Modelo Predictivo para la Gestión Eficiente del Arribo de Embarcaciones - Waypoint LLC

1. Descripción general del problema

Contexto

Waypoint LLC es una empresa que posee una destacada trayectoria en logística y apoyo a misiones de barcos, habiendo participado en más de 8000 proyectos a nivel mundial. La compañía se especializa en soluciones logísticas en puertos y lugares de difícil acceso para el gobierno de EE. UU. en entornos desafiantes. Con una licencia global obtenida en 2020, Waypoint está habilitada para atender navíos de la Armada de Estados Unidos en 30 regiones del mundo hasta 2030, por lo que compite con otras empresas para la obtención de licitaciones de atención a barcos. En este modelo, gana el licitante que pueda entregar todos los servicios requeridos en el tiempo requerido al menor costo. Estas ofertas pueden tener un tiempo de preparación que varía desde el mismo día de la licitación.

Waypoint LLC busca ganar más licitaciones en su unidad de negocio de atención a barcos (GMAC). Para lograrlo, necesita hacer inversiones estratégicas que le permitan tener una capacidad de reacción rápida y disponer de los insumos suficientes para cubrir las demandas de sus clientes en el tiempo y con la calidad esperados. Con este fin, Waypoint desea predecir dónde será la próxima oferta. Para ello, comparte un dataset de más de 8000 ofertas de distintos barcos y lugares alrededor del mundo en los periodos de 2021 - 2024, lo que permite evaluar diversos escenarios y planificar con más tiempo.

Pregunta de negocio y alcance del proyecto

El equipo de tomadores de decisión de Waypoint tiene como objetivo resolver la siguiente cuestión: **“¿Cómo puede Waypoint predecir a qué puerto llegará un barco seleccionado por el usuario?”.** Para lograr esto, se plantea abordar las siguientes necesidades:

1. Análisis Descriptivo del Barco: Basado en información histórica que incluye:
    - Tiempo en puerto
    - Tiempo en altamar
    - Tiempo interarribo
    - Puertos más transitados
2. Predicción de Arribo utilizando ML:
    - Implementar herramientas de Machine Learning para predecir el puerto de arribo del barco seleccionado.
    - Proporcionar un porcentaje de certidumbre en la predicción realizada.

3. Descripción de los datos

El dataset se encuentra en Github, estará dispuesto para ser versionado, este consta de un conjunto de 8837 muestras con 16 dimensiones listadas a continuación:

Contenido del dataset

Dimensión	Tipo de Valor	Descripción
RTOP	String	Request Task Order Proposal o identificador único de oferta
Project #	String	Identificador único de proyectos ganados usados al interior de la compañía
Combo	String	Campo que agrupa cuando varias ofertas se licitan en un único conjunto
Status	String	Status de la oferta manejado internamente por la compañía, este puede ser: - New Bid=Oferta nueva sin trabajar - Build=Oferta en proceso de diseño - Submitted=Oferta sometida y en espera de recibir respuesta - Not Submitted=Oferta que no fue licitada por parte de la empresa - Won=Oferta ganada por la compañía - Lost=Oferta pérdida por la compañía - Cancelled=Oferta de la que el gobierno se retracta de licitar
Ship - Name	String	Nombre del barco
Ship ID - Ship Classification	String	Tipo de estructura de barco
Responsible	String	Personas responsables en ejecutar la oferta a nivel interno
Region Name	String	Region a la cual pertenece la oferta
Country	String	País a la cual pertenece la oferta
Business Unit	String	Unidad de negocio a la cual pertenece la oferta (GMAC, WEXMAC, PROVISIONS)
Awarded Company	String	Compañía que gana la licitación
Arrival	Date	Llegada del barco a puerto
Departure	Date	Salida del barco a puerto
Deadline(mmddyyyy)	Date	Fecha máxima para poder licitar
Time to Execute	String	Tiempo entre Arrival y Departure

Basados en las consideraciones de los Champion del proyecto, se reduce la dimensionalidad del dataset a campos fundamentales, por ende, se conservaron las siguientes variables: Ship Name, Vessel type, Region Name, Country, Location, Arrival, Departure y Time to Execute

