Pedro Saturio Toledano Montes

MLOps

Memoria – Selección de modelo_empleabilidad_recien_titulados

1. Introducción

En este proyecto, el objetivo ha sido aplicar los conocimientos adquiridos sobre MLFlow en un escenario más realista. MLFlow es una herramienta clave para gestionar y hacer seguimiento de experimentos de machine learning, y en este caso, se ha utilizado para entrenar y evaluar modelos de clasificación sobre un dataset relacionado con la empleabilidad de los estudiantes.

Esta práctica busca predecir la empleabilidad de los estudiantes en función de sus habilidades, rendimiento académico y experiencia laboral. Para lograrlo, se han entrenado y comparado dos modelos de clasificación utilizando MLFlow para registrar todas las ejecuciones, parámetros y métricas asociadas a cada experimento.

2. Descripción del Dataset

El conjunto de datos utilizado para este proyecto proviene de Kaggle y contiene información sobre el rendimiento académico, formación y estado laboral de los estudiantes. El objetivo es predecir la empleabilidad de los estudiantes, es decir, si cuando terminan sus estudios salen con un puesto de trabajo o no.

Las principales columnas en el dataset incluyen:

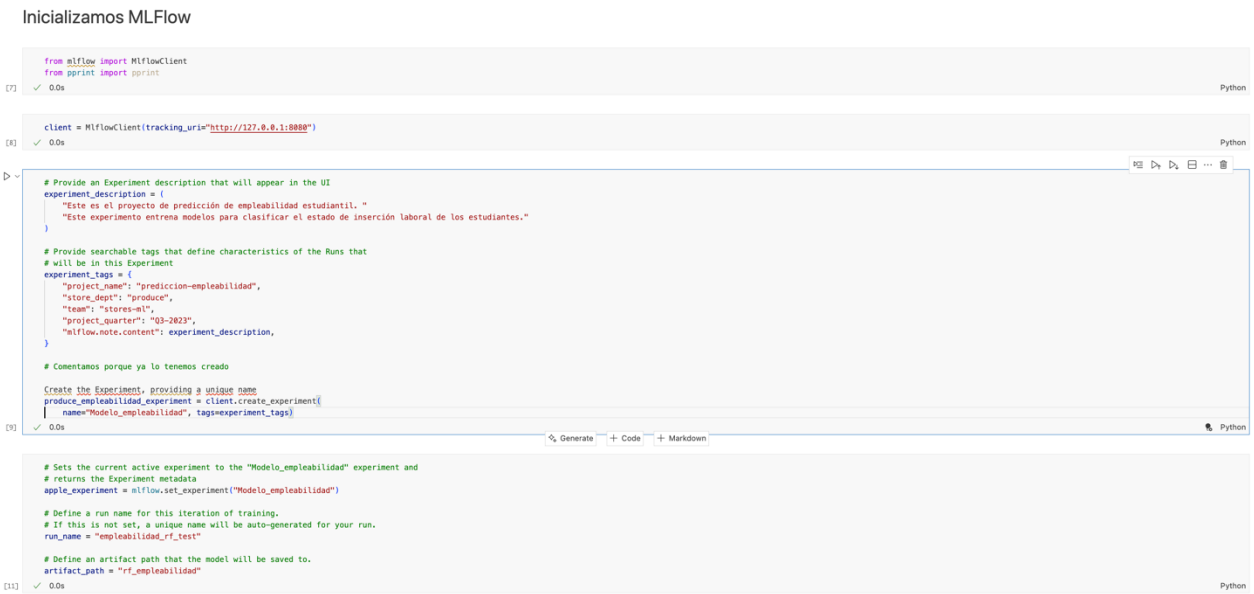
- **CGPA**: Promedio general de calificaciones obtenido por el estudiante.
- **Internships**: Número de prácticas realizadas por el estudiante.
- **Projects**: Número de proyectos en los que el estudiante ha trabajado.
- **Workshops/Certifications**: Cursos y certificaciones obtenidas por el estudiante en línea.
- **AptitudeTestScore**: Puntaje en pruebas de aptitud utilizadas en procesos de reclutamiento.

