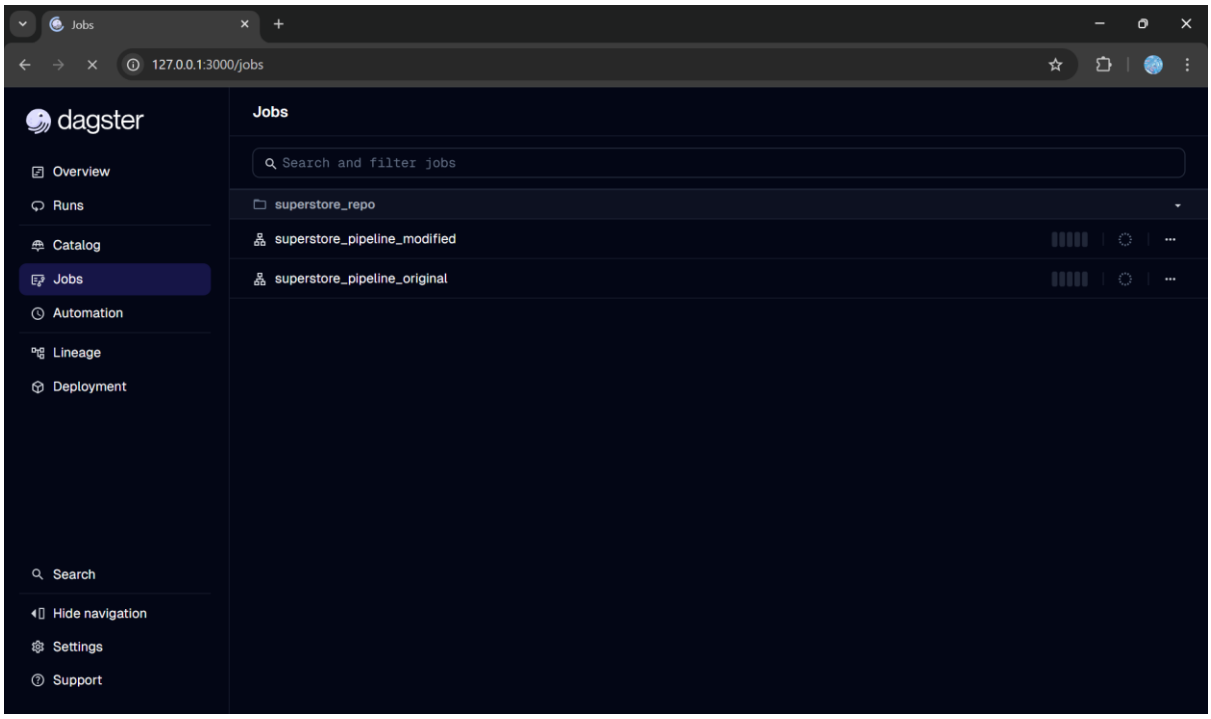
Report: Time Savings and Reproducibility Using Dagster ML Pipeline

