Proyecto 1: Técnicas Clasificación

1st Bryan Campos Castro
Alajuela, Costa Rica
bryancampos20@gmail.com

2nd Miguel David Sánchez Sánchez
Heredia, Costa Rica
miguelsanchez712000@gmail.com

***Abstract*—Este proyecto tiene como objetivo aplicar técnicas de clasificación binaria utilizando dos algoritmos populares de Machine Learning: Regresión Logística y K-Nearest Neighbors (KNN). Se utilizan dos conjuntos de datos: el Pima Indians Diabetes Database, que busca predecir si un paciente tiene diabetes basándose en características como niveles de glucosa y presión arterial, y un segundo dataset quirúrgico que clasifica la aparición de complicaciones postoperatorias en pacientes. Ambos datasets fueron explorados y preprocesados, y la división de los datos en conjuntos de entrenamiento (70%), validación (15%) y prueba (15%).**

Los modelos fueron entrenados y evaluados en ambos conjuntos de datos, ajustando sus hiperparámetros para optimizar el rendimiento. Se realizaron evaluaciones utilizando métricas como accuracy, precision, recall, y la matriz de confusión para comparar el desempeño de los modelos. Los resultados obtenidos muestran que tanto la Regresión Logística como KNN son eficaces para la clasificación binaria, aunque el rendimiento varía entre los datasets y dependiendo del ajuste de hiperparámetros. Finalmente, se realiza una comparación detallada entre los modelos para analizar su efectividad en problemas médicos de clasificación binaria.

I. Introducción

En el campo de la inteligencia artificial, las técnicas de clasificación binaria se han vuelto fundamentales para resolver problemas en diversas áreas, especialmente en la medicina, donde se busca predecir el estado de salud de los pacientes a partir de características observadas [1]. Este proyecto se enfoca en aplicar algoritmos de Machine Learning para la clasificación binaria, con el objetivo de predecir la presencia de condiciones médicas específicas en los pacientes.

Se han seleccionado dos conjuntos de datos representativos para abordar este problema: el Pima Indians Diabetes Database [2] y un segundo dataset quirúrgico que clasifica la aparición de complicaciones postoperatorias en pacientes. Ambos conjuntos de datos permiten realizar análisis predictivos basados en una serie de variables que incluyen aspectos clínicos y demográficos de los pacientes.

Para este proyecto, se utilizarán dos algoritmos de clasificación ampliamente conocidos: Regresión Logística y K-Nearest Neighbors (KNN) [3]. La Regresión Logística es un modelo lineal que estima la probabilidad de ocurrencia de un evento binario, mientras que KNN es un algoritmo basado en instancias que clasifica nuevos ejemplos en función de las distancias a los puntos de datos más cercanos. Ambos algoritmos se ajustarán y evaluarán en los dos conjuntos de datos seleccionados.

II. Dataset Surgical binary classification

El *Surgical Binary Classification Dataset* es una excelente opción para compararlo con el *Pima Indians Diabetes Database* por las siguientes razones:

A. Contexto Médico y de Salud

Ambos datasets se centran en temas médicos, lo que permite comparar dos áreas críticas de la salud: complicaciones quirúrgicas y diabetes. El *Pima Indians Diabetes Database* se enfoca en los factores de riesgo para desarrollar diabetes, mientras que el *Surgical Binary Classification Dataset* aborda la aparición de complicaciones postquirúrgicas. Esta similitud en el contexto médico facilita la aplicación de modelos de clasificación binaria en ambos casos.

B. Clasificación Binaria

En ambos datasets, el objetivo principal es resolver problemas de clasificación binaria. En el caso del *Pima Indians Diabetes Database*, la variable objetivo es *Outcome*, que indica si un paciente tiene o no diabetes (0 o 1). Por otro lado, en el *Surgical Binary Classification Dataset*, se elige la columna *complication* como la variable objetivo, que también tiene dos clases: presencia (1) o ausencia (0) de complicaciones postquirúrgicas. Esta coherencia en el formato de las variables objetivo hace que ambos datasets sean comparables en términos de aplicación de los mismos algoritmos de clasificación binaria, como *Regresión Logística* y *KNN*.

