Predicción de Diabetes Tipo 2 utilizando Perceptrón Multicapa (MLP)

Práctica 1 – Implementación de un Perceptrón Multicapa (MLP)

Dataset: Diabetes – Kaggle
Autores: Noelia Blanco, María Eugenia Puchkariov, Gonzalo Del Priore **Fecha:** 06/08/2025
Materia: Deep Learning
Institución: Universidad de Montevideo

---

1. Resumen del problema

El objetivo es **predecir la presencia de diabetes** (binario: 0 = no, 1 = sí) a partir de variables clínicas (glucosa, IMC, edad, etc.).
Implementamos un **Perceptrón Multicapa (MLP)** y seguimos el pipeline completo: carga y descripción del dataset, preparación (limpieza + normalización), **split 80/20**, definición de arquitectura, entrenamiento, evaluación con métricas, visualizaciones, y conclusiones.

---

2. Transformación del problema

Este problema puede ser abordado como una tarea de clasificación binaria, donde la variable objetivo es si la paciente tiene diabetes (Outcome: 0 o 1). El objetivo es construir un modelo de red neuronal capaz de aprender patrones complejos no lineales a partir de las variables de entrada, y predecir correctamente la presencia de diabetes. Para ello, se transforma el problema en una función que mapea las variables clínicas a la probabilidad de un diagnóstico positivo.

Tabla de variables

- Pregnancies: Número de embarazos
- Glucose: Concentración de glucosa en plasma a las 2h
- BloodPressure: Presión arterial diastólica (mm Hg)
- SkinThickness: Espesor del pliegue cutáneo del tríceps (mm)
- Insulin: Insulina sérica a las 2h (mu U/ml)
- BMI: Índice de masa corporal (peso/altura²)
- DiabetesPedigreeFunction: Función de pedigrí de diabetes
- Age: Edad (años)
- Outcome: 1 si tiene diabetes, 0 en caso contrario

---

3. Descripción del dataset

Este dataset proviene del National Institute of Diabetes and Digestive and Kidney Diseases y tiene como objetivo predecir la aparición de diabetes tipo 2 en mujeres Pima mayores de 21 años, a partir de diversas mediciones clínicas.

Características del dataset

- Cantidad de registros: 768
- Cantidad de variables: 9 columnas (8 variables predictoras y 1 variable objetivo)

Variables

Variable	Descripción	Tipo
Pregnancies	Número de veces que la paciente ha estado embarazada	Numérica (entero)
Glucose	Concentración de glucosa en plasma 2 horas después de una prueba oral	Numérica
BloodPressure	Presión arterial diastólica (mm Hg)	Numérica
SkinThickness	Espesor del pliegue cutáneo del tríceps (mm)	Numérica
Insulin	Insulina sérica en 2 horas (mu U/ml)	Numérica
BMI	Índice de masa corporal (peso en kg / (altura en m)^2)	Numérica
DiabetesPedigreeFunction	Función de pedigrí (riesgo hereditario de diabetes)	Numérica
Age	Edad de la paciente (en años)	Numérica
Outcome	Resultado: 0 = No diabética, 1 = Diabética	Categórica binaria

---

4. Preparación y división de los datos

En esta sección se lleva a cabo la limpieza, transformación y preparación del dataset para el entrenamiento del modelo de clasificación.

Reemplazo de valores inválidos

En varias columnas clínicas, como Glucose, BloodPressure, SkinThickness, Insulin y BMI, aparecen valores iguales a cero (0), los cuales no son posibles en un contexto fisiológico real y se consideran datos faltantes o inválidos.

Para solucionar esto:

1. Se reemplazan los ceros por valores nulos (NaN).
2. Se imputan estos valores nulos utilizando la *media de cada columna*, preservando la distribución de los datos sin eliminar filas.

---

5. Arquitectura del modelo MLP

Se construirán dos modelos de Perceptrón Multicapa (MLP):

Uno utilizando scikit-learn

Otro con Keras, incorporando una técnica adicional de regularización Ambos modelos serán entrenados con los mismos datos, y se comparará su desempeño en el conjunto de prueba.

