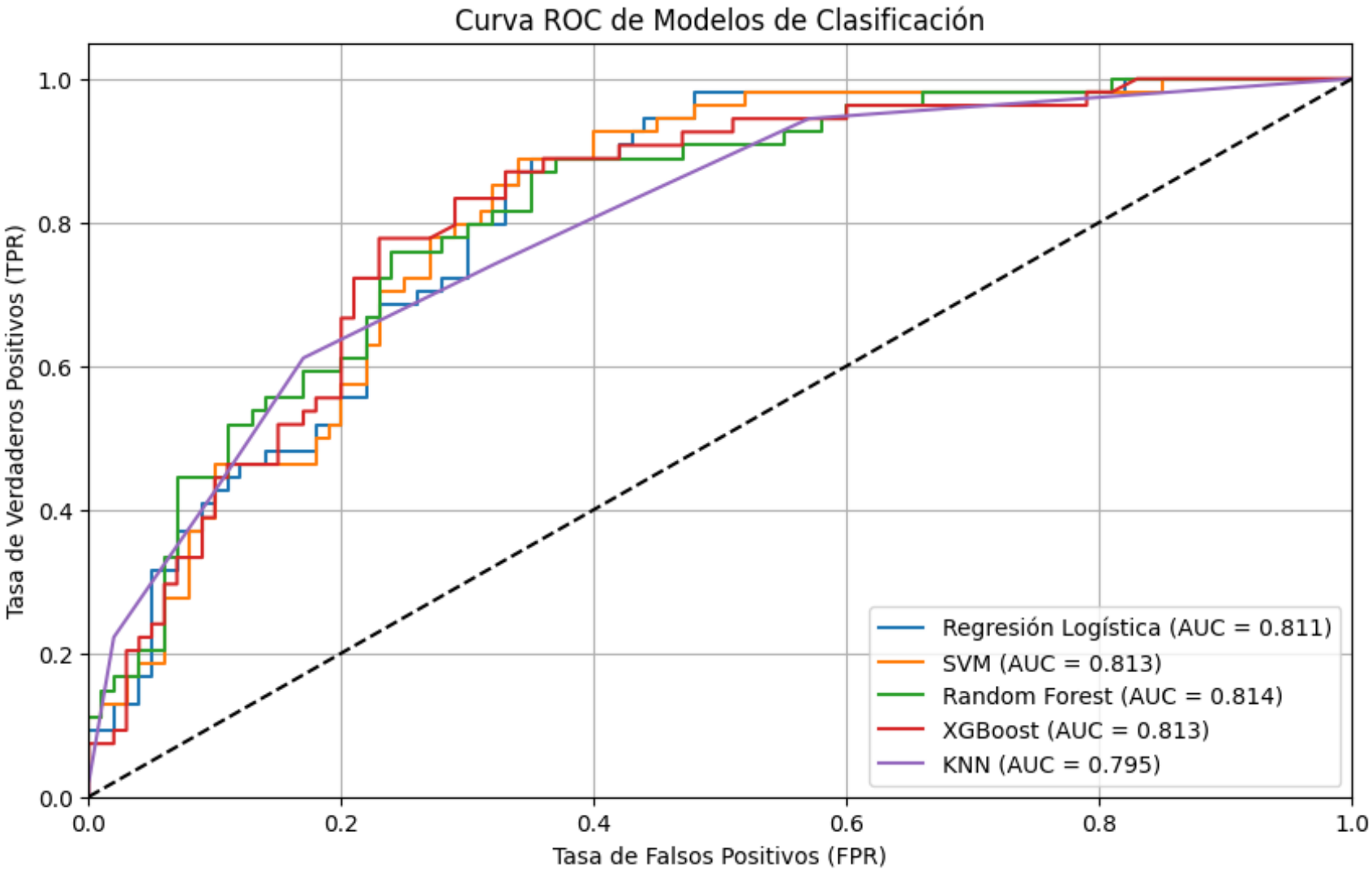



Figure: Curva ROC

# Resultados

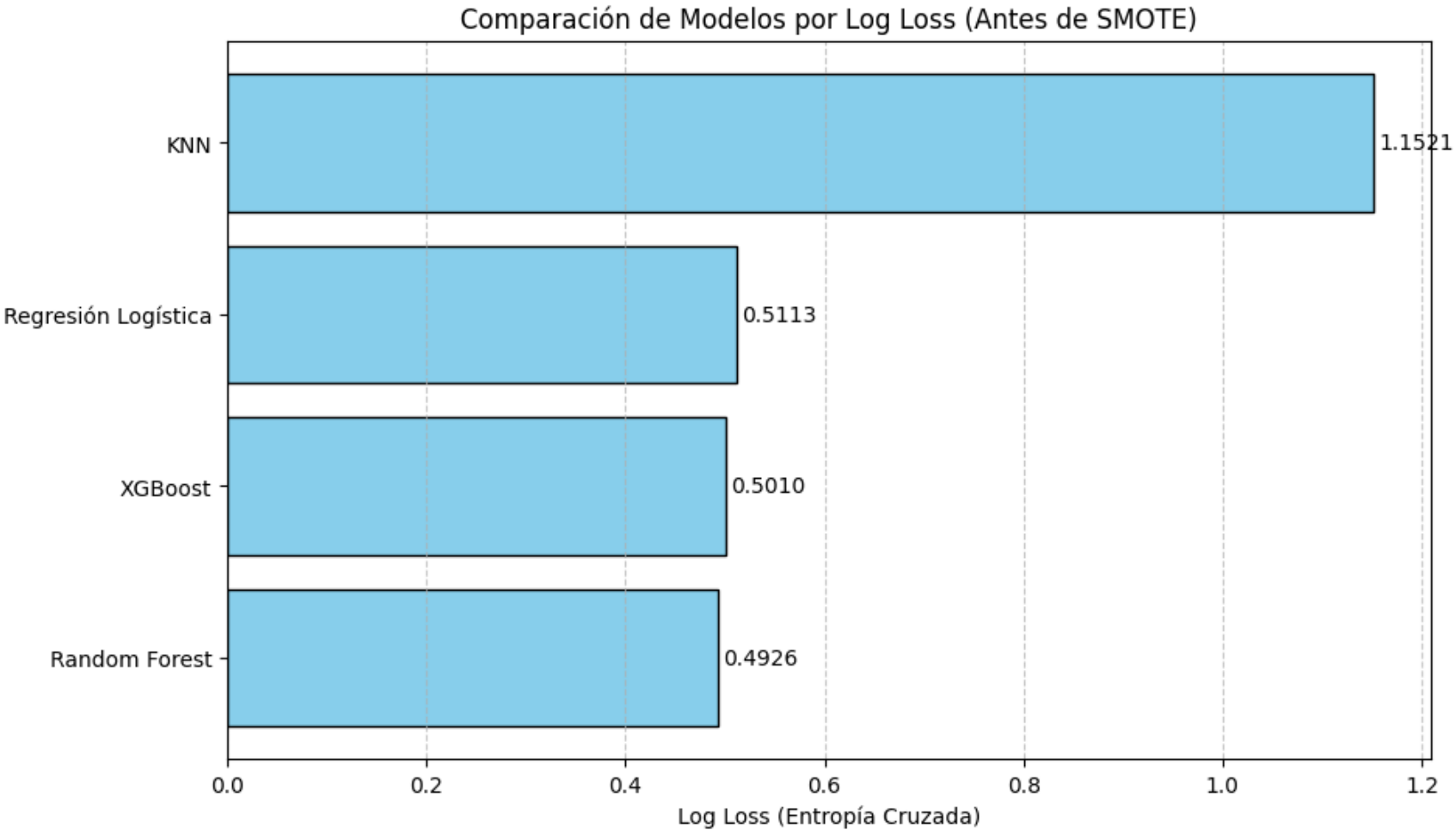



Figure: Entropía cruzada

# Resultados

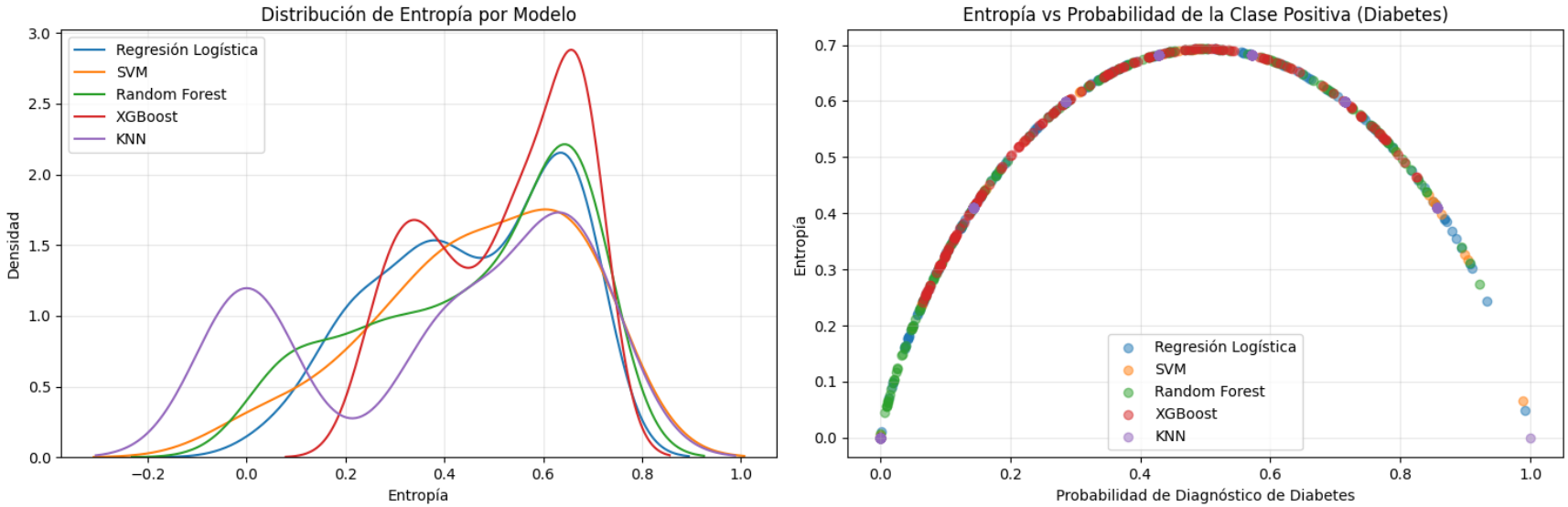



Figure: Distribución y probabilidad

# Desequilibrio de Clases

- Clase 0 (No Diabetes): 65%
- Clase 1 (Diabetes): 35%
- El desequilibrio afecta la precisión en la detección de casos positivos.
- Los modelos tienden a predecir mayoritariamente la clase mayoritaria.

# ¿Qué es SMOTE?

- **SMOTE (Synthetic Minority Over-sampling Technique):**
  - Genera muestras sintéticas de la clase minoritaria.
  - Utiliza vecinos más cercanos (KNN) para crear instancias sintéticas.
- **Objetivo:** Balancear la distribución de clases para mejorar la sensibilidad del modelo.

# Impacto de SMOTE en el Dataset

- Aumento en el número de observaciones de la clase minoritaria.
- Mejor equilibrio en la distribución de clases.
- Reducción de falsos negativos en la detección de diabetes.

