Universidad Tecnológica de Panamá

Maestría en Analítica de Datos

Curso:

Modelos Predictivos

Proyecto Final

“Relación entre el estilo de vida y
marcadores clínicos en pacientes con diabetes tipo 2”

Facilitador:

Juan Marcos Castillo, PhD

Estudiante:

De León, Stephanie 8-856-600

Año

2025

Tabla de Contenido

1. Introducción

La diabetes es una enfermedad metabólica crónica caracterizada por niveles elevados de glucosa en sangre (o azúcar en sangre), que con el tiempo conduce a daños graves en el corazón, los vasos sanguíneos, los ojos, los riñones y los nervios.

La más común es la diabetes tipo 2, generalmente en adultos, que ocurre cuando el cuerpo se vuelve resistente a la insulina o no produce suficiente insulina. En las últimas tres décadas, la prevalencia de la diabetes tipo 2 ha aumentado drásticamente en países de todos los niveles de ingresos. La diabetes tipo 1, una vez conocida como diabetes juvenil o diabetes insulinodependiente, es una afección crónica en la que el páncreas produce poca o ninguna insulina por sí mismo. Para las personas que viven con diabetes, el acceso a un tratamiento asequible, incluida la insulina, es fundamental para su supervivencia.

Este proyecto de investigación analiza la base de datos “Diabetes Dataset”, disponible en Kaggle, que recopila información de pacientes mujeres de al menos 21 años de ascendencia indígena Pima. Los informes con datos clínicos y estilos de vida de pacientes con esta enfermedad permiten identificar patrones de riesgo antes que la enfermedad se manifieste, permitiendo a los profesionales de la salud priorizar pacientes en riesgo y realizar pruebas confirmatorias con mayor anticipación. Se aplicarán diversos modelos predictivos que ayudan a identificar patrones en los marcadores clínicos a fin de permitir la detección temprana de la enfermedad.

2. Justificación

Según cifras de la Organización Panamericana de la Salud, 112 millones de adultos (de 18 años o más) viven con diabetes en las Américas; esta cifra se ha triplicado en la Región desde 1990. La prevalencia ha aumentado más rápidamente en los países de ingresos bajos y medianos que en los de ingresos altos.

Su detección temprana es crucial para prevenir complicaciones graves, como enfermedades cardiovasculares, daño renal y neuropatías. Sin embargo, los modelos predictivos tradicionales suelen presentar limitaciones en su capacidad para identificar correctamente a los pacientes en riesgo, especialmente en términos de falsos negativos, lo que puede retrasar intervenciones médicas oportunas.

El punto de partida para vivir bien con diabetes es un diagnóstico temprano: cuanto más tiempo viva una persona con diabetes no diagnosticada y no tratada, es probable que sus resultados de salud sean peores. Por lo tanto, la justificación de esta investigación está dirigida a evaluar diversos modelos predictivos basados en marcadores clínicos y variables relacionadas con el estilo de vida, que permitan optimizar la detección temprana de la enfermedad y reduzcan los erros críticos, conocidos como falsos negativos (FN) que ocurre cuando se asigna incorrectamente a un paciente como “no enfermo” (negativo), cuando en realidad sí tiene la enfermedad (positivo).

3. Antecedentes

La diabetes tipo 2 es un problema de salud global que afecta a más de 400 millones de personas en el mundo, según la Organización Mundial de la Salud (OMS). En la Región, la diabetes (incluyendo la enfermedad renal relacionada a diabetes) causa al año la pérdida aproximadamente ocho millones de años de vida por muerte prematura.

El aumento expansivo de la epidemia de diabetes va de la mano con el incremento de sus factores de riesgo. Las Américas es la región con más sobrepeso/obesidad e inactividad física del mundo: 68 de cada 100 adultos tiene sobrepeso u obesidad y 36 de cada 100 personas tienen un nivel de actividad física insuficiente.

Al analizar grandes cantidades de datos clínicos (como niveles de glucosa, presión arterial, historial médico, hábitos de vida), los modelos predictivos pueden identificar patrones que podrían no ser evidentes a simple vista. Esto permite que los profesionales de la salud actúen de manera preventiva, antes de que surjan problemas graves.

En esta investigación se evaluaron cuatro modelos:

- Regresión Logística básica y optimizada: para definir un punto de referencia.
- Random Forest estándar y optimizado: para obtener un mejor equilibrio entre precisión (exactitud de las predicciones correctas) y sensibilidad del modelo (capacidad para identificar correctamente los casos positivos).

Se llevó a cabo una comparación entre los modelos para determinar cuál ofrecía el mejor rendimiento en términos de predicción y generalización en el conjunto de datos. La optimización de los modelos incluyó ajustes en hiperparámetros clave, lo que permitió mejorar su capacidad para manejar los datos de manera eficiente y reducir el sobreajuste, asegurando que los resultados fueran lo más confiables y aplicables posible en un contexto real.

4. Definición del problema

La diabetes tipo 2 es una enfermedad prevenible y manejable si se detecta a tiempo, pero muchos pacientes no son diagnosticados hasta que presentan complicaciones graves.

Esta investigación tiene como objetivo encontrar la relación entre el estilo de vida y marcadores clínicos en pacientes con diabetes tipo 2, utilizando modelos de machine learning que ayuden a incrementar la detección de casos reales, con bajos niveles de falsos niveles.

Para esto, se evaluó el desempeño de cuatro modelos:

1. **Regresión Logística**: método tradicional y fácil de interpretar, pero que tiene limitaciones para modelar relaciones complejas.
2. **Regresión Logística Optimizada**: incrementa la detección de casos reales, pero reduce la precisión global del modelo, lo que podría incrementar la cantidad de falsos positivos.
3. **Random Forest:** se obtiene un equilibrio más favorable, aumentando detección de casos reales.
4. **Random Forest Optimizado**: Maximiza la cantidad de casos reales detectados y aumenta la presión del modelo, brindando los mejores resultados en la investigación.

