

Fase1 Avance de Proyecto

Carlos - A01066264

Eduardo - A01796827

Jorge Acevedo - A00565936

Miguel Marines - A01705317

César Tirado - A01795845

Link video

<https://youtu.be/rYL3GA8cOVI>

01

Contexto

02

Análisis de
Requerimientos

03

Manipulación y
preparación de datos

04

Exploración y
procesamiento de datos

05

Modelado y evaluación










06

Siguientes pasos

Agenda

Machine Learning Canvas

<https://docs.google.com/document/d/16SzESTBqJna5VncSqPsJ9dns-QjF8p5nc1l8Xmcx7kc/edit?tab=t.0#heading=h.1h1dkvi2yfu9>

<p>PREDICTION TASK </p> <p>Type of task: Regression.</p> <p>Entity of prediction: Individual employee.</p> <p>Possible outcomes: Number of hours absent (continuous variable, e.g. 0–40 hours per week).</p> <p>When outcomes are observed: After each work period (weekly/monthly attendance records).</p>	<p>DECISIONS </p> <p>How predictions become actions:</p> <ul style="list-style-type: none"> HR and supervisors receive predicted absence hours per employee. If predicted absence > threshold (e.g. 8 hours/week), system suggests preventive actions: <ul style="list-style-type: none"> Adjust work schedules. Reassign workloads. Contact employees for early support (health, logistics). <p>Parameters:</p> <ul style="list-style-type: none"> Prediction confidence intervals. Absence thresholds configurable by HR. Integration into the workforce management system. 	<p>VALUE PROPOSITION </p> <p>Beneficiaries: HR teams, managers, company operations.</p> <p>Pain points solved:</p> <ul style="list-style-type: none"> Unplanned staff shortages. Loss of productivity. Reactive instead of proactive HR planning. <p>Integration & Interface:</p> <ul style="list-style-type: none"> Embedded in the HR dashboard (web-based). Automated alerts via email/slack for high-risk employees. 	<p>DATA COLLECTION </p> <p>Initial sourcing:</p> <ul style="list-style-type: none"> HR database extracts (attendance, demographics, health reports). Manual input from surveys (transport, wellbeing). <p>Continuous update strategy:</p> <ul style="list-style-type: none"> Weekly batch ingestion from the attendance system. Monthly health survey updates. Cost control via incremental database queries and scheduled ETL jobs. 	<p>DATA SOURCES </p> <p>Internal:</p> <ul style="list-style-type: none"> HR attendance table. Employee demographics (HR system). Health reports database. <p>External:</p> <ul style="list-style-type: none"> Public transportation delay APIs. Regional health statistics.
<p>IMPACT SIMULATION </p> <p>Cost/gain values:</p> <ul style="list-style-type: none"> Correct prediction (absence anticipated): Saves cost of emergency replacements (~\$200–\$500/employee/week). Incorrect prediction (absence not anticipated): Uncovered shifts, productivity loss. False alarm (predicting absence that doesn't occur): Overstaffing cost, but less critical. <p>Pre-deployment simulation:</p> <ul style="list-style-type: none"> Historical HR records (3+ years). Backtesting predictions against real absences. <p>Deployment criteria:</p> <ul style="list-style-type: none"> MAE < 2 hours/week. Stable performance across employee subgroups. <p>Fairness constraints:</p> <ul style="list-style-type: none"> Avoid bias toward specific age groups, genders, or health conditions. 	<p>MAKING PREDICTIONS </p> <p>Mode: Batch predictions.</p> <p>Frequency: Weekly (ahead of planning schedules).</p> <p>Time budget: < 5 minutes per batch run (featureization + model scoring).</p> <p>Compute resources:</p> <ul style="list-style-type: none"> Cloud VM (medium tier). Scalable if HR database grows. 		<p>BUILDING MODELS </p> <p>Models needed:</p> <ul style="list-style-type: none"> Primary regression model (hours absent). Benchmark baseline model (e.g. linear regression). <p>Update frequency: Quarterly retraining or when prediction drift is detected.</p> <p>Time budget: < 1 hour training on full dataset.</p> <p>Compute resources:</p> <ul style="list-style-type: none"> Cloud ML service (e.g. AWS SageMaker / GCP Vertex AI). GPU not required (tree-based regressors sufficient). 	<p>FEATURES </p> <p>Entity representation: Employee profile vector (age, education, distance, transport, health history).</p> <p>Transformations:</p> <ul style="list-style-type: none"> Normalize continuous variables (distance, age). Encode categorical variables (transport type, education level). Aggregations: past average absences (last 3 months). Derived features: distance/time ratio, health risk score.