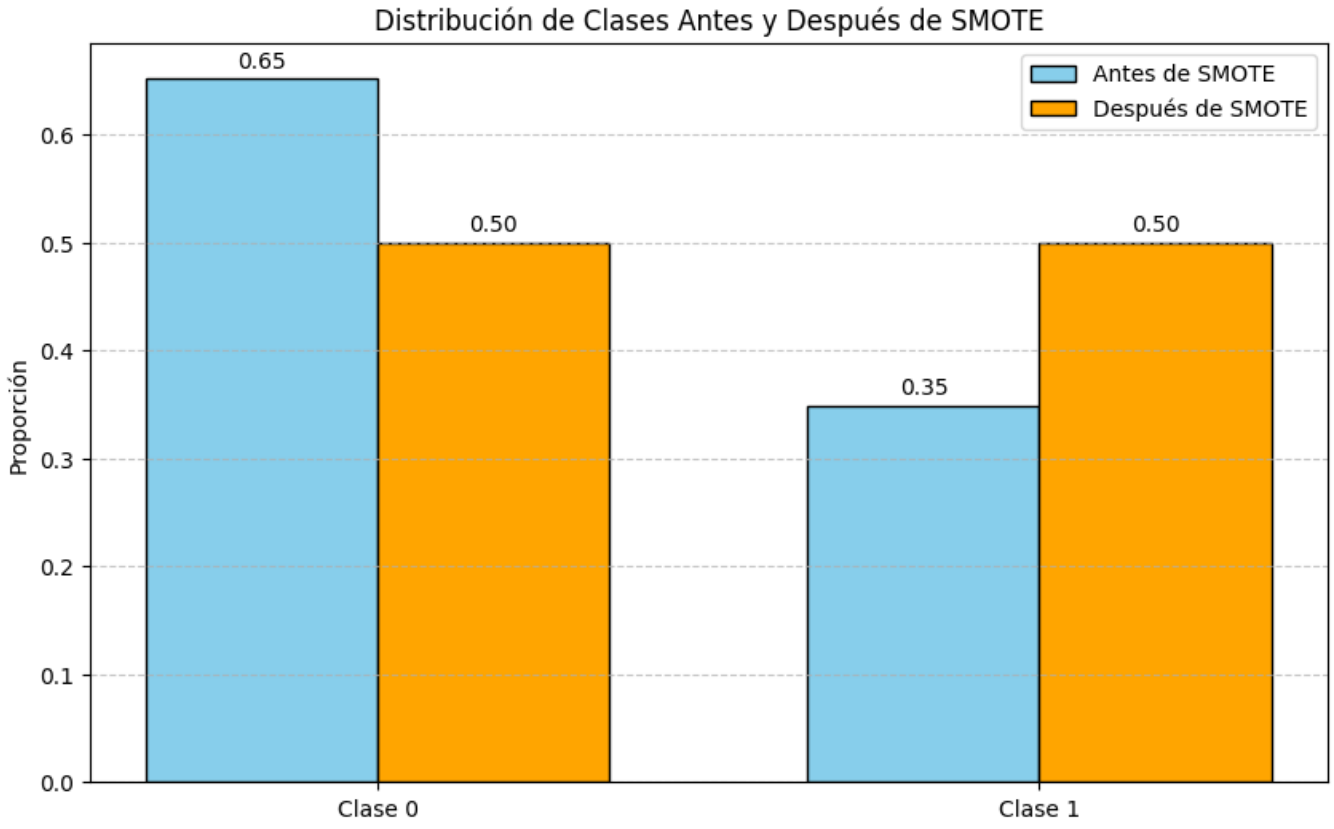



Figure: Impacto de SMOTE en la Distribución de Clases

# Justificación Técnica de SMOTE

- Aumenta la sensibilidad (Recall) en la clase minoritaria.
- Mejora la capacidad de generalización del modelo.
- Evita el sobreajuste al generar datos sintéticos en lugar de replicar muestras.

# Conclusiones de SMOTE

- Mejora en el Recall Weighted sin afectar significativamente la precisión.
- Mayor estabilidad en las métricas de evaluación.
- Beneficios en la detección temprana de diabetes.

