

AVANCE | FASE I DE PROYECTO

Integrantes:

Andrea Monserrat Ruiz Gomez - **A01794631**

Daniel Acevedo Sainos - **A01795496**

Luis Alejandro Aguilar - **A01795362**

Juan Manuel Rodríguez Mateos – **A01794890**

Marcos Eduardo García Ortiz - **A01276213**

Héctor Raúl Peraza Alvarez - **A01795125**



CONTENIDO

1. Introducción
2. Manipulación y preparación de los datos
3. Exploración y Preprocesamiento de Datos
4. Versionado de Datos
5. Construcción, Ajuste y Evaluación de Modelos
6. Aplicación de Mejores Prácticas
7. Conclusiones

1. INTRODUCCIÓN

Análisis Introductorio

- El objetivo principal de este proyecto es desarrollar un modelo de Machine Learning capaz de predecir la presencia de diabetes en pacientes a partir de un conjunto de datos que contiene características clínicas. Para lograr este objetivo, se aplicarán técnicas de análisis exploratorio de datos (EDA), preprocesamiento de datos, selección de modelos, entrenamiento y evaluación, así como mejores prácticas de MLOps para asegurar la reproducibilidad, el despliegue eficiente y la monitorización del modelo.

Objetivos

- Desarrollar un modelo de Machine Learning con alta precisión para la predicción de diabetes.
- Identificar las variables clínicas que son más relevantes para la predicción de la diabetes.
- Automatizar el pipeline de Machine Learning utilizando MLOps.
- Implementar un sistema de monitorización del modelo para asegurar su rendimiento a lo largo del tiempo.

2. MANIPULACIÓN Y PREPARACIÓN DE LOS DATOS

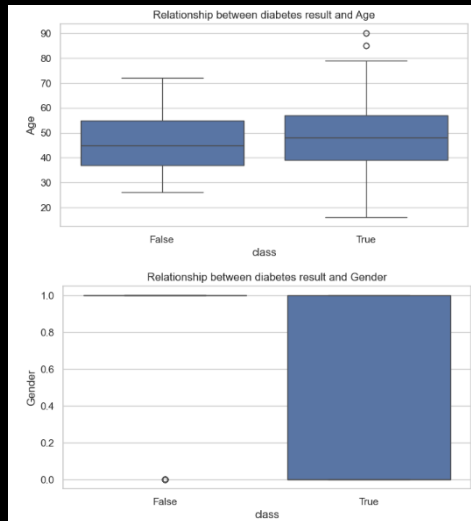
- **Importación de datos:** Se utilizó la biblioteca pandas para importar el dataset de diabetes, el cual contiene **520 registros y 17 columnas**.
- **Limpieza de datos:** Se manejaron valores nulos y se convirtieron a variables categóricas a booleanas para facilitar el procesamiento.
- **Transformaciones necesarias:** Se realizaron transformaciones utilizando pipelines para escalar los datos numéricos y codificar las variables categóricas.

	Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis
0	40	Male	No	Yes	No	Yes	No	No	No	Yes	No	Yes	No
1	58	Male	No	No	No	Yes	No	No	Yes	No	No	No	Yes
2	41	Male	Yes	No	No	Yes	Yes	No	No	Yes	No	Yes	No
3	45	Male	No	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No
4	60	Male	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes

muscle stiffness	Alopecia	Obesity	class
Yes	Yes	Yes	Positive
No	Yes	No	Positive
Yes	Yes	No	Positive
No	No	No	Positive
Yes	Yes	Yes	Positive

3. EXPLORACIÓN Y PREPROCESAMIENTO DE DATOS

Análisis Exploratorio de Datos (EDA)



Se realizaron visualizaciones y estadísticas descriptivas para identificar patrones y relaciones importantes en los datos clínicos