Introducción y Contexto

Operaciones de Aprendizaje Automático (Machine Learning Operations)

Propósito del Proyecto

El objetivo es anticipar patrones de ausentismo laboral, impulsando la productividad y el bienestar en la organización.

Calidad y Preparación de Datos

Se asegura la calidad y adecuada preparación de los datos antes de crear el modelo predictivo.

Colaboración Multidisciplinaria

El trabajo colaborativo de Data Engineer, Data Scientist, ML Engineer, Software Engineer y DevOps fue clave en la ejecución.

Automatización de Procesos

DevOps automatizó los procesos con herramientas modernas como Git y DVC, garantizando eficiencia y trazabilidad.



Análisis de Requerimientos

Definición estratégica del problema y propuesta de valor

Impacto del ausentismo laboral

El ausentismo laboral eleva los costos operativos y dificulta la planeación eficiente, afectando la productividad empresarial.

Predicción personalizada de ausencias

Se busca anticipar las horas de ausencia de cada colaborador según perfil, entorno y comportamiento para mejorar la gestión.

Optimización y prevención

La predicción permite optimizar recursos, implementar estrategias de bienestar y prevenir interrupciones, elevando la productividad.

Manipulación y Preparación de Datos

Garantizar datos confiables para decisiones inteligentes

Eliminación de datos inconsistentes

Se eliminaron duplicados, valores nulos y registros inconsistentes para mejorar la calidad del conjunto de datos.

Estandarización y corrección de formatos

Se estandarizaron los formatos y se corrigieron valores fuera de rango, logrando uniformidad en las variables.

Automatización en la limpieza de datos

Se implementó un pipeline automatizado de limpieza para mantener la consistencia en futuras actualizaciones de datos.

Resultado: un conjunto de datos limpio, estandarizado y validado para modelado.

Manipulación y Preparación de Datos

Removed Variables

Variable	Reason for Removal
ID	It's a unique identifier with no predictive value.
mixed_type_col	Contained mixed or inconsistent values that didn't provide useful information to the model.

These columns were removed at the beginning of the pipeline to reduce noise and prevent errors during later processing.

Categorical Variables

Variable	Applied Transformation	Justification
Reason for absence	Values outside the 0-28 range were replaced with 0 (Unknown).	Ensures consistency with the defined code mapping and avoids errors from corrupt values.
Month of absence	Values outside the 0-12 range were replaced with 0 (Unknown).	Ensures consistency with the defined code mapping and avoids errors from corrupt values.
Day of the week	Values outside the 2-6 range were imputed with the mode.	Only weekdays exist in this dataset; imputing with the most frequent value prevents distortion.
Seasons	Values outside the 1-4 range were imputed with the mode.	Limited to 4 seasons; imputing the mode avoids data loss.
Education	Values outside the 1-4 range were imputed with the mode.	Only valid education levels are retained; the mode preserves general distribution.
Disciplinary failure	Values other than 0 or 1 were imputed with the mode.	This binary variable must be 0 or 1; errors were corrected without introducing bias.
Social drinker	Values other than 0 or 1 were imputed with the mode.	Ensures consistency in this binary variable.
Social smoker	Values other than 0 or 1 were imputed with the mode.	Maintains integrity of binary data.

