Presentación Final

Titanic Dataset

Equipo 4:

- A00836286 - Esteban Sierra Baccio
- A00837527 - Diego Esparza Ruíz
- A01722667 - Javier Jorge Hernández Verduzco
- A01285193 - Sergio Omar Flores García
- A01613878 - Sergio Aarón Hernández Orta

Der Untergang der "Titanic"
por Willy Stöwer

Problema y motivación

5
Semanas

891
Pasajeros

3
Hipótesis

Nuestro reto

- Nuestra pregunta de investigación era: “¿qué diferenció a los que sobrevivieron de aquellos que no, y qué nos dice de las circunstancias del desastre?”

- Durante cinco semanas y cuatro entregas, ideamos y respondimos preguntas de interés, relacionadas a la igualdad o disparidad enfrentada por diversos grupos a bordo del Titanic al momento de desembarcar cuando sucedió el accidente.

- Formulamos tres hipótesis a partir de un análisis exploratorio, y encontramos las respuestas a través de estadística y machine learning.

Metodología

Entrega 1

Exploración y Formulación de Hipótesis:

- Primera exposición a los datos, encontramos los patrones más obvios/superficiales y los graficamos:
  - Diferencia de supervivencia por clase.
  - Distribución de las tarifas.
  - Proporción de diferentes grupos (clase, familias, puerto de embarque)
- A partir de estos patrones formulamos hipótesis, preguntas que pensábamos que serían interesantes de resulver.

Entrega 2

Diseño Metodológico y Preprocesamiento:

- Diseñamos el pipeline que procesa el dataset original para agregar las transformaciones que consideramos importantes para nuestros futuros análisis con modelos de machine learning, y que imputan de manera sensata los datos faltantes, considerandos sesgos en la documentación de los datos.

Metodología

Entrega 3

Modelado e Interpretación:

- Usando el set de datos preprocesado, entrenamos cuatro modelos diferentes, cada uno optimizado probando decenas de combinaciones de hiperparámetro.
- Haciendo uso de pruebas específicas para cada modelo, extrajimos los insights que encontraron cada modelo sobre las relaciones escondidas entre los datos que transformamos.
- No solo analizamos las features que mejor predecían la supervivencia, sino que exploramos los casos en los que había más incertidumbre o con un resultado contradictorio a la expectativa del modelo.

Entrega 4

Análisis Crítico y Consideraciones Éticas

- Como cierre de nuestro proyecto, analizamos las implicaciones éticas, los sesgos del equipo y de los algorítmos que entrenamos, las limitaciones de análisis como el que hicimos, y el impacto social que de retroactivamente analizar datos en situaciones como esta.

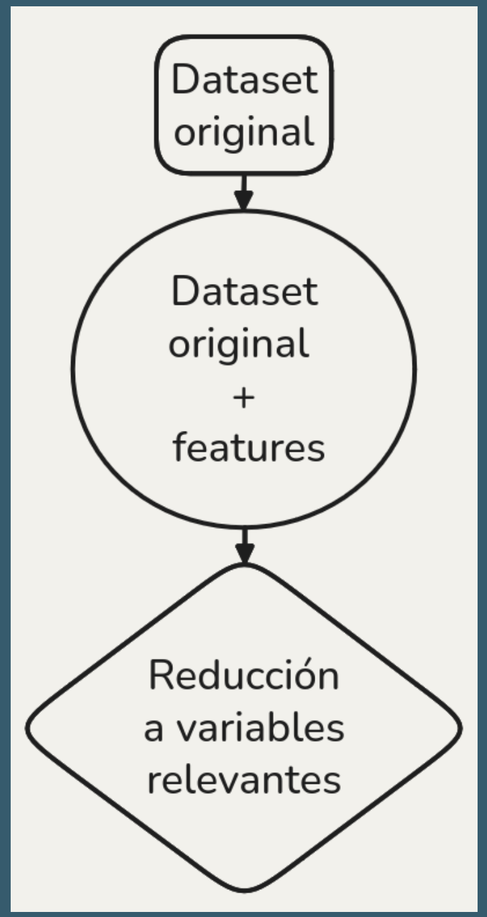
Hipótesis

1. Los pasajeros con hijos menores de edad tuvieron mayor probabilidad de supervivencia que los pasajeros sin hijos menores de edad.