Este enfoque no solo resuelve el problema de clasificación, sino que también proporciona a los médicos una herramienta más confiable para identificar pacientes en riesgo, facilitando intervenciones tempranas y mejorando los resultados de salud.

5. Análisis Predictivo

5.1 Determinación de la base de datos

El conjunto de datos utilizado para este análisis predictivo proviene de Kaggle (Diabetes Dataset), el cual contiene registros médicos de 768 pacientes mujeres de al menos 21 años de ascendencia indígena Pima, con y sin diagnóstico de diabetes pacientes La base de datos incluye 9 variables:

Tabla 1. Descripción de variables

Nombre de la Variable	Tipo de Dato	Tipo de Variable	Descripción
Pregnancies	Entero	Discreta	Número de embarazos
Glucose	Entero	Continua	Concentración de glucosa en sangre (mg/dL)
BloodPressure	Entero	Continua	Presión arterial diastólica (mm Hg)
SkinThickness	Entero	Continua	Grosor del pliegue cutáneo (mm)
Insulin	Entero	Continua	Nivel de insulina en sangre (muU/mL)
BMI	Flotante	Continua	IMC (peso en kg / altura² en m).
DiabetesPedigreeFunction	Flotante	Continua	Riesgo genético de diabetes
Age	Entero	Discreta	Edad del paciente (años)
Outcome	Entero	Binaria	Presencia de Diabetes: 1 (sí), 0 (no)

5.2 Pre-procesamiento y limpieza

Antes de aplicar los modelos predictivos, se realizaron las siguientes etapas de preprocesamiento:

1. **Manejo de valores nulos o inconsistentes**:
    - Se verificó la presencia de valores faltantes o ceros en variables críticas (como glucosa, presión arterial, IMC), los cuales fueron reemplazados por la mediana según correspondiera. Esta técnica es recomendada para datos médicos, dado que la mediana es menos sensible a outliers comparado con la media.

2. **Normalización y escalado**:
    - Dado que algunas variables tienen diferentes escalas (ej. glucosa vs. IMC), se aplicó estandarización (StandardScaler) para mejorar el rendimiento de los modelos.

5.3 Análisis Descriptivo

Luego de la limpieza de los datos, se realizó un análisis exploratorio para entender la distribución general de los datos y las correlaciones entre las variables.

- El 65% de los pacientes no tienen la enfermedad (Outcome = *0),* mientras que el restante 35% sí la tiene presente (*Outcome = 1*).
- La glucosa y el IMC son las dos variables que muestran una mayor correlación con la presencia de diabetes.
- Se utilizaron histogramas y gráficos de caja para mostrar la distribución de las variables y la presencia de outliers.

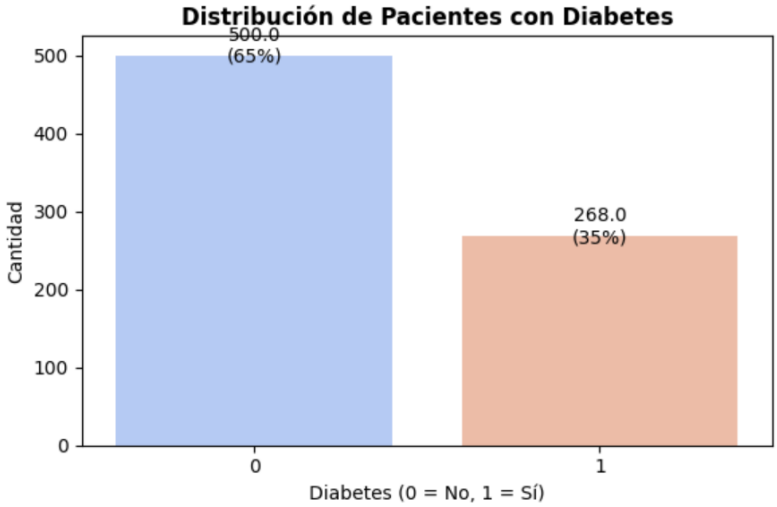
Tabla 2. Resumen Estadístico Descriptivo

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
count	768	768	768	768	768	768	768	768	768
mean	4	121.66	72.39	29.11	140.67	32.46	0.47	33	0
std	3	30.44	12.10	8.79	86.38	6.88	0.33	12	0
min	0	44.00	24.00	7.00	14.00	18.20	0.08	21	0
0.25	1	99.75	64.00	25.00	121.50	27.50	0.24	24	0
0.5	3	117.00	72.00	29.00	125.00	32.30	0.37	29	0
0.75	6	140.25	80.00	32.00	127.25	36.60	0.63	41	1
max	17	199.00	122.00	99.00	846.00	67.10	2.42	81	1

El análisis descriptivo indica que los registros de pacientes con diabetes poseen indicadores clínicos de glucosa media elevada (121.66 mg/dL) y están en el rango de obesidad, mostrando un IMC promedio de 32.46.