# Resultados después de SMOTE, Matriz de Confusión

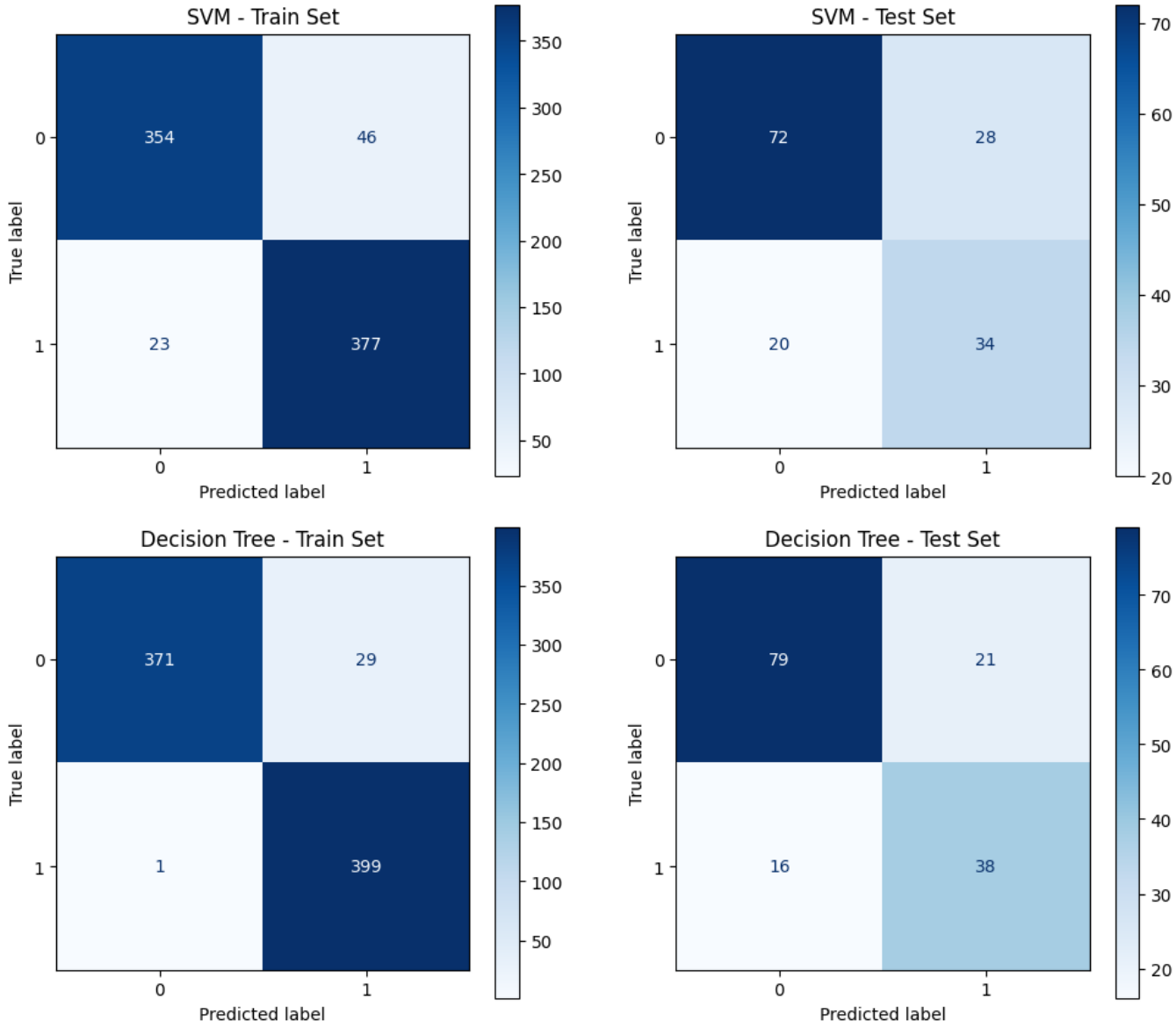



Figure: Matrices de Confusión para Modelos 1 y 2

# Resultados después de SMOTE, Matriz de Confusión

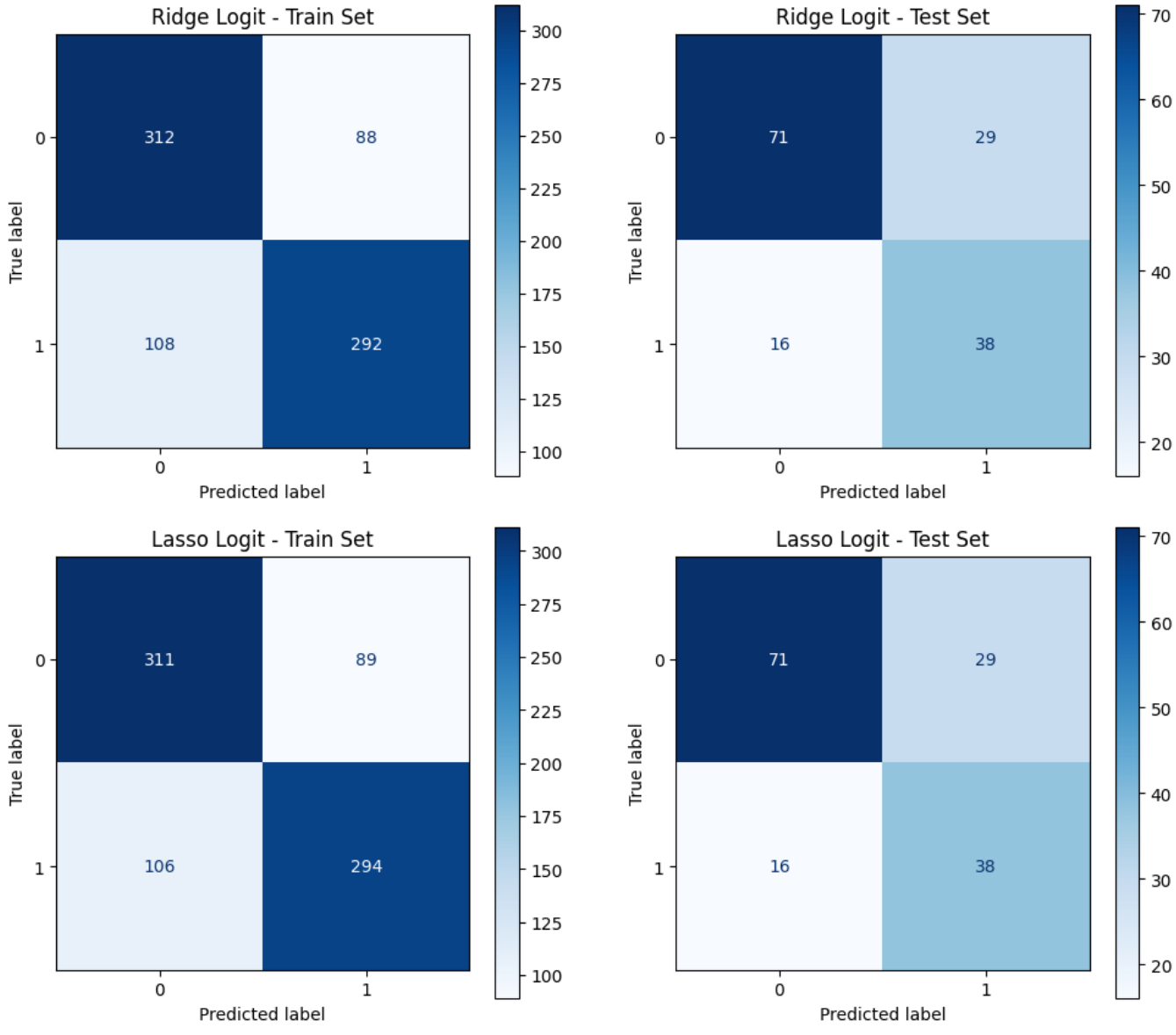



Figure: Matrices de Confusión para Modelos 3 y 4

# Resultados después de SMOTE, Matriz de Confusión

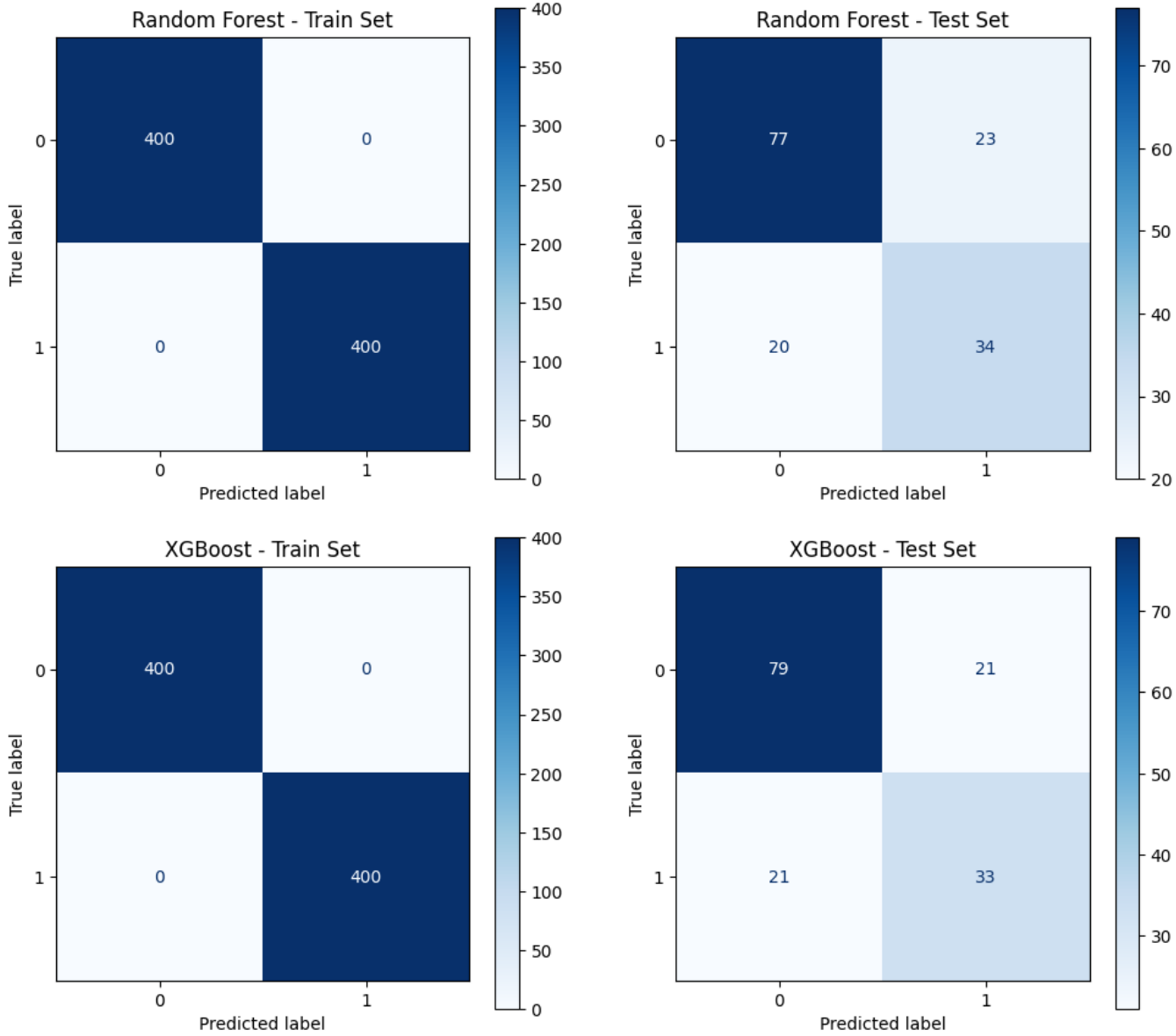



Figure: Matrices de Confusión para Modelos 5 y 6

# Resultados después de SMOTE, Matriz de Confusión

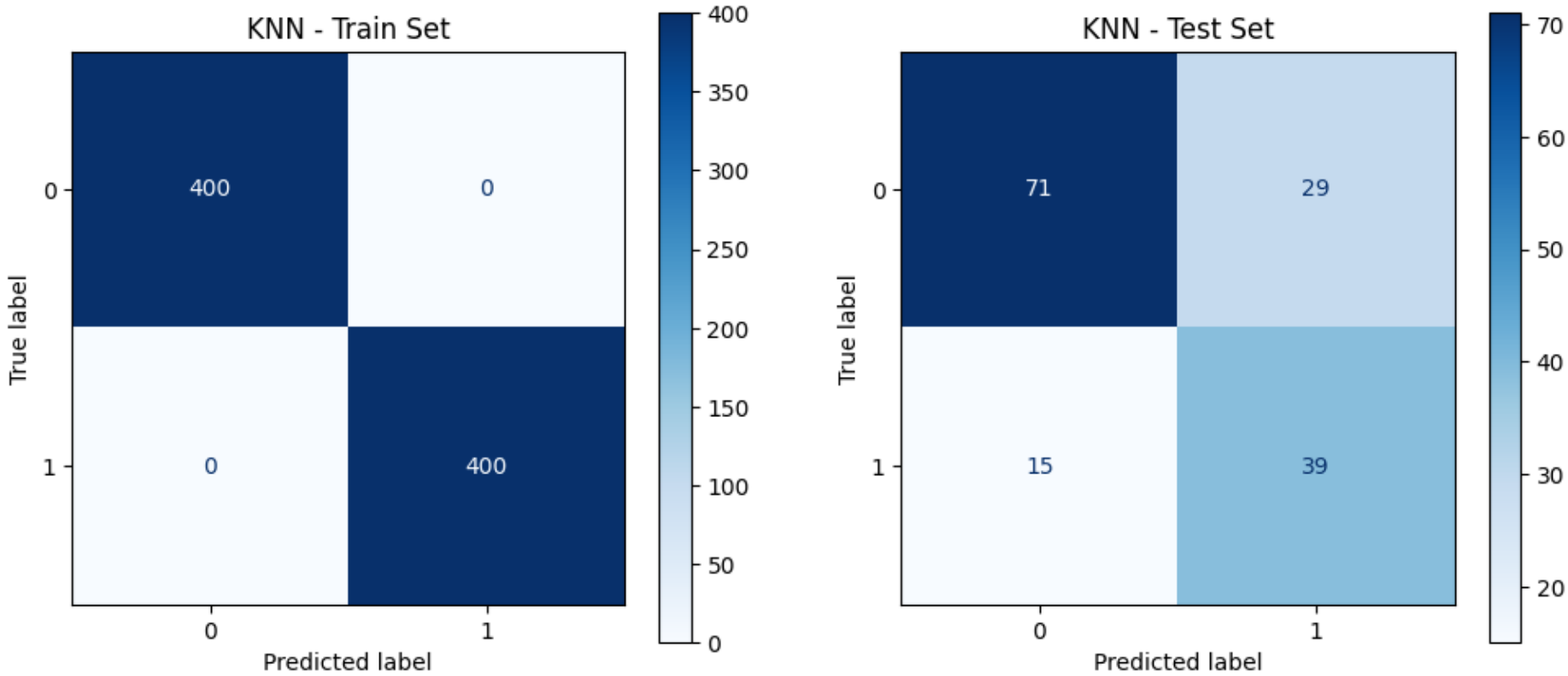



Figure: Matrices de Confusión para Modelo 7

# Resultados después de SMOTE

- Decision Tree mostró la mejor capacidad de generalización después de SMOTE.
- XGBoost presentó el mejor desempeño en términos de F1-Score.
- Se observó una mejora en Recall Weighted en todos los modelos.

