Proyecto 1: Clasificación de datos aplicados.

Edgar André Araya Vargas
Ingeniería en Computación
Instituto Tecnológico de Costa Rica
Cartago, Costa Rica
andrearaya1234@estudiantec.cr

Jose Daniel Rojas Calderon
Ingeniería en Computación.)
Instituto Tecnologico de Costa Rica
Cartago, Costa Rica
josedanielrojas5@estudiantec.cr

***Abstract*—Este estudio aplica los modelos K-Nearest Neighbors (KNN) y regresión logística a dos conjuntos de datos: predicción de diabetes y rotación de empleados. Se utilizó preprocesamiento avanzado, balanceo de clases y selección de características. Ambos modelos mostraron un rendimiento competitivo, destacando KNN por su flexibilidad y la regresión logística por su simplicidad. El balanceo de datos mejoró significativamente el recall. La elección del modelo depende de las necesidades específicas de cada aplicación.**

***Index Terms*—machine learning, k-nearest Neighbors, regresión logística, balanceo de clases, predicción de diabetes, rotación de empleados, selección de características, preprocesamiento de datos.**

I. Introducción

El campo del Machine Learning ha experimentado un crecimiento exponencial en los últimos años, ofreciendo soluciones innovadoras a problemas complejos en diversas áreas. Este proyecto se centra en la aplicación de técnicas de clasificación de datos a dos conjuntos de datos distintos, con el objetivo de explorar y comparar diferentes herramientas de Machine Learning, contribuyendo así al desarrollo del conocimiento a través de la investigación aplicada.

El primer conjunto de datos, proporcionado por el Instituto Nacional de Diabetes y Enfermedades Digestivas y Renales, se enfoca en la predicción diagnóstica de diabetes. Este dataset incluye diversas mediciones médicas y demográficas de pacientes, como el número de embarazos, nivel de glucosa, presión arterial, grosor del pliegue cutáneo, insulina, índice de masa corporal (BMI), función de pedigrí de diabetes y edad. El objetivo es desarrollar un modelo capaz de predecir con precisión si un paciente tiene diabetes basándose en estas variables.

El segundo conjunto de datos, seleccionado por nuestro equipo, se relaciona con la rotación de empleados (attrition) de IBM. Este dataset contiene una amplia gama de variables relacionadas con los empleados, incluyendo edad, tasa de salario, departamento, distancia desde el hogar, educación, satisfacción laboral, ingresos mensuales y años en la empresa, entre otros. El análisis de este conjunto de datos busca identificar factores que puedan influir en la retención de empleados, un tema de gran relevancia en la gestión de recursos humanos y la estrategia empresarial.

La elección de estos dos conjuntos de datos permite abordar problemas de clasificación en contextos muy diferentes: uno en el ámbito de la salud y otro en el ámbito empresarial. Esta diversidad enriquece el proyecto, permitiendo explorar la versatilidad y eficacia de las técnicas de Machine Learning en escenarios distintos.

II. Metodología

Para abordar los objetivos del proyecto, se ha diseñado una metodología estructurada que abarca desde la exploración inicial de los datos hasta la evaluación comparativa de los modelos de clasificación. A continuación, se detallan los pasos principales:

A. Exploración y Preprocesamiento de Datos

Para ambos conjuntos de datos (Diabetes e IBM Attrition), se realizará un análisis exploratorio exhaustivo:

1) **Análisis estadístico básico:** Se examinarán las distribuciones de las variables, medidas de tendencia central y dispersión.
2) **Manejo de valores faltantes:** Se identificarán y tratarán los valores faltantes mediante reemplazo por la mediana.
3) **Detección y tratamiento de outliers:** Se identificarán valores atípicos y se decidirá cómo manejarlos basándose en el contexto de cada variable.
4) **Análisis de balance de clases:** Se evaluará la distribución de las clases objetivo (diabetes/no diabetes, attrition/no attrition) utilizando visualizaciones con matplotlib o seaborn.
5) **Justificación de observaciones:** Se documentarán y justificarán todas las decisiones tomadas durante el preprocesamiento.

B. Preparación de Datos para Modelado

1) **División del conjunto de datos:**
   - 70% para entrenamiento
   - 15% para validación
   - 15% para testing
2) Normalización/Estandarización de características numéricas si es necesario.
3) Codificación de variables categóricas (especialmente relevante para el conjunto de datos de IBM Attrition).

C. Desarrollo y Evaluación de Modelos

Se implementarán dos modelos de clasificación para cada conjunto de datos:

1) Regresión Logística

2) K-Nearest Neighbors (KNN)

Para cada modelo:

- Se realizarán múltiples ejecuciones con diferentes hiperparámetros.
- Se utilizará el conjunto de validación para verificar la convergencia y evitar el sobreajuste.
- Se mantendrán constantes los conjuntos de entrenamiento y prueba para asegurar la comparabilidad entre experimentos.

