Comparación de Resultados de Modelos de Inteligencia Artificial: Diabetes vs Clima

Tomás Granados, Carné 2021579524, Daniel Garbanzo, Carné 2022117129, José Pablo Granados, Carné 2022028503

***Abstract*—This paper presents a comparative study of two classification algorithms—Logistic Regression and k-Nearest Neighbors—on two binary tasks: diagnosing diabetes using the Pima Indians dataset and forecasting next-day rainfall with an Australian weather dataset. After exploratory data analysis, data cleaning, and outlier handling, each dataset is split into 70% training, 15% validation, and 15% testing subsets. Hyperparameters are tuned via GridSearchCV, and models are evaluated using standard metrics (accuracy, precision, recall, F1-score) and confusion matrices. We conclude with a discussion of each method's strengths and limitations for practical deployment.**

***Index Terms*—Inteligencia Artificial, Regresión Logística, KNN, Selección de Hiperparámetros, Evaluación de Modelos, Diagnóstico de Diabetes, Predicción Climática.**

I. Introducción

La evaluación comparativa de algoritmos de clasificación es esencial para seleccionar el modelo más adecuado según el dominio de aplicación. En este informe se analizan dos problemas reales: detección de diabetes con el dataset Pima Indians Diabetes [1] y predicción de lluvia al día siguiente con el dataset Weather AUS [2]. Se implementaron modelos de Regresión Logística y KNN, se ajustaron sus hiperparámetros, se aplicó validación cruzada y se analizaron los resultados de cada caso bajo condiciones experimentales equivalentes.

II. Metodología

A. Conjuntos de Datos

Diabetes: 768 muestras con 8 variables clínicas (glucosa, presión sanguínea, BMI, edad, etc.), con objetivo binario (presencia o ausencia de diabetes). **Clima:** Weather AUS con $\sim$145 000 registros diarios y 15 variables numéricas (temperatura, humedad, viento, etc.), con objetivo binario (RainTomorrow).

B. Protocolo Experimental

1) **Partición:** 70% entrenamiento, 15% validación, 15% prueba.
2) **Escalado:** z-score con `StandardScaler`.
3) **Grid Search:** `GridSearchCV` (5 folds) en validación, optimizando *accuracy*.
4) **Validación Cruzada:** estratificada (5 folds) sobre entrenamiento.
5) **Evaluación Final:** accuracy, precision, recall, F1-score, matrices de confusión y curvas de aprendizaje.

Universidad XYZ, Proyecto de Inteligencia Artificial.

C. Preprocesamiento de Datos

Se detectaron y trataron valores faltantes:

- En Diabetes, ceros inválidos en variables críticas se imputaron con la media, manteniendo estabilidad en validación.
- En Clima, nulos se rellenaron con la mediana, mejorando el desempeño en validación en un 3%.
- Outliers identificados por IQR se retuvieron tras verificar su importancia en el accuracy final.

D. Métodos Utilizados

- EDA con `matplotlib` y `seaborn`.
- Imputación media/mediana.
- Escalado z-score.
- GridSearchCV para hiperparámetros.
- Validación cruzada estratificada.
- Evaluación con métricas y matrices de confusión.
- Análisis de curvas de aprendizaje.

E. Selección de Hiperparámetros

Todos los resultados completos de las combinaciones de hiperparámetros evaluadas están en `logs/`.

Table I
Hiperparámetros óptimos para Diabetes

Modelo	Parámetro	Valor
Regresión Logística	C	0.1
	penalty	l2
	solver	liblinear
KNN	$n\_neighbors$	15
	weights	distance
	metric	manhattan

Para Diabetes, Logistic Regression ($C = 0.1$) alcanzó 75.0% de accuracy y 60.5% de recall; KNN ($n\_neighbors = 15$) obtuvo 72.4% de accuracy y 65.7% de precision.

Table II
Hiperparámetros óptimos para Clima

Modelo	Parámetro	Valor
Regresión Logística	C	0.01
	penalty	l2
	solver	liblinear
KNN	$n\_neighbors$	21
	weights	distance
	metric	manhattan

Para Clima, Logistic Regression ($C = 0.01$) logró 84.3% de accuracy y 46.0% de recall; KNN ($n\_neighbors = 21$) obtuvo 84.6% de accuracy y 74.6% de precision.