2. La probabilidad de supervivencia en primera clase es al menos 50% mejor que la de tercera clase, para todas las edades y géneros.

3. La posibilidad de supervivencia disminuye en cuestión de ser de tercera vinculados a varios familiares provenientes de Southampton.

Metodología



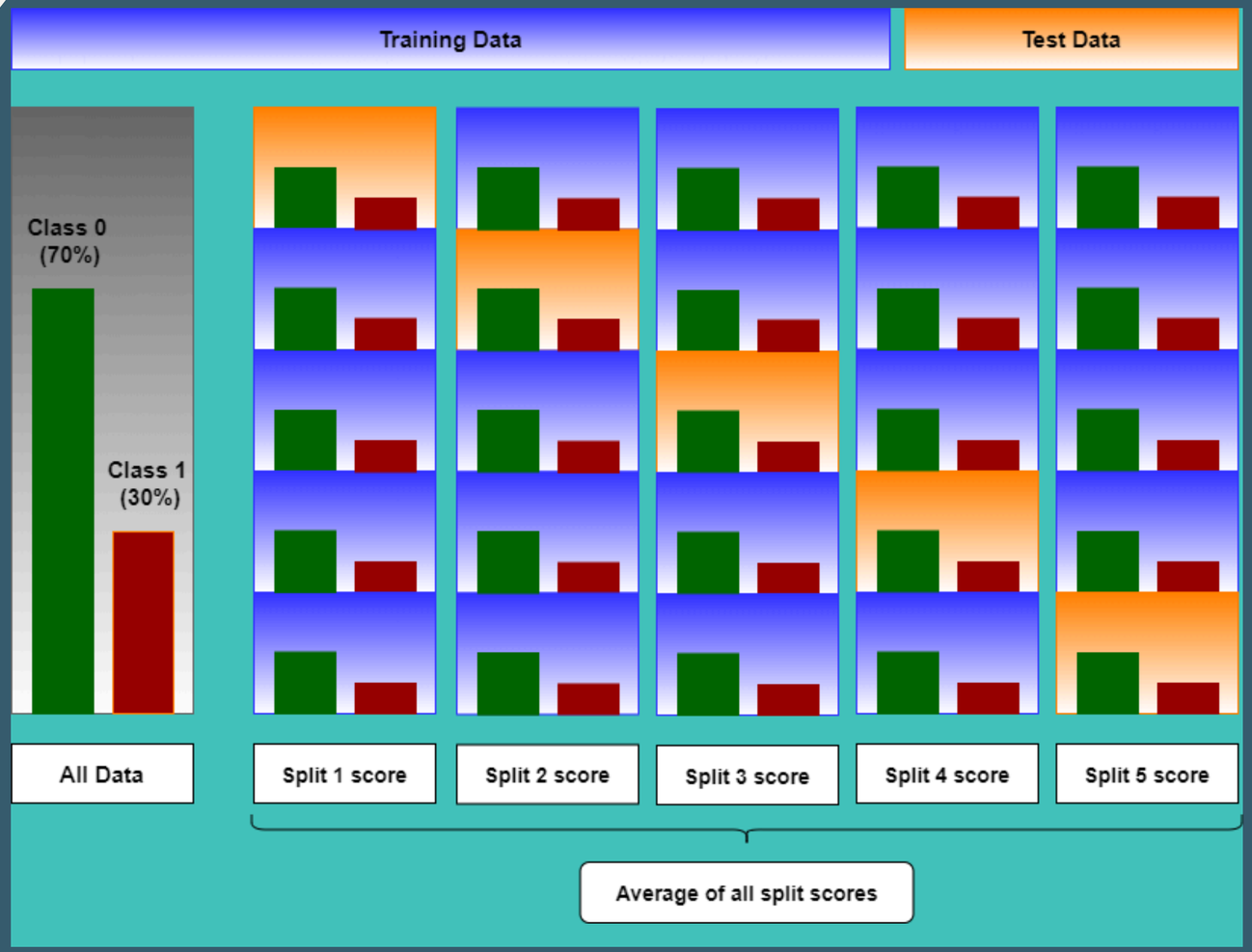
Preprocesamiento

- Inicialmente, el dataset tiene 12 columnas
  - Categóricas (ej. Pclass)
  - Numéricas (ej. Fare)
  - Binarias (ej. Survived)