Cabe resaltar que en esta fase del proyecto se complementó el proceso de ETL, por lo que a continuación se presenta los hallazgos y consideraciones realizadas:

- Aquellas muestras que contengan valores nulos de Arrival, Departure, Ship Name y location inicialmente serán limpiados del modelo
- Los Champion del proyecto también informan que ofertas con dias en ejecución menores a 0 y mayores a 60 serán desechadas debido a un posible error humano
- Todas las ofertas cuyo Status sea Cancelled, no serán tomadas en cuenta

Debido a esto se limpiaron 2135 ofertas que representan 24% de los datos

4. Modelos desarrollados y evaluación

Se utilizaron tres técnicas de clasificación de aprendizaje automático, evaluados en DataBricks al barco USS Trenton:

1. Regresión Logística
2. Random Forest
3. Gradient Boosting

Para el modelo más simple de regresión logística, se realizó una optimización mediante validación cruzada. A partir del mejor modelo obtenido, se realizó la predicción.

En el caso del modelo de Random Forest, se empleó una técnica empírica de ajuste de hiperparámetros, donde se evaluaron varios modelos con distintas combinaciones de:

- Número de estimadores
- Profundidad del árbol
- Número máximo de variables

Para el último modelo evaluado, Gradient Boosting, se utilizó una técnica similar a la de Random Forest, iterando sobre modelos con diversos hiperparámetros:

- Número de estimadores
- Profundidad máxima
- Tasa de aprendizaje

Los criterios de selección homogéneos para todos fueron Accuracy, Precision, recall y F1_scores extraídos con la librería de Scikit learn.

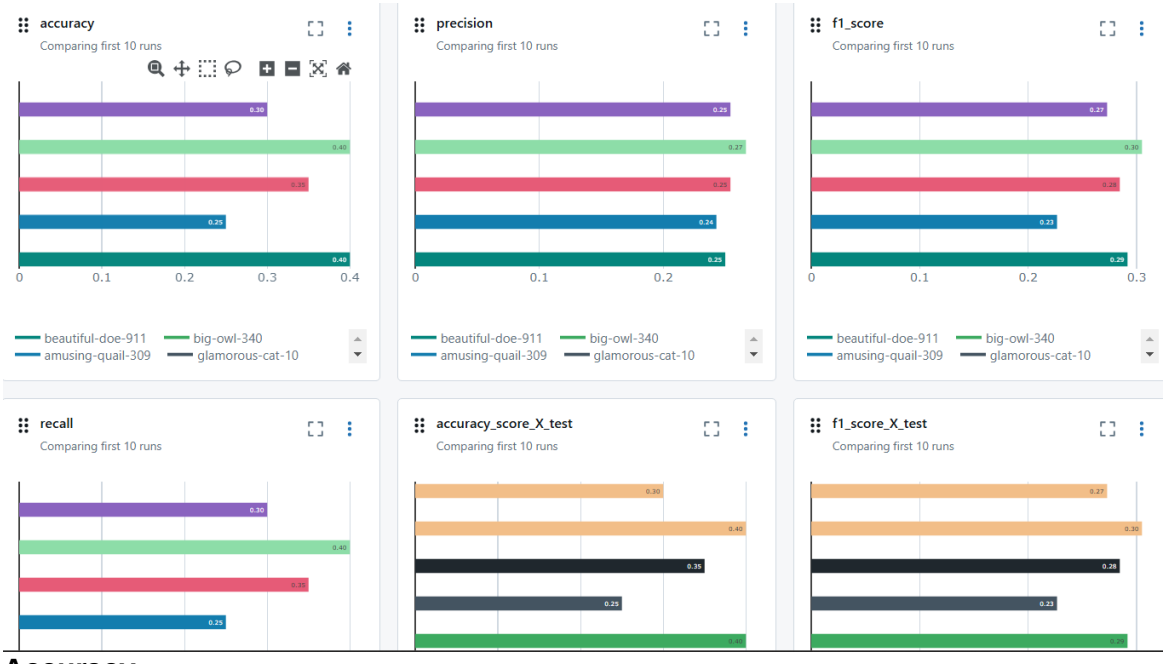
Run Name	Created	Dataset	Duration	Source	Models
beautiful-doe-911	5 minutes ago	-	6.8s	Gradient...	sklearn
big-owl-340	5 minutes ago	dataset (1f38dfe7) Eval , dataset (51e7...	35.1s	Gradient...	sklearn
amusing-quail-309	6 minutes ago	-	6.5s	Gradient...	sklearn
glamorous-cat-10	6 minutes ago	dataset (1f38dfe7) Eval , dataset (51e7...	27.7s	Gradient...	sklearn
unruly-dove-620	6 minutes ago	-	5.8s	Gradient...	sklearn
mercurial-boar-469	7 minutes ago	dataset (1f38dfe7) Eval , dataset (51e7...	23.9s	Gradient...	sklearn
unequaled-shrew-662	7 minutes ago	-	6.2s	Gradient...	sklearn
peaceful-finch-589	7 minutes ago	dataset (1f38dfe7) Eval , dataset (51e7...	26.7s	Gradient...	sklearn
polite-snipe-798	7 minutes ago	-	5.0s	Gradient...	sklearn
hilarious-bat-749	7 minutes ago	dataset (1f38dfe7) Eval , dataset (51e7...	10.0s	Gradient...	sklearn
amusing-ox-704	8 minutes ago	-	6.2s	Gradient...	sklearn
valuable-ray-907	8 minutes ago	dataset (1f38dfe7) Eval , dataset (51e7...	13.6s	Gradient...	sklearn
omniscient-shoat-572	8 minutes ago	-	5.8s	Gradient...	sklearn
sneaky-elk-862	8 minutes ago	dataset (1f38dfe7) Eval , dataset (51e7...	15.4s	Gradient...	sklearn
intrigued-moth-326	8 minutes ago	-	5.0s	Gradient...	sklearn