- **SoftSkillRating**: Calificación de las habilidades blandas del estudiante, como la comunicación y trabajo en equipo.
- **ExtraCurricularActivities**: Actividades extracurriculares en las que participa el estudiante.
- **PlacementTraining**: Programas de formación recibidos por los estudiantes para prepararse para la búsqueda de empleo.
- **SSC_Marks**: Calificaciones obtenidas en la educación secundaria.
- **HCS_Marks**: Calificaciones obtenidas en bachillerato.
- **PlacementStatus**: Columna objetivo, con dos categorías: "Placed" (empleado) y "Not Placed" (no empleado).

3. Inicialización de MLFlow

Para poder registrar y hacer un seguimiento adecuado de los experimentos de Machine Learning realizados, se utilizó **MLFlow**, una plataforma de código abierto que facilita la gestión del ciclo de vida de los modelos de machine learning.

Se configuró un servidor local con la IP http://127.0.0.1:8080 y se creó un experimento denominado **'Modelo_empleabilidad'** dónde guardaremos todas las ejecuciones de cada modelo.

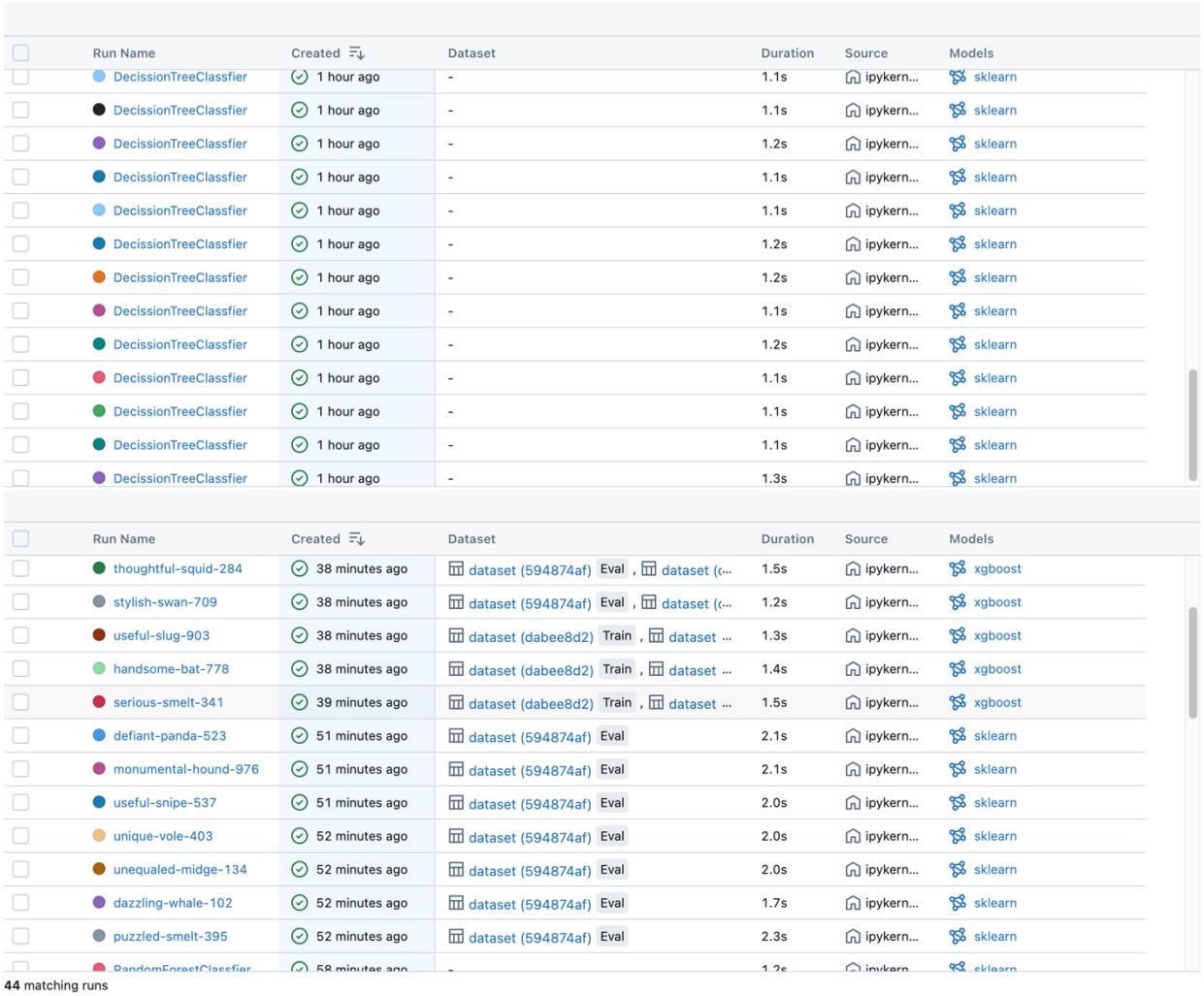


4. Preprocesamiento y análisis de los datos

En esta parte, analizamos los datos y nos dimos cuenta de que algunas variables tenían escalas muy diferentes. Esto podría hacer que los modelos no funcionaran de la mejor manera, por lo que decidimos estandarizar los datos. De esta forma, los modelos pueden trabajar mejor con ellos.

5. Modelos de clasificación

Para la clasificación, utilizamos cuatro modelos: DecisionTreeClassifier, RandomForestClassifier, KNeighborsClassifier (KNN) y XGBoostClassifier. Realizamos entre 10 y 15 ejecuciones por modelo, variando los hiperparámetros con el objetivo de obtener el mayor accuracy posible. Todas las ejecuciones fueron registradas en MLFlow; en los dos primeros modelos, el registro se realizó manualmente, mientras que en los otros dos, se hizo de manera automática.



Run Name	Created	Dataset	Duration	Source	Models