D. Tratamiento de Valores Atípicos

En este estudio, se aplicaron dos técnicas principales para el manejo de valores atípicos: la corrección de valores no válidos y el tratamiento de outliers utilizando el rango intercuartílico (IQR).

Primero, se identificaron y trataron los valores fuera de los límites biológicamente plausibles para cada característica del dataset. Estos límites se establecieron basándose en rangos típicos para cada variable. Por ejemplo, se definieron límites como $(0, 10)$ para el número de embarazos, y $(70, 300)$ para el nivel de glucosa. Los valores que caían fuera de estos rangos fueron reemplazados por el valor de la mediana de los datos válidos, lo que ayuda a evitar la distorsión del dataset debido a datos erróneos o implausibles.

A continuación, se aplicó un tratamiento más conservador a los outliers dentro de los límites biológicos utilizando el rango intercuartílico (IQR). Para cada característica, se calcularon los cuartiles primero (Q1) y tercero (Q3), y se determinó el IQR como la diferencia entre Q3 y Q1. Los límites inferiores y superiores para los valores atípicos se establecieron como $Q1 - 1.5 \times IQR$ y $Q3 + 1.5 \times IQR$, respectivamente. Sin embargo, estos límites fueron ajustados para no exceder los rangos biológicos previamente definidos.

Esta metodología asegura que los datos utilizados para el modelado sean representativos y estén libres de valores extremos que puedan influir negativamente en el rendimiento del modelo, mientras se preserva la información útil contenida en los datos no atípicos.

E. Evaluación de Modelos

La evaluación de los modelos se realizará utilizando el conjunto de datos de prueba. Se calcularán las siguientes métricas:

- Accuracy
- Precision
- Recall
- Matriz de confusión

Además, se realizará una comparación detallada entre los modelos de regresión logística y KNN para cada conjunto de datos, analizando sus fortalezas y debilidades en el contexto específico de cada problema.

III. Análisis Exploratorio de "Diabetes"

A. Descripción del Conjunto de Datos

El conjunto de datos de diabetes contiene información sobre 768 pacientes y consta de 9 variables. Utilizando la función `info()` de pandas, obtuvimos la información sobre la estructura del dataset y observamos que no hay valores nulos en ninguna de las columnas, lo cual es positivo para nuestro análisis. Sin embargo, esto no garantiza la ausencia de valores atípicos o incorrectos.

B. Análisis Visual de la Distribución de Datos

Utilizamos boxplots y histogramas para visualizar la distribución de las variables numéricas y detectar posibles valores atípicos.

C. Distribución de Características Numéricas

Para obtener una visión más detallada de la distribución de cada variable numérica, generamos histogramas utilizando la función `hist()` de pandas:

```
df1.hist(bins=30, figsize=(12, 10))
plt.suptitle('Distribucin de las Caractersticas
    Numricas')
plt.show()
```

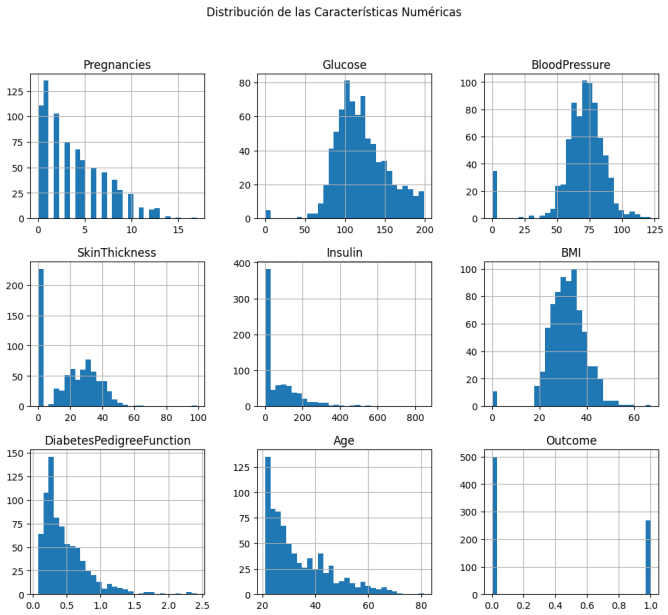

Fig. 1. Distribución de las Características Numéricas

La Figura 1 nos proporciona información valiosa sobre la forma de las distribuciones de nuestras variables:

- **Pregnancies**: Muestra una distribución asimétrica positiva (cola a la derecha), con la mayoría de los valores concentrados en el rango inferior.
- **Glucose**: Presenta una distribución aproximadamente normal, con un ligero sesgo hacia la derecha.
- **BloodPressure**: Exhibe una distribución casi simétrica, con una concentración de valores alrededor de 70-80 mm Hg.
- **SkinThickness**: Muestra una distribución multimodal, lo que podría indicar la presencia de subgrupos en la población o problemas en la medición de esta variable.

