Evaluación Final

Aplicación de Aprendizaje Automático Supervisado y No Supervisado

TRIANA MARTINEZ LEONEL
GARCIA VARGAS JESUS
SANTIAGO CABALLERO ALEJANDRO

Profesor: Eduardo Zurek, Ph.D.
Programa de Ingeniería de Sistemas
Universidad del Norte

2024

Contents

1 Introducción

Este proyecto desarrolla una evaluación completa de técnicas de aprendizaje automático supervisado y no supervisado aplicadas al conjunto de datos *Bank Marketing*, proveniente del repositorio UCI Machine Learning Repository.

El objetivo es analizar, limpiar, modelar y evaluar los datos con distintos algoritmos. Adicionalmente, se implementa un algoritmo KNN en lenguaje C para demostrar comprensión algorítmica a bajo nivel.

Todo el proceso incluye: exploración de datos (EDA), preprocesamiento, reducción de dimensionalidad con PCA, clustering, clasificación con modelos supervisados, validación cruzada, ajuste de hiperparámetros y comparación de resultados.

2 Objetivo General

Aplicar técnicas de aprendizaje automático para analizar y modelar el dataset *Bank Marketing*, utilizando tanto métodos supervisados como no supervisados, e implementar desde cero un modelo KNN en lenguaje C para evaluar su desempeño y comprender los fundamentos del algoritmo.

3 Objetivos Específicos

- Realizar un análisis exploratorio del dataset.
- Preprocesar adecuadamente las variables numéricas y categóricas.
- Aplicar reducción de dimensionalidad mediante PCA.
- Implementar técnicas de clustering como K-Means.
- Entrenar modelos supervisados de clasificación.
- Comparar el rendimiento de los modelos con métricas estándar.
- Implementar un modelo KNN en lenguaje C y evaluar su desempeño.

4 Justificación del Dataset

El dataset *Bank Marketing* fue seleccionado debido a su amplia presencia en estudios de minería de datos, su aplicación real en campañas bancarias y su riqueza en variables heterogéneas.

Contiene 41,188 observaciones y 20 variables predictoras mixtas (categóricas, numéricas y temporales), lo que permite evaluar técnicas avanzadas tanto de aprendizaje supervisado como no supervisado.

5 Hipótesis

Se plantea que la variable `duration` (duración de la última llamada realizada al cliente) tendrá el mayor peso predictivo sobre la variable objetivo `y`, que indica si el cliente acepta o no la campaña de depósito a término.

Asimismo, se considera que variables macroeconómicas (`euribor3m`, `emp.var.rate`, `cons.conf.idx`) influirán en la decisión del cliente.

6 Exploración Inicial del Dataset

6.1 Distribución de la Edad

En la Figura 1 se muestra la distribución de la edad, evidenciando un sesgo hacia adultos entre 30 y 50 años.

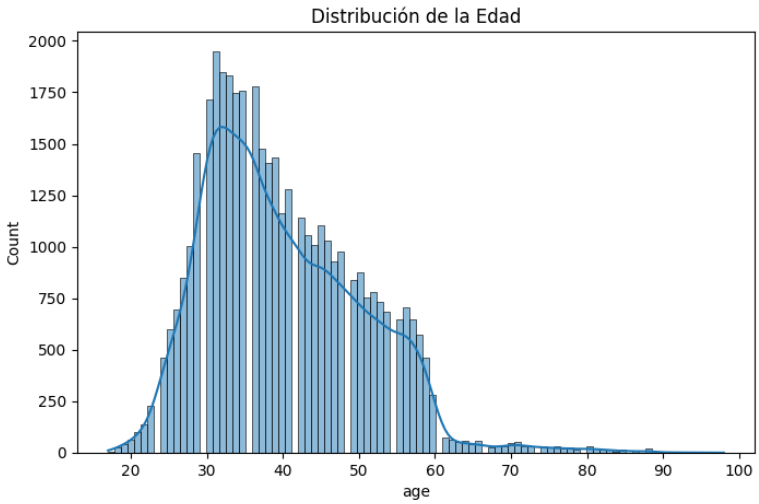

Figure 1: Distribución de la Edad

6.2 Correlación Numérica

La Figura 2 muestra la matriz de correlación. Se destaca una correlación fuerte entre las variables macroeconómicas, mientras que la duración de las llamadas (`duration`) es la variable con mayor correlación hacia la etiqueta.