Numerical Variables

Variable	Applied Transformations	Justification
Transportation expense	IQR Winsorization + Median Imputation + Final Rounding	Controls outliers, fills in missing values, and formats as integers for modeling.
Distance from Residence to Work	IQR Winsorization + Median Imputation + Final Rounding	Normalizes distribution and ensures numerical integrity.
Service time	IQR Winsorization + Median Imputation + Final Rounding	Improves consistency in employment-related data.
Age	IQR Winsorization + Median Imputation + Final Rounding	Ensures a logical age range and complete data.
Work load Average/day	IQR Winsorization + Median Imputation + Final Rounding	Reduces impact of extreme values and fills in missing data.
Hit target	IQR Winsorization + Median Imputation + Final Rounding	Adjusts the variable to a realistic and coherent range.
Son	IQR Winsorization + Median Imputation + Final Rounding	Though discrete, extreme and missing values are treated.
Pet	IQR Winsorization + Median Imputation + Final Rounding	Kept as an integer while controlling for outliers.
Weight	IQR Winsorization + Median Imputation + Final Rounding	Ensures values fall within a physiologically plausible range.
Height	IQR Winsorization + Median Imputation + Final Rounding	Controls for logical range and completes data.
Body mass index	IQR Winsorization + Median Imputation + Final Rounding	Improves consistency of this index derived from weight and height.
Absenteeism time in hours	IQR Winsorization + Median Imputation + Final Rounding	Target variable; cleaned to avoid bias and prediction errors.

Pipeline

```
[17] > preprocessing_pipeline = Pipeline([
    1 ('drop_columns', FunctionTransformer(drop_columns)),
    2 ('strip_objects', FunctionTransformer(strip_object_columns)),
    3 ('safe_round', FunctionTransformer(safe_round_to_int_df)),
    4 ('fix_invalids', FunctionTransformer(fix_invalid_values)),
    5 ('winsorize', FunctionTransformer(winsorize_1qr)),
    6 ('fillna', FunctionTransformer(fillna_with_medians)),
    7 ('final_int', FunctionTransformer(final_int_conversion))
    8 ])

Terminal X
/content/drive/MyDrive/MLops/MLops_Project# git log --oneline
c4593a2 (HEAD -> master) Dataset V2 - DVC
dc4fc98 Dataset V1 - DVC
05701a4 Initialize DVC
2558233 (origin/master, origin/HEAD) Project Dataset
/content/drive/MyDrive/MLops/MLops
/content/drive/MyDrive/MLops/MLops_Pr
/content/drive/MyDrive/MLops/MLops_Project#
```

Clean Dataset

```
[18] > df_clean = preprocessing_pipeline.fit_transform(df.copy())

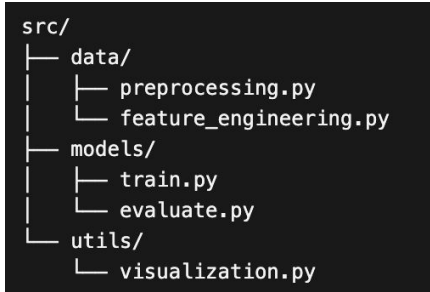
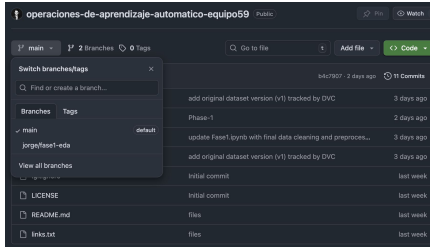
[18] > df_clean.to_csv(os.path.join(os.path.abspath(DATA_PATH), "work_absenteeism.csv"), index=
```

Script Manipulación y Preparación de Datos:

https://github.com/Cpano98/operaciones-de-aprendizaje-automatico-equipos59/blob/main/AbsenteeismAtWork/Phase-1/data_preparation.ipynb

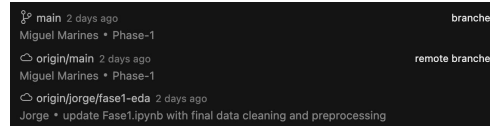
Flujo de trabajo e Ingeniería de Software

