Presentado por Jhonatan Segura- keepcoding mlops

Para este ejercicio de Machine Learning, hemos elegido un proyecto diseñado específicamente para clasificar documentos en cuatro tipos distintos. El proceso inicia con la lectura de documentos de clasificación que contienen múltiples muestras, seguida de una extracción detallada de caracteres mediante OCR.

Una vez obtenido el texto, se procede a eliminar las "stop words" y a realizar la lematización, limpiando así el corpus de cada documento. Posteriormente, los datos se dividen en conjuntos de entrenamiento y prueba. Finalmente, se entrenan cuatro algoritmos diferentes para determinar cuál ofrece el mejor rendimiento en la clasificación.

Naive Bayes

- SVM
- Random Forest
- Logistic Regression
- Naive Bayes

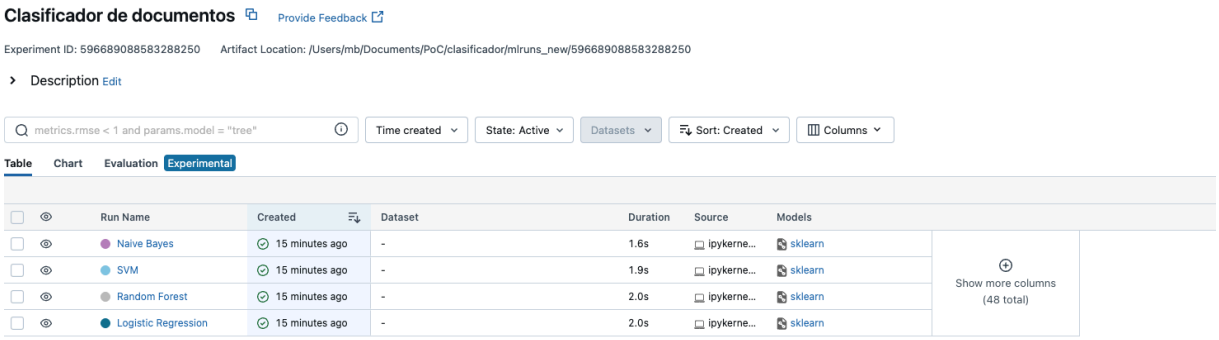
Hemos integrado MLflow, una plataforma de código abierto, para optimizar la gestión del ciclo de vida de nuestros modelos de machine learning. Esta integración nos permite registrar y visualizar las métricas de rendimiento de cada modelo de manera sistemática.

Con MLflow, podemos:

- **Realizar un seguimiento detallado de los experimentos:** Capturar automáticamente parámetros, métricas y artefactos de cada ejecución, facilitando la reproducibilidad y el análisis comparativo.
- **Comparar modelos de forma eficiente:** Analizar las métricas de diferentes modelos en una interfaz centralizada, lo que nos permite identificar el modelo con el mejor rendimiento para cada caso de uso.
- **Gestionar el control de versiones de los modelos:** Almacenar y organizar las

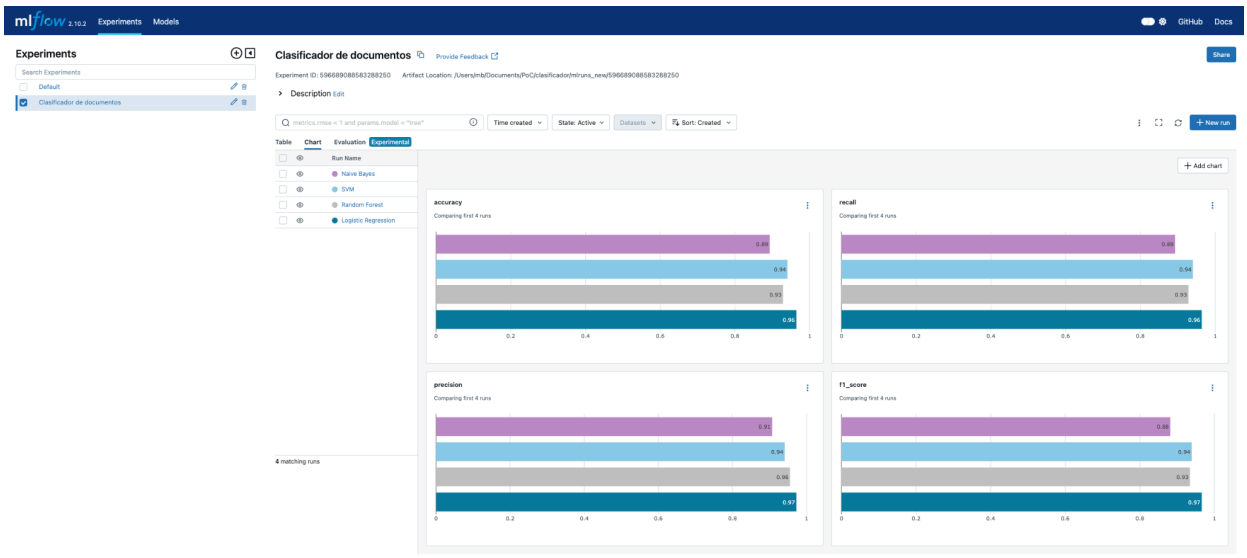
versiones de nuestros modelos en un registro centralizado, lo que simplifica la implementación y el despliegue.