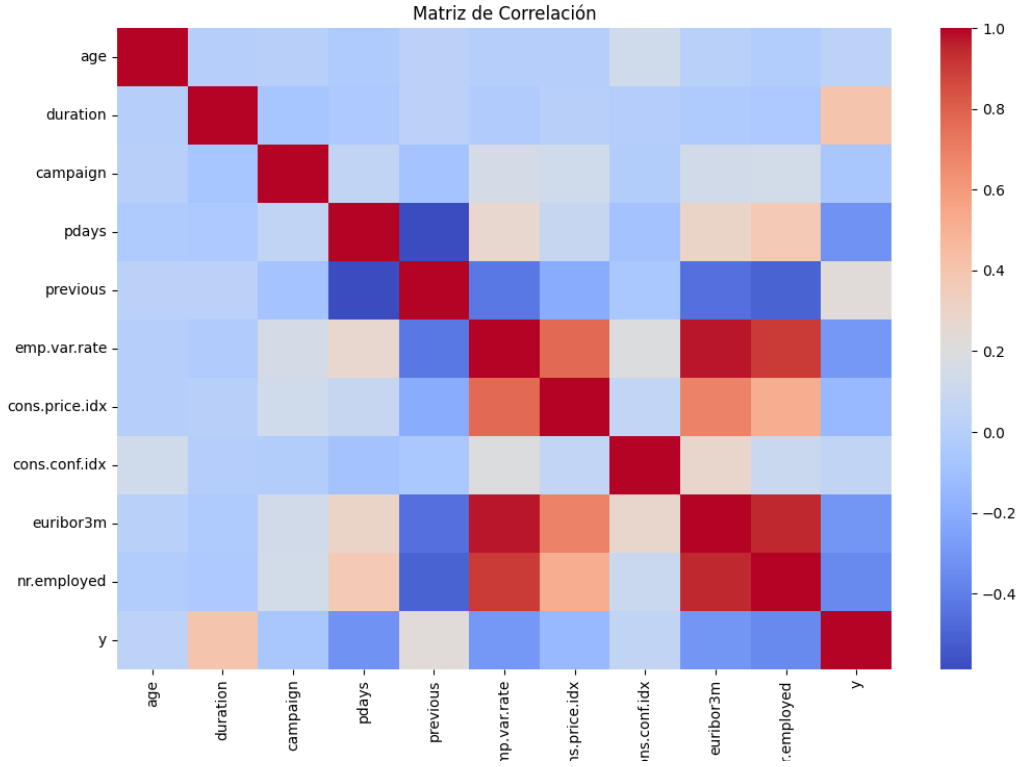

Figure 2: Matriz de Correlación Numérica

7 Preprocesamiento

Para preparar el dataset:

- Se convirtió la variable objetivo `y` a valores binarios.
- Se aplicó *one-hot encoding* a todas las variables categóricas.
- Se normalizaron las variables numéricas con StandardScaler.

El conjunto final resultó en 53 variables.

8 Reducción de Dimensionalidad con PCA

Se utilizó PCA reteniendo el 95% de la varianza. La Figura 3 muestra la varianza explicada de cada componente principal.