- **Insulin**: Presenta una distribución muy asimétrica positiva, con una gran cantidad de valores bajos y algunos valores extremadamente altos.
- **BMI**: Muestra una distribución ligeramente asimétrica positiva, con la mayoría de los valores entre 20 y 40.
- **DiabetesPedigreeFunction**: Exhibe una distribución muy asimétrica positiva, con la mayoría de los valores concentrados en el rango inferior y una cola larga hacia la derecha.
- **Age**: Presenta una distribución asimétrica positiva, con una mayor concentración de individuos jóvenes y adultos de mediana edad en la muestra.

D. Análisis de Valores Únicos y Valores Atípicos

Examinamos la cantidad de valores únicos en cada variable:

```
Pregnancies 17
Glucose 136
BloodPressure 47
SkinThickness 51
Insulin 186
BMI 248
DiabetesPedigreeFunction 517
Age 52
Outcome 2
```

Destacamos que la variable objetivo 'Outcome' tiene 2 valores únicos, lo que confirma que estamos ante un problema de clasificación binaria.

Examinamos la presencia de valores atipicos mediante un boxplot:

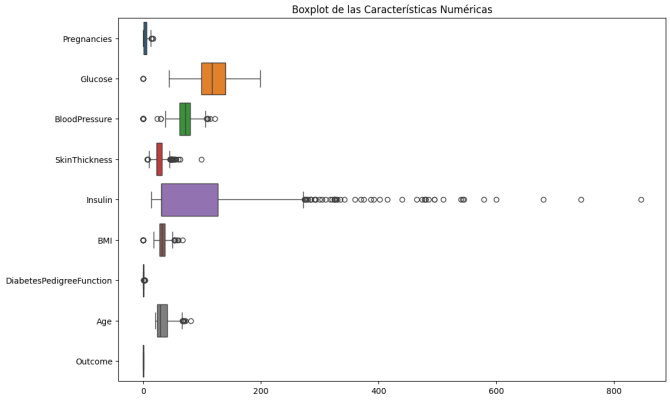

Fig. 2. Boxplot de las Características Numéricas

El boxplot (Figura 2) revela la presencia de valores atípicos en varias variables, especialmente en 'Insulin' y 'DiabetesPedigreeFunction'. Estos outliers podrían tener un impacto significativo en nuestros modelos y requerirán un análisis más detallado.

Notamos gran cantidad de valores 0 en el boxplot (Figura 2) en ciertas características:

```
Glucose 5
BloodPressure 35
SkinThickness 227
Insulin 374
BMI 11
```

'SkinThickness' e 'Insulin' contienen muchos valores 0, lo cual es fisiológicamente improbable y comprometeria los resultados de nuestro modelo.

E. Análisis de la Variable Objetivo

Examinamos la distribución de la variable objetivo 'Outcome' para evaluar el balance de las clases:

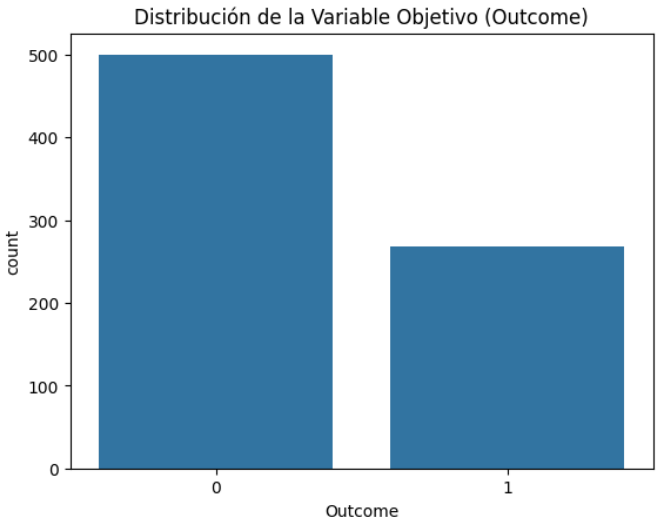

Fig. 3. Distribución de la Variable Objetivo (Outcome)

La Figura 3 muestra un desequilibrio en las clases, con una mayor proporción de casos negativos (sin diabetes) que positivos (con diabetes). Este desequilibrio deberá ser considerado durante la fase de modelado para evitar sesgos en nuestras predicciones.

IV. Análisis Exploratorio de "Desgaste de Empleados"

El set de datos de IBM HR Analytics Employee Attrition Performance, tiene 1471 registros de empleados con un total de 35 variables. Entre estas el valor mas importante para el analisis como lo es "Attrition".

A. Analisis visual de los Datos

Para el analisis visual de los datos se realizaron graficos para comparar y relacionar ciertas variables que se consideraron son las mas importantes. Para la seleccion de las caracteristicas se utilizo el commando list(dataSet.columns) para poder visualizar las caracteristicas ya que el son 35 caracteristicas en total.