Esta aproximación no solo mejora la transparencia y la trazabilidad de nuestros procesos de desarrollo de modelos, sino que también nos habilita para tomar decisiones basadas en datos sólidos al seleccionar e implementar los modelos más adecuados.



Inicialmente, se consideró implementar esto en un contenedor Docker; sin embargo, surgieron numerosos problemas con el mapeo de volúmenes, donde se almacenan los modelos. También se identificó que la plataforma funciona óptimamente con almacenamiento virtual en la nube, como buckets.

Interpretación de la grafica:



Interpretación de Resultados del Experimento: "Clasificador de Documentos"

1. Resumen General

La imagen muestra el panel de resultados de un experimento realizado en la plataforma **MLflow**. El objetivo de este experimento fue comparar el rendimiento de cuatro modelos de Machine Learning diferentes para una tarea de **clasificación de documentos**. El modelo **Regresión Logística** demostró ser superior en todas las métricas evaluadas.

2. Modelos y Métricas Evaluadas

Se compararon los siguientes cuatro algoritmos:

- **Naive Bayes** (barra morada)
- **SVM** (Máquina de Vectores de Soporte - barra celeste)
- **Random Forest** (Bosque Aleatorio - barra verde azulado)
- **Logistic Regression** (Regresión Logística - barra azul)

El rendimiento se midió con base en cuatro métricas clave:

- **Accuracy (Exactitud):** Mide el porcentaje total de predicciones correctas. Un valor alto indica que el modelo es acertado en general.
- **Recall (Sensibilidad):** Mide la capacidad del modelo para encontrar todas las instancias positivas reales. Un recall alto es crucial cuando es importante no pasar por alto ningún caso relevante (minimizar falsos negativos).
- **Precision (Precisión):** De todas las predicciones positivas que hizo el modelo, ¿cuántas eran realmente correctas? Una precisión alta es importante cuando el costo de un falso positivo es alto.
- **F1-Score:** Es la media armónica entre Precision y Recall. Ofrece una métrica balanceada que es útil cuando se busca un equilibrio entre no clasificar mal los casos positivos y no omitir ningún caso.

3. Análisis Detallado de los Resultados

Al analizar los gráficos de barras, se pueden extraer las siguientes conclusiones:

- **Modelo Ganador:** El modelo de **Regresión Logística (Logistic Regression)** es el claro ganador del experimento. Consistentemente obtiene los puntajes más altos en las cuatro métricas, con valores de **0.96 en Accuracy y Recall**, y **0.97 en Precision y F1-Score**. Esto indica que no solo es el más exacto, sino también el mejor balanceado.

- **Ranking de Rendimiento:** El orden de efectividad de los modelos es el siguiente:

    1. **Regresión Logística:** Rendimiento sobresaliente.
    2. **SVM:** Sólido segundo lugar, con un rendimiento estable y alto en todas las métricas (todas en **0.94**).

3. **Random Forest:** Ocupa el tercer lugar. Aunque su precisión es alta (**0.96**), su recall (**0.93**) y F1-Score (**0.93**) son ligeramente inferiores a los de SVM.
4. **Naive Bayes:** Es el modelo con el rendimiento más bajo del grupo. Aunque sus resultados son decentes (entre **0.88 y 0.90**), no es competitivo en comparación con los otros tres modelos evaluados.

4. Conclusión y Recomendación para el Informe

Para la tarea de clasificación de documentos, el experimento demuestra de manera concluyente que el modelo de **Regresión Logística** es la opción más robusta y eficaz. Se recomienda seleccionar este modelo para su implementación o para las siguientes fases del proyecto, ya que ofrece la mayor probabilidad de éxito en la clasificación de nuevos documentos, minimizando tanto los errores de omisión (falsos negativos) como los de clasificación incorrecta (falsos positivos).