C. Importancia Clínica

Ambos problemas tienen un alto impacto en la salud pública y en la medicina preventiva. El *Pima Indians Diabetes Database* permite prever la probabilidad de que una persona desarrolle diabetes, una enfermedad crónica de gran relevancia a nivel mundial. Del mismo modo, el *Surgical Binary Classification Dataset* se centra en la predicción de complicaciones postquirúrgicas, que son críticas para mejorar los resultados de los pacientes y reducir el costo y la duración de las hospitalizaciones. Comparar estos dos datasets permite estudiar cómo los modelos de clasificación pueden apoyar la toma de decisiones médicas en distintos escenarios.

D. Caracterización de los Pacientes

Tanto en el dataset quirúrgico como en el de diabetes, los pacientes se caracterizan por una serie de factores de riesgo o condiciones previas que influyen en el resultado final (complicaciones o desarrollo de diabetes). Esta similitud en la

estructura del dataset permite aplicar técnicas de exploración de datos y modelos predictivos de manera coherente en ambos casos, proporcionando una comparación sólida entre los factores de riesgo en diferentes contextos médicos.

E. Aplicación de Modelos Predictivos

Ambos datasets permiten la aplicación de los mismos modelos de clasificación, como *K-Nearest Neighbors* y *Regresión Logística*. Estos modelos son adecuados para resolver problemas binarios y analizar cómo diferentes factores contribuyen al desarrollo de complicaciones o enfermedades. Al aplicar los mismos modelos, se puede comparar cómo las características del paciente influyen en los resultados en cada contexto (diabetes vs. complicaciones quirúrgicas), lo que permite generar un análisis robusto.

III. Pipeline

El algoritmo fue programado en *Python* y se utilizaron las librerías de *Pandas* para la manipulación de datos, *Matplotlib* y *Seaborn* para la visualización de los resultados, y *Scikit-learn* para el uso de algoritmos de clasificación como *Regresión Logística* y *K-Nearest Neighbors (KNN)*.

El proceso comienza con la carga y preprocesamiento de los datos de dos conjuntos de datos médicos: *Pima Indians Diabetes Database* y *Surgical Binary Classification Dataset*. En ambos casos, la variable objetivo es binaria, lo que permite realizar una clasificación.

IV. Carga y limpieza de datos

Se utilizan las funciones de *Pandas* para cargar los datos y manejar valores nulos o inconsistencias en los conjuntos de datos.

```
[5 rows x 25 columns]
       Pregnancies     Glucose  BloodPressure  SkinThickness     Insulin  \
count   768.000000  768.000000     768.000000     768.000000  768.000000
mean      3.845052  120.894531      69.105469      20.536458   79.799479
std       3.369578   31.972618      19.355807      15.952218  115.244002
min       0.000000    0.000000       0.000000       0.000000    0.000000
25%       1.000000   99.000000      62.000000       0.000000    0.000000
50%       3.000000  117.000000      72.000000      23.000000   30.500000
75%       6.000000  140.250000      80.000000      32.000000  127.250000
max      17.000000  199.000000     122.000000      99.000000  846.000000

              BMI  DiabetesPedigreeFunction         Age     Outcome
count  768.000000                768.000000  768.000000  768.000000
mean    31.992578                  0.471876   33.240885    0.348958
std      7.884160                  0.331329   11.760232    0.476951
min      0.000000                  0.078000   21.000000    0.000000
25%     27.300000                  0.243750   24.000000    0.000000
50%     32.000000                  0.372500   29.000000    0.000000
75%     36.600000                  0.626250   41.000000    1.000000
max     67.100000                  2.420000   81.000000    1.000000
```