Evaluación del MLP con Scikit-learn

Una vez entrenado el modelo MLPClassifier, se realiza la predicción sobre el conjunto de prueba y se evalúa su rendimiento utilizando métricas estándar y la matriz de confusión.

Reporte de clasificación

El reporte incluye métricas clave para cada clase:

- *Precisión (precision)*: proporción de verdaderos positivos entre los casos predichos como positivos.
- *Recall*: proporción de verdaderos positivos entre los casos realmente positivos.
- *F1-score*: media armónica entre precisión y recall.
- *Support*: número de muestras reales por clase.

Resultados obtenidos:

Clase	Precision	Recall	F1-score	Soporte
0 (No diabética)	0.81	0.79	0.80	100
1 (Diabética)	0.62	0.65	0.64	54

- *Accuracy global*: 0.74 (74%)
- *Promedios macro y ponderado*: 0.72–0.74

Esto indica que el modelo tiene un desempeño razonable, con mayor efectividad al clasificar correctamente a pacientes no diabéticas (clase 0), y una leve disminución de rendimiento al predecir la clase 1 (diabéticas), posiblemente debido al desbalance de clases.

Matriz de confusión

La matriz de confusión permite visualizar los verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos:

- El modelo acertó en *79 casos de clase 0* y *35 casos de clase 1*.
- Cometió *21 falsos positivos* (predijo 1 cuando era 0) y *19 falsos negativos* (predijo 0 cuando era 1).

La visualización confirma que el modelo es más conservador al predecir la clase positiva (1), y muestra un rendimiento aceptable pero mejorable, especialmente en la detección de pacientes diabéticas.

Este análisis sugiere que podrían aplicarse técnicas de mejora como balanceo de clases, ajuste de hiperparámetros o regularización para optimizar el desempeño.

6. Visualizaciones

Se grafican la evolución de la precisión y la función de pérdida durante el entrenamiento del modelo en Keras. Esto permite evaluar visualmente la convergencia del modelo y detectar posibles problemas como sobreajuste o subajuste.

---

7. Conclusiones

Este proyecto tuvo como objetivo aplicar redes neuronales de tipo Perceptrón Multicapa (MLP) para resolver un problema de clasificación binaria, utilizando un conjunto de datos clínicos de pacientes propensos a diabetes tipo 2. A lo largo del trabajo, se abordaron las etapas de análisis exploratorio, limpieza, normalización, entrenamiento de modelos y evaluación comparativa.

Se implementaron dos modelos principales:

- *MLP con Scikit-learn: permitió una implementación rápida, eficiente y con buenos resultados sin necesidad de mayor configuración. Alcanzó un **accuracy del 74%* y mostró un buen desempeño general, especialmente al clasificar correctamente a pacientes no diabéticos.
- *MLP con Keras y regularización con Dropout: ofreció mayor flexibilidad en la definición de la arquitectura y la posibilidad de aplicar técnicas modernas de regularización. Alcanzó una *precisión del 73.38%, con una función de pérdida final de *0.5271*, y mostró buen comportamiento de generalización a lo largo del entrenamiento.

Aspectos clave observados

- La *normalización previa* de los datos resultó esencial para garantizar un entrenamiento eficiente de las redes neuronales.
- El tratamiento de *valores faltantes o inválidos*, mediante imputación con la media, fue fundamental para evitar sesgos.
- La *división estratificada* del conjunto de datos permitió mantener una proporción adecuada de clases.
- La visualización de métricas de entrenamiento y validación fue clave para *detectar problemas de sobreajuste o subajuste*, los cuales no se observaron en este caso.

Ambos modelos poseen buena capacidad de predecir si un paciente es o no diabètico, aunque es necesario maximizar el recall, ya que, en aquellos casos donde el paciente sea diabético y nuestro sistema prediga que no lo es, no se va a asignar un tratamiento perpetuando el riesgos a su salud. Una alternativa sería probar modelos especializados como XGBoost que permiten un control más fino del balance entre precision y recall mediante la manipulación de sus hiperparámetros.