Repo GitHub: <https://github.com/Cpano98/operaciones-de-aprendizaje-automatico-equipo59>



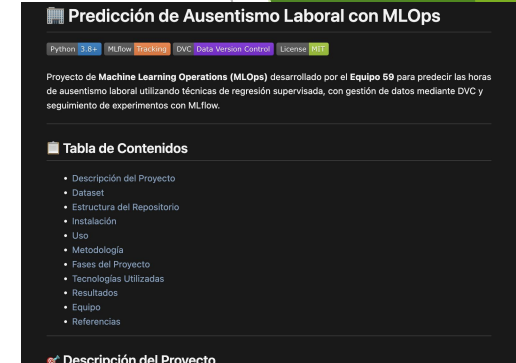
Repositorio de GitHub

Diseño de estructura para scripts, data y documentación.



Flujo colaborativo

PR's, Ramas.

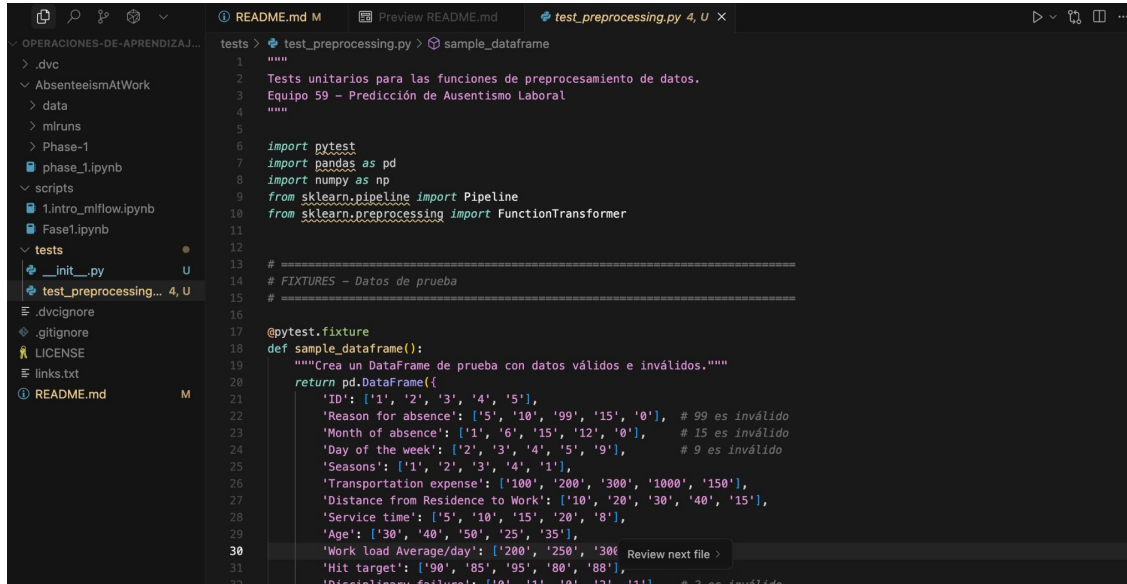


Documentación

Repositorio documentado para instalación de dependencias.

Flujo de trabajo e Ingeniería de Software

Repo GitHub: <https://github.com/Cpano98/operaciones-de-aprendizaje-automatico-equipo59>



The screenshot shows a code editor with a file explorer on the left and a code editor on the right. The file explorer shows a project structure with folders like 'data', 'scripts', and 'tests'. The code editor displays the content of 'test_preprocessing.py'. The code includes imports for 'pytest', 'pandas', 'numpy', 'sklearn.pipeline', and 'sklearn.preprocessing'. It defines a fixture 'sample_dataframe' that creates a pandas DataFrame with various data points, including some invalid values. The code is written in Python and uses standard indentation for function definitions and loops.