Fig. 1. Resumen datos diabetes

```
[5 rows x 25 columns]
       Pregnancies     Glucose  BloodPressure  SkinThickness     Insulin  \
count   768.000000  768.000000     768.000000     768.000000  768.000000
mean      3.845052  120.894531      69.105469      20.536458   79.799479
std       3.369578   31.972618      19.355807      15.952218  115.244002
min       0.000000    0.000000       0.000000       0.000000    0.000000
25%       1.000000   99.000000      62.000000       0.000000    0.000000
50%       3.000000  117.000000      72.000000      23.000000   30.500000
75%       6.000000  140.250000      80.000000      32.000000  127.250000
max      17.000000  199.000000     122.000000      99.000000  846.000000

              BMI  DiabetesPedigreeFunction         Age     Outcome
count  768.000000                768.000000  768.000000  768.000000
mean    31.992578                  0.471876   33.240885    0.348958
std      7.884160                  0.331329   11.760232    0.476951
min      0.000000                  0.078000   21.000000    0.000000
25%     27.300000                  0.243750   24.000000    0.000000
50%     32.000000                  0.372500   29.000000    0.000000
75%     36.600000                  0.626250   41.000000    1.000000
max     67.100000                  2.420000   81.000000    1.000000

                bmi           Age    asa_status  baseline_cancer  \
count  14635.000000  14635.000000  14635.000000     14635.000000
mean      31.295642     63.205268      0.632320         0.262316
std        8.152709     18.088191      0.539952         0.439909
min        2.150000      6.100000      0.000000         0.000000
25%       26.510000     51.500000      0.000000         0.000000
50%       28.980000     59.700000      1.000000         0.000000
75%       35.295000     74.700000      1.000000         1.000000
max       92.590000     90.000000      2.000000         1.000000

       baseline_charlson  baseline_cvd  baseline_dementia  baseline_diabetes  \
count       14635.000000  14635.000000       14635.000000       14635.000000
mean            0.977520      0.620294           0.004851           0.120875
std             1.758355      0.485330           0.069485           0.325993
min             0.000000      0.000000           0.000000           0.000000
25%             0.000000      0.000000           0.000000           0.000000
50%             0.000000      1.000000           0.000000           0.000000
75%             2.000000      1.000000           0.000000           0.000000
max            13.000000      1.000000           1.000000           1.000000

       baseline_digestive  baseline_osteoart  ...  complication_rsi  \
count        14635.000000       14635.000000  ...      14635.000000
mean             0.189546           0.342740  ...         -0.699044
std              0.391955           0.474642  ...          1.339394
min              0.000000           0.000000  ...         -4.720000
25%              0.000000           0.000000  ...         -1.970000
50%              0.000000           0.000000  ...         -0.580000
75%              0.000000           1.000000  ...          0.000000
max              1.000000           1.000000  ...         12.560000

                dow        gender          hour         month     moonphase  \
count  14635.000000  14635.000000  14635.000000  14635.000000  14635.000000
mean       1.606970      0.548890     10.171613      5.915408      1.187086
std        1.497738      0.497621      2.659881      3.239825      1.158357
min        0.000000      0.000000      6.070000      0.000000      0.000000
25%        0.000000      0.000000      7.820000      3.000000      0.000000
50%        1.000000      1.000000      9.120000      7.000000      1.000000
75%        3.000000      1.000000     12.050000      8.000000      2.000000
max        4.000000      1.000000     18.920000     11.000000      3.000000
```