- Desglosamos las variables categóricas y agregamos las features que creamos:
  - Name → Title
  - PArch + SibSp → FamilySize → IsAlone

- Al preprocesar, eliminamos todas aquellas que no contribuyen al entendimiento de nuestra pregunta de investigación. (ej. LastName)

Metodología



Cada combinación de hiperparámetros se entrenó 5 veces

Modelos implementados

- Implementamos cuatro modelos:
  - Regresión Logística
  - Random Forest
  - XGBoost (Gradient Boosting)
  - Support Vector Machine

- Todos los modelos los entrenamos con el mismo set de datos procesados, pero usando validación cruzada estratificada (5-fold).

- Optimizamos estos modelos con GridSearchCV, exhaustivamente probando todas las combinaciones de hiperparámetros.

Resultados

Feature	Importance
Title_Master	1.377055
Title_Mr	1.244537
Sex_female	1.150582
CabinDeck_E	0.826376
CabinDeck_D	0.811097
CabinDeck_Unknown	0.806421
Pclass_3	0.737894
Title_Mrs	0.517542
FamilySize	0.507862
Sex_male	0.436385

Mejor modelo

- El mejor modelo después de nuestra optimización fue el de regresión logística, con 81% de precisión en identificar a supervivientes.
- Curiosamente, todos los modelos eran mejores identificando quienes no sobrevivian que quienes si.
- Tiene sesgos muy interesantes. No es perfecto, como ningún modelo.

```
--- Logistic Regression (l1/l2) ---
              precision    recall  f1-score   support

           0       0.85      0.89      0.87       110
           1       0.81      0.75      0.78        69

    accuracy                           0.84       179
   macro avg       0.83      0.82      0.83       179
weighted avg       0.84      0.84      0.84       179

[[98 12]
 [17 52]]
ROC-AUC: 0.8778656126482214
```

Resultados

```
Métricas desagregadas por 'Sex_male':
  Grupo '1.0':
    - Accuracy: 0.8475
    - Precision: 0.8750
    - Recall: 0.2917
    - F1-Score: 0.4375
    - ROC-AUC: 0.7575
    - PR-AUC: 0.5959
    - Balanced Accuracy: 0.6405
    - MCC: 0.4500
  Grupo '0.0':
    - Accuracy: 0.8197
    - Precision: 0.8036
    - Recall: 1.0000
    - F1-Score: 0.8911
    - ROC-AUC: 0.8389
    - PR-AUC: 0.9249
    - Balanced Accuracy: 0.6562
    - MCC: 0.5011
```

Insight

- Uno de los sesgos más reveladores, lo encontramos en como predice el modelo en base al sexo.
- En el caso de las mujeres, el modelo casi asume supervivencia, con un recall de 100% comparado al 29% en hombres. Para casi todos los pasajeros hombres que sobrevivieron, predice que hubieran fallecido.
- Aún así, la precisión de los hombres es mayor: en los hombres que sí predice como sobrevivientes, casi siempre está en lo correcto.
- Al haber sobrevivido tan pocos hombres, el modelo prácticamente asume que no sobrevivirán a menos que las otras variables lo hagan confiar en lo contrario.
- Casi todos los modelos presentaron este sesgo de una forma u otra.