67 matching runs

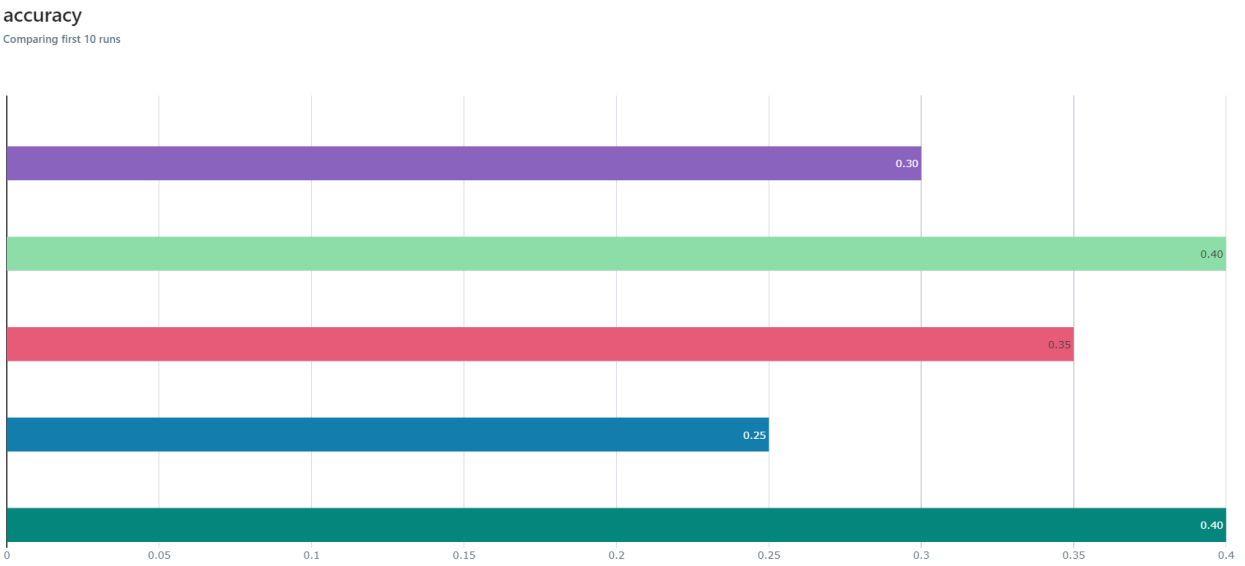
NOTA: Debido a que la cantidad de datos es heterogénea para distintos barcos, los valores pueden variar y los modelos a usar también.