Fig. 2. Primeras filas del dataset quirúrgico

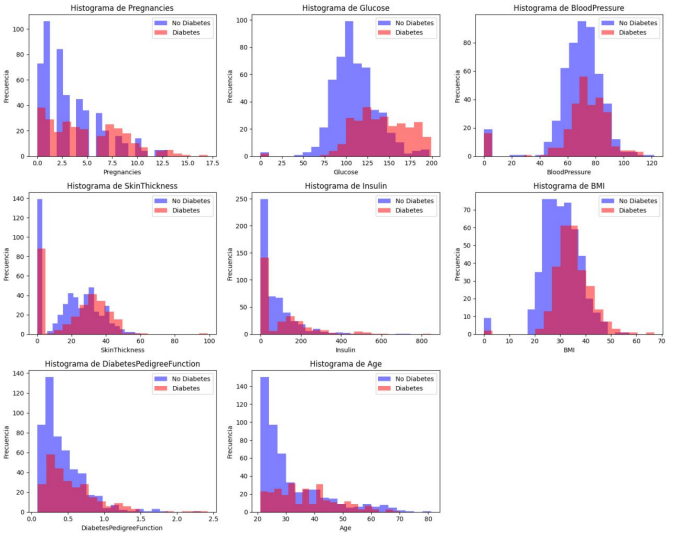

Fig. 3. Histogramas diabetes

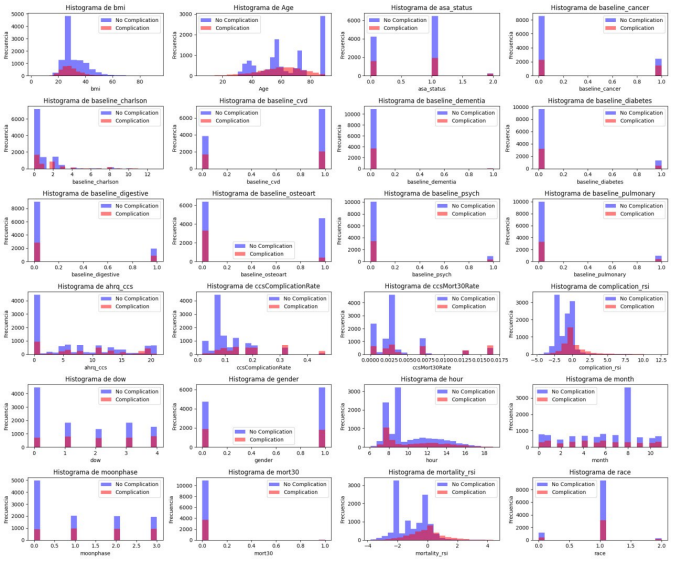

Fig. 4. Histograma surgical

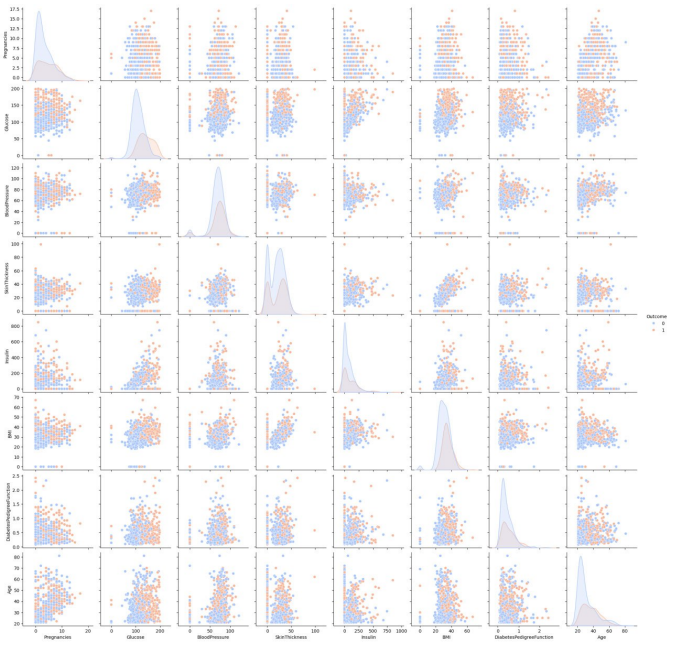

Fig. 5. Comparacion de datos dos dimenciones diabetes

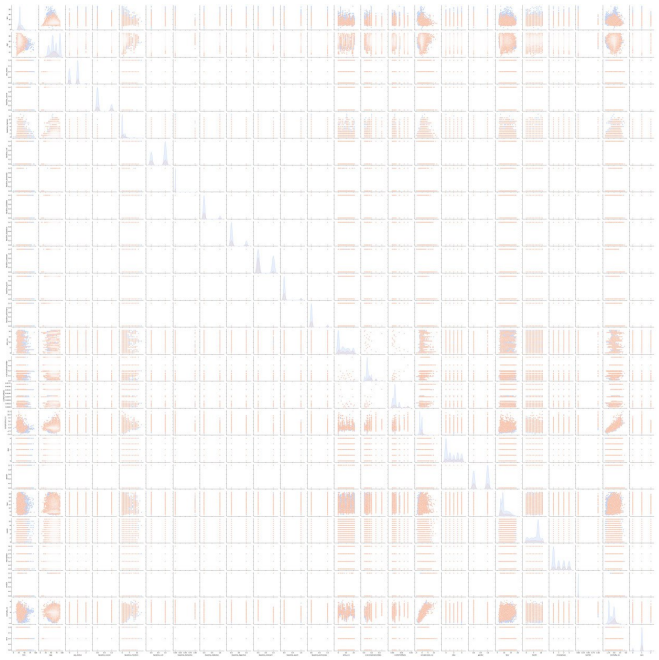

Fig. 6. Comparacion de datos dos dimenciones surgical

V. División de los datos

Se separan los datos en conjuntos de entrenamiento, validación y prueba en una proporción de 70% para entrenamiento, 15% para validación y 15% para prueba. En el dataset de diabetes, se utiliza la columna *Outcome*, mientras que en el dataset quirúrgico se utiliza *complication*.

