

Proyecto final: Spaceship Titanic

Aplicaciones de la matemática en ingeniería - MAT281

Renata Córdova

Universidad Técnica Federico Santa María

Lunes, 4 de diciembre de 2023

Índice

- 1 Descripción del proyecto
 - Desafío: Spaceship Titanic
 - Métrica
 - Datos
 - Formato submission
- 2 Tratamiento del Train Set
 - Tratamiento Atributos Numéricos
 - Tratamiento Atributos Categóricos
- 3 Selección de Modelo
 - Modelos
 - Mejor modelo
- 4 Métricas y Análisis de resultados
 - Puntajes
- 5 Resultados Kaggle
- 6 Conclusion
- 7 Referencias

Index

- 1 Descripción del proyecto
 - Desafío: Spaceship Titanic
 - Métrica
 - Datos
 - Formato submission
- 2 Tratamiento del Train Set
 - Tratamiento Atributos Numéricos
 - Tratamiento Atributos Categóricos
- 3 Selección de Modelo
 - Modelos
 - Mejor modelo
- 4 Métricas y Análisis de resultados
 - Puntajes
- 5 Resultados Kaggle
- 6 Conclusion
- 7 Referencias

Desafío: Spaceship Titanic

La nave espacial Titanic fue un transatlántico de pasajeros interestelar lanzado con casi 13.000 pasajeros a bordo.

Mientras rodeaba Alpha Centauri en ruta hacia su primer destino, el tórrido 55 Cancri E, la desprevenida nave espacial Titanic chocó con una anomalía del espacio-tiempo escondida dentro de una nube de polvo. Aunque la nave permaneció intacta, ¡casi la mitad de los pasajeros fueron transportados a una dimensión alternativa!

El desafío consiste en predecir qué pasajeros fueron transportados por la anomalía, utilizando los registros recuperados del sistema informático dañado de la nave espacial.

Métrica

Las predicciones se evalúan en función de su *accuracy* de clasificación, es decir, el porcentaje de etiquetas predichas que son correctas.

$$Accuracy = \frac{\text{Número de predicciones correctas}}{\text{Número total de predicciones}}$$

Datos

Los datos proporcionados por la competencia Kaggle son:

- **train.csv:** ~8700 datos.
 - Un identificador único para cada pasajero: `PassengerId`
 - **Atributos:** `HomePlanet`, `CryoSleep`, `Cabin`, `Destination`, `Age`, `VIP`, `RoomService`, `FoodCourt`, `ShoppingMall`, `Spa`, `VRDeck`, `Name`.
 - **Target:** `Transported`.
- **test.csv:** ~4300 datos.

Formato submission

La forma correcta de presentar la predicción y cargarla en Kaggle es en un archivo .csv con las columnas:

- `PassengerId`: Identificación para cada pasajero en el conjunto de pruebas.
- `Transported`: El target. Para cada pasajero, predecir si es Verdadero o Falso (tipo de dato booleano).

Index

- 1 Descripción del proyecto
 - Desafío: Spaceship Titanic
 - Métrica
 - Datos
 - Formato submission
- 2 **Tratamiento del Train Set**
 - **Tratamiento Atributos Numéricos**
 - **Tratamiento Atributos Categóricos**
- 3 Selección de Modelo
 - Modelos
 - Mejor modelo
- 4 Métricas y Análisis de resultados
 - Puntajes
- 5 Resultados Kaggle
- 6 Conclusion
- 7 Referencias

Tratamiento del Train Set

Primera etapa: `PassengerId`

Cada `Id PassengerId` toma la forma *gggg_pp*, donde *gggg* indica el grupo con el que el pasajero está viajando y *pp* es su número dentro del grupo. Las personas en un grupo suelen ser miembros de la familia, pero no siempre. Entonces, **se crean dos nuevas características numéricas:** `Grupo`, `Numero`.

Luego, la columna `PassengerId` **se convierte en el índice del Train Set**

Segunda etapa: `Split`

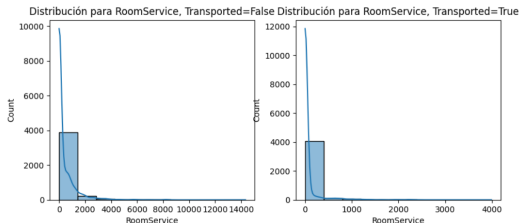
Se aplica un `train_set_split`, con el objetivo de realizar mediciones y análisis de resultados. Se forma con un 20 % de los datos de entrenamiento un nuevo conjunto de pruebas (etiquetado).

Tratamiento Atributos Numéricos

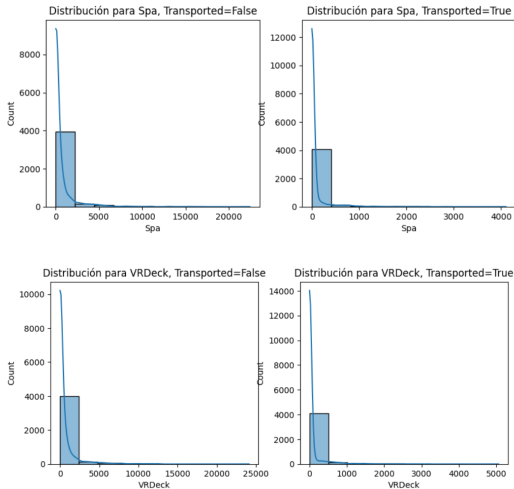
Tercera etapa: Estadística y visualización descriptiva

Para cada **atributo numérico** se realizan **gráficos de la distribución de los datos, separados por clase**.