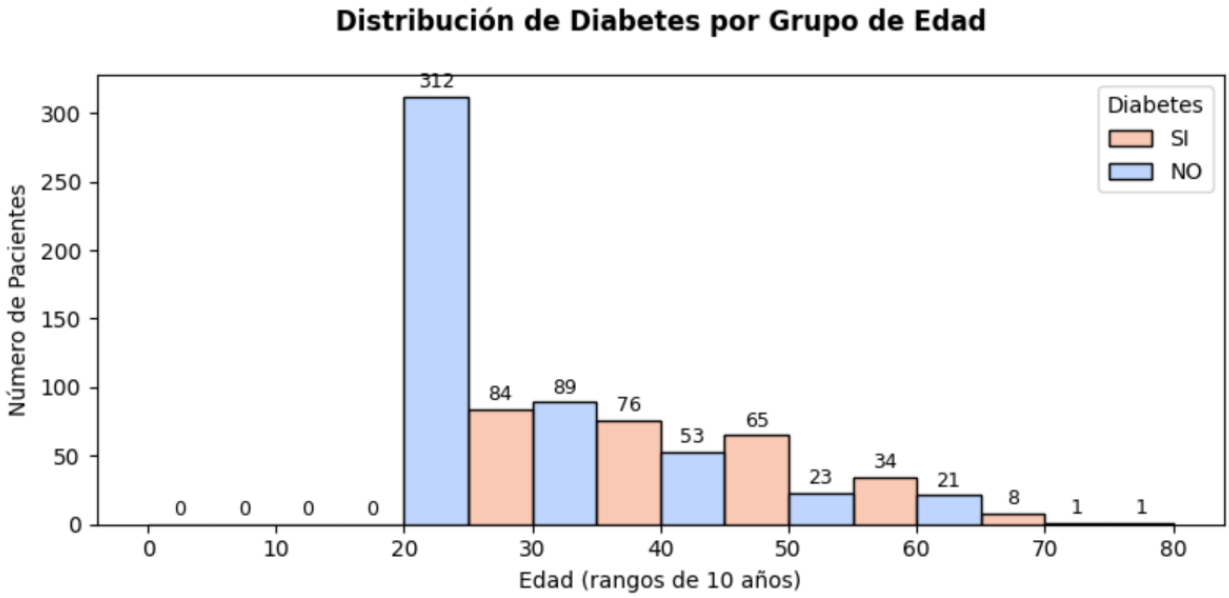
La amplia desviación estándar en insulina (86.38) sugiere alta variabilidad entre pacientes, principalmente en jóvenes menores a 29 años.

Gráfica 1. Distribución de Pacientes con Diabetes



El 65% de los registros del dataset provienen de pacientes que no tienen diabetes, mientras que el 35% restante sí presenta la enfermedad.

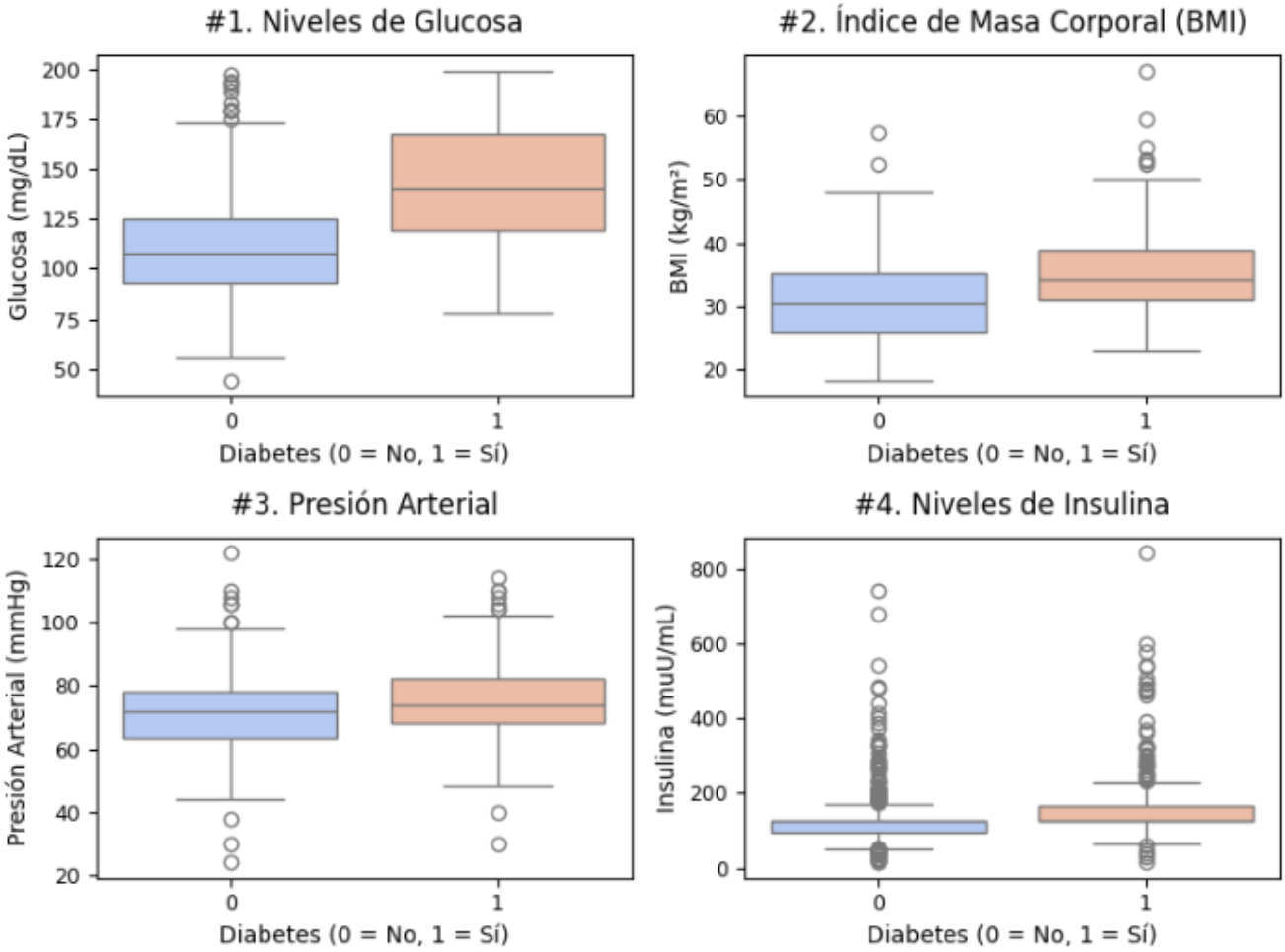
Gráfica 2. Distribución por edad de pacientes con Diabetes**.**



A mayor edad, mayor probabilidad de tener diabetes. Esto coincide con lo que se sabe médicamente: la diabetes tipo 2 es más frecuente en adultos mayores debido a factores como resistencia a la insulina, estilo de vida, etc.