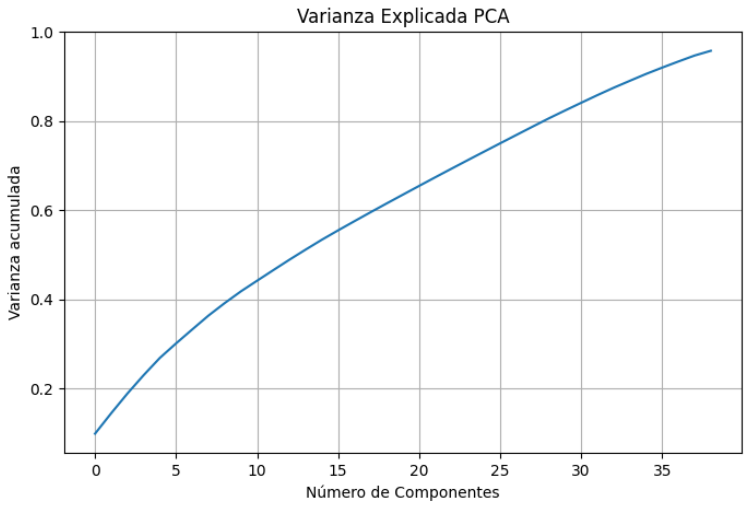

Figure 3: Varianza Explicada por PCA

Los componentes generados capturan adecuadamente la estructura del dataset.

9 Clustering con K-Means

Para determinar el número óptimo de clusters, se aplicó el método del codo (Figura 4). Se seleccionó $k = 3$.

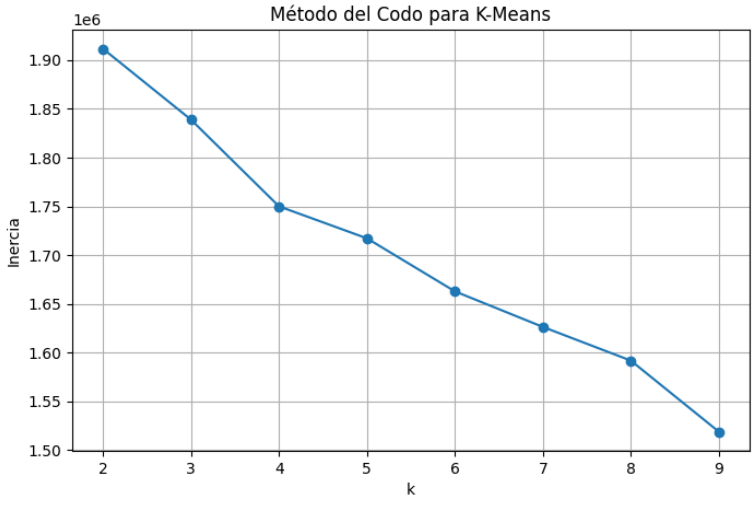

Figure 4: Método del Codo

Los clusters fueron proyectados en los dos primeros componentes principales (Figura 5).

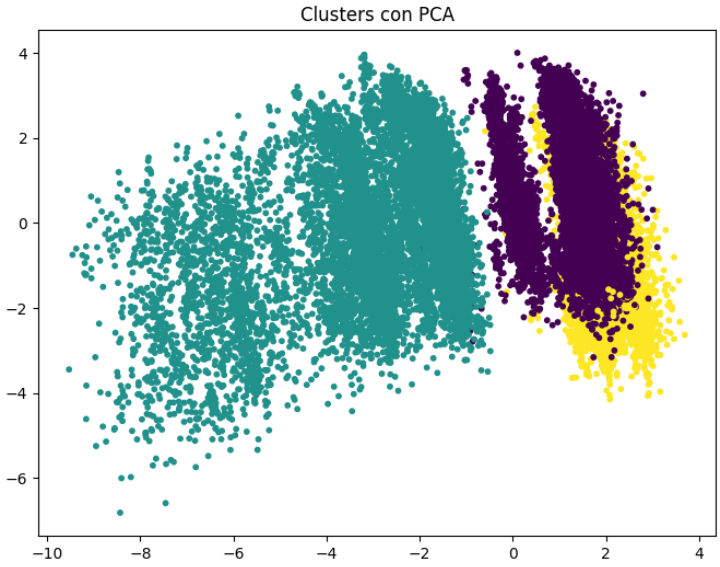

Figure 5: Clusters Proyectados con PCA

10 Modelos Supervisados

Se entrenaron tres modelos:

- Regresión Logística
- Random Forest
- SVM

Los resultados fueron:

- **Regresión Logística:** 91.21% de accuracy.
- **Random Forest:** 91.72% de accuracy.
- **SVM:** 90.59% de accuracy.

Las matrices de confusión se muestran en las Figuras 6, 7 y 8.