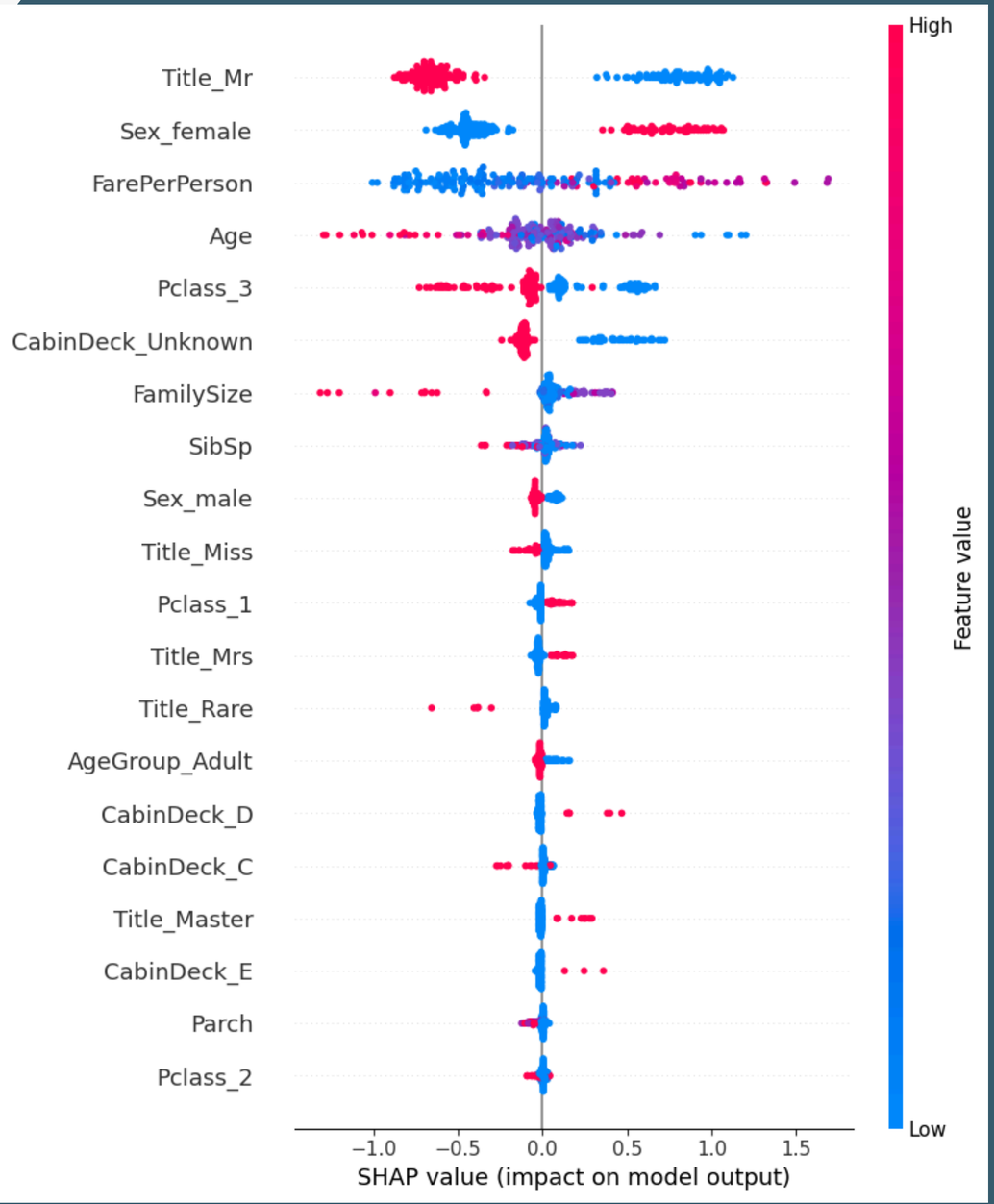
```
Métricas desagregadas por 'AgeGroup_Child':
  Grupo '0.0':
    - Accuracy: 0.7818
    - Precision: 0.7407
    - Recall: 0.6452
    - F1-Score: 0.6897
    - ROC-AUC: 0.8541
    - PR-AUC: 0.8314
    - Balanced Accuracy: 0.7546
    - MCC: 0.5256
  Grupo '1.0':
    - Accuracy: 0.9286
    - Precision: 0.8750
    - Recall: 1.0000
    - F1-Score: 0.9333
    - ROC-AUC: 0.8776
    - PR-AUC: 0.8260
    - Balanced Accuracy: 0.9286
    - MCC: 0.8660

Métricas desagregadas por 'AgeGroup_Adolescent/Teenager':
  Grupo '0.0':
    - Accuracy: 0.7818
    - Precision: 0.7321
    - Recall: 0.6613
    - F1-Score: 0.6949
    - ROC-AUC: 0.8425
    - PR-AUC: 0.8049
    - Balanced Accuracy: 0.7578
    - MCC: 0.5274
  Grupo '1.0':
    - Accuracy: 0.9286
    - Precision: 1.0000
    - Recall: 0.8571
    - F1-Score: 0.9231
    - ROC-AUC: 1.0000
    - PR-AUC: 1.0000
    - Balanced Accuracy: 0.9286
    - MCC: 0.8660
```

