

# Entrega Final de Proyecto

## Inteligencia Artificial

Código: 2508401

Grupo: 01

Fecha: 12/ 11/ 2022

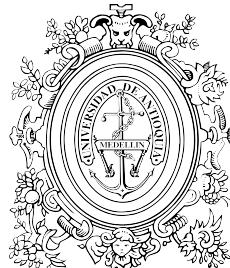
**Elaborado por**

*Sofía Gutiérrez Tangarife*, CC: 1001576929

*Melissa Galeano Ruiz*, CC: 1000416463

**Profesor**

*Raúl Ramos Pollán*



**UNIVERSIDAD  
DE ANTIOQUIA**

1 8 0 3

Facultad de Ingeniería

Departamento de Ingeniería Mecánica

Universidad de Antioquia

## Contents

<b>1</b>	<b>Introducción</b>	<b>2</b>
<b>2</b>	<b>Exploración descriptiva del Dataset</b>	<b>2</b>
2.1	Preprocesado de los datos . . . . .	3
2.2	Modelo Supervisado . . . . .	3
2.2.1	Analisis de Datos - Matriz de Correlación . . . . .	4
2.2.2	Matriz de Confusión . . . . .	5
2.2.3	Evaluación del Modelo Supervisado . . . . .	5
2.3	Modelo No Supervisado . . . . .	6
<b>3</b>	<b>Resultados, Métricas y Curvas de Aprendizaje</b>	<b>8</b>
<b>4</b>	<b>Retos y Consideraciones de Despliegue</b>	<b>10</b>
<b>5</b>	<b>Referencias</b>	<b>11</b>

# 1 Introducción

El notebook desarrollado para este proyecto esta divido en cuatro secciones principales:

1. Cleaning Dataset
2. Dataset Preparation
3. Dataset Analysis
4. Model Evaluation

En cada una de ellas se realiza y se describe detalladamente el análisis llevado a cabo con los datos de entrenamiento para evaluar la información que se tiene del problema, y de esta manera, generar un modelo predictivo de datos funcional que permita predecir cuáles pasajeros podrían ser enviados a la dimensión alterna por la anomalía.

## 2 Exploración descriptiva del Dataset

En el análisis y procesamiento de datos para modelos de aprendizaje supervisado se debe hacer una buena gestión de la información que se tiene para poder obtener modelos con un alto porcentaje de acierto en las predicciones.

En esta primera sección del notebook, se hace una análisis cualitativo de los datos en donde se observa la información de todas las posibles variables junto con sus valores. Además, en este punto cabe resaltar que se evalúa la cantidad de datos nulos y no nulos, ya que la falta de información se debe ajustar para poder realizar un análisis cuantitativo posterior que permita generar predicciones significativas.

Una vez se determinan cuantos valores nulos y no nulos se tienen en cada una de las columnas del dataset, y a qué tipos de datos corresponden (Discretos o Continuos), se comienza el proceso de reparación de