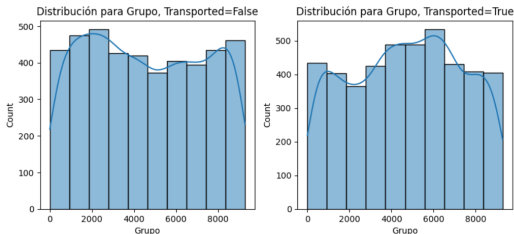
A continuación se presentan los atributos con mayores diferencias de distribución observadas entre clases.



Tratamiento Atributos Numéricos



Tratamiento Atributos Numéricos



Resultados

Se observan diferencia en el rango de valores par las variables RoomService, Spa y VRDeck.

Para Grupo, se observa curvas de distribución diferentes: la clase transportada prefiere valores medianos y la clase no transportada los valores extremos.

Tratamiento Atributos Numéricos

Cuarta etapa: Preprocesamiento

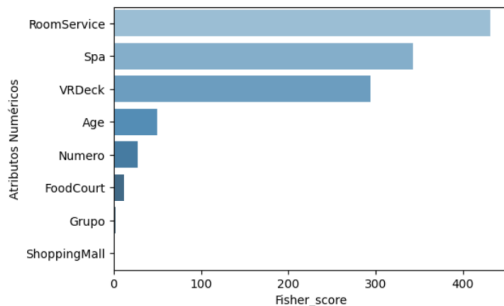
Se sustituyen los **datos nulos por la desviación estándar**, para mantener la variabilidad de cada atributo.

Luego, se aplica el **escalamiento** `MinMaxScaler`, para mantener la distribución de los datos.

Quinta etapa: Fisher-score

A cada atributo se mide su puntuación **Fisher-score** respecto la etiqueta.

Tratamiento Atributos Numéricos



Sexta etapa: Recorte de atributos

De acuerdo al análisis estadístico anterior y los resultados de las puntuaciones Fisher-score, **eliminaremos las variables** Age, FoodCourt, ShoppingMall, Grupo y Numero.

Tratamiento Atributos Categóricos

Séptima etapa: Suma y recorte de atributos

Cada dato de `Cabin` corresponde al número de camarote donde se aloja el pasajero. Tiene la forma cubierta/número/lado, donde cubierta y lado son una letra mayúscula. Entonces, **se crean tres nuevas características:** `Cabin1`, `Cabin2` y `Cabin3`.

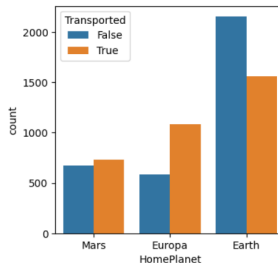
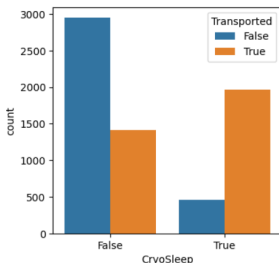
Observamos que `Cabin1` tiene 8 categorías; `Cabin3`, 2 categorías y `Cabin2` corresponde a una variable numérica de 1756 datos diferentes. Por lo tanto, **eliminamos las columnas** `Cabin` y `Cabin2`.

Tratamiento Atributos Categóricos

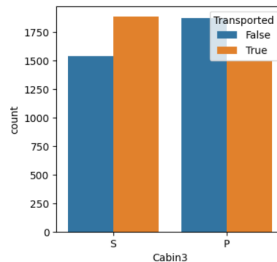
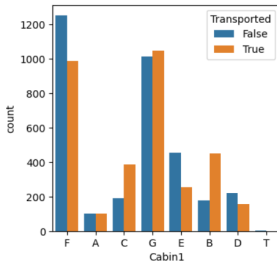
Octava etapa: Estadística y visualización descriptiva

Para cada **atributo categórico** se realizan **gráficos de la distribución de los datos, separados por clase**.

A continuación se presentan los atributos con mayores diferencias de distribución observadas entre clases.



Tratamiento Atributos Categóricos



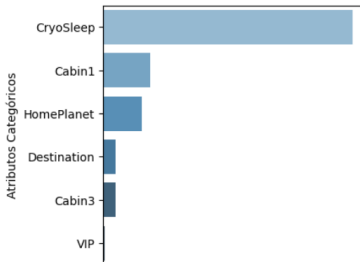
Tratamiento Atributos Categóricos

Novena etapa: Datos nulos

Se sustituyen los **datos nulos por la moda**

Décima etapa: Información mutua

Se mide la **Información mutua** entre el target y cada atributo.



Tratamiento Atributos Categóricos

Undécima etapa

De acuerdo al análisis estadístico anterior y los resultados de Información Mutua, **eliminaremos las variables** `Destination`, `Cabin3` y `VIP`.