Resultados

Insight

- El otro gran sesgo es de la edad, de manera similar al sesgo favorable para las mujeres, los jóvenes también tienen un sesgo positivo.
- El modelo presenta un sesgo favorable hacia los niños. Su Recall es de 1.0000, lo que significa que identificó a cada niño que sobrevivió.
- Similar a los niños, el modelo es casi perfecto con los adolescentes. Su Precisión de 1.0000 nos muestra que cada vez que predijo que un adolescente sobreviviría, la predicción fue correcta. Un ROC-AUC de 1.0000 demuestra que tiene una capacidad de discriminación perfecta para este grupo.

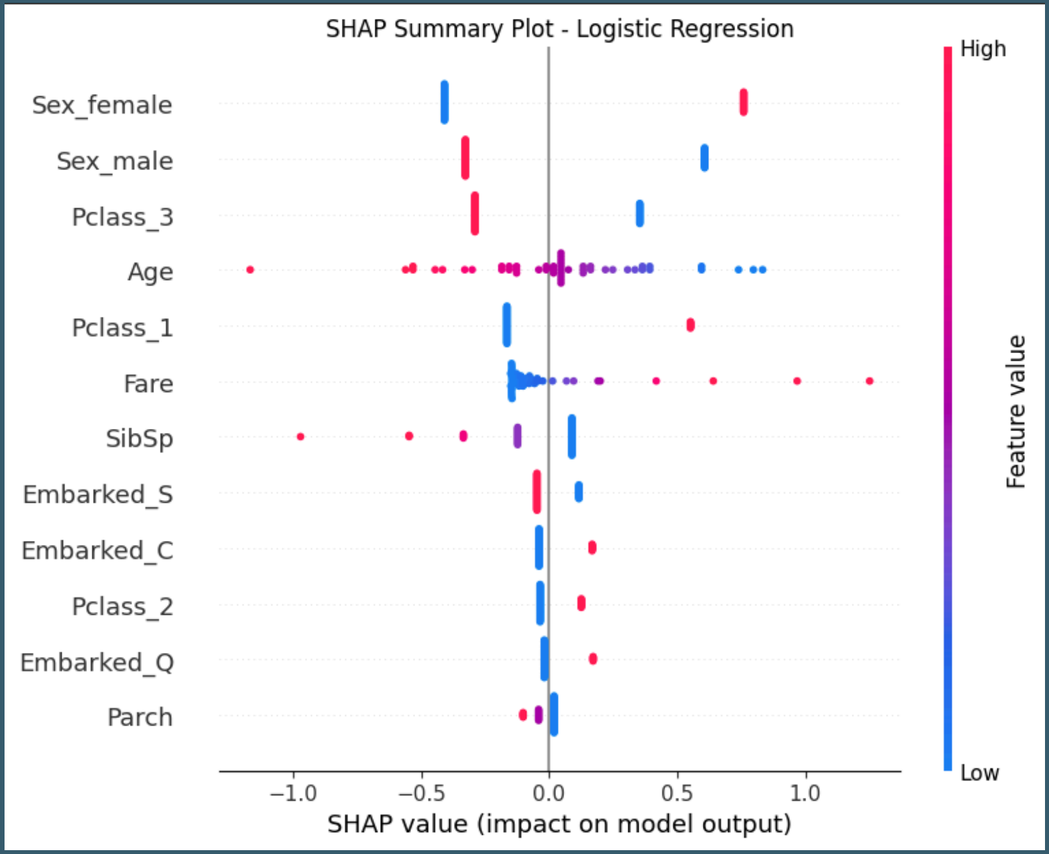
Resultados



Insight

- Los sesgos anteriores se presentaron de diferentes maneras en todos los modelos, pero en esta gráfica se puede ver el efecto general de las variables:
  - Ser un hombre adulto, reduce significantemente la probabilidad de supervivencia (Title_Mr es ambas).
  - Ser una mujer de cualquier edad aumenta la probabilidad de supervivencia.
  - Pagar más por el boleto por persona, probablemente por su correlación con la clase de boleto, indica mayor supervivencia.