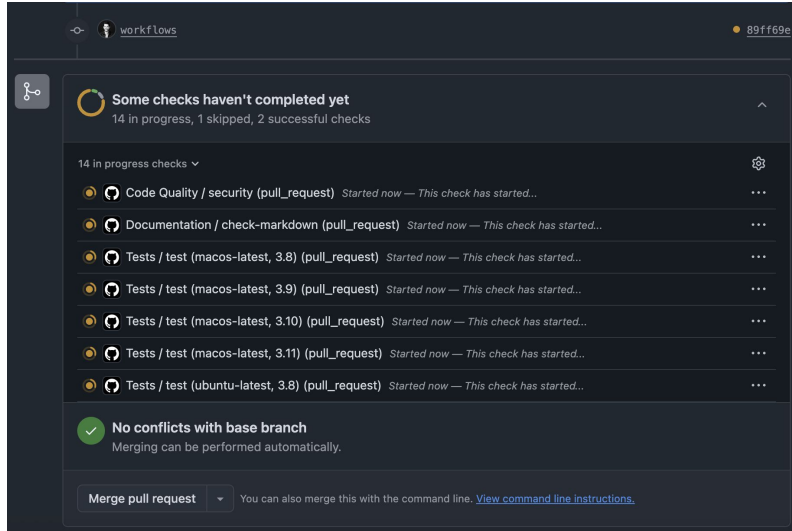
```
tests > test_preprocessing.py > sample_dataframe
1
2 Tests unitarios para las funciones de preprocesamiento de datos.
3 Equipo 59 - Predicción de Ausentismo Laboral
4
5
6 import pytest
7 import pandas as pd
8 import numpy as np
9 from sklearn.pipeline import Pipeline
10 from sklearn.preprocessing import FunctionTransformer
11
12
13 # =====
14 # FIXTURES - Datos de prueba
15 # =====
16
17 @pytest.fixture
18 def sample_dataframe():
19     """Crea un DataFrame de prueba con datos válidos e inválidos."""
20     return pd.DataFrame({
21         'ID': ['1', '2', '3', '4', '5'],
22         'Reason for absence': ['15', '10', '99', '15', '0'], # 99 es inválido
23         'Month of absence': ['1', '6', '15', '12', '0'], # 15 es inválido
24         'Day of the week': ['2', '3', '4', '5', '9'], # 9 es inválido
25         'Seasons': ['1', '2', '3', '4', '1'],
26         'Transportation expense': ['100', '200', '300', '1000', '150'],
27         'Distance from Residence to Work': ['10', '20', '30', '40', '15'],
28         'Service time': ['5', '10', '15', '20', '8'],
29         'Age': ['30', '40', '50', '25', '35'],
30         'Work load Average/day': ['200', '250', '300', '350', '400'],
31         'Hit target': ['90', '85', '95', '80', '88'],
32         'Disciplinary failure': ['0', '1', '0', '1', '1'] # 2 es inválido
33     })
```

Pruebas unitarias

Desarrollo de pruebas unitarias en el Procesamiento de datos.

Flujo de trabajo e Ingeniería de Software

Repo GitHub: <https://github.com/Cpano98/operaciones-de-aprendizaje-automatico-equipo59>



Workflows

Desarrollo de workflows antes de un merge a la rama principal.

Exploración y Preprocesamiento de Datos

Transformar información en conocimiento accionable



Identificación de patrones clave

El análisis exploratorio permitió descubrir patrones y correlaciones importantes, ayudando a revelar causas principales del ausentismo laboral.

Preprocesamiento de datos

Se realizó normalización, codificación de atributos y reducción de dimensionalidad, preparando una base sólida para análisis posteriores.

Correlaciones significativas

Las variables de salud, transporte y entorno familiar mostraron las correlaciones más relevantes con el ausentismo.

Insight: la correlación más significativa proviene de variables de salud, transporte y entorno familiar.



Versionado de Datos

Asegurar trazabilidad y control total del ciclo de vida de datos

Historial Completo de Datasets

DVC permite mantener versiones originales y depuradas de los datos, facilitando la gestión histórica y la comparación eficiente.

Trazabilidad y Colaboración

La integración con plataformas como GitHub garantiza la trazabilidad de cambios, colaboración entre equipos y auditoría transparente.