5. Observaciones y conclusiones de los modelos

A continuación, se muestra un ejemplo de las gráficas extraídas los mejores resultados de los tres modelos en el entrenamiento:

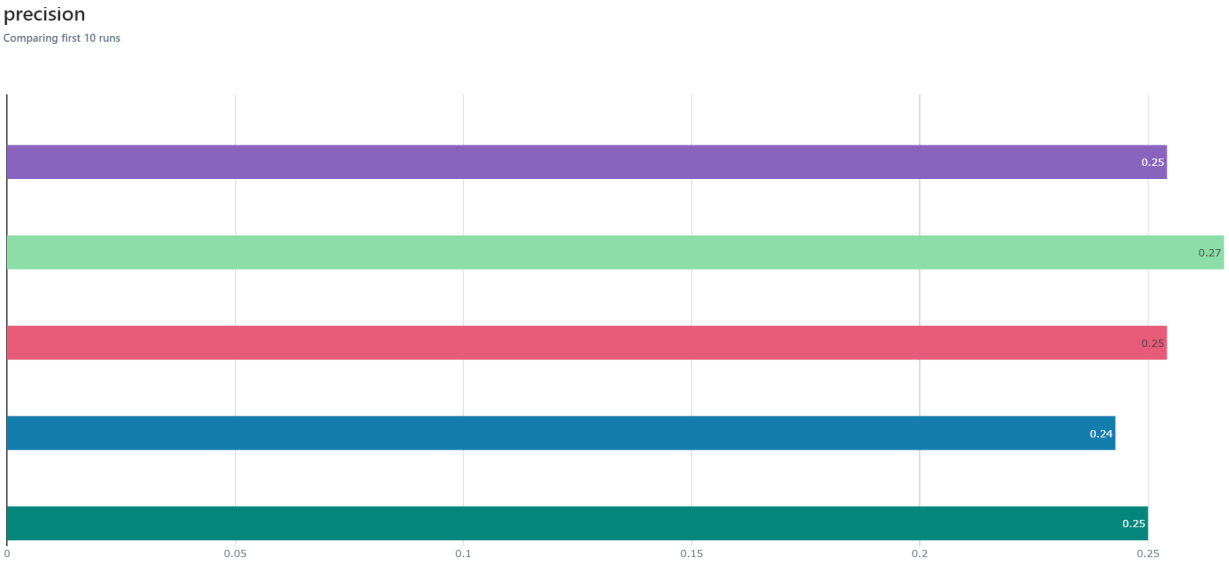


Accuracy



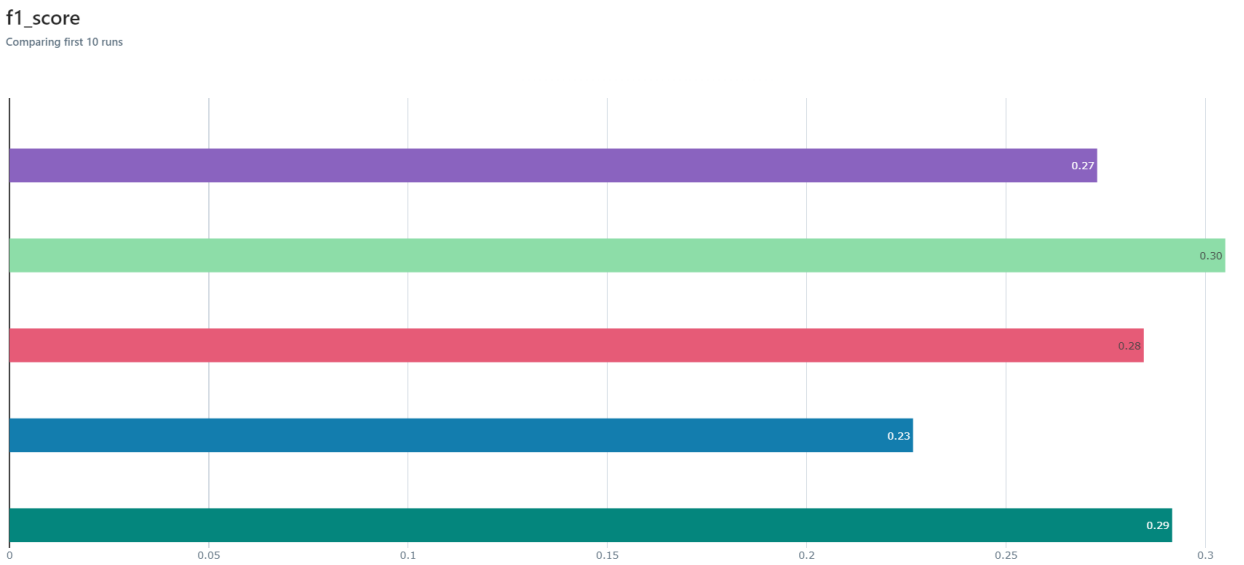
La proporción de predicciones correctas sobre el total de predicciones. Un accuracy de 0.4, modesto para los resultados esperados, mejor modelo Gradient Boosting.