Preprocesamiento

```
Out[11]:
```

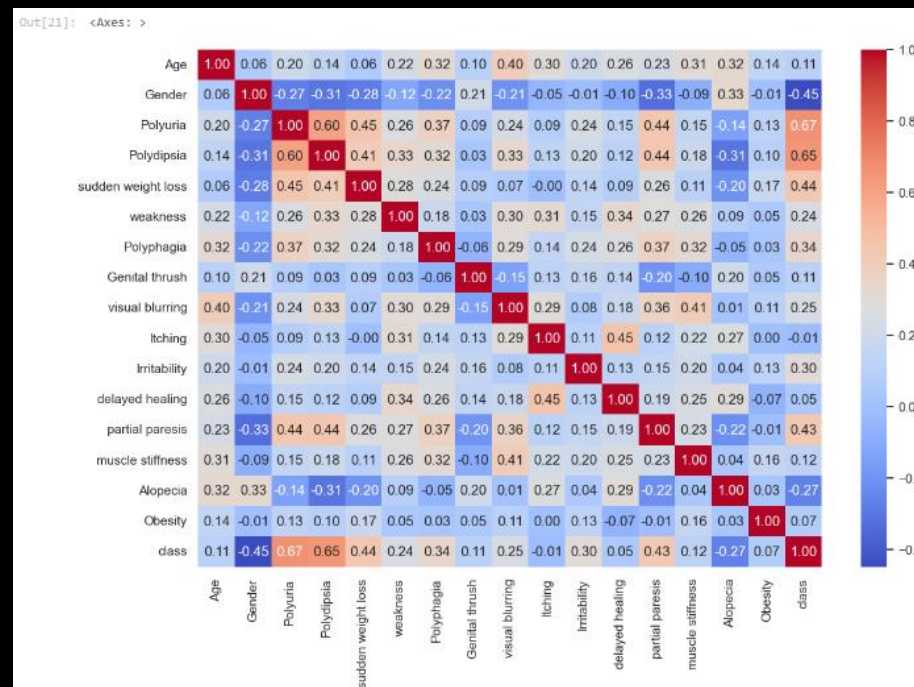
	count	mean	std	min	25%	50%	75%	max
Age	520.0	48.028846	12.151466	16.0	39.0	47.5	57.0	90.0

Transformación de variables escalares a booleanas

Se aplicaron técnicas de normalización y codificación de variables categóricas para mejorar la calidad de Dataset