La presencia de la enfermedad es dominante en grupos de edad mayores a 30 años (especialmente 30-50).

Gráfica 3. Gráficas de Caja de las variables numéricas



#1. Niveles de glucosa en sangre:

- Pacientes sin diabetes (0): La mediana está en un rango "normal" de glucosa (ej. 70–110 mg/dL en ayunas).
- Pacientes con diabetes (1): La mediana será significativamente más alta (ej. >126 mg/dL en ayunas).
- En pacientes con diables, el rango de los datos está mayormente ubicado hacia arriba, indicando que algunos pacientes tienen niveles extremadamente altos de glucosa.

#2. Índice de Masa Corporal:

- Pacientes sin diabetes (0): La mayoría tiene un IMC en rango normal (18.5–24.9) o sobrepeso (25–29.9). Hay pocos casos con IMC ≥30 (obesidad), lo que sugiere que la obesidad no es común en este grupo.
- Pacientes con diabetes (1): los datos se ubican hacia más altos (obesidad, ≥30), con presencia de IMC >35, lo que indica obesidad mórbida.

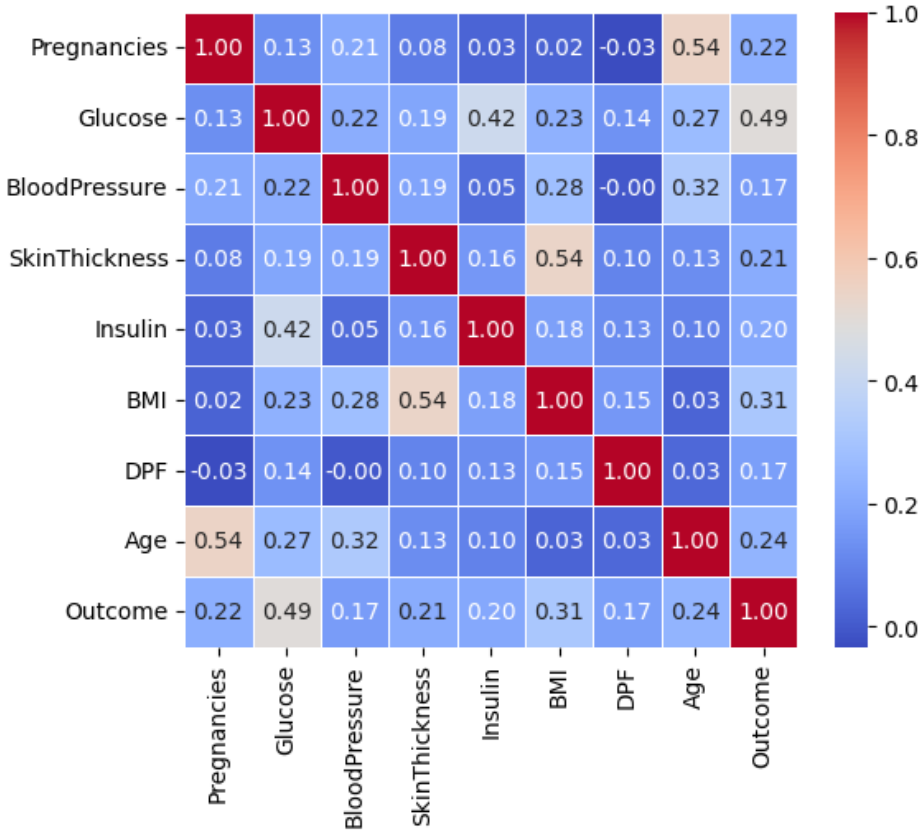
#3. Presión Arterial

- Pacientes sin diabetes (0): presión arterial en rangos normales (120/80 mmHg) o prehipertensión (120–139/80–89 mmHg).
- Pacientes con diabetes (1): mayor dispersión hacia presión arterial (hipertensión ≥140/90 mmHg). Coincide con la evidencia médica, que indica que la resistencia a la insulina y la inflamación crónica dañan los vasos sanguíneos, aumentando la presión arterial.

#4. Niveles de Insulina

- Pacientes con diabetes (1): variabilidad en los niveles de insulina, mostrando casos con insulina elevada (resistencia) y otros son insulina baja (déficit en la producción).
- Pacientes sin diabetes (0): presentan niveles de insulina normales, con una distribución más estrecha, indicando menor variabilidad. Algunas personas sin diabetes (0) pueden mostrar insulina alta lo que podría indicar estados prediabéticos no diagnosticados.

Gráfica 4. Diagrama de Correlación entre las variables



- Variables de mayor impacto en Diabetes: glucosa (0.49), IMC (0.31) y edad (0.24). Estas tres variables son las que más contribuyen al diagnóstico de diabetes en este dataset.
- Correlaciones entre variables independientes: embarazados y edad (0.54), IMC y grosor de piel (0.54) y glucosa e insulina (0.42).
- Variables con baja influencia en Diabetes: Presión arterial (0.17) y DPF (0.17).