En la figura 4 podemos notar como analizamos el numero de empleados que salio de la empresa segun el departamento. Como podemos notar, el demartamento de desarrollo fue uno de los que mas salidas tuvo, sin embargo hay que tomar en cuenta que la empresa trabaja en el area de tecnologia y desarrollo por lo tanto podriamos decir que la mayor cantidad de empleados pertenecen a este departamento. Ademas de que este departamento es el departamento en el que mas personas salieron, es el que mas personal por lo tanto es algo esperable.

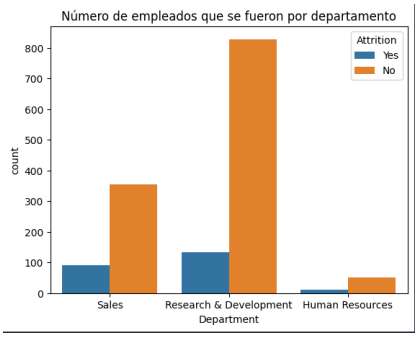

Fig. 4. Ejemplo de cantidad de empleados que salieron por departamento.

En segundo lugar esta el departamento de ventas y por ultimo recursos humanos.

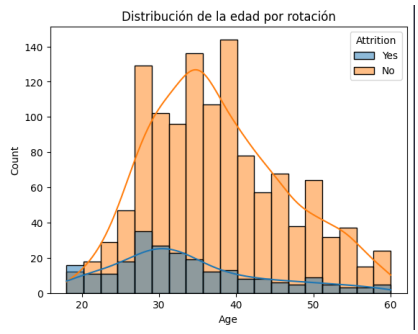

Fig. 5. Ejemplo de cantidad de empleados que salieron segun la edad.

La figura 5, nos muestra la distribucion de la edad del personal que si salio de la empresa. Se aprecia como la mayoria de los empleados ronda entre los 27a los 40 años. Sin embargo la mayor cantidad de empleados que salieron son los que tenia entre 27 a 30. Podria indicar que son personas que buscan una experiencia diferente o algun puesto en otra empresa para seguir desarrollandose profesionalmente, en contraste con los de mayor edad, estos parece que buscan mantenerse en la empresa por mas tiempo.

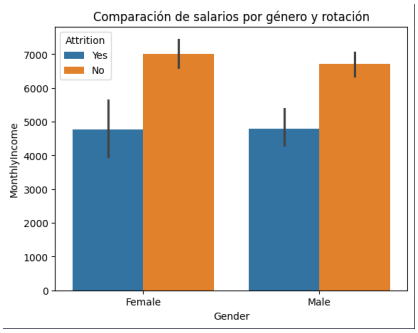

Fig. 6. Ejemplo de comparascion entre la rotacion con los salarios por genero

En la figura 6 podemos ver el analisis del genero de los empleados que salieron o no de la empresa en relacion al salario que ganan, muestra que es bastante parejo tanto para hombres como para mujeres aun que las mujeres estan levemente por arriba, se podria decir que no es algo que cambie mucho en esta empresa. Y los salarios parecen bastante equilibrados comparandolos por genero.

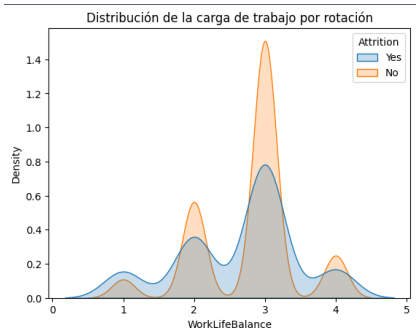

Fig. 7. Ejemplo de rotacion de empleados y su percepcion de la carga de trabajo.

En la figura 7 podemos ver una de las caracteristicas mas interesantes. Se analiza la distribucion de la rotacion en relacion a la carga de trabajo. Podemos notar que aun cuando se considera que se tiene un buen balance, ya que vemos que los que son calificados con 3 o 'Better' es donde mas empleados se quedan en la empresa, es tambien el punto donde mas empleados han abandonado seguido por 2 o 'Good'. Se podria pensar que si se considera que se tienen un balance tan bueno pueda llevar a mantener mas tiempo a la persona en el puesto de trabajo.

WorkLifeBalance 1 'Bad' 2 'Good' 3 'Better' 4 'Best'

V. Resultados de "Diabetes"

En este análisis, nos centraremos en la evaluación y comparación de dos modelos de aprendizaje automático aplicados a nuestro conjunto de datos: el modelo K-Nearest Neighbors (KNN) y el modelo de Regresión Logística. Ambos modelos son ampliamente utilizados en tareas de clasificación, pero difieren en sus enfoques y características.

A continuación, presentaremos los resultados obtenidos para cada modelo, seguidos de un análisis comparativo que nos ayudará a tomar decisiones informadas sobre la selección del modelo más apropiado para nuestras necesidades.