3. EXPLORACIÓN Y PREPROCESAMIENTO DE DATOS

Reducción de dimensionalidad

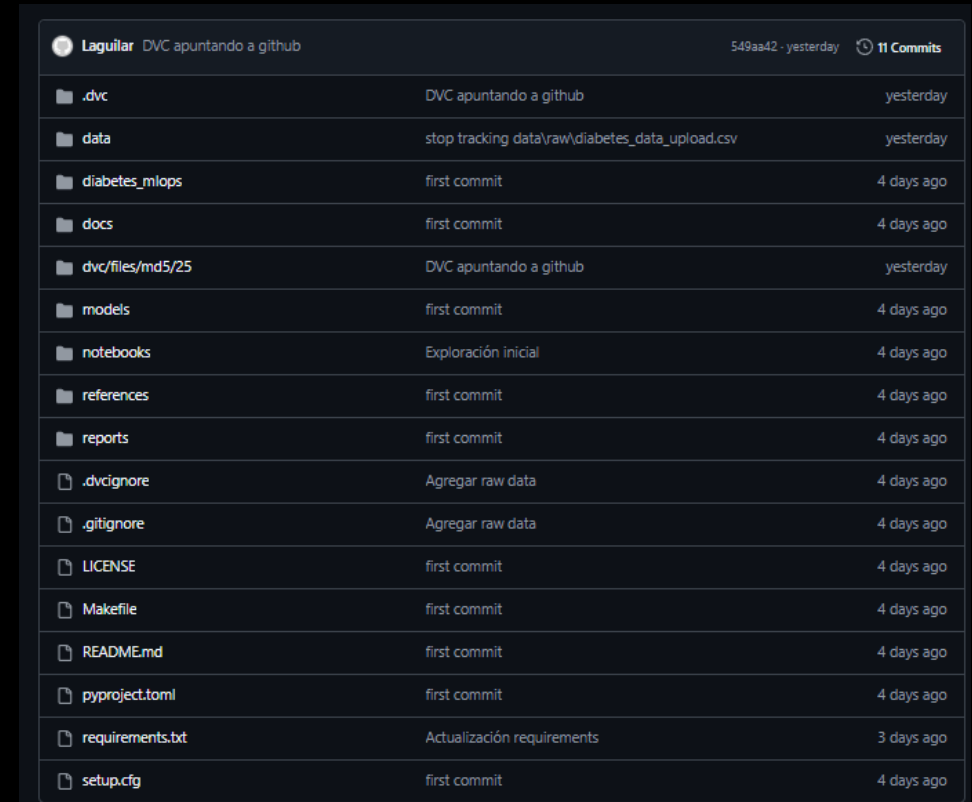


Se evaluó la correlación entre las variables para optimizar el modelado

4. VERSIONADO DE DATOS

Gestión de Cambios y Reproducibilidad

- **Control de versiones:** se utilizó un repositorio en GitHub y DVC para gestionar los cambios en los datos y así asegurar la reproducibilidad del proyecto.
- **Registro de modificaciones:** cada cambio en los datos fue documentado, permitiendo mantener un historial detallado para auditoría.
- **Reproducibilidad garantizada:** el uso de herramientas de versionado asegura que los experimentos pueden ser reproducidos y los datos rastreados de manera precisa.



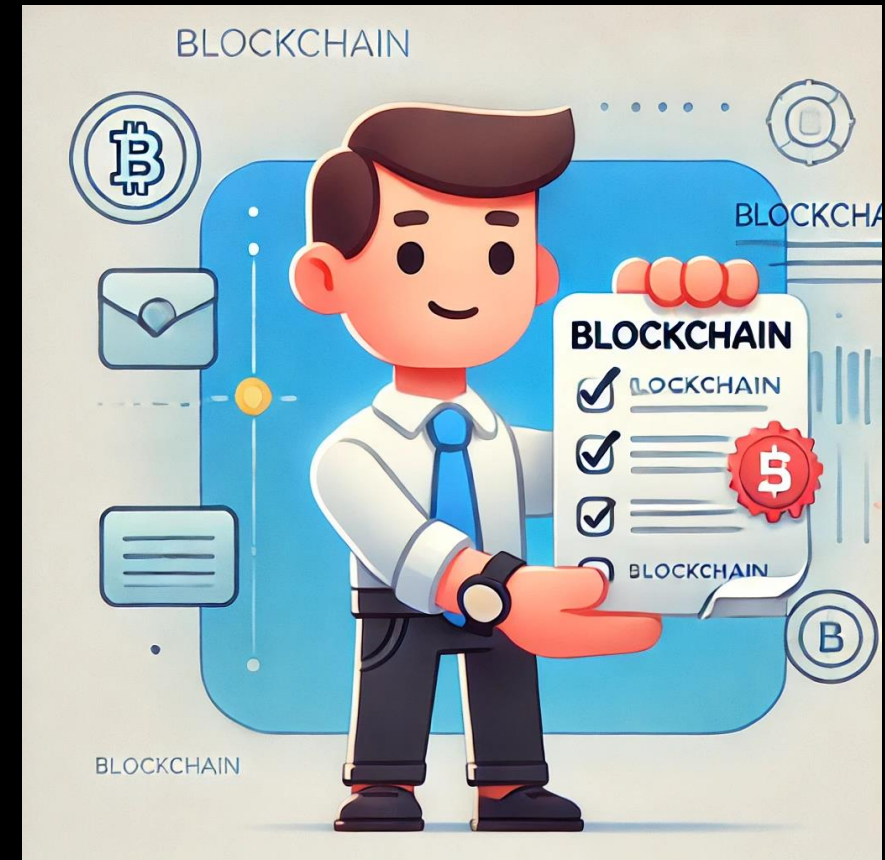
The screenshot shows a DVC repository interface for a project named 'Laguilar'. The interface displays a list of files and folders tracked by DVC, along with their commit history and timestamps. The repository is pointing to a GitHub repository.