DecissionTreeClassfier	1 hour ago	-	1.1s	ipykern...	sklearn
DecissionTreeClassfier	1 hour ago	-	1.1s	ipykern...	sklearn
DecissionTreeClassfier	1 hour ago	-	1.2s	ipykern...	sklearn
DecissionTreeClassfier	1 hour ago	-	1.1s	ipykern...	sklearn
DecissionTreeClassfier	1 hour ago	-	1.1s	ipykern...	sklearn
DecissionTreeClassfier	1 hour ago	-	1.2s	ipykern...	sklearn
DecissionTreeClassfier	1 hour ago	-	1.2s	ipykern...	sklearn
DecissionTreeClassfier	1 hour ago	-	1.1s	ipykern...	sklearn
DecissionTreeClassfier	1 hour ago	-	1.2s	ipykern...	sklearn
DecissionTreeClassfier	1 hour ago	-	1.1s	ipykern...	sklearn
DecissionTreeClassfier	1 hour ago	-	1.1s	ipykern...	sklearn
DecissionTreeClassfier	1 hour ago	-	1.1s	ipykern...	sklearn
DecissionTreeClassfier	1 hour ago	-	1.3s	ipykern...	sklearn

Run Name	Created	Dataset	Duration	Source	Models
thoughtful-squid-284	38 minutes ago	dataset (594874af) Eval , dataset (...	1.5s	ipykern...	xgboost
stylish-swan-709	38 minutes ago	dataset (594874af) Eval , dataset (...	1.2s	ipykern...	xgboost
useful-slug-903	38 minutes ago	dataset (dabee8d2) Train , dataset ...	1.3s	ipykern...	xgboost
handsome-bat-778	38 minutes ago	dataset (dabee8d2) Train , dataset ...	1.4s	ipykern...	xgboost
serious-smelt-341	39 minutes ago	dataset (dabee8d2) Train , dataset ...	1.5s	ipykern...	xgboost
defiant-panda-523	51 minutes ago	dataset (594874af) Eval	2.1s	ipykern...	sklearn
monumental-hound-976	51 minutes ago	dataset (594874af) Eval	2.1s	ipykern...	sklearn
useful-snipe-537	51 minutes ago	dataset (594874af) Eval	2.0s	ipykern...	sklearn
unique-vole-403	52 minutes ago	dataset (594874af) Eval	2.0s	ipykern...	sklearn
unequaled-midge-134	52 minutes ago	dataset (594874af) Eval	2.0s	ipykern...	sklearn
dazzling-whale-102	52 minutes ago	dataset (594874af) Eval	1.7s	ipykern...	sklearn
puzzled-smelt-395	52 minutes ago	dataset (594874af) Eval	2.3s	ipykern...	sklearn
RandomForestClassfier	58 minutes ago	-	1.2s	ipykern...	sklearn