| Index | Recall Weighted | Accuracy | Precision | Recall | F1-Score | Best Params |
|---|---|---|---|---|---|---|
| Decision Tree | 0.7597 | 0.7597 | 0.7658 | 0.7597 | 0.7620 | {'max_depth': 10, 'min_samples_split': 2} |
| KNN | 0.7143 | 0.7143 | 0.7372 | 0.7143 | 0.7199 | {'n_neighbors': 3, 'weights': 'distance'} |
| Lasso Logit | 0.7078 | 0.7078 | 0.7288 | 0.7078 | 0.7133 | {'C': 1, 'penalty': 'l1'} |
| Random Forest | 0.7208 | 0.7208 | 0.7246 | 0.7208 | 0.7224 | {'max_depth': None, 'n_estimators': 200} |
| Ridge Logit | 0.7078 | 0.7078 | 0.7288 | 0.7078 | 0.7133 | {'C': 1, 'penalty': 'l2'} |
| SVM | 0.6883 | 0.6883 | 0.7005 | 0.6883 | 0.6926 | {'C': 10, 'kernel': 'rbf'} |
| XGBoost | 0.7273 | 0.7273 | 0.7273 | 0.7273 | 0.7273 | {'learning_rate': 0.3, 'max_depth': 10, 'n_estimators': 50} |