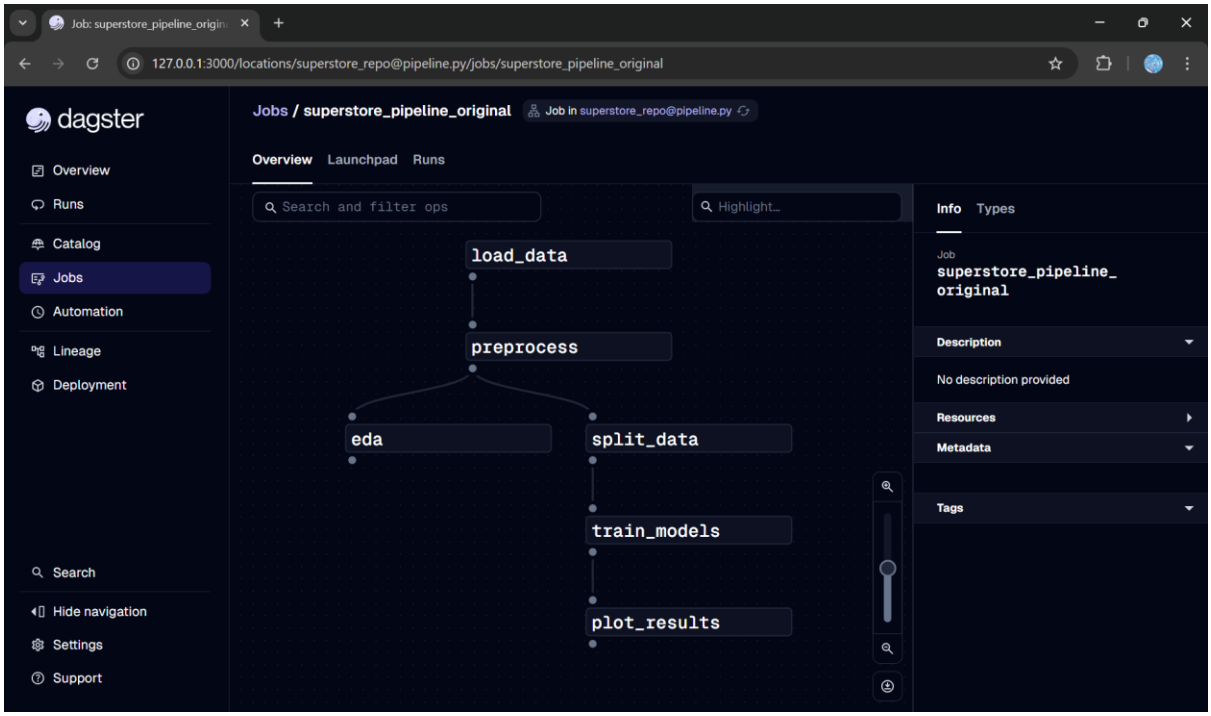
Conventional machine learning workflows run in Jupyter notebooks can struggle with reproducibility because of the non-linear execution of cells, modifications of code, and changes made to the versions of data. Often, these problems lead to inconsistent results and hinder effective tracking of experiments. To mitigate these challenges, this project investigates the use of Dagster as a data orchestration framework for building a structured and reproducible machine learning pipeline.

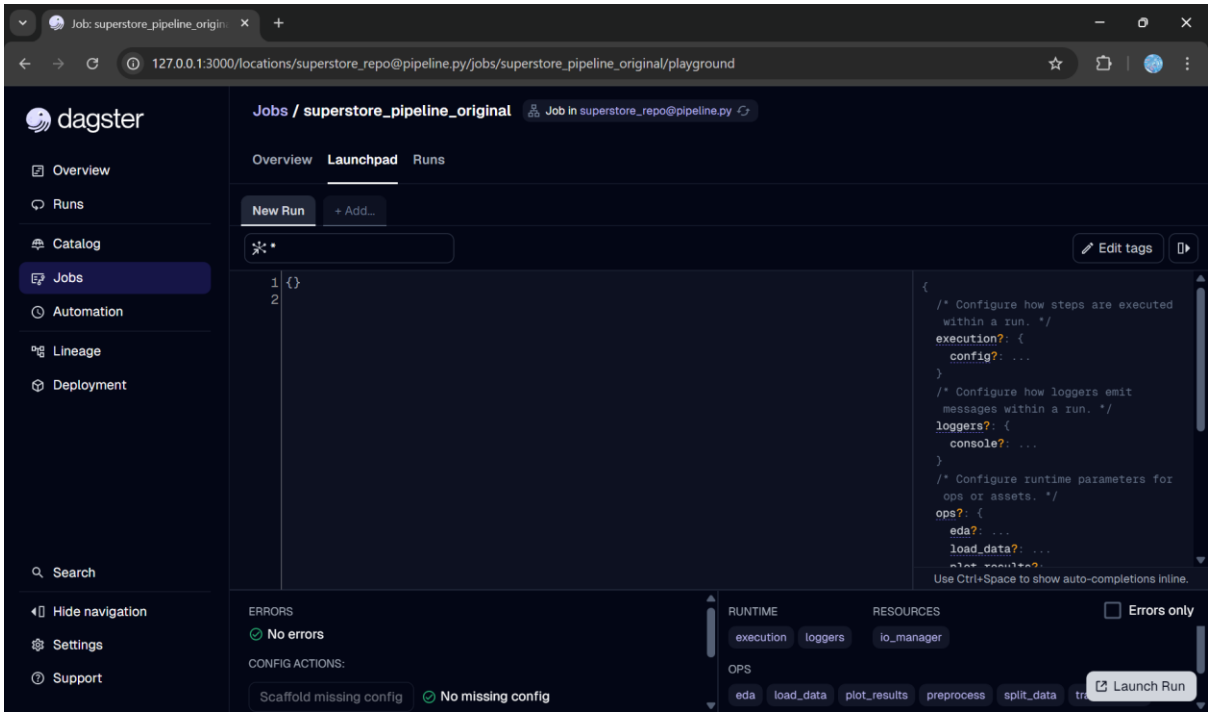
The pipeline was built using the Superstore dataset and goes through all the primary steps of a standard machine learning workflow: data ingestion and preprocessing, model training and evaluation, and visualization of results. Three different types of classification models (Decision Tree, Random Forest, and Logistic Regression) were built and evaluated. Dagster guarantees that there are explicit dependencies established between different stages of the pipeline, which will ensure consistent execution orders and reproducible data flows. Reproducibility of both runs was demonstrated by running the same pipeline against the original and a modified dataset without changing code in either version of the pipeline. While both runs of the pipelines were tracked by Dagster as different runs, the timing difference between the two pipelines was measurable. While the workflow was executed in a traditional Jupyter Notebook, the original dataset took **2.80 seconds** to execute, while the modified dataset required only **2.33 seconds** to complete; in contrast, the Dagster-based pipeline took **1 minute 14 seconds** for the original dataset and **1 minute 12 seconds** for the modified dataset. While the notebook-based approach was substantially faster for this dataset size than the Dagster-based pipeline, there are clear advantages to using Dagster with respect to execution control, reproducibility, and workflow organization, all of which make it a better fit for machine learning pipelines scaled and deployed into production environments where reliability and traceability are critically important.