Implementamos un clasificador K-Nearest Neighbors (KNN) con diferentes configuraciones y evaluamos su rendimiento en un conjunto de datos balanceado. Los resultados para tres configuraciones diferentes son los siguientes:

A. *Configuración KNN 1: 3 características, 10 vecinos*

En esta configuración, se seleccionaron las características `Glucose`, `BMI` y `Age`. El modelo KNN utilizó 10 vecinos. Los resultados se muestran en la Tabla I, y la matriz de confusión correspondiente se presenta en la Figura 8.

En esta configuración, el modelo alcanzó una *accuracy* de 0.76 en el conjunto de prueba. La *precision* y el *recall* fueron ambos de 0.7722, lo que indica un buen equilibrio entre la identificación de casos positivos y negativos. El modelo clasificó correctamente 53 casos negativos y 61 casos positivos, con 18 falsos negativos y 18 falsos positivos.

TABLE I
RESULTADOS DE KNN CON 3 CARACTERÍSTICAS Y 10 VECINOS

Métrica	Validación	Prueba
Accuracy	0.70	0.76
Precision	0.6753	0.7722
Recall	0.7222	0.7722

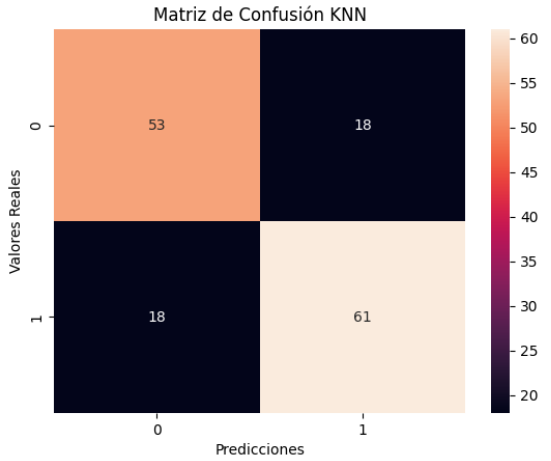

Fig. 8. Matriz de Confusión para KNN con 3 características y 10 vecinos

B. Configuración KNN 2: 4 características, 24 vecinos

En esta configuración, se añadieron `SkinThickness` a las características `Glucose`, `BMI` y `Age`. El modelo KNN utilizó 24 vecinos. Los resultados se resumen en la Tabla II, y la matriz de confusión correspondiente se muestra en la Figura 9.

TABLE II
RESULTADOS DE KNN CON 4 CARACTERÍSTICAS Y 24 VECINOS

Métrica	Validación	Prueba
Accuracy	0.7267	0.74
Precision	0.6867	0.7381
Recall	0.7917	0.7848

El modelo obtuvo una *accuracy* de 0.74 en el conjunto de prueba. La *precision* fue de 0.7381, mientras que el *recall* fue de 0.7848. El modelo clasificó correctamente 49 casos negativos y 62 positivos, con 22 falsos positivos y 17 falsos negativos.

C. Configuración KNN 3: 3 características, 64 vecinos

En esta configuración, se volvieron a utilizar las características `Glucose`, `BMI` y `Age`, pero se incrementó el número de vecinos a 64. Los resultados obtenidos se presentan en la Tabla III, y la matriz de confusión en la Figura 18.

El modelo mostró una *accuracy* de 0.7133 en el conjunto de prueba, con una *precision* de 0.7143 y un *recall* de 0.7595. Clasificó correctamente 46 casos negativos y 69 casos positivos, con 25 falsos positivos y 10 falsos negativos.

Luego implementamos el modelo de Regresión Logística con dos configuraciones distintas de características seleccionadas. Para cada configuración, se realizó un balanceo de clases, y se evaluaron las métricas de *accuracy*, *precision* y *recall* tanto en el conjunto de validación como en el de prueba.

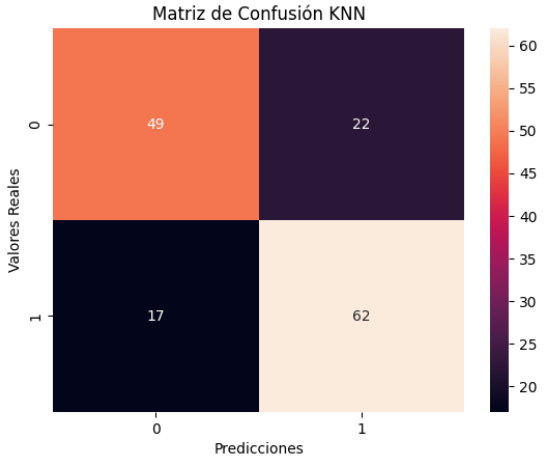

Fig. 9. Matriz de Confusión para KNN con 4 características y 24 vecinos

TABLE III
RESULTADOS DE KNN CON 3 CARACTERÍSTICAS Y 64 VECINOS

Métrica	Validación	Prueba
Accuracy	0.7267	0.7133
Precision	0.6914	0.7143
Recall	0.7778	0.7595

D. Configuración Regresión Logística 1: 4 Características Seleccionadas

En la primera configuración, se seleccionaron las siguientes características: `Glucose`, `SkinThickness`, `BMI` y `Age`. Los resultados se muestran en la Tabla IV, y la matriz de confusión correspondiente se presenta en la Figura 11.