Figure: Resultados después de aplicar SMOTE

# Resultados después de SMOTE

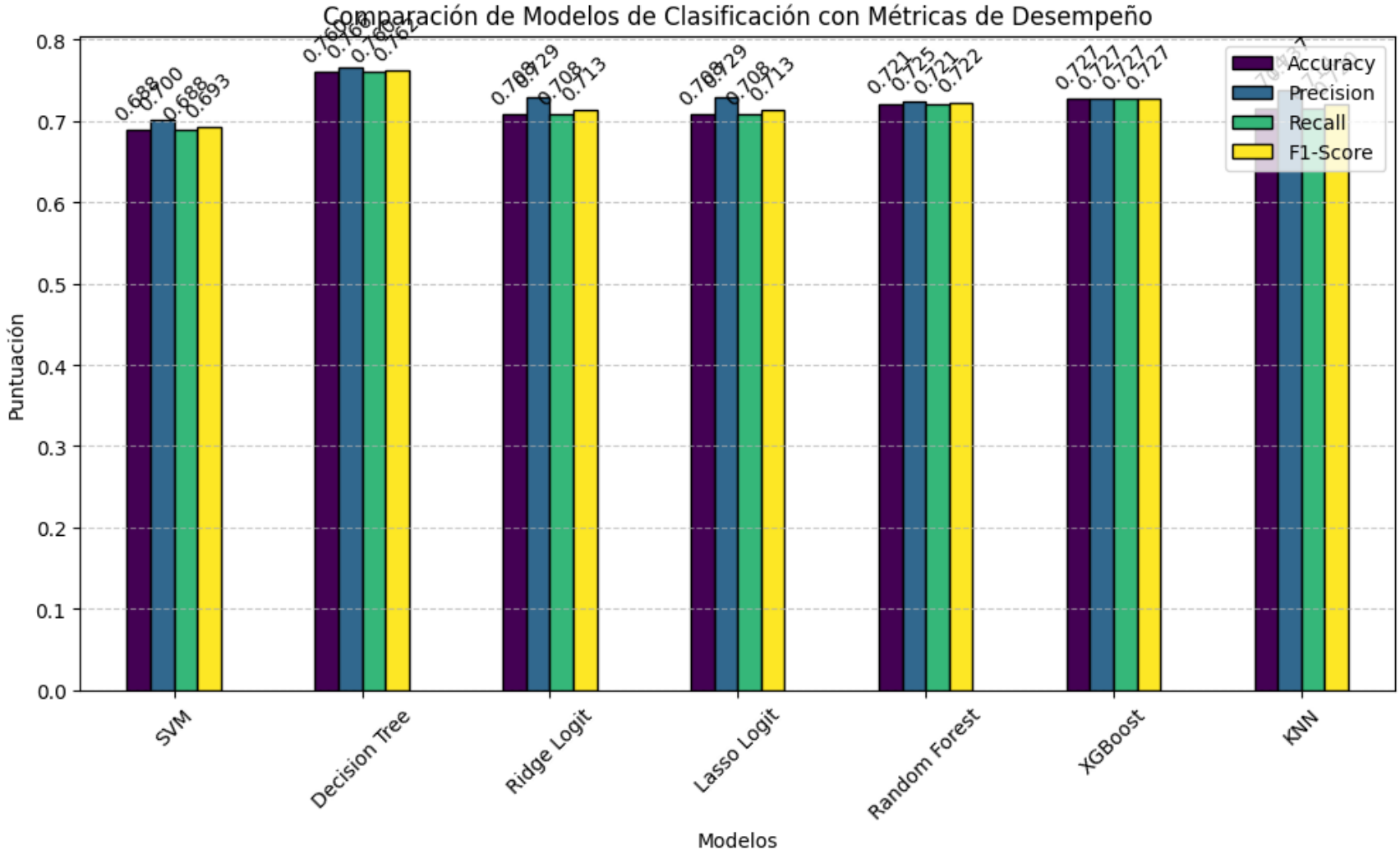



Figure: Resultados SMOTE gráfico de barras

# Curva ROC y AUC (Después de SMOTE)

- Mejora en AUC en la mayoría de los modelos después de aplicar SMOTE.
- XGBoost y Decision Tree mostraron el mayor aumento en AUC.
- La mejora en AUC indica una mejor capacidad de clasificación.