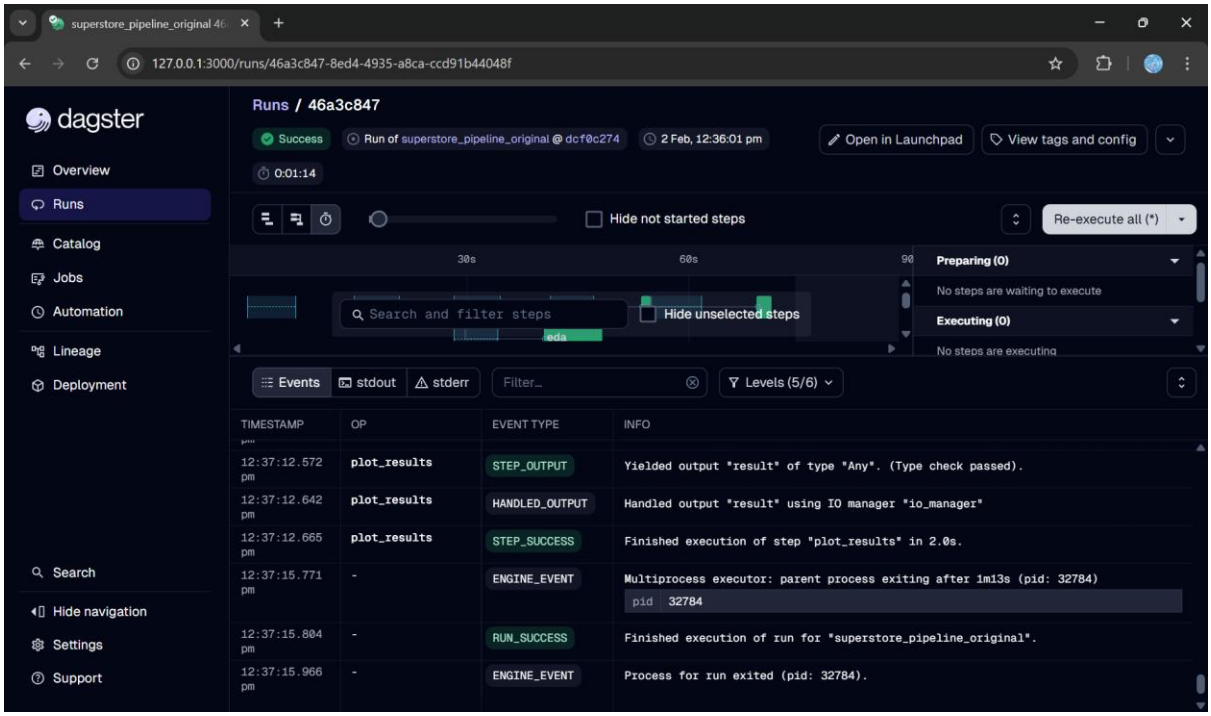
Dagster:



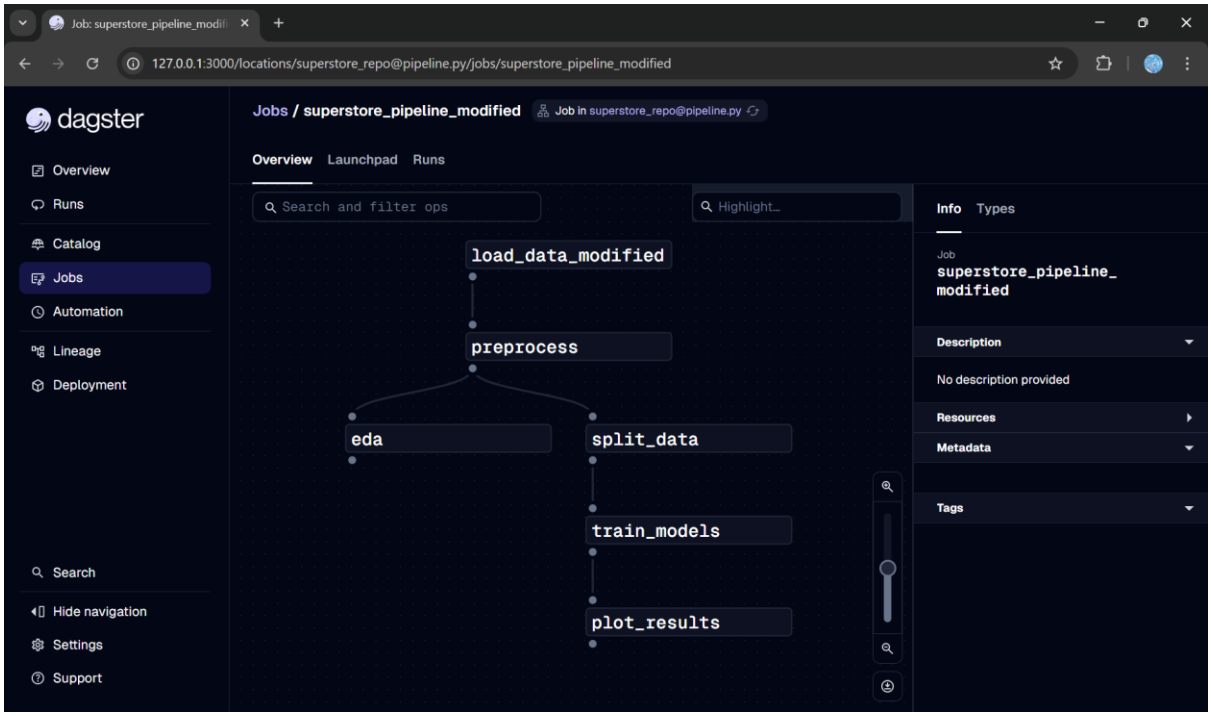
Dagster pipeline (Original Dataset):