5.4 Selección de variables

Al analizar el mapa de correlación se tomó la decisión de aplicar dos corridas en los modelos predictivos: la primera considerando todas las variables y la segunda filtrando las variables que tienen muy baja influencia, como el historial genético (DFP).

5.5 Modelos Predictivos

5.5.1 Regresión Logística (data imputada)

Fue utilizado por ser el modelo comúnmente empleado para clasificación binaria (diabetes: 0 vs 1) y por tener correlaciones claras entre las variables (por ejemplo, glucosa y diabetes).

La configuración utilizada fue la siguiente:

- División datos de prueba y de entrenamiento: 30% de los datos se reservan para prueba (X_test, y_test), que evalúan el rendimiento con datos nunca vistos durante el entrenamiento y 70% se usa para entrenamiento (X_train, y_train), para que el modelo aprenda los patrones.
- Semilla fija para reproducibilidad (random_state): 50 (misma división cada vez que se ejecuta el código).
- Estandarización de variables utilizando fit_transform(X_train) para calcular la media/desviación del train y transform(X_test) para aplicar la misma transformación sin recalcular parámetros.
- Los datos fueron escalados, evitando que algunas variables dominen el modelo solo por tener valores más grandes.
- El máximo de iteraciones fue 1000, para permitir que el algoritmo tenga suficientes iteraciones para alcanzar convergencia, que se refiere al momento en el que el algoritmo de optimización encuentra una solución estable donde los parámetros del modelo (coeficientes) ya no cambian significativamente entre iteraciones.

Las métricas de evaluación aplicadas fueron las siguientes:

- **Precisión:** proporción de predicciones correctas (exactitud).
- **Matriz de confusión**: desglose en Verdaderos Positivos/Negativos y Falsos Positivos/Negativos.
- **precisión**: capacidad de no clasificar falsos positivos.
- **Recall**: capacidad de detectar todos los positivos reales.
- **F1-score**: media armónica de precisión y recall (balance para clases desequilibradas).

5.5.1.1 Resultados

Precisión: 0.76 (el modelo acierta el 76% de las veces)

Tabla 3. Matriz de Confusión RL

Real	Predicción: 0 (No Diabetes)	Predicción: 1 (Diabetes)
No (0)	133 (Verdaderos Negativos)	15 (Falsos Positivos)
Si (1)	**41 (Falsos Negativos)**	42 (Verdaderos Positivos)

- **Falsos Positivos (15)**: Personas sin diabetes predichas como diabéticas.
- **Falsos Negativos (41)**: Personas con diabetes predichas como no diabéticas (considerado como error grave).

Tabla 4. Reporte de Clasificación RL

Real	Precisión	Recall	F1-score	Soporte
No (0)	0.76	0.90	0.83	148
Si (1)	0.74	0.51	0.60	83
Precisión			0.76	231
Promedio	0.75	0.70	0.71	231
Promedio ponderado	0.75	0.76	0.74	231

- De todos los que el modelo predijo como "No diabetes", el 76% realmente no la tenían.
- De todos los casos reales de "No diabetes", el modelo identificó el 90%.
- De todos los que el modelo predijo como "Diabetes", el 74% realmente la tenían.
- De todos los casos reales de diabetes, el modelo solo identificó el 51%.

Análisis:

- El modelo es mejor prediciendo "No diabetes" (0) que "Diabetes" (1).
- Al tener un recall bajo para diabetes (51%), hace que el modelo falle en detectar el 48% de los casos reales de diabetes (falsos negativos).
- Estos resultados pueden estar ocasionados por un desbalance de clases (más casos de "No diabetes" en los datos).

Tabla 5. Coeficientes de Correlación RL

Variable	Coeficiente
Glucose	1.196562
BMI	0.742767
Pregnancies	0.409174
DiabetesPedigreeFunction	0.233021
Age	0.196053
SkinThickness	0.014958
Insulin	-0.155298
BloodPressure	-0.217999

- Glucose y BMI son los factores de riesgo más fuertes (coeficientes > 0.8).
- BloodPressure tiene coeficiente negativo, lo que puede significar que hay valores atípicos en la data.

5.5.2 Regresión Logística Optimizada

Se utilizó la técnica de class_weight='balanced' para corregir el desbalance en los datos. Cada clase recibe un peso inversamente proporcional a su frecuencia en los datos, a fin de equilibrar la atención del modelo para que no ignore la clase minoritaria.

Adicionalmente se removieron las variables SkinThickness, Insulin y BloodPressure basado en su baja contribución.

- SkinThickness (Coef: -0.015): casi cero impacto en la predicción
- Insulin (Coef: -0.155): aunque la insulina está biológicamente ligada a la diabetes, su coeficiente es bajo y negativo. Adicional muestra redundancia con Glucose (correlación alta, ya que la glucosa depende de la insulina).
- BloodPressure (Coef: -0.218): clínicamente contradictorio (la hipertensión suele asociarse a diabetes).

5.5.2.1 Resultados

Precisión: 0.73 (el modelo acierta el 73% de las veces)

Tabla 6. Matriz de Confusión RL Optimizada

Real	Predicción: 0 (No Diabetes)	Predicción: 1 (Diabetes)
No (0)	114 (Verdaderos Negativos)	34 (Falsos Positivos)
Si (1)	**28 (Falsos Negativos)**	55 (Verdaderos Positivos)

- **Falsos Positivos (34)**: Personas sin diabetes predichas como diabéticas.
- **Falsos Negativos (28)**: Personas con diabetes predichas como no diabéticas (considerado como error grave).

