

DataHub Personal – Presentación Documentada

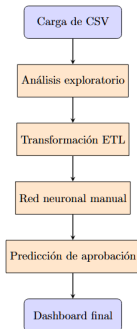
Francisco Pinto

UTEM – INFB6052

Julio 2025

1. Contexto del Proyecto

- Proyecto de ciencia de datos aplicado al análisis educativo.
- Uso de Streamlit, MongoDB, Python y Machine Learning desde cero.
- Flujo estructurado en cinco etapas documentadas.



2. Carga y análisis exploratorio

- Carga de CSV y exploración de columnas clave.
- Distribuciones y estadísticas iniciales.

```
import streamlit as st
import matplotlib.pyplot as plt

st.set_page_config(page_title="Análisis Exploratorio", layout="wide")
st.title("🐡 Análisis exploratorio")

if "data_uploaded" not in st.session_state or not st.session_state["data_uploaded"]:
    st.warning("Por favor sube un archivo en la página principal primero.")
    st.stop()

df = st.session_state["df"]


st.subheader("Vista previa")
st.dataframe(df.head())

st.write(f"Registros: {len(df)}")
st.write(f"Columnas: {list(df.columns)}")

if 'math score' in df.columns:
    st.subheader("Distribución de puntajes de matemáticas")
    fig, ax = plt.subplots()
    ax.hist(df['math score'], bins=20)
    st.pyplot(fig)
```

3. Proceso ETL

- Limpieza de datos y generación de nuevas columnas.
- Promedio de notas y cálculo de aprobación.
- Exportación a MongoDB.

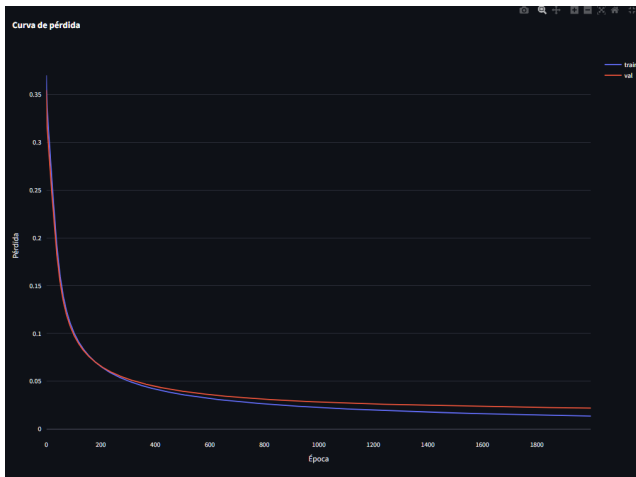
 **Simulación de flujo ETL**

✓ ETL completado y datos transformados.

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score	average_score	passed
0	male	group B	master degree	free/reduced	completed	49	71	56	58.6667	0
1	male	group C	bachelor degree	standard	none	56	66	56	59.3333	0
2	male	group B	master degree	standard	completed	4	24	32	20	0
3	female	group C	some high school	free/reduced	completed	30	99	22	50.3333	0
4	male	group B	associate degree	free/reduced	completed	67	17	37	40.3333	0

4. Red neuronal desde cero

- Entrenamiento de una red MLP simple en Python puro.
- Visualización de curva de pérdida y pesos.
- Interfaz para predecir aprobaciones.



5. Dashboard Final

- Métricas agregadas y visualización por grupos.
- Lectura directa de predicciones guardadas.



6. Conclusiones

- Proyecto completo, funcional y replicable.
- Útil como base para sistemas de recomendación académica.
- Estructura modular con alta mantenibilidad.