Además, realizamos una reducción de datos en el *Surgical Binary Classification Dataset* basada en un análisis visual de los histogramas de dos dimensiones. Observamos que varias características compartían distribuciones muy similares, lo que indicaba redundancia en la información aportada. A criterio visual, decidimos eliminar las características cuya distribución era prácticamente indistinguible de otras, manteniendo aquellas con diferencias claras en su comportamiento. Esta reducción de características permitió simplificar el modelo y mejorar la eficiencia en el procesamiento de datos, sin comprometer la calidad de las predicciones.

La figura 6 muestra el histograma bidimensional de las características seleccionadas en el *Surgical Binary Classification Dataset*, donde se identificaron aquellas cuya distribución era indistinguible de otras. Este análisis visual fue clave para decidir qué características eliminar y cuáles mantener para mejorar la eficiencia del modelo.

VI. Aplicación de algoritmos

A. Regresión Logística y KNN

Se ejecutan los algoritmos de *Regresión Logística* y *KNN* en ambos conjuntos de datos para diferentes combinaciones de características, evaluando cómo cambian las métricas de rendimiento con distintas configuraciones.

B. Evaluación de rendimiento

Para cada combinación, se calculan las métricas de *Accuracy*, *Precision*, *Recall*, *F1-Score* y se genera la *Matriz de Confusión* para evaluar la calidad de las predicciones.

a) Por qué utilizamos el F1-Score: El F1-Score es una métrica fundamental en tareas de clasificación, especialmente cuando se trabaja con datasets desbalanceados o cuando la importancia de los falsos positivos y falsos negativos varía. En nuestro caso, el dataset quirúrgico puede presentar un desequilibrio en las clases, lo que haría que métricas como la precisión o el recall por separado no capturen adecuadamente el rendimiento del modelo. El F1-Score combina tanto la precisión como el recall en una única métrica, proporcionando un balance entre estos dos aspectos. De este modo, es particularmente útil cuando las clases no están balanceadas, ya que garantiza que tanto los positivos predichos correctamente como los falsos positivos y falsos negativos se consideren en la evaluación. Al centrarnos en el F1-Score, podemos seleccionar combinaciones de características que ofrezcan un rendimiento más equilibrado y robusto en nuestro modelo, haciendo que el sistema sea más confiable.

1) Resultados de la Regresión Lineal para el Dataset de Diabetes: Al comparar los resultados de regresión lineal y KNN en los datasets de Diabetes y Quirúrgico, se observa que KNN generalmente ofrece un mejor rendimiento en términos de precisión (*Precision*) y F1-Score, especialmente en el dataset quirúrgico, donde alcanza una **Precision** cercana al 93%. Por otro lado, la regresión lineal tiende a ser más consistente en el recall, mostrando valores más equilibrados entre *Precision* y *Recall*, aunque su F1-Score es menor en ambos datasets. En términos generales, KNN se destaca por capturar mejor la relación entre las variables en ambos casos, pero la regresión lineal ofrece un enfoque más balanceado entre la sensibilidad y la especificidad de las predicciones.

a) Top 5 combinaciones de características ordenadas por F1-Score:

1) **Combinación**: ('Pregnancies', 'Glucose', 'BloodPressure', 'BMI')
   Accuracy: 0.7739
   Precision: 0.6486
   Recall: 0.6486
   F1-Score: 0.6486
   Matriz de Confusión en Validación:
$$\begin{bmatrix} 65 & 13 \\ 13 & 24 \end{bmatrix}$$
2) **Combinación**: ('Pregnancies', 'Glucose', 'BloodPressure', 'Insulin', 'BMI')
   Accuracy: 0.7739
   Precision: 0.6486
   Recall: 0.6486
   F1-Score: 0.6486
   Matriz de Confusión en Validación:
$$\begin{bmatrix} 65 & 13 \\ 13 & 24 \end{bmatrix}$$
3) **Combinación**: ('Pregnancies', 'Glucose', 'BloodPressure', 'BMI', 'DiabetesPedigreeFunction')
   Accuracy: 0.7739
   Precision: 0.6486
   Recall: 0.6486
   F1-Score: 0.6486
   Matriz de Confusión en Validación:
$$\begin{bmatrix} 65 & 13 \\ 13 & 24 \end{bmatrix}$$
4) **Combinación**: ('Pregnancies', 'Glucose', 'BMI', 'DiabetesPedigreeFunction')
   Accuracy: 0.7826
   Precision: 0.6765
   Recall: 0.6216
   F1-Score: 0.6479
   Matriz de Confusión en Validación:
$$\begin{bmatrix} 67 & 11 \\ 14 & 23 \end{bmatrix}$$
5) **Combinación**: ('Pregnancies', 'Glucose', 'Insulin', 'BMI', 'DiabetesPedigreeFunction')
   Accuracy: 0.7826
   Precision: 0.6765
   Recall: 0.6216
   F1-Score: 0.6479
   Matriz de Confusión en Validación:
$$\begin{bmatrix} 67 & 11 \\ 14 & 23 \end{bmatrix}$$