44 matching runs

5. Comparación de resultados

Para la comparación de los resultados, comenzamos seleccionando todas las ejecuciones de un mismo modelo y comparando las métricas obtenidas en cada una. De entre todas esas ejecuciones, elegimos la que mostró las mejores métricas. Una vez que seleccionamos la mejor ejecución de cada modelo, comparamos los resultados entre los cuatro modelos utilizados. Finalmente, decidimos cuál de ellos ofrecía el mejor rendimiento global.

DecissionTreeClassifier

Run details

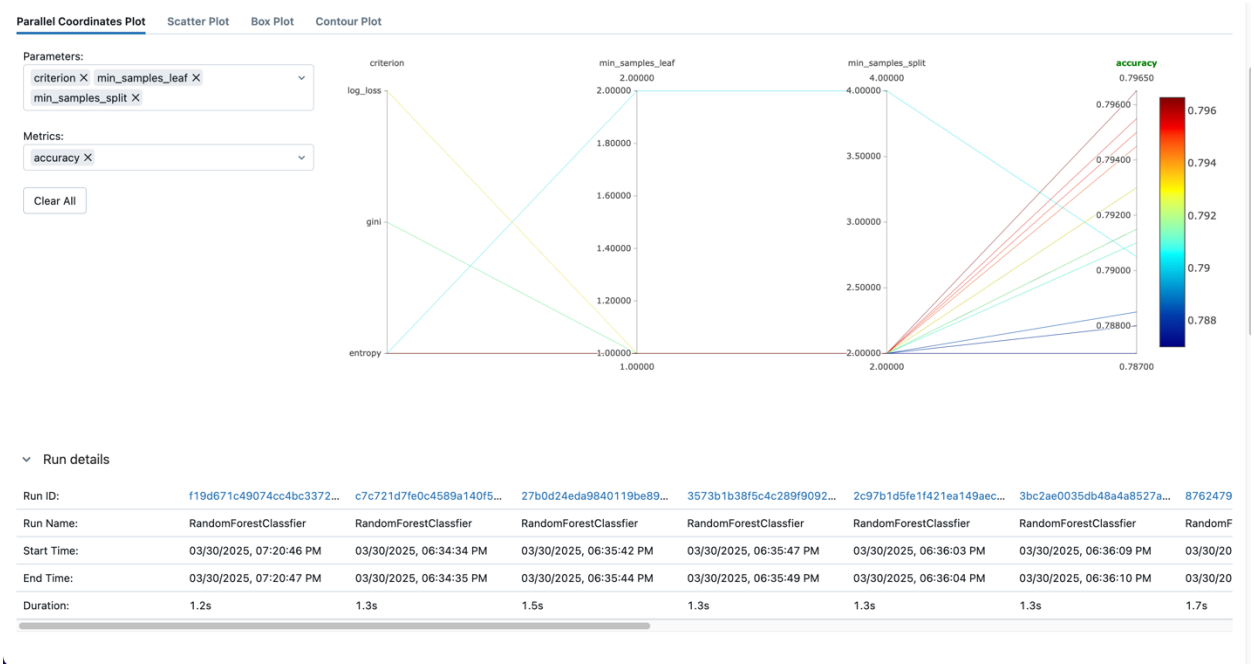
Run ID:	d4fa8f2067224749ac39e7...	380d4f2fb23d4aa184566b...	0e4ed01d49634011b0111...	a2d9e0d4d864472f8545e1...	cd89e00561d4469e949a0...	a4ac11f0be3949c8854a7f...	a5e750b
Run Name:	DecissionTreeClassfier	DecissionTreeClassfier	DecissionTreeClassfier	DecissionTreeClassfier	DecissionTreeClassfier	DecissionTreeClassfier	Decission
Start Time:	03/30/2025, 06:29:38 PM	03/30/2025, 06:33:55 PM	03/30/2025, 06:33:50 PM	03/30/2025, 06:33:37 PM	03/30/2025, 06:33:13 PM	03/30/2025, 06:33:02 PM	03/30/20
End Time:	03/30/2025, 06:29:39 PM	03/30/2025, 06:33:56 PM	03/30/2025, 06:33:51 PM	03/30/2025, 06:33:38 PM	03/30/2025, 06:33:14 PM	03/30/2025, 06:33:03 PM	03/30/20
Duration:	1.3s	1.1s	1.1s	1.1s	1.2s	1.1s	1.1s