Precisión



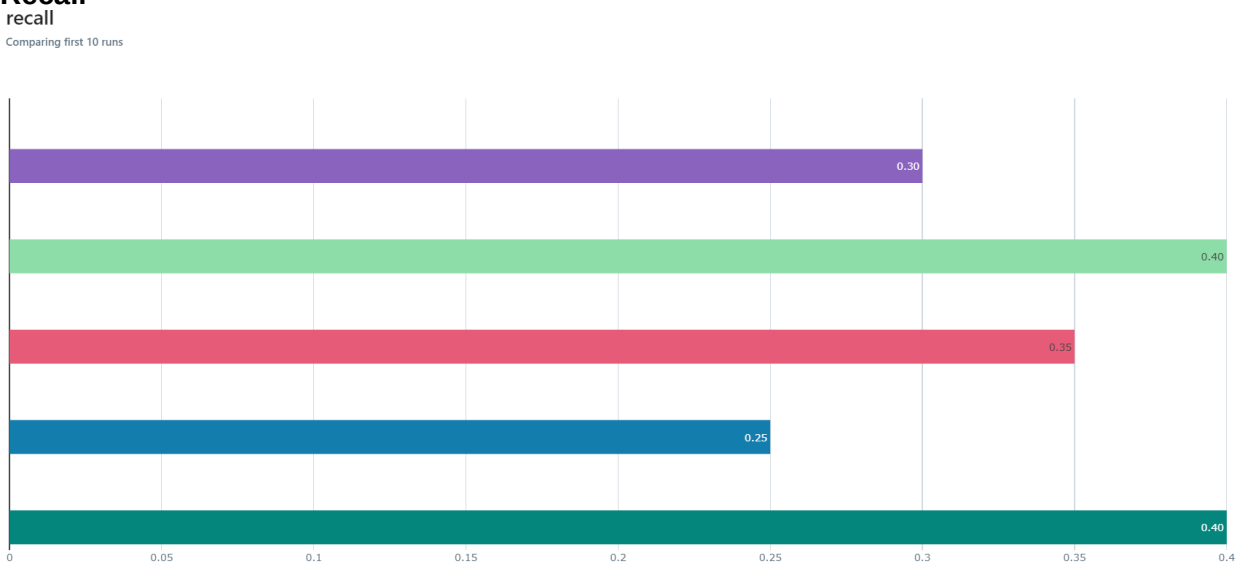
La proporción de verdaderos positivos entre todas las instancias que fueron clasificadas como positivas. Una precisión de 0.27 nos muestra un resultado con bajos VP, mejor modelo Gradient Boosting

F1 Score



Es la medida de armonía entre precision y recall. El mejor resultado es nuevamente de Gradient Boosting con un F1 de 0.3, muy modesto si tomamos en cuenta que su escala va de 0 –1

Recall



Es la proporción de verdaderos positivos entre los que debieron ser positivos. Este nuevamente lo encabeza el gradient boosting con un resultado modesto de 0.4.