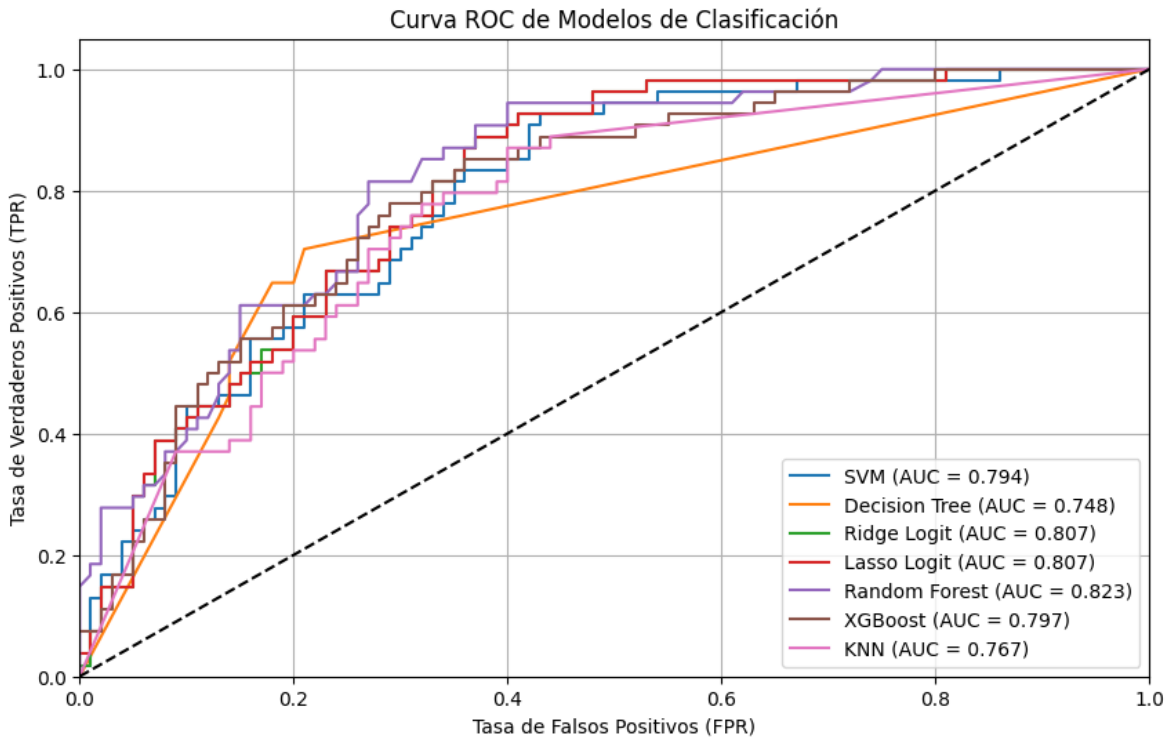



Figure: Curva ROC después de SMOTE

# Entropía Cruzada SMOTE

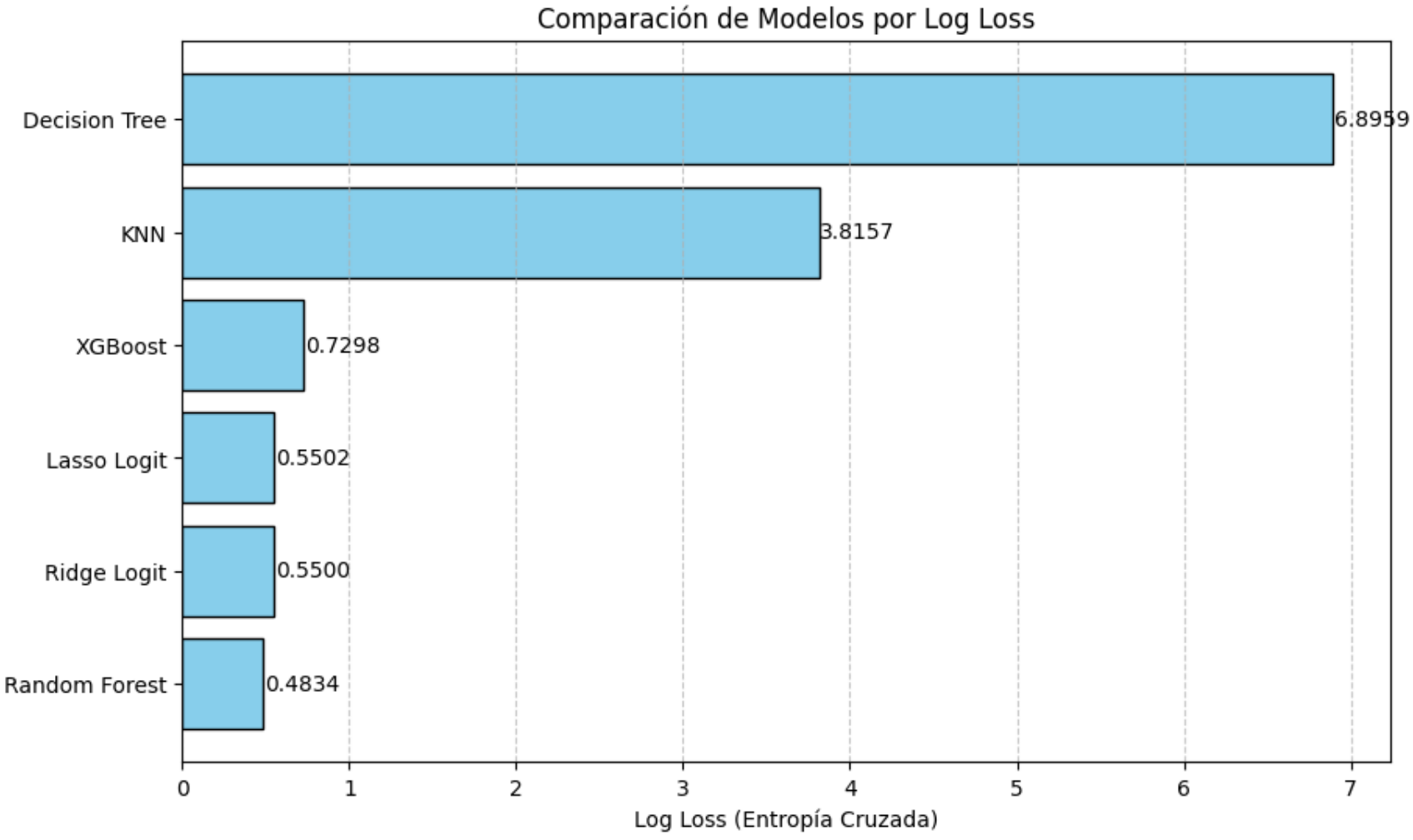



Figure: Entropía cruzada

# Distribución de Entropía

- Análisis de entropía para evaluar la incertidumbre en las predicciones.
- KNN y Decision Tree mostraron la menor entropía después de SMOTE.
- XGBoost mostró mayor certeza en las predicciones de la clase positiva.

# Distribución de Entropía SMOTE

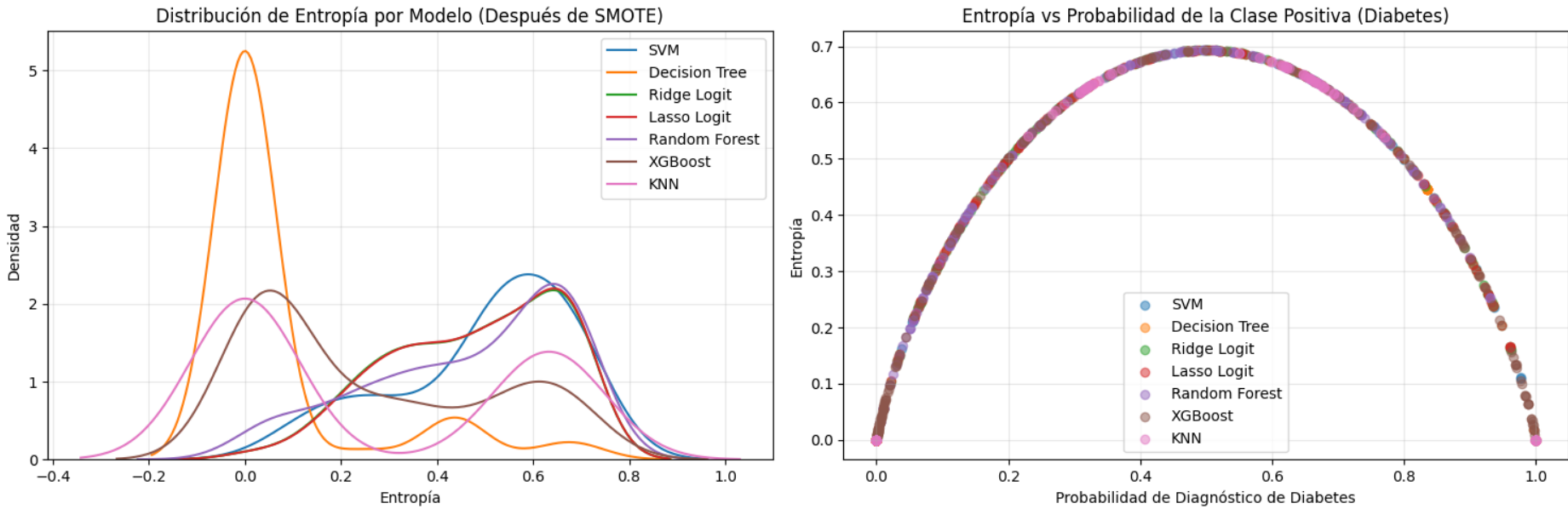



Figure: Distribución de Entropía por Modelo

# Conclusiones Clave

- SMOTE mejoró significativamente el Recall Weighted y el AUC.
- Decision Tree y XGBoost mostraron el mejor desempeño después de SMOTE.
- La entropía cruzada reveló mayor certeza en las predicciones.
- La combinación de métricas permitió una evaluación más completa.