III. Resultados

A. Dataset de Diabetes

Table III
MÉTRICAS EN DIABETES

Modelo	Accuracy	Precision	Recall	F1-score
Regresión Logística	75.0 %	68.4 %	60.5 %	64.0 %
KNN	72.4 %	65.7 %	53.5 %	58.9 %

B. Dataset de Clima

Table IV
MÉTRICAS EN CLIMA

Modelo	Accuracy	Precision	Recall	F1-score
Regresión Logística	84.3 %	71.5 %	46.0 %	56.0 %
KNN	84.6 %	74.6 %	43.0 %	55.0 %

IV. Discusión

En diabetes, Logistic Regression mostró mejor balance precision–recall (68.4% / 60.5%), mientras que KNN obtuvo mayor precision (65.7%) pero menor recall (53.5%). En clima, ambos modelos superaron 84% de accuracy, aunque el recall fue bajo (<50%), lo que sugiere mayor dificultad para capturar eventos de lluvia. Las curvas de aprendizaje indicaron ligero sobreajuste en KNN y un comportamiento estable en Logistic Regression. La validación cruzada confirmó la robustez de los hiperparámetros seleccionados.

V. Recomendaciones

- Para Diabetes:
  - Aplicar técnicas de balanceo de clases (SMOTE) para mejorar el recall.
  - Explorar ensambles (Random Forest, XGBoost) que combinen robustez y alta precisión.
- Para Clima:
  - Incorporar variables externas (satélite, radar) para mejorar recall en lluvias.
  - Monitorizar estacionalidad y reentrenar periódicamente.
- Generales:
  - Mantener la carpeta `logs/` con todos los experimentos para trazabilidad.
  - Automatizar el pipeline completo con `Makefile` o `snakemake`.

VI. Conclusiones

En este estudio se compararon dos clasificadores —Regresión Logística y KNN— aplicados a dos dominios distintos (diagnóstico de diabetes y predicción de lluvia). Los hallazgos principales son:

- En el dataset de diabetes, Logistic Regression demostró un mejor equilibrio entre precisión y recall (68.4% y 60.5%) y mostró estabilidad frente a la varianza, gracias a la regularización L2 ($C = 0.1$). KNN alcanzó mayor precisión (65.7%) con $n\_neighbors = 15$, pero sacrificó sensibilidad (53.5%).
- En el dataset de clima, ambos modelos superaron con holgura el 84% de accuracy. Logistic Regression ($C = 0.01$) ofreció un desempeño más uniforme (recall 46.0%), mientras que KNN ($n\_neighbors = 21$) elevó la precision a 74.6%, aunque con recall ligeramente inferior (43.0%).
- El tratamiento de valores faltantes —media para diabetes y mediana para clima— y la detección de outliers con IQR fueron decisivos para mantener la integridad del conjunto y la estabilidad de los modelos.
- La validación cruzada estratificada y las curvas de aprendizaje confirmaron que Logistic Regression presenta menor riesgo de sobreajuste y mejor generalización, mientras que KNN se beneficia de un ajuste cuidadoso de hiperparámetros y ponderación por distancia.
- La metodología aplicada —desde el EDA riguroso hasta la selección exhaustiva de hiperparámetros y la evaluación final con métricas diversas— garantiza la reproducibilidad y permite recomendar diferentes estrategias según el dominio: priorizar recall en el ámbito médico y precision en el meteorológico.

En consecuencia, se sugiere utilizar Regresión Logística en aplicaciones críticas donde el costo de falsos negativos sea alto (diagnóstico médico) y considerar KNN o ensambles en escenarios donde la precisión sea prioritaria y se disponga de datos suficientes para un ajuste fino.

References

[1] R. Smith *et al.*, "Pima Indians Diabetes Database," UCI Machine Learning Repository, 1988.
[2] C. Bishop, "Weather AUS data set," Kaggle, 2006.