Mejor modelo

El modelo con mejores resultados fue con gradient boosting basados en los siguientes hiper-parametros:

Parameters (4)

Search parameters

Parameter	Value
learning_rate	0.01
maxdepth	2
num_trees	500
predicted location	Augusta Bay

Metrics (4)

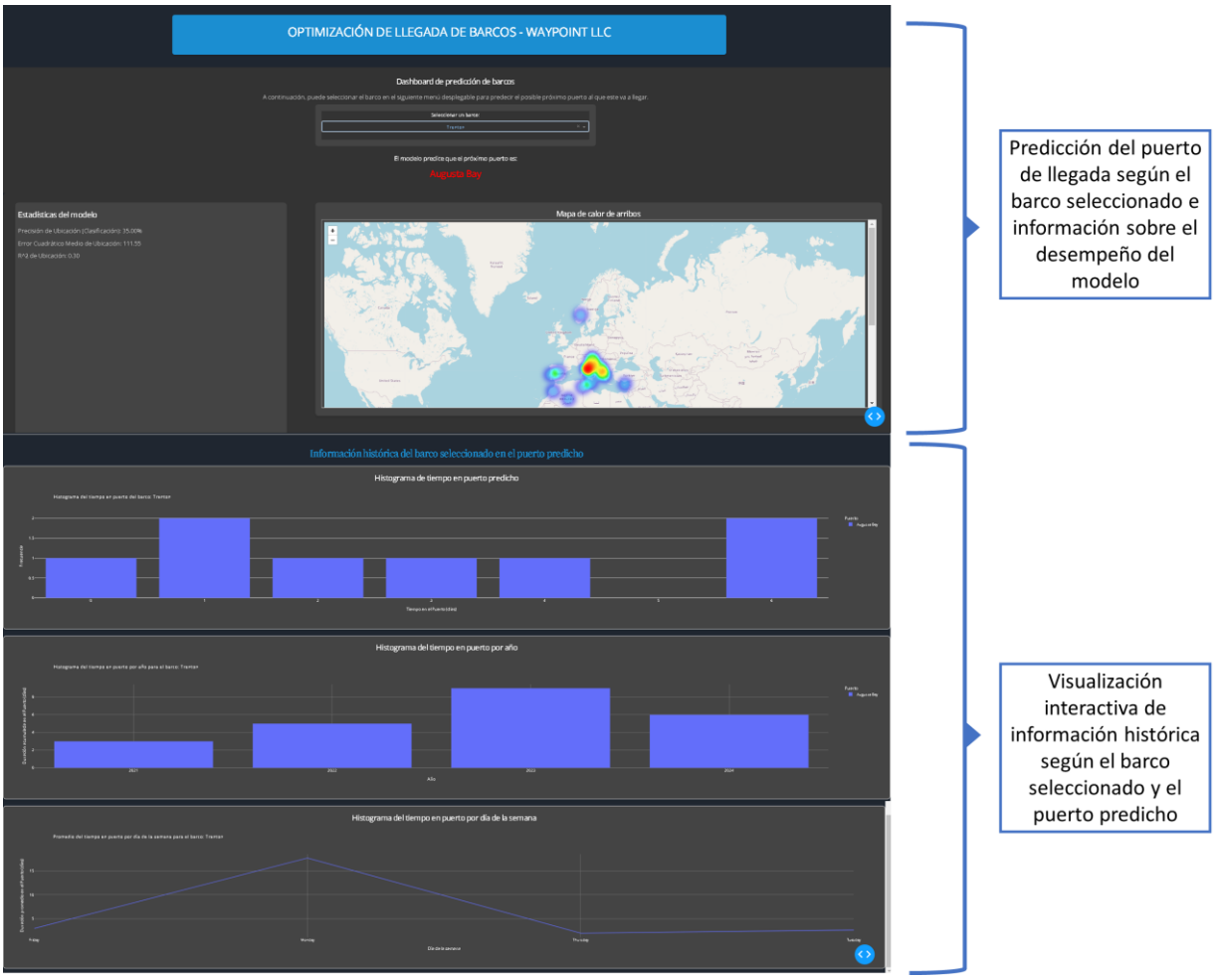
Search metrics

Metric	Latest
accuracy	0.4
f1_score	0.305
precision	0.26666666666666666
recall	0.4

Cabe resaltar que los scripts utilizados para los modelos se encuentran en la carpeta “models” en el repositorio de GitHub.

6. Tablero interactivo

El tablero desarrollado tiene como principal objetivo predecir el próximo puerto de llegada de los barcos en función de datos históricos, permitiendo la toma de decisiones informadas sobre las operaciones logísticas que puede realizar Waypoint LLC para licitaciones de los barcos, así como proporcionar información histórica del barco en el puerto predicho.