TABLE IV
RESULTADOS DE REGRESIÓN LOGÍSTICA CON 4 CARACTERÍSTICAS SELECCIONADAS

Métrica	Validación	Prueba
Accuracy	0.75	0.74
Precision	0.74	0.77
Recall	0.72	0.72

Como se observa, el modelo logra una *accuracy* de 0.74 en el conjunto de prueba, lo que indica un desempeño moderado. La *precision* fue ligeramente superior en el conjunto de prueba con un valor de 0.77, mientras que el *recall* se mantuvo en 0.72, lo que refleja una capacidad del modelo para identificar correctamente las instancias positivas (casos con diabetes) en un 72% de las veces.

E. Configuración Regresión Logística 2: 8 Características Seleccionadas

En la segunda configuración, se seleccionaron un total de 8 características: `Pregnancies`, `Glucose`, `BloodPressure`, `SkinThickness`, `Insulin`, `BMI`, `DiabetesPedigreeFunction` y `Age`. Los resultados

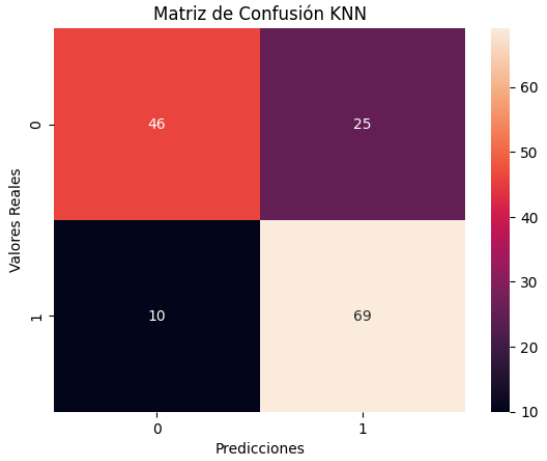

Fig. 10. Matriz de Confusión para KNN con 3 características y 64 vecinos

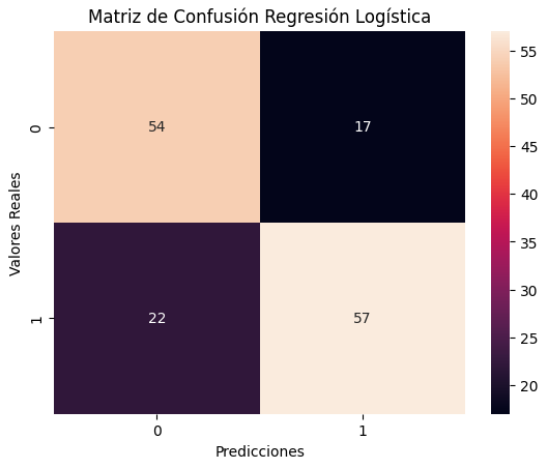

Fig. 11. Matriz de confusión de la Regresión Logística con 4 características seleccionadas

se resumen en la Tabla V, y la matriz de confusión correspondiente se muestra en la Figura 12.

TABLE V
RESULTADOS DE REGRESIÓN LOGÍSTICA CON 8 CARACTERÍSTICAS SELECCIONADAS

Métrica	**Validación**	**Prueba**
Accuracy	0.73	0.74
Precision	0.71	0.79
Recall	0.75	0.68

En esta configuración, se puede observar una mejora en la *precision*, alcanzando un valor de 0.79 en el conjunto de prueba. Sin embargo, el *recall* disminuyó a 0.68, lo que indica que el modelo fue capaz de identificar correctamente solo el 68% de las instancias positivas. La *accuracy* se mantuvo en 0.74, similar a la primera configuración.

F. Discusión

En las tres configuraciones de KNN, la primera con 3 características y 10 vecinos fue la que mejor equilibró *precision* y *recall*, ambos con un valor de 0.7722. Aunque agregar una característica en la segunda configuración mejoró un poco el *recall*, la precisión disminuyó y la *accuracy* apenas cambió. La tercera configuración, con 64 vecinos, mostró la menor *accuracy*, aunque mantuvo un *recall* alto.