Tabla 7. Reporte de Clasificación RL Optimizada

Real	Precisión	Recall	F1-score	Soporte
No (0)	0.80	0.77	0.79	148
Si (1)	0.62	0.66	0.64	83
Precisión			0.73	231
Promedio	0.71	0.72	0.71	231
Promedio ponderado	0.74	0.73	0.73	231

- De todos los que el modelo predijo como "No diabetes", el 80% realmente no la tenían.
- De todos los casos reales de "No diabetes", el modelo identificó el 77%.
- De todos los que el modelo predijo como "Diabetes", el 62% realmente la tenían.
- De todos los casos reales de diabetes, el modelo identificó el 66%.

Análisis:

- El modelo es mejor prediciendo "No diabetes" (0) que "Diabetes" (1).
- El recall para diabetes del 66%, hace que el modelo falle en detectar el 34% de los casos reales de diabetes (falsos negativos).

Tabla 8. Coeficientes de Correlación RL Optimizada

Variable	Coeficiente
Glucose	0.987362
BMI	0.537756
Pregnancies	0.311427
DiabetesPedigreeFunction	0.155466
Age	0.111975

- Por cada aumento unitario en glucosa (escalada), el log-odds de diabetes incrementa 0.9449.
- Glucose es el factor de riesgo más fuerte (coeficiente > 0.9).

Conclusiones

- Detecta más casos reales, con un aumento de Recall de 48% a 67%.
- Menos variables, lo que lo convierte más simple e interpretable.
- La precisión bajó solo un punto, de 74% a 73%
- Precisión para diabetes bajó al 62%: más personas sanas serán evaluadas innecesariamente.

5.5.3 Random Forest (Data imputada)

Este modelo suele funcionar mejor que la regresión logística cuando hay desbalance de clases y relaciones no lineales entre variables.

Tabla 9. Parámetros clave RF

Parámetro	Valor Usado	Análisis
class_weight='balanced'	-	Equilibra automáticamente el peso de las clases (diabetes vs no diabetes).
n_estimators=200	200	Más árboles = mejor rendimiento (pero más lento). Ideal entre 100-300.
max_depth=5	5	Evita árboles muy complejos (controla overfitting).
min_samples_split=10	10	Obliga a que cada nodo tenga al menos 10 muestras para dividirse (más robusto).

Resultados esperados

1. **Mejor recall para diabetes**: suele capturar mejor las relaciones no lineales.
2. **Posible mejora en F1-score**: Mejor equilibrio entre precisión y recall que la regresión logística.

5.5.3.1 Resultados

Precisión: 0.77 (el modelo acierta el 77% de las veces)

Tabla 10. Matriz de Confusión RF

Real	Predicción: 0 (No Diabetes)	Predicción: 1 (Diabetes)
No (0)	120 (Verdaderos Negativos)	28 (Falsos Positivos)
Si (1)	**25 (Falsos Negativos)**	58 (Verdaderos Positivos)

- **Falsos Positivos (28)**: Personas sin diabetes predichas como diabéticas.
- **Falsos Negativos (29)**: Personas con diabetes predichas como no diabéticas (considerado como error grave).

Tabla 11. Reporte de Clasificación RF

Real	Precisión	Recall	F1-score	Soporte
No (0)	0.83	0.81	0.82	148
Si (1)	0.67	0.70	0.69	83
Precisión			0.77	231
Promedio	0.75	0.75	0.75	231
Promedio ponderado	0.77	0.77	0.77	231

- De todos los que el modelo predijo como "No diabetes", el 83% realmente no la tenían.
- De todos los casos reales de "No diabetes", el modelo identificó el 81%.
- De todos los que el modelo predijo como "Diabetes", el 67% realmente la tenían.
- De todos los casos reales de diabetes, el modelo identificó el 70%.

Tabla 12. Coeficientes de Correlación RF

Variable	Coeficiente
Glucose	0.418295
BMI	0.227236
Age	0.179241
DiabetesPedigreeFunction	0.100013
Pregnancies	0.075214

- Glucose es el factor de riesgo más fuerte (coeficiente > 0.4).

Conclusiones:

- Precisión Global (Accuracy): 77%.
- El modelo detecta 67% de los casos reales de diabetes.

5.5.4 Random Forest Optimizado (Data imputada)

El principal objetivo es detectar más casos reales de diabetes mientras se controla el overfitting.

Consideraciones:

- Probar múltiples combinaciones de hiperparámetros (n_estimators, max_depth,min_samples_split) para encontrar la configuración óptima.
- Usa validación cruzada (5 folds) para asegurar que el modelo generalice bien.

5.5.4.1 Resultados

Precisión: 0.77 (el modelo acierta el 77% de las veces)

Tabla 13. Matriz de Confusión RF Optimizado

Real	Predicción: 0 (No Diabetes)	Predicción: 1 (Diabetes)
No (0)	119 (Verdaderos Negativos)	29 (Falsos Positivos)
Si (1)	**23 (Falsos Negativos)**	60 (Verdaderos Positivos)

- **Falsos Positivos (29)**: Personas sin diabetes predichas como diabéticas.
- **Falsos Negativos (23)**: Personas con diabetes predichas como no diabéticas (considerado como error grave).

Tabla 14. Reporte de Clasificación RL Optimizado

Real	Precisión	Recall	F1-score	Soporte
No (0)	0.84	0.80	0.82	148
Si (1)	0.67	0.72	0.70	83
Precisión			0.77	231
Promedio	0.76	0.76	0.76	231
Promedio ponderado	0.78	0.77	0.78	231

- De todos los que el modelo predijo como "No diabetes", el 84% realmente no la tenían.
- De todos los casos reales de "No diabetes", el modelo identificó el 80%.