En esta entrega se realizaron varios avances en el desarrollo del tablero, como se observa en la figura anterior. Por lo que las características del tablero son:

1. **Interactividad**: El usuario puede seleccionar un barco desde un menú desplegable para obtener predicciones personalizadas y visualizar gráficos de acuerdo con sus selecciones. Los resultados y las visualizaciones se actualizan automáticamente al cambiar el barco seleccionado.

2. **Diseño Moderno y Funcional**: Con una interfaz limpia y moderna, el tablero tiene un diseño intuitivo con controles fáciles de usar, que incluyen tarjetas de control para interactuar con el modelo y ajustar las visualizaciones.
3. **Tecnologías usadas:** El tablero ha sido desarrollado en Python, empleando principalmente las bibliotecas Dash y Plotly. Para optimizar la organización del repositorio, se ha implementado un script que agrupa todas las funciones de los gráficos utilizados en el tablero, mejorando la estructura y facilitando el mantenimiento del código.

En cuanto a las funcionalidades que el tablero ofrece, se resalta:

1. **Predicción del Puerto de Llegada**: A través de un modelo predictivo, el tablero estima el próximo puerto de llegada de un barco seleccionado por el usuario. Esta predicción se realiza a partir de un conjunto de datos históricos y se presenta con las métricas de desempeño, tales como la precisión de la ubicación (porcentaje de aciertos), el error cuadrático medio (MSE) y el R^2, lo que permite evaluar el rendimiento del modelo predictivo en tiempo real.

2. **Visualización Interactiva de Información Histórica**:
    a. **Mapa de Calor**: Se muestra un mapa interactivo que ilustra la frecuencia de arribos de barcos a diferentes puertos, proporcionando información visual para el análisis espacial.
    b. **Histogramas y gráficos**: El tablero presenta 2 histogramas y 1 gráfico de líneas que permiten al usuario explorar patrones de tiempo en puerto predicho:
        i. Tiempo en puerto para el barco seleccionado.
        ii. Tiempo en puerto por año.
        iii. Tiempo en puerto por día de la semana.

Esta información proporciona detalles precisos sobre las operaciones logísticas previstas para el barco al llegar al puerto predicho, y facilita un seguimiento continuo de estas. Por ejemplo, al mostrar un histograma con el tiempo que el barco ha permanecido en el puerto en ocasiones anteriores, Waypoint puede anticipar los recursos necesarios y realizar la preparación logística de manera más eficiente, optimizando el tiempo de respuesta y asegurando que los recursos estén disponibles con antelación.
Para el desarrollo del tablero se utilizaron los scripts app.py y scripts/dashboard.py, principalmente, los cuales pueden ser consultados en el repositorio de GitHub.

7. Reporte de trabajo en equipo

Nombre y Rol	Responsabilidad

Karine Terán: Despliegue de maquina en EC2, Diseñadora del tablero	- Despliegue de la máquina virtual en EC2 de AWS incluyendo la configuración del ambiente y las dependencias necesarias. - Colaboración en el desarrollo y diseño del tablero interactivo utilizando Dash y Plotly principalmente,para contribuir a la visualización de datos históricos que permitan una mejor interpretación de la información.
Diego Tibaduisa: Ingeniero de Datos e Infraestructura	- Monitoreo del adecuado versionamiento del código y la correcta organización del repositorio. - Apoyo en el análisis para comparar los modelos, así como en la compilación y documentación detallada de los desarrollos realizados
Andrés Arcila: Analista de Datos, modelado Predictivo y diseño del tablero	- Implementación y gestión de modelos mediante MLFlow sobre Databricks, asegurando un control eficiente del ciclo de vida del modelo y la experimentación. - Participación en la creación del tablero interactivo, integrando el mejor modelo predictivo del posible puerto al que va a arribar próximamente el barco.

Entregables

Los demás entregables y evidencias solicitadas en esta entrega se encuentran en los siguientes repositorios:

Link GitHub: CLICK AQUÍ
Link DataBricks: CLICK AQUÍ
Link tablero: http://3.95.24.84:8050/