Duodécima etapa

Se aplica el **escalamiento** `OneHotEncoder`, con la función `pd.get_dummies`.

Index

- 1 Descripción del proyecto
 - Desafío: Spaceship Titanic
 - Métrica
 - Datos
 - Formato submission
- 2 Tratamiento del Train Set
 - Tratamiento Atributos Numéricos
 - Tratamiento Atributos Categóricos
- 3 Selección de Modelo
 - Modelos
 - Mejor modelo
- 4 Métricas y Análisis de resultados
 - Puntajes
- 5 Resultados Kaggle
- 6 Conclusion
- 7 Referencias

Modelos

Se lleva a cabo una búsqueda en cuadrícula (**Grid Search**) en la que se varían diversos hiperparámetros para los modelos de **Logistic Regression, SVC, Random Forest Classifier y Decision Tree Classifier**.

Cada búsqueda en cuadrícula se configura con 'accuracy' como métrica de evaluación, con 5 pliegues en la validación cruzada cv y n_jobs igual a -1.

Mejor modelo

En base a los resultados obtenidos en cada búsqueda en cuadrícula, el mejor modelo es **RandomForestClassifier ('max_depth': 10, 'n_estimators': 100)**, con un accuracy de 0.79.

Index

- 1 Descripción del proyecto
 - Desafío: Spaceship Titanic
 - Métrica
 - Datos
 - Formato submission
- 2 Tratamiento del Train Set
 - Tratamiento Atributos Numéricos
 - Tratamiento Atributos Categóricos
- 3 Selección de Modelo
 - Modelos
 - Mejor modelo
- 4 **Métricas y Análisis de resultados**
 - **Puntajes**
- 5 Resultados Kaggle
- 6 Conclusion
- 7 Referencias

Puntajes

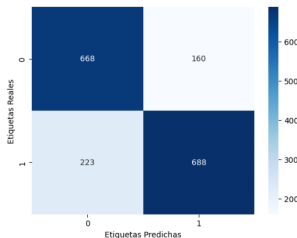
Se aplica el mismo tratamiento al conjunto de pruebas creado durante el split. Posteriormente, se evalúa una serie de puntajes del mejor modelo encontrado, aplicándolos al conjunto de pruebas procesado.

- **Accuracy:** 0.7798
- **Precision:** 0.8113
- **Recall:** 0.7552
- **F1 Score:** 0.7823

Se observan valores similares en todos los puntajes, por lo tanto, el modelo no tiende más un tipo de error.

Matriz de confusión

Por medio de la matriz de confusión podemos visualizar el comportamiento del modelo.




Se observa que el modelo en la mayoría de los casos predice correctamente. También, se ve que el modelo tiende a no cometer más un tipo de error.

Index


- 1 Descripción del proyecto
 - Desafío: Spaceship Titanic
 - Métrica
 - Datos
 - Formato submission
- 2 Tratamiento del Train Set
 - Tratamiento Atributos Numéricos
 - Tratamiento Atributos Categóricos
- 3 Selección de Modelo
 - Modelos
 - Mejor modelo
- 4 Métricas y Análisis de resultados
 - Puntajes
- 5 **Resultados Kaggle**
- 6 Conclusion
- 7 Referencias

Resultados Kaggle

 Getting Started Prediction Competition


Spaceship Titanic

Predict which passengers are transported to an alternate dimension

 Kaggle · 2,512 teams · Ongoing

[Overview](#) [Data](#) [Code](#) [Models](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Team](#) [Submissions](#) [Submit Predictions](#) [...](#)

Submissions

<div>All Successful Errors</div>		Recent ▾
Submission and Description		Public Score ⓘ
<div> submission_space_titanic.csv Complete · now</div>		0.78512

Index

- 1 Descripción del proyecto
 - Desafío: Spaceship Titanic
 - Métrica
 - Datos
 - Formato submission
- 2 Tratamiento del Train Set
 - Tratamiento Atributos Numéricos
 - Tratamiento Atributos Categóricos
- 3 Selección de Modelo
 - Modelos
 - Mejor modelo
- 4 Métricas y Análisis de resultados
 - Puntajes
- 5 Resultados Kaggle
- 6 Conclusion**
- 7 Referencias

Conclusiones

El accuracy obtenido fue de 0.78512, indicando un buen desempeño en el trabajo realizado. No obstante, aún hay margen para mejoras.

La desventaja de recortar variables mediante **métodos de filtrado individual (como Fisher-score e Información mutua)** radica en su incapacidad para detectar redundancias o complementariedades entre atributos. Por lo tanto, una posible mejora sería la exploración de métodos de filtrado multivariado o la consideración de algoritmos de reducción de dimensionalidad (PCA, t-SNE, etc.).

Otra observación relevante es que, al abordar inicialmente los atributos numéricos, **se eliminó la variable Cabin2 sin evaluar su impacto en la capacidad predictiva.**

Referencia

- Kaggle. (2023). Spaceship Titanic Competition. Recuperado de <https://www.kaggle.com/competitions/spaceship-titanic/>