File/Folder	Commit Message	Commit Time
.dvc	DVC apuntando a github	yesterday
data	stop tracking data/raw/diabetes_data_upload.csv	yesterday
diabetes_mlops	first commit	4 days ago
docs	first commit	4 days ago
dvc/files/md5/25	DVC apuntando a github	yesterday
models	first commit	4 days ago
notebooks	Exploración inicial	4 days ago
references	first commit	4 days ago
reports	first commit	4 days ago
.dvcignore	Agregar raw data	4 days ago
.gitignore	Agregar raw data	4 days ago
LICENSE	first commit	4 days ago
Makefile	first commit	4 days ago
README.md	first commit	4 days ago
pyproject.toml	first commit	4 days ago
requirements.txt	Actualización requirements	3 days ago
setup.cfg	first commit	4 days ago

5. CONSTRUCCIÓN, AJUSTE Y EVALUACIÓN DE MODELOS

Desarrollo y Optimización del proyecto

- **Selección del algoritmo:** Se eligió la regresión logística para la predicción de diabetes debido a su capacidad para clasificar de manera eficiente.
- **Ajuste de hiperparámetros:** Se optimizaron los hiperparámetros del modelo usando validación cruzada para mejorar la precisión.
- **Evaluación del modelo:** Se utilizó la matriz de confusión y el reporte de clasificación para medir el rendimiento, alcanzando un 97% de precisión.



6. APLICACIÓN DE MEJORES PRÁCTICA

Optimización del Pipeline de Modelado

- **Automatización del pipeline:** Se implementó un pipeline que automatiza las tareas de preprocesamiento, entrenamiento y evaluación del modelo.
- **Reproducibilidad:** Las prácticas de MLOps como el versionado de datos y el seguimiento de experimentos garantizan que el proyecto sea reproducible.
- **Eficiencia y escalabilidad:** El pipeline optimizado mejora la eficiencia y escalabilidad, asegurando que el modelo pueda adaptarse a nuevos datos y escenarios.

```
In [53]: report = classification_report(y_test, y_pred, output_dict=True)
report_df = pd.DataFrame(report).transpose()
print(report_df)
```

	precision	recall	f1-score	support
False	0.903226	0.848485	0.875000	33.000000
True	0.931507	0.957746	0.944444	71.000000
accuracy	0.923077	0.923077	0.923077	0.923077
macro avg	0.917366	0.903116	0.909722	104.000000
weighted avg	0.922533	0.923077	0.922409	104.000000

Mejoras del modelo

```
In [54]: scores = cross_val_score(model, X, y, cv=10)
print("Average accuracy with CV:", np.mean(scores))
```

Average accuracy with CV: 0.9269230769230768

```
In [37]: with mlflow.start_run() as run:
mlflow.sklearn.log_model(model, "model")
mlflow.log_params(model.get_params())
```

6. APLICACIÓN DE MEJORES PRÁCTICA

Integración de DVC en el Pipeline de Machine Learning

1. Gestión de versiones :

1. Se implementó DVC para controlar las versiones de los datos y modelos, lo que permite tener un historial claro de los cambios y facilitar la colaboración.

2. Rastreo de datos :

1. DVC permite rastrear los conjuntos de datos utilizados, asegurando que todos los miembros del equipo que trabajen con la misma versión de los datos.

3. Manejo de grandes conjuntos de datos :

1. DVC ayuda a gestionar conjuntos de datos grandes sin sobrecargar el sistema, lo que mejora el rendimiento y la eficiencia del proyecto.

4. Reproducibilidad :

1. Gracias a DVC, los experimentos se pueden reproducir fácilmente en cualquier entorno, lo que garantiza consistencia en los resultados.

5. Remoto de montaje :

1. Se utilizó DVC para almacenar los conjuntos de datos de manera eficiente en un repositorio remoto, facilitando el acceso y la colaboración.

7. CONCLUSIONES

- En conclusión, el desarrollo de este modelo de predicción de diabetes ha demostrado ser eficaz para identificar a pacientes en riesgo a partir de datos clínicos básicos. La combinación de un enfoque basado en Machine Learning con prácticas robustas de MLOps ha permitido no solo crear un modelo preciso, sino también asegurar su despliegue y mantenimiento a largo plazo.
- Además, la identificación de las variables más relevantes puede proporcionar información valiosa a los profesionales de la salud para la toma de decisiones informadas, mejorando la detección temprana de la diabetes y contribuyendo así a la prevención de complicaciones graves.
- Este proyecto sienta las bases para futuros desarrollos y mejoras, como la integración de datos adicionales o la optimización de los algoritmos, lo que podría aumentar aún más la precisión y utilidad clínica del modelo.