- De todos los que el modelo predijo como "Diabetes", el 67% realmente la tenían.
- De todos los casos reales de diabetes, el modelo identificó el 72%.

Tabla 15. Coeficientes de Correlación RL Optimizado

Variable	Coeficiente
Glucose	0.422408
BMI	0.231349
Age	0.175163
DiabetesPedigreeFunction	0.096644
Pregnancies	0.074437

- Glucose es el factor de riesgo más fuerte (coeficiente > 0.4).

5.5.5 Regresión Logística (Datos Filtrados)

Al identificar la presencia de valores ceros en variables críticas (como glucosa, presión arterial, IMC) se decidió imputar los valores, reemplazando los ceros por la mediana correspondiente. Al ver que los porcentajes de precisión en los diversos modelos predictivos aplicado todavía se encontraban bajos, se tomó la decisión de replicar el modelo de Regresión Logística, pero filtrando los registros con valores cero en vez de imputarlos; lo cual generó mejores resultados en el modelo.

5.5.5.1 Resultados

Precisión: 0.84 (el modelo acierta el 84% de las veces)

Tabla 16. Matriz de Confusión RL (datos filtrados)

Real	Predicción: 0 (No Diabetes)	Predicción: 1 (Diabetes)
No (0)	75 (Verdaderos Negativos)	9 (Falsos Positivos)
Si (1)	**10 (Falsos Negativos)**	24 (Verdaderos Positivos)

- **Falsos Positivos (9)**: Personas sin diabetes predichas como diabéticas.
- **Falsos Negativos (10)**: Personas con diabetes predichas como no diabéticas (considerado como error grave).

Tabla 17. Reporte de Clasificación RL (datos filtrados)

Real	Precisión	Recall	F1-score	Soporte
No (0)	0.88	0.89	0.89	84
Si (1)	0.73	0.71	0.72	34
Precisión			0.84	118
Promedio	0.80	0.80	0.80	118
Promedio ponderado	0.84	0.84	0.84	118

- De todos los que el modelo predijo como "No diabetes", el 88% realmente no la tenían.
- De todos los casos reales de "No diabetes", el modelo identificó el 89%.

- De todos los que el modelo predijo como "Diabetes", el 73% realmente la tenían.
- De todos los casos reales de diabetes, el modelo identificó el 71%.

Tabla 18. Coeficientes de Correlación RL (datos filtrados)

Variable	Coeficiente
Glucose	1.094017
Age	0.422807
DiabetesPedigreeFunction	0.322549
BMI	0.272442
Pregnancies	0.159321
SkinThickness	0.149535
BloodPressure	-0.017029
Insulin	-0.052020

- Glucose es el factor de riesgo más fuerte (coeficiente > 1).

5.6 Comparación de resultados de los modelos predictivos

Tabla 19. Comparación de resultados de las principales métricas de los modelos predictivos

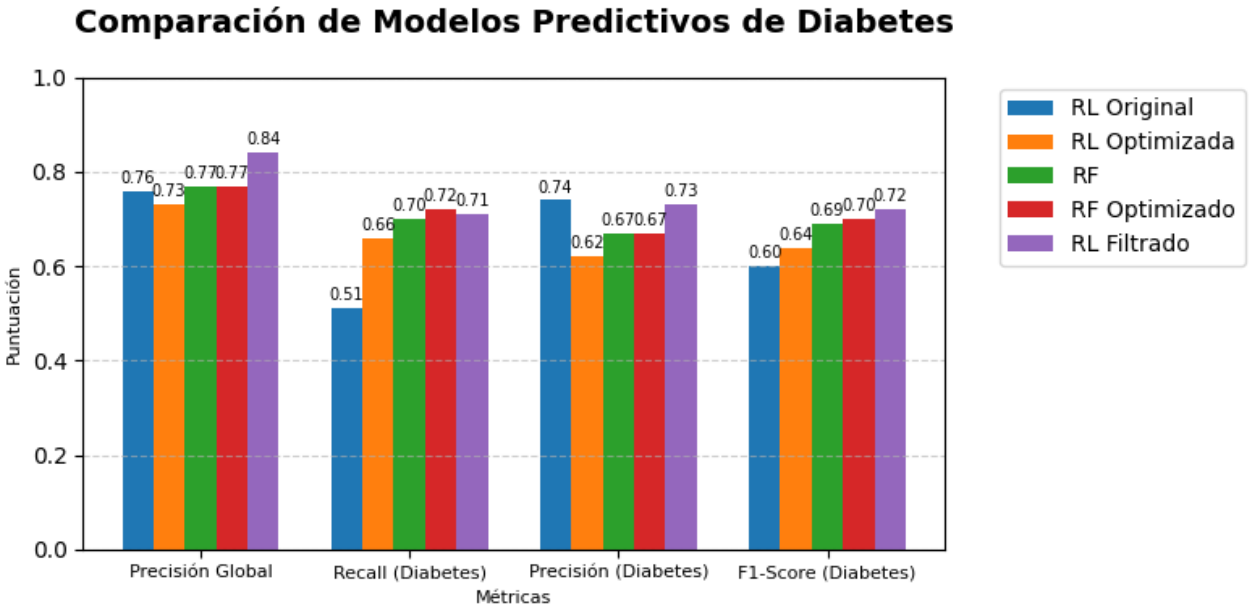
	Data Imputada				Data filtrada
Métrica	**Reg.Log**	**Reg.Log Optimizada**	**Random Forest**	**RF Optimizado**	**Reg.Log**
Precisión Global	76%	73%	77%	77%	**84%**
Recall (Diabetes)	51%	66%	70%	72%	**71%**
Precisión (Diabetes)	74%	62%	67%	67%	**73%**
F1-Score (Diabetes)	60%	64%	69%	70%	**72%**
Falsos Negativos	41	28	25	23	**10**

Random Forest supera los resultados de la Regresión Logística en métricas clave como Recall (sensibilidad) y F1-Score para la clase "Diabetes", lo que indica un mejor equilibrio entre precisión y capacidad para detectar casos positivos. Un Recall bajo significa muchos falsos negativos, lo que es inaceptable en un modelo de diagnóstico.