Ecosistema Reproducible MLOps

El versionado efectivo brinda un entorno reproducible, cumpliendo estándares corporativos y facilitando flujos de trabajo confiables en MLOps.



Modelado y Evaluación

Entrenamiento y Selección de Modelos

Modelos de Machine Learning se entrenan con datos confiables y se seleccionan algoritmos óptimos para el mejor desempeño.

Calibración de Hiperparámetros

Los hiperparámetros del modelo se calibran cuidadosamente para mejorar la precisión y la eficacia en los resultados.

Evaluación y Comparación de Métricas

Se evalúan los modelos usando métricas como MAE, RMSE y R^2 para comparar y tomar decisiones fundamentadas.

Integración a la Plataforma

El modelo auditado y funcional se prepara para su integración final en la plataforma digital.

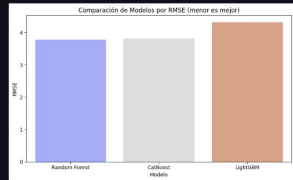
- LightGBM Regressor

Preprocesamiento:

- `StandardScaler` → variables numéricas
- `OneHotEncoder` → categóricas

- MAE, RMSE, R^2

	Model	MAE	RMSE	R2
0	Random Forest	2.612297	3.775265	0.34412
1	CatBoost	2.708630	3.815941	0.32991
2	LightGBM	2.992765	4.315723	0.14289



Ajuste de Hiper Parámetros y Registro en ML flow:

Técnica de optimización: `RandomizedSearchCV`

Parámetros ajustados:

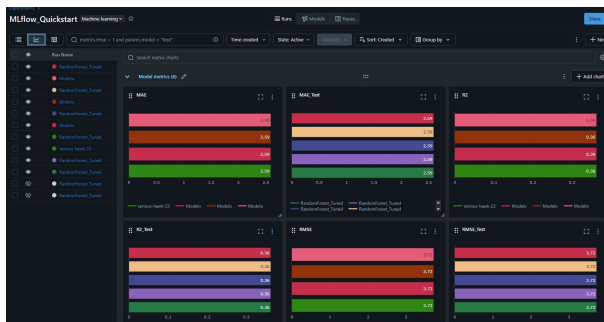
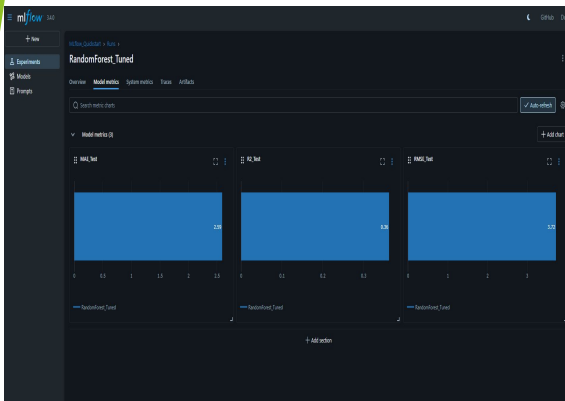
- `n_estimators`, `max_depth`, `min_samples_split`, `min_samples_leaf`

Mejora lograda:

- ↓ RMSE en un ~8–10%
- ↑ R^2 en +0.05 puntos

Registro del modelo optimizado con MLflow:

- Run: `RandomForest_Tuned`
- Tracking URI: `mlruns/MLflow_Quickstart`
- Logs de métricas y parámetros



Resultados clave y próximos pasos

Ecosistema de datos confiable

Se estableció un ecosistema de datos sólido y versionado que asegura la integridad y trazabilidad de la información.

Automatización del preprocesamiento

Se implementó un pipeline automatizado y reutilizable para preprocesar datos, optimizando tiempos y reduciendo errores manuales.

Colaboración técnica fortalecida

El uso de MLOps mejoró la colaboración entre los roles técnicos, permitiendo flujos de trabajo más ágiles y coordinados.

Preparación para IA en talento

El equipo está listo para modelar e implementar IA, impulsando la gestión estratégica del talento en la organización.