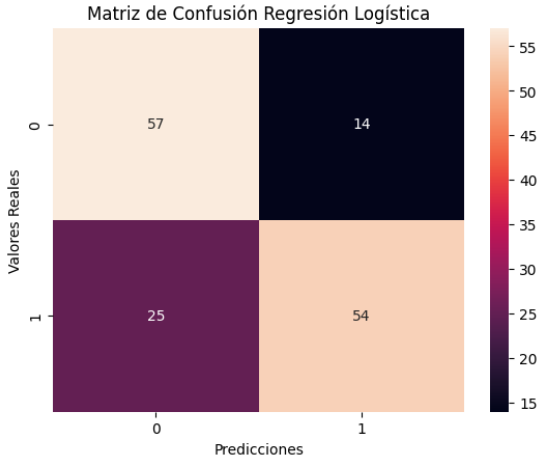

Fig. 12. Matriz de confusión de la Regresión Logística con 8 características seleccionadas

Este resultado sugiere que usar 10 vecinos y las tres características (`Glucose`, `BMI` y `Age`) es la mejor opción, ya que ofrece un buen equilibrio entre las métricas.

Por otro lado, la Regresión Logística tuvo un desempeño similar en ambas configuraciones, con una *accuracy* de 0.74. La primera configuración tuvo mejor *recall* (0.72), lo que indica que fue más efectiva en identificar casos positivos. Sin embargo, la segunda configuración fue más precisa, es decir, cometió menos errores al clasificar como positivos.

Comparando ambos modelos, KNN ofrece más flexibilidad en ajustar el número de vecinos y características, lo que permite mejorar el balance entre precisión y sensibilidad según el objetivo. La Regresión Logística es más estable, pero no necesariamente mejora con más características, por lo que elegir el mejor modelo depende de las necesidades de la tarea, ya sea minimizar falsos positivos o identificar más casos positivos.

VI. RESULTADOS DE "DESGASTE DE EMPLEADOS"

Para los resultados, se realizaron 2 conjuntos de pruebas por cada uno de los modelos(KNN y Regresion Logistica). En el primer conjunto tenemos las corridas de los algoritmos con los datos balanceados. Tambien se implemento la funcion SelectKBest(chi2, k=columns) de la biblioteca Scikit-learn, esta lo que hace es basado en la funcion chi2, que nos ayuda a determinar si existe una diferencia significativa entre lo que observamos en nuestros datos y lo que esperábamos encontrar. En este caso nos va a ayudar ya que va a analizar varios features categoricos.

A. Resultados KNN

Para los resultados de la primera prueba se seleccionaron 15 caracteristicas y 10 vecinos para el algoritmo.

- **Accuracy**: 0.80 (Validación), 0.77 (Prueba)
- **Precision**: 0.74 (Validación), 0.72 (Prueba)
- **Recall**: 0.90 (Validación), 0.87 (Prueba)

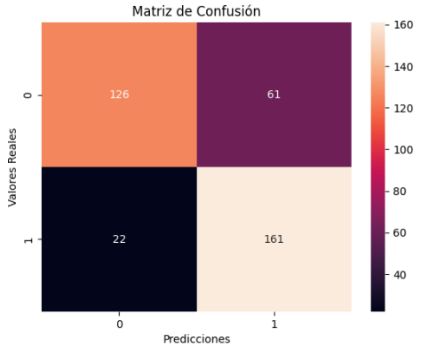

Fig. 13. Matriz de Confusión para KNN con 15 características y 10 vecinos

Para los resultados de la segunda prueba se seleccionaron 25 caracteristicas y 5 vecinos para el algoritmo.

- **Accuracy**: 0.83 (Validación), 0.82 (Prueba)
- **Precision**: 0.75 (Validación), 0.74 (Prueba)
- **Recall**: 0.97 (Validación), 0.98 (Prueba)

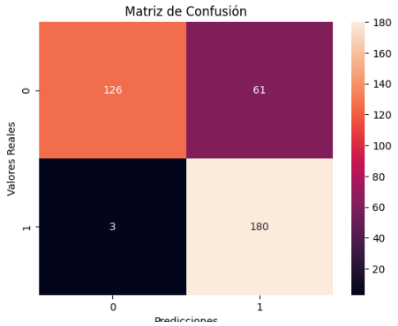

Fig. 14. Matriz de Confusión para KNN con 25 características y 5 vecinos

Para los resultados de la tercera prueba se seleccionaron 15 caracteristicas y 20 vecinos para el algoritmo.

- **Accuracy**: 0.84 (Validación), 0.83 (Prueba)
- **Precision**: 1 (Validación), 1 (Prueba)
- **Recall**: 0.15 (Validación), 0.11 (Prueba)

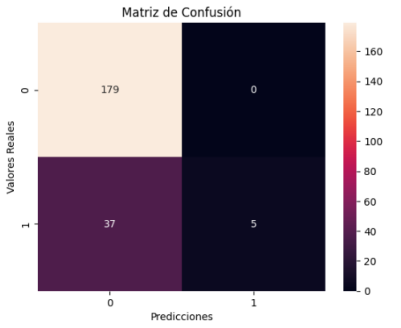

Fig. 15. Matriz de Confusión para KNN con 15 características y 20 vecinos

Para los resultados de la cuarta prueba se seleccionaron 10 caracteristicas y 12 vecinos para el algoritmo.