Los resultados muestran que las características más relevantes para predecir diabetes son **Glucose**, **BMI**, y **Pregnancies**, presentes en todas las combinaciones de mayor F1-Score. Otros factores como **BloodPressure**, **Insulin**, y **DiabetesPedigreeFunction** también contribuyen, pero en menor medida. En conjunto, estos atributos permiten un buen balance entre precisión y sensibilidad en el modelo de regresión lineal.

2) Resultados de la Regresión Lineal para el Dataset Quirúrgico:

a) Top 5 combinaciones de características quirúrgicas ordenadas por F1-Score:

1) **Combinación**: `('bmi', 'Age', 'ccsComplicationRate', 'complication_rsi', 'mortality_rsi')`
   Accuracy: 0.8024
   Precision: 0.6920
   Recall: 0.3537
   F1-Score: 0.4681
   Matriz de Confusión en Validación:
$$\begin{bmatrix} 1571 & 85 \\ 349 & 191 \end{bmatrix}$$
2) **Combinación**: `('bmi', 'Age', 'ccsComplicationRate', 'ccsMort30Rate', 'complication_rsi', 'mortality_rsi')`
   Accuracy: 0.8024

Precision: 0.6920
Recall: 0.3537
F1-Score: 0.4681
Matriz de Confusión en Validación:

$$\begin{bmatrix} 1571 & 85 \\ 349 & 191 \end{bmatrix}$$

3) **Combinación**: `('bmi', 'Age', 'ccsComplicationRate', 'complication_rsi', 'hour', 'mortality_rsi')`
Accuracy: 0.8024
Precision: 0.6920
Recall: 0.3537
F1-Score: 0.4681
Matriz de Confusión en Validación:

$$\begin{bmatrix} 1571 & 85 \\ 349 & 191 \end{bmatrix}$$

4) **Combinación**: `('bmi', 'Age', 'ccsComplicationRate', 'ccsMort30Rate', 'complication_rsi', 'hour', 'mortality_rsi')`
Accuracy: 0.8024
Precision: 0.6920
Recall: 0.3537
F1-Score: 0.4681
Matriz de Confusión en Validación:

$$\begin{bmatrix} 1571 & 85 \\ 349 & 191 \end{bmatrix}$$

5) **Combinación**: `('bmi', 'Age', 'baseline_charlson', 'ccsComplicationRate', 'complication_rsi', 'mortality_rsi')`
Accuracy: 0.8010
Precision: 0.6833
Recall: 0.3556
F1-Score: 0.4677
Matriz de Confusión en Validación:

$$\begin{bmatrix} 1567 & 89 \\ 348 & 192 \end{bmatrix}$$

Los resultados indican que las características más relevantes para predecir complicaciones quirúrgicas son **bmi**, **Age**, y **ccsComplicationRate**, presentes en todas las combinaciones con mejor F1-Score. Factores como **complication_rsi** y **mortality_rsi** también contribuyen significativamente, mientras que otros como **ccsMort30Rate** y **baseline_charlson** tienen una influencia menor. Estas características permiten capturar la complejidad de los resultados quirúrgicos, con un balance entre precisión y sensibilidad moderado.

3) Resultados del KNN para el Dataset de Diabetes:

a) Top 5 combinaciones de características de KNN Diabetes ordenadas por F1-Score:

1) **Combinación**: ('Pregnancies', 'Glucose', 'BMI')
Accuracy: 0.7652
Precision: 0.6316
Recall: 0.6486
F1-Score: 0.64
Matriz de Confusión en Validación:

$$\begin{bmatrix} 64 & 14 \\ 13 & 24 \end{bmatrix}$$

2) **Combinación**: ('Pregnancies', 'Glucose', 'BMI', 'DiabetesPedigreeFunction')
Accuracy: 0.7652
Precision: 0.6316
Recall: 0.6486
F1-Score: 0.64
Matriz de Confusión en Validación:

$$\begin{bmatrix} 64 & 14 \\ 13 & 24 \end{bmatrix}$$

3) **Combinación**: ('Glucose', 'SkinThickness', 'DiabetesPedigreeFunction')
Accuracy: 0.7739
Precision: 0.6571
Recall: 0.6216
F1-Score: 0.6389
Matriz de Confusión en Validación:

$$\begin{bmatrix} 66 & 12 \\ 14 & 23 \end{bmatrix}$$

4) **Combinación**: ('Pregnancies', 'Glucose', 'BMI', 'DiabetesPedigreeFunction', 'Age')
Accuracy: 0.7478
Precision: 0.5952
Recall: 0.6757
F1-Score: 0.6329
Matriz de Confusión en Validación:

$$\begin{bmatrix} 61 & 17 \\ 12 & 25 \end{bmatrix}$$

5) **Combinación**: ('Pregnancies', 'Glucose', 'BloodPressure', 'BMI', 'Age')
Accuracy: 0.7565
Precision: 0.6154
Recall: 0.6486
F1-Score: 0.6316
Matriz de Confusión en Validación:

$$\begin{bmatrix} 63 & 15 \\ 13 & 24 \end{bmatrix}$$

Los resultados muestran que las características más importantes para el modelo KNN en la predicción de diabetes son **Pregnancies**, **Glucose**, y **BMI**, que aparecen en todas las combinaciones de mayor F1-Score. Factores adicionales como **DiabetesPedigreeFunction** y **Age** también mejoran el rendimiento del modelo, mientras que **BloodPressure** y **SkinThickness** tienen una menor influencia. En conjunto, estas características permiten capturar un buen equilibrio entre precisión y sensibilidad en el modelo KNN.

4) Resultados del KNN para el Dataset Quirúrgico:

a) Top 5 combinaciones de características de KNN Quirúrgico ordenadas por F1-Score:

1) **Combinación**: `('bmi', 'Age', 'baseline_diabetes', 'complication_rsi')`
   Accuracy: 0.8975
   Precision: 0.9301
   Recall: 0.6349
   F1-Score: 0.7546
   Matriz de Confusión en Validación:
$$\begin{bmatrix} 1624 & 26 \\ 199 & 346 \end{bmatrix}$$
2) **Combinación**: `('bmi', 'Age', 'baseline_diabetes', 'ccsMort30Rate', 'complication_rsi')`
   Accuracy: 0.8975
   Precision: 0.9301
   Recall: 0.6349
   F1-Score: 0.7546
   Matriz de Confusión en Validación:
$$\begin{bmatrix} 1624 & 26 \\ 199 & 346 \end{bmatrix}$$
3) **Combinación**: `('bmi', 'Age', 'baseline_diabetes', 'ccsComplicationRate', 'complication_rsi')`
   Accuracy: 0.8970
   Precision: 0.9299
   Recall: 0.6330
   F1-Score: 0.7533
   Matriz de Confusión en Validación:
$$\begin{bmatrix} 1624 & 26 \\ 200 & 345 \end{bmatrix}$$
4) **Combinación**: `('bmi', 'Age', 'baseline_diabetes', 'ccsComplicationRate', 'ccsMort30Rate', 'complication_rsi')`
   Accuracy: 0.8970
   Precision: 0.9299
   Recall: 0.6330
   F1-Score: 0.7533
   Matriz de Confusión en Validación:
$$\begin{bmatrix} 1624 & 26 \\ 200 & 345 \end{bmatrix}$$
5) **Combinación**: `('bmi', 'Age', 'ccsComplicationRate', 'complication_rsi')`
   Accuracy: 0.8966
   Precision: 0.9297
   Recall: 0.6312
   F1-Score: 0.7519
   Matriz de Confusión en Validación:
$$\begin{bmatrix} 1624 & 26 \\ 201 & 344 \end{bmatrix}$$

Los resultados indican que las características más relevantes para predecir complicaciones quirúrgicas utilizando KNN son **bmi**, **Age**, y **baseline_diabetes**, que aparecen en todas las combinaciones con mayor F1-Score. Además, factores como **complication_rsi**, **ccsComplicationRate**, y **ccsMort30Rate** también tienen un impacto considerable en el rendimiento del modelo. Estas características permiten un buen balance entre precisión y sensibilidad en la predicción de complicaciones quirúrgicas.