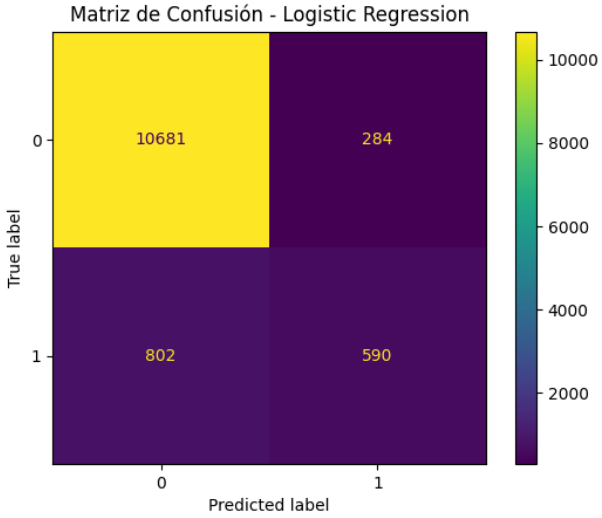

Figure 6: Matriz de Confusión - Regresión Logística

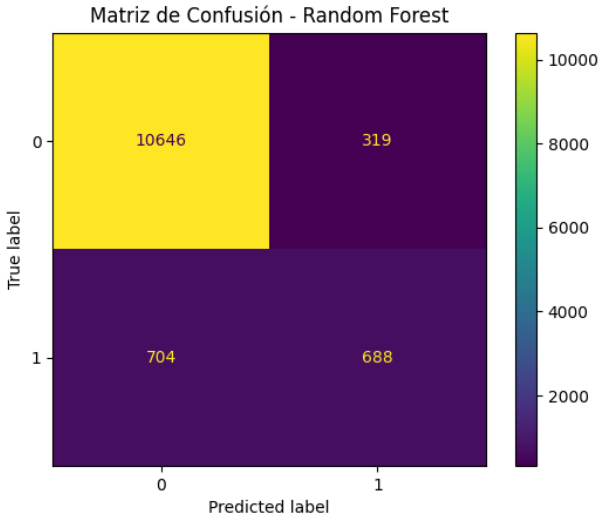

Figure 7: Matriz de Confusión - Random Forest

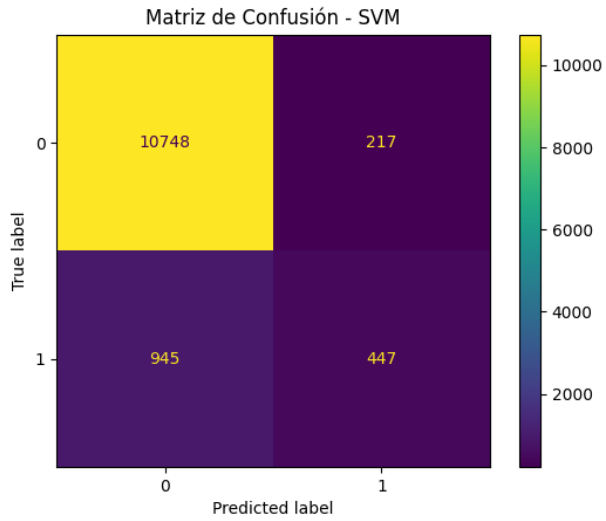

Figure 8: Matriz de Confusión - SVM

11 Validación Cruzada y Ajuste de Hiperparámetros

Se aplicó validación cruzada de 5 particiones, obteniéndose un promedio de 91.35% de accuracy.

Además, se realizó Grid Search sobre Random Forest, encontrando como mejor configuración:

- `n_estimators = 300`
- `max_depth = 20`

12 Implementación en C del Algoritmo KNN

Se implementó desde cero un modelo KNN usando las variables `age`, `duration` y `y`. El modelo logró un 77.80% de accuracy, lo que confirma la simplicidad del modelo y la pérdida de información por utilizar pocas variables.

13 Conclusiones

- El modelo con mejor desempeño fue Random Forest.
- El dataset está desbalanceado, lo que afecta la detección de la clase positiva.
- PCA mejoró la representación multidimensional del dataset.
- La implementación en C permitió comprender cálculos internos del algoritmo KNN.
- Clustering no coincide claramente con la variable objetivo, pero permite descubrir perfiles generales.