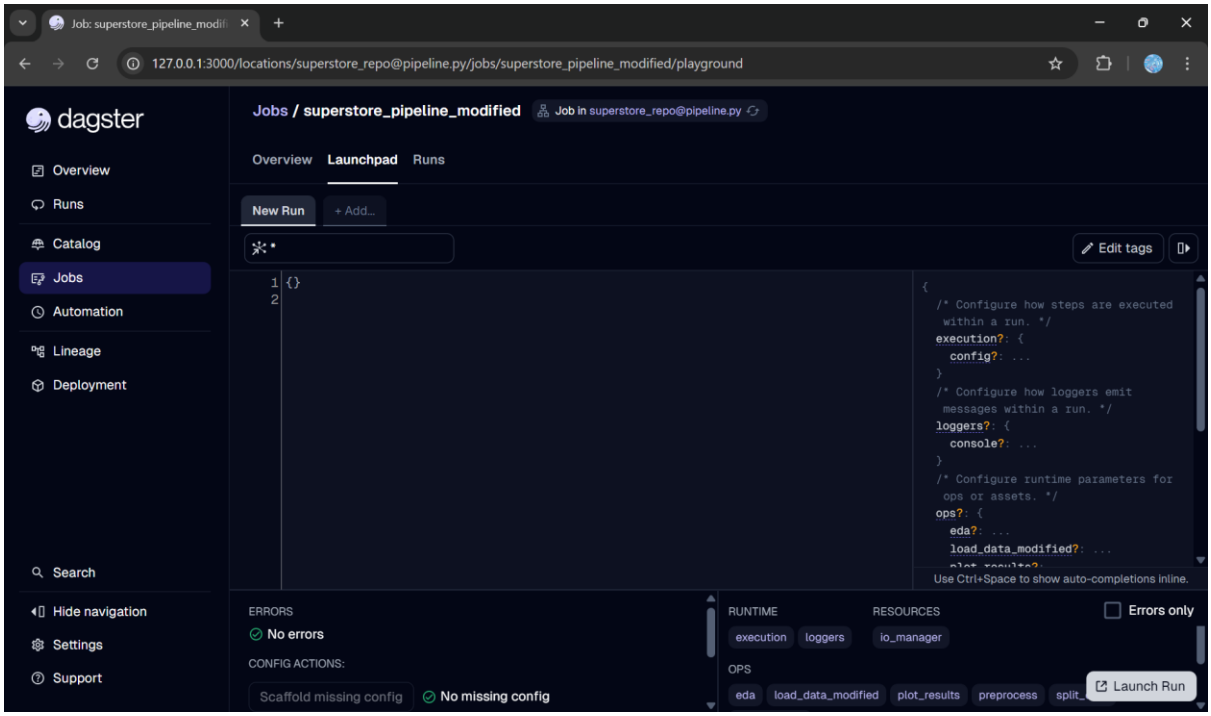


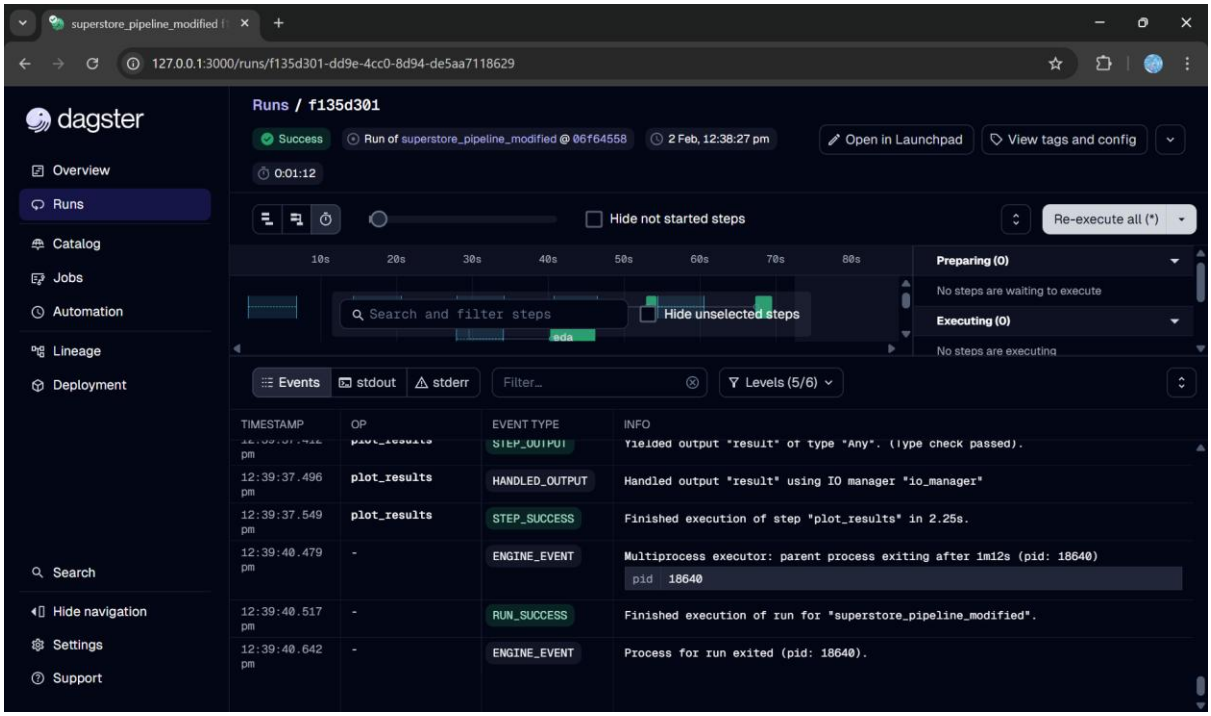




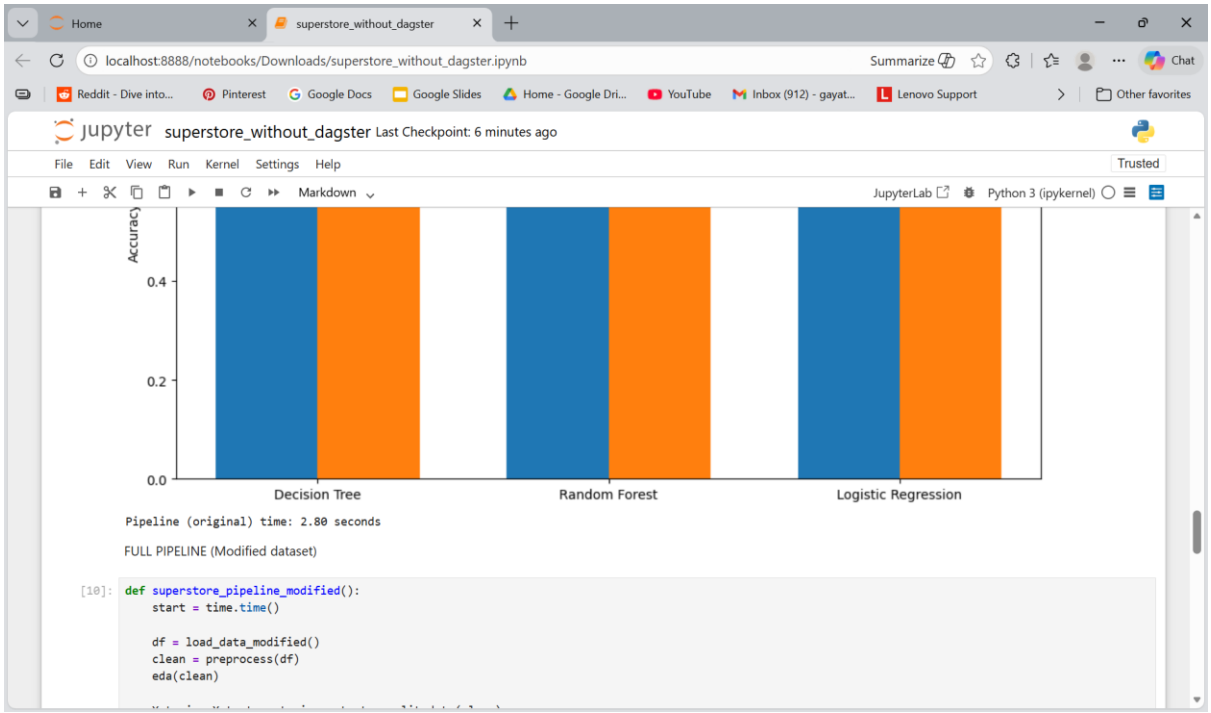
Dagster pipeline (Modified Dataset):


Job: superstore_pipeline_modifi
127.0.0.1:3000/locations/superstore_repo@pipeline.py/jobs/superstore_pipeline_modified
dagster
Jobs / superstore_pipeline_modified
Job in superstore_repo@pipeline.py
Overview
Launchpad
Runs
Search and filter ops
Highlight...
load_data_modified
preprocess
eda
split_data
train_models
plot_results
Info
Types
Job
superstore_pipeline_modified
Description
No description provided
Resources
Metadata
Tags
Search
Hide navigation
Settings
Support