Los mejores hiperparámetros para este modelo fueron los siguientes:

criterion	gini
max_depth	4
min_samples_leaf	1
min_samples_split	2
random_state	42

accuracy	0.788
f1	0.741
precision	0.759
recall	0.723

RandomForestClassifier

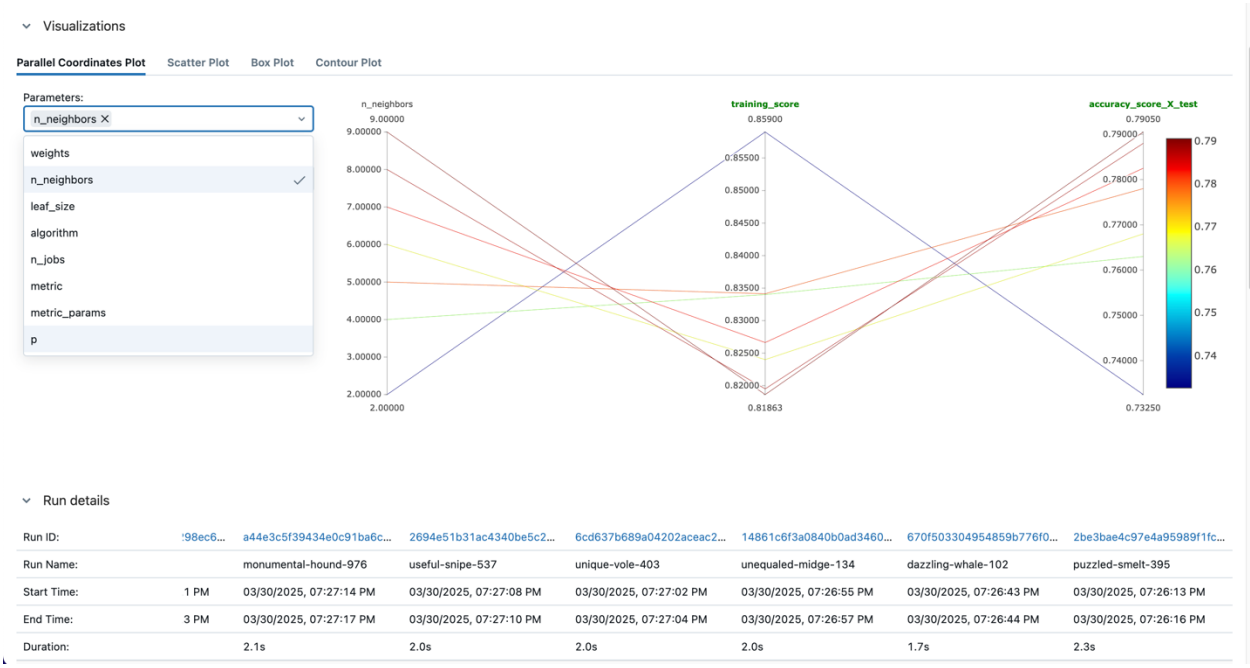


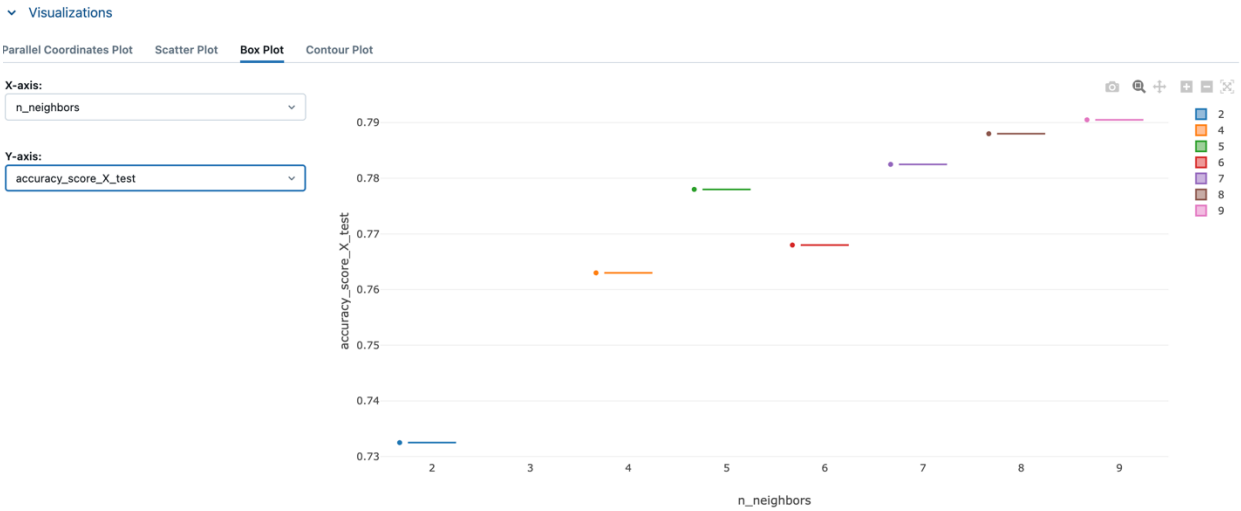
Los mejores hiperparámetros para este modelo fueron los siguientes:

criterion	entropy
min_samples_leaf	1
min_samples_split	2
n_estimators	5