La Regresión Logística con datos filtrados logra la mayor precisión global (84%), lo que sugiere que la limpieza de datos puede ser crucial para mejorar el rendimiento.

La optimización mejora significativamente el Recall (sensibilidad) de 51% a 71%, reduciendo drásticamente falsos negativos (de 41 a 10), lo cual es clave para evitar retrasos en el diagnóstico y tratamiento.

Gráfica 5. Comparación de resultados de los Modelos Predictivos



6. Conclusiones

Para los estudios de pacientes con diabetes, reducir los casos de fasos negativos (no se diagnostica a tiempo la enfermedad) es crucial porque se evita que el paciente desarrolle otras complicaciones a partir de la diabetes. Además, si la diabetes no se detecta a tiempo, el paciente va a necesitar de tratamientos más complejos y costosos en el futuro, lo que impacta directamente los sistemas de salud pública.

Los modelos Random Forest Optimizado (Recall 72%) o la Regresión Logística con datos filtrados (Recall 71%) son opciones más seguras que una Regresión Logística estándar con Recall bajo (51%). La Regresión Logística con datos filtrados proporcionó los mejores resultados de cantidad de falsos negativos, generando solo 10 casos en la predicción, frente a 23–41 en los otros modelos.
El modelo más robusto es la Regresión Logística con datos filtrados, al lograr la mayor precisión global (84%), un F1-Score competitivo (72%) y un Recall aceptable (71%).

Filtrar los registros que tenían datos incompletos tuvo un mayor impacto comparado con cambiar de algoritmo o ajustar hiperparámetros. La calidad de los datos demostró ser más determinante que la elección del algoritmo o la optimización, demostrando que un modelo simple con datos limpios puede ser más eficiente que un modelo complejo con datos ruidosos. Al aplicar la imputación de datos se pueden introducir distorsiones ocasionados por correlaciones falsas, mientras que si se filtra el dataset para que solamente incluya información consistente y confiable mejora la capacidad del modelo para aprender patrones reales.

7. Recomendaciones y futuros estudios

Dado que la calidad de los datos demostró ser más determinante que la elección del modelo predictivo o su optimización se recomienda tener validadores en el proceso de recolección de datos, que restrinjan el ingreso de valores cero en variables críticas.
Por otro lado, se recomienda ampliar el estudio, incluyendo más registros y a adicionalmente, agregar otras variables que mejoren las predicciones de los modelos, como lo son detalles del estilo de vida (sedentarismo, dieta, tabaquismo).
El estudio puede expandirse con datos de otros países o grupos étnicos que contribuyan a la generalización del modelo.

El estudio muestra un comportamiento común de Machine Learning llamado "garbage in, garbage out" (basura entra, basura sale), que explica que no importa que tan bueno o complejo sea el algoritmo utilizado para las predicciones, si lo datos no están lo suficientemente limpios y con una estructura robusta, las predicciones no serán lo suficientemente acertadas.

Se puede explotar el uso de otros modelos predictivos a fin de buscar resultados aún más precisos, como lo son:

- Extreme Gradient Boosting (XGBOOST): maneja eficazmente grandes conjuntos de datos con un gran número de variables. Comienza ajustando un modelo inicial a los datos. A continuación, se construye un segundo modelo que se centra en predecir con exactitud las observaciones que el primer modelo predijo mal. Se espera que la combinación de estos dos modelos sea mejor que cada uno de ellos. Este proceso de refuerzo se repite varias veces, y cada modelo sucesivo intenta corregir las deficiencias del conjunto refuerzo combinado que contiene todos los modelos anteriores.
- Red Neuronal Simple (MLP): son un tipo de algoritmo de aprendizaje profundo, aptas para analizar datos secuenciales, están organizadas en capas o nodos, a través de los cuales se transmite la información de manera simultánea en pasos discretos de tiempo.

8. Bibliografía

- Mosquera Ruiz, A. (2023). Experimentos con redes neuronales recurrentes LSTM para la predicción del nivel de glucosa de pacientes con diabetes. *Revista Ontare, 11*, (páginas). DOI.

- XLMiner. (s. f.). *XGBOOST (Extreme Gradient Boosting)*. XLSTAT. Recuperado el 7 de abril de 2025, de: https://www.xlstat.com/es/soluciones/funciones/extreme-gradient-boosting-xgboost

- Mathchi. (2024). *Diabetes dataset* [Data set]. Kaggle. Recuperado de: https://www.kaggle.com/datasets/mathchi/diabetes-data-set/data

- DataCamp Team**.** (n.d.). *Understanding Logistic Regression in Python* [Tutorial]. DataCamp. Recuperado de: https://www.datacamp.com/es/tutorial/understanding-logistic-regression-python

- DataCamp Team. (n.d.). *Random Forests Classifier in Python* [Tutorial]. DataCamp. Recuperado de: https://www.datacamp.com/tutorial/random-forests-classifier-python

- Ciencia de Datos**.** (n.d.). *Random Forest con Python* [Guía]. Ciencia de Datos. Recuperado de: https://cienciadedatos.net/documentos/py08_random_forest_python

9. Anexo

Repositorio de Github:
https://github.com/StephDL-utp/modelos_predictivos/tree/main