# Implicaciones en Diagnóstico de Diabetes

- Mejor detección de casos positivos reduce falsos negativos.
- Ayuda en la detección temprana de diabetes tipo 2.
- Aplicable en contextos clínicos para apoyar la toma de decisiones.

# Participación del Equipo

- Carlos Castillo: Preprocesamiento y análisis de resultados, Implementación de modelos y tuning de hiperparámetros
- Ángela Torres: Comparación con Paper.
- Juan Saldaña: Documentación, presentación y análisis de métricas.

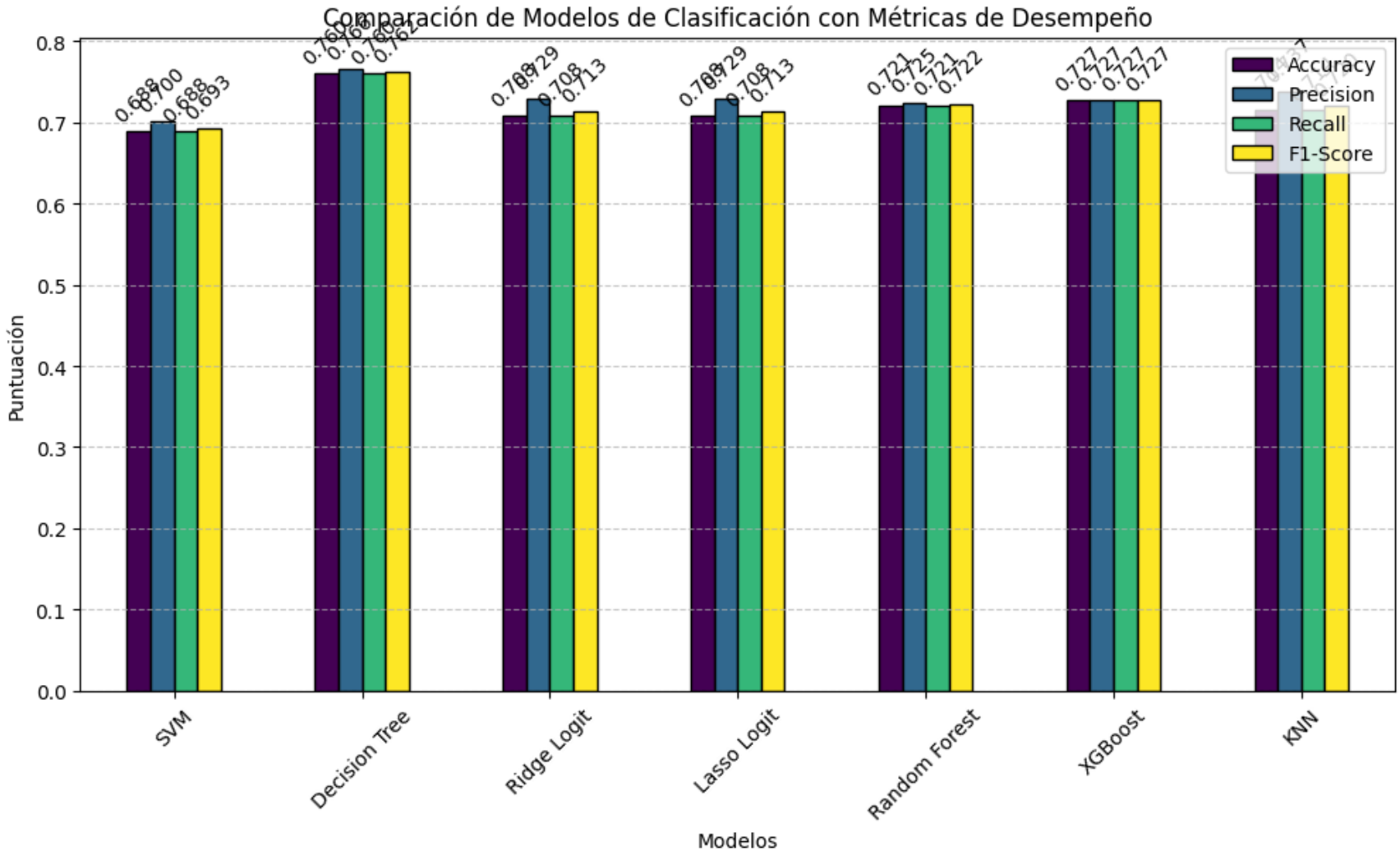
# Conclusión: Desempeño en Clasificación

- El dataset de Pima Indians Diabetes demostró ser adecuado para tareas de clasificación en la detección de diabetes tipo 2.
- Los modelos de clasificación, especialmente **Random Forest** y **XGBoost**, mostraron un buen equilibrio entre precisión y recall.
- La aplicación de **SMOTE** mejoró significativamente el recall en la clase minoritaria, lo que indica una mejor capacidad de detección de casos positivos.
- La entropía mostró que los modelos son consistentes en sus predicciones con baja incertidumbre.
- La curva ROC sugiere que los modelos son efectivos para distinguir entre clases, con AUC cercanas a 0.81.
- Por lo tanto, el dataset es útil para el diagnóstico temprano de diabetes utilizando enfoques de clasificación.

# Conclusión: Desempeño en Regresión

- Al utilizar la variable **Glucose** como objetivo en regresión, los modelos lograron predecir valores con un error moderado.
- Los modelos lineales como **Lasso** y **Ridge** mostraron un rendimiento consistente con bajos valores de MAE, indicando una buena capacidad de generalización.
- Sin embargo, el MSE fue relativamente alto debido a la amplia dispersión en los valores de glucosa.
- La métrica de **$R^2$ Ajustado** reveló una moderada capacidad explicativa, lo que sugiere que no todas las variables predictoras contribuyen de manera significativa.
- El **Random Forest Regressor** mostró un rendimiento menos consistente, indicando que la variable objetivo presenta patrones complejos que no son completamente capturados.
- En resumen, el dataset es adecuado para análisis exploratorios y predictivos de niveles de glucosa, aunque su capacidad para regresión puede mejorar con ingeniería de características.

# Conclusión: Desempeño en Regresión




Resultados SMOTE gráfico de barras

# QA

¿Preguntas?