accuracy	0.796
f1	0.75
precision	0.769
recall	0.732

KNeighborsClassifier



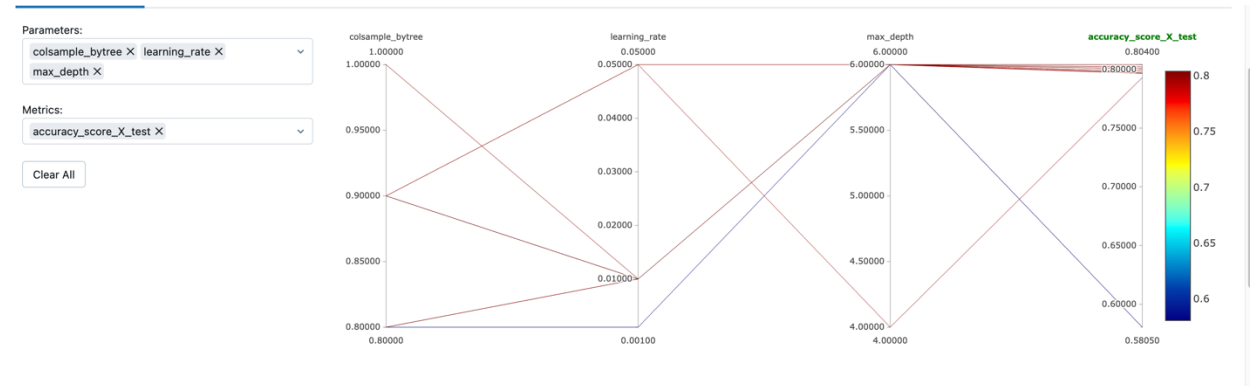


Los mejores hiperparámetros para este modelo fueron los siguientes:

n_neighbors	9

accuracy_score_X_test	0.791
f1_score_X_test	0.746
precision_score_X_test	0.759
recall_score_X_test	0.734
training_accuracy_score	0.819
training_f1_score	0.818
training_log_loss	0.373
training_precision_score	0.818
training_recall_score	0.819