VII. Comparación con los datos de testing

a) Evaluación en el conjunto de Test para Regresión Lineal en Diabetes: **Combinación**: `('Pregnancies', 'Glucose', 'BloodPressure', 'BMI')`
Accuracy: 0.7414
Precision: 0.6667
Recall: 0.6047
F1-Score: 0.6341
Matriz de Confusión en Test:
$$\begin{bmatrix} 60 & 13 \\ 17 & 26 \end{bmatrix}$$

b) Evaluación en el conjunto de Test para Regresión Lineal en Quirúrgico: **Combinación**: `('bmi', 'Age', 'ccsComplicationRate', 'complication_rsi', 'mortality_rsi')`
Accuracy: 0.8024
Precision: 0.6920
Recall: 0.3537
F1-Score: 0.4681
Matriz de Confusión en Test:
$$\begin{bmatrix} 1571 & 85 \\ 349 & 191 \end{bmatrix}$$

c) Evaluación en el conjunto de Test para KNN en Diabetes: **Combinación**: `('Pregnancies', 'Glucose', 'BMI')`
Accuracy: 0.6638
Precision: 0.5476
Recall: 0.5349
F1-Score: 0.5412
Matriz de Confusión en Test:
$$\begin{bmatrix} 54 & 19 \\ 20 & 23 \end{bmatrix}$$

d) Evaluación en el conjunto de Test para KNN en Quirúrgico: **Combinación**: `('bmi', 'Age', 'baseline_diabetes', 'complication_rsi')`
Accuracy: 0.9057
Precision: 0.9237
Recall: 0.6722
F1-Score: 0.7781
Matriz de Confusión en Test:
$$\begin{bmatrix} 1626 & 30 \\ 177 & 363 \end{bmatrix}$$

VIII. Conclusión

Los resultados obtenidos muestran un rendimiento variable entre los distintos modelos y datasets. Para la regresión lineal en el dataset de Diabetes, la combinación de características que incluye `'Pregnancies', 'Glucose', 'BloodPressure', 'BMI'` alcanzó una precisión de 66.67% y un F1-Score de 63.41%, lo que indica un balance aceptable entre precisión y *recall*, aunque con margen de mejora. Por otro lado, la regresión lineal en el dataset quirúrgico mostró una mayor precisión (69.20%) pero un F1-Score más bajo (46.81%) debido a un *recall* reducido, lo que sugiere que el modelo puede predecir correctamente una mayor proporción de resultados positivos, aunque a expensas de detectar una menor cantidad de verdaderos positivos.

En el caso de KNN, los resultados en el dataset de Diabetes fueron algo inferiores, con una precisión de 54.76% y un F1-Score de 54.12%. Sin embargo, KNN en el dataset quirúrgico mostró un desempeño notablemente mejor, con una precisión de 92.37% y un F1-Score de 77.81%, lo que sugiere que este modelo se ajusta mejor a este tipo de datos y ofrece un equilibrio adecuado entre la capacidad de predicción y la precisión en los resultados.

En resumen, el rendimiento de los modelos varía significativamente según el tipo de datos, y KNN parece ser más adecuado para el dataset quirúrgico, mientras que la regresión lineal ofrece un desempeño más consistente en el dataset de Diabetes.

IX. Rúbrica

Conjunto de datos de Diabetes		
Criterios	Puntuación máxima	Puntuación obtenida
Análisis del conjunto de datos y features	5	
Análisis de regresión logística	15	
Análisis de KNN	15	
Comparación de modelos	10	
Conjunto de datos seleccionado		
Criterios	Puntuación máxima	Puntuación obtenida
Análisis del conjunto de datos y features	5	
Análisis de regresión logística	15	
Análisis de KNN	15	
Comparación de modelos	10	
Aspectos Generales		
Criterios	Puntuación máxima	Puntuación obtenida
Completitud de entregables	5	
Estructura de artículo científico	5	
Aspectos Extra (SRE)		
Criterio	Puntuación máxima	Puntuación obtenida
Usar Gitflow como proceso de colaboración y utilizar tags de versionamiento (main branch). Deben participar los 2 integrantes.	5	
Total	105	

Fig. 7.

References

[1] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

[2] Dua, D., & Graff, C. (2019). Pima Indians Diabetes Database. UCI Machine Learning Repository. Retrieved from https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes

[3] Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.