Resultados



Insight

Interpretamos según SHAP, PDP y LIME: el modelo prioriza señales de género, luego de clase y después de edad; estas variables empujan con fuerza la predicción, mientras que fare, familia y embarque aportan poco y de forma inconstante, lo que explica tanto la brecha estructural entre grupos como las excepciones que se entienden caso por caso con explicaciones locales.

Resultados

Después de validar las hipótesis con nuestro modelo concluimos que:

- **La Hipótesis 1 no se sostiene:** tener hijos menores no mejora la supervivencia y el efecto es marginal e inconsistente.
- **La Hipótesis 2 se confirma con fuerza:** la clase social es un factor central, y la 1ª clase ofrece una ventaja de supervivencia muy superior a la 3ª, más marcada en hombres.
- **La Hipótesis 3 se apoya parcialmente:** la desventaja de la 3ª clase persiste, pero el peso adicional de la familia grande o el puerto de embarque es limitado.

De estos resultados se desprenden tres grandes aprendizajes:

- La predicción puede simplificarse, ya que género y clase concentran la mayor parte de la señal.
- La estructura social emerge como factor dominante, por encima de la edad o composición familiar.
- No se observa una protección sistemática por lazos familiares, lo que limita la idea de altruismo preferencial.

Dashboard

Lecciones Éticas

1

Este proyecto y las conclusiones que hacemos, son en retrospectiva. Es difícil definir lineamientos sobre quien “merece” o no sobrevivir en un desastre del que no pueden todos salir vivos. No podemos culpar a aquellos que valoran su vida y quisieran conservarla existiendo fuera de ellos.

2

Posibles injusticias y decisiones o acciones sistemáticas solo pueden ser **respaldadas** por datos, no inferidas de ellos. Para entender que pasó en el Titanic se necesitan de testimonios, no solo teorías basadas en un dataset como el nuestro.

Limitaciones Clave

Datos de Cabina

- La mayoria de los datos de cabina, específicamente de la segunda y tercera clase, están incompletos o son inexistentes.
- Estos datos pudieron haberse perdido en el desastre, ser registrados manualmente y descartados por inconsistencias, entre muchas posibilidades de por qué no se encuentran.
- De haberse incluido más datos, hubiera sido posible analizar de forma más profunda la distribución espacial relativa a los botes o las salidas a las cubiertas superiores, posiblemente descubriendo más insights a lo que definió la probabilidad de sobrevivir de aquellos en segunda y tercera clase.

Datos Post-Desastre

- El dataset fue recopilado después del desastre, de múltiples fuentes.
- Esto significa que hay más y mejores datos para aquellos que sobrevivieron, y que falta de datos se correlaciona con la no-supervivencia.
- El dataset fue creado como una reconstrucción histórica, no científica. Puede que algunos de los sesgos observados en nuestros resultados, incluso los más claros como clase o género, fueran reforzados por una narrativa de los medios en esa época.

Conclusiones

Trabajo Futuro

- Sabiendo que el dataset fue formado de diferentes fuentes con diferentes variables y grados de calidad y cantidad de datos, hacer múltiples estudios más pequeños sobre cada fuente podría brindar una perspectiva más completa.
- Posiblemente, reimaginación de algunas hipótesis, buscando otros patrones más interesante o con una imagen más completa.

Contribuciones Principales

- Hicimos un análisis que responde de forma objetiva las hipótesis que nos propusimos.
- Creamos un pipeline que hace la metodología completamente repetible.
- Un análisis de los sesgos en los datos, además de reconocer las limitaciones de nuestro análisis en ellos.
- Creamos y posteamos un Dashboard con el que cualquiera puede explorar y jugar con los datos de este reto.