XGBoost

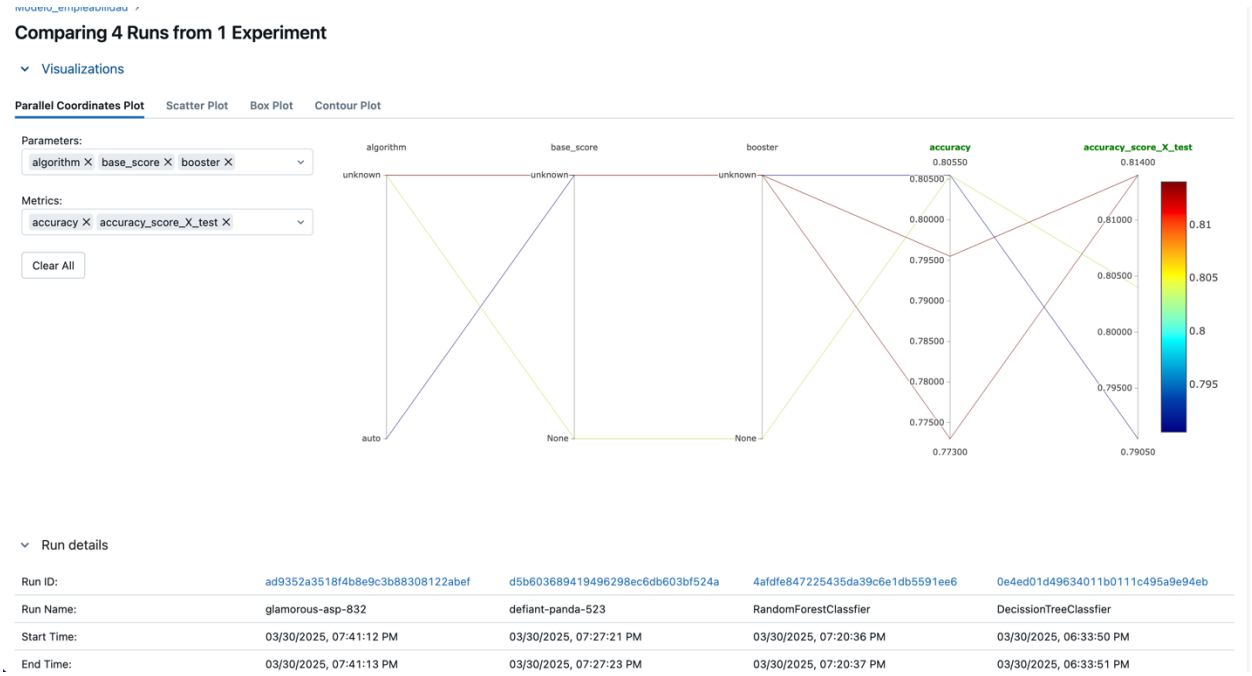


Los mejores hiperparámetros para este modelo fueron los siguientes:

colsample_bytree	0.9
learning_rate	0.01
max_depth	6
num_boost_round	100

accuracy_score_X_test	0.804
f1_score_X_test	0.755
precision_score_X_test	0.793
recall_score_X_test	0.721

Comparación de las mejores ejecuciones de cada modelo



Run ID:	ad9352a3518f4b8e9c3b88308122abef	d5b603689419496298ec6db603bf524a	4afdfe847225435da39c6e1db5591ee6	0e4ed01d49634011b0111c495a9e94eb
Run Name:	glamorous-asp-832	defiant-panda-523	RandomForestClassfier	DecissionTreeClassfier
Start Time:	03/30/2025, 07:41:12 PM	03/30/2025, 07:27:21 PM	03/30/2025, 07:20:36 PM	03/30/2025, 06:33:50 PM
End Time:	03/30/2025, 07:41:13 PM	03/30/2025, 07:27:23 PM	03/30/2025, 07:20:37 PM	03/30/2025, 06:33:51 PM

6. Conclusión

Nos decidimos por el mejor modelo, en este caso, la ejecución identificada como (ad9352a3518f4b8e9c3b88308122abef) – glamorous-asp-832, que alcanzó un accuracy de 0.804. Esta ejecución corresponde al modelo de XGBoost, el cual resultó ser el más eficiente en cuanto a rendimiento.