- **Accuracy**: 0.84 (Validación), 0.83 (Prueba)
- **Precision**: 1 (Validación), 0.8 (Prueba)
- **Recall**: 0.15 (Validación), 0.09 (Prueba)

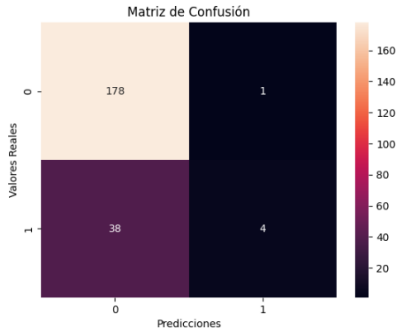

Fig. 16. Matriz de Confusión para KNN con 10 características y 12 vecinos

B. Resultados Regresion Logistica

En la primera prueba con regresion logistica con 10 caracteristicas tuvimos los siguientes resultados.

- **Accuracy**: 0.84 (Validación), 0.84 (Prueba)
- **Precision**: 1 (Validación), 1 (Prueba)
- **Recall**: 0.15 (Validación), 0.11 (Prueba)

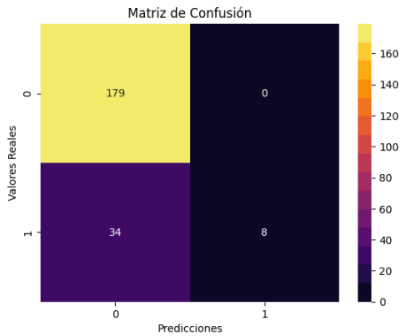

Fig. 17. Matriz de Confusión para Regresion Logistica con 10 características

En la segundo prueba con regresion logistica con 28 caracteristicas tuvimos los siguientes resultados.

- **Accuracy**: 0.87 (Validación), 0.84 (Prueba)
- **Precision**: 0.78 (Validación), 0.78 (Prueba)
- **Recall**: 0.39 (Validación), 0.26 (Prueba)

C. Discusión

Balanceo de datos: El balanceo de datos parece mejorar el rendimiento en general, especialmente en términos de recall. Aumentar el número de características de 15 a 25 mejoró ligeramente el rendimiento del modelo KNN, pero no hubo cambios significativos en los otros casos.

Enfocandonos en los resultados de los modelos parecen tener resultados similares, Sin embargo, la elección del mejor

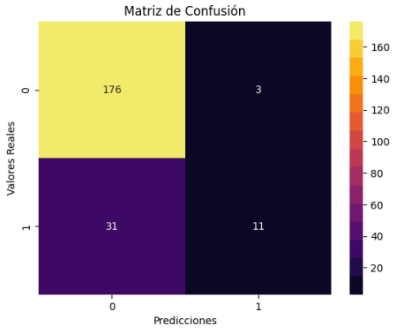

Fig. 18. Matriz de Confusión para Regresion Logistica con 28 características

modelo dependerá de las prioridades específicas de la aplicación (precisión, recall, etc.).El alto rendimiento en el conjunto de validación y prueba podría ser una señal de sobreajuste, especialmente en los casos donde la precisión es muy alta y el recall es bajo. Considera técnicas de regularización o aumentar el tamaño del conjunto de datos para mitigar el sobreajuste.

VII. Conclusiones

En conclusión, el presente trabajo ha demostrado que tanto el algoritmo K-Nearest Neighbors (KNN) como la regresión logística ofrecen un desempeño competente en la clasificación de datos, con resultados comparables en términos de precisión y capacidad predictiva. El tratamiento robusto de los datos, incluyendo la normalización y el manejo de valores atípicos, resultó crucial para mejorar la calidad del modelado.

El balanceo de clases mediante técnicas como SMOTE fue esencial para mitigar el impacto de los conjuntos de datos desbalanceados, lo cual se reflejó en mejores métricas de recall, especialmente en los modelos KNN. Asimismo, la selección de características tuvo un efecto notable en el desempeño de los modelos, destacando la importancia de una selección cuidadosa basada en la relevancia estadística y la validación cruzada.

A pesar de que ambos modelos presentaron métricas similares, la elección del modelo óptimo depende de las necesidades específicas de la aplicación. Si se prioriza una mayor sensibilidad para identificar correctamente las clases positivas, KNN con un número adecuado de vecinos puede ser la opción preferida. Sin embargo, para escenarios donde la simplicidad y la interpretabilidad del modelo son más importantes, la regresión logística puede ser más adecuada, manteniendo un balance entre precisión y recall.

Finalmente, se recomienda explorar técnicas de regularización para evitar el sobreajuste, así como expandir el conjunto de datos para asegurar la generalización de los modelos a escenarios más amplios y complejos.

References

[1] https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset
[2] https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database?resource=download