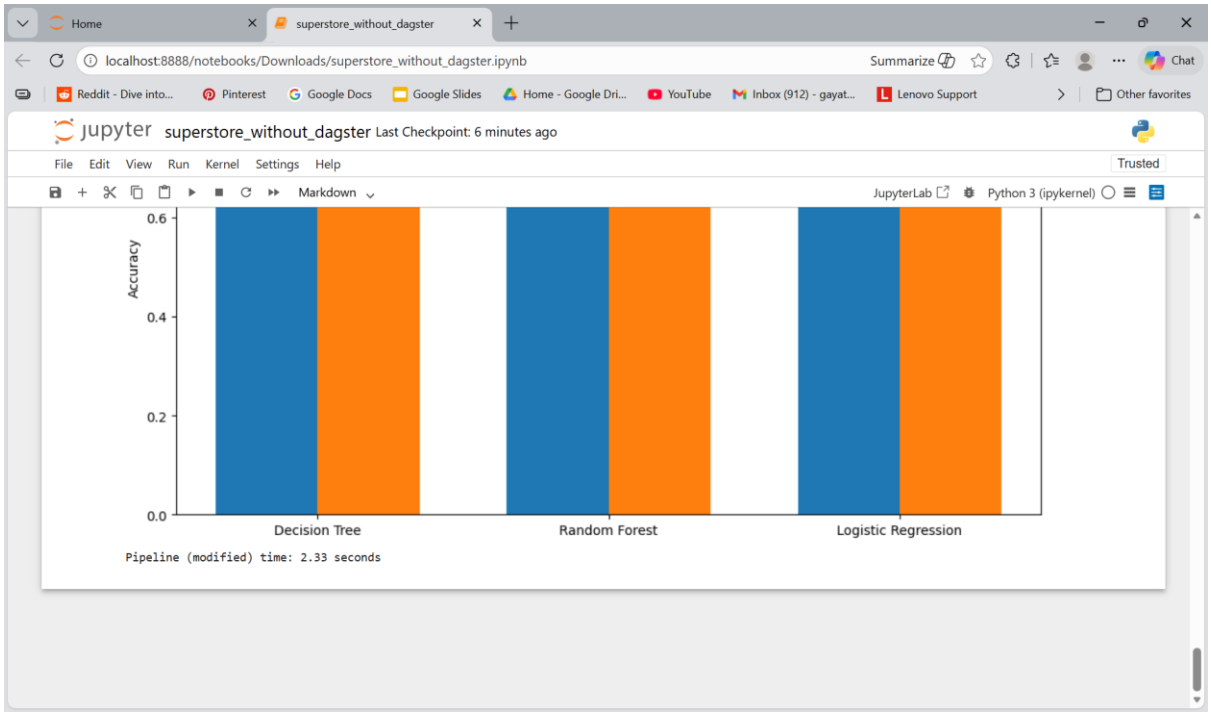

Job: superstore_pipeline_modifi
127.0.0.1:3000/locations/superstore_repo@pipeline.py/jobs/superstore_pipeline_modified/playground
dagster
Jobs / superstore_pipeline_modified
Job in superstore_repo@pipeline.py
Overview
Launchpad
Runs
New Run
+ Add...
Edit tags
{}
/* Configure how steps are executed within a run. */
execution?: {
config?: ...
}
/* Configure how loggers emit messages within a run. */
loggers?: {
console?: ...
}
/* Configure runtime parameters for ops or assets. */
ops?: {
eda?: ...
load_data_modified?: ...
Use Ctrl+Space to show auto-completions inline.
ERRORS
No errors
CONFIG ACTIONS:
Scaffold missing config
No missing config
RUNTIME
RESOURCES
execution
loggers
io_manager
Errors only
OPS
eda
load_data_modified
plot_results
preprocess
Launch Run
Search
Hide navigation
Settings
Support



Jupyter Notebook pipeline (Original Dataset):



Jupyter Notebook pipeline (Modified Dataset):


Home
superstore_without_dagster
localhost:8888/notebooks/Downloads/superstore_without_dagster.ipynb
Summarize
Chat
Reddit - Dive into...
Pinterest
Google Docs
Google Slides
Home - Google Dri...
YouTube
Inbox (912) - gayat...
Lenovo Support
Other favorites
jupyter
superstore_without_dagster Last Checkpoint: 6 minutes ago
File Edit View Run Kernel Settings Help
Trusted
Markdown
JupyterLab
Python 3 (ipykernel)
Accuracy
0.0
0.2
0.4
0.6
Decision Tree
Random Forest
Logistic Regression
Pipeline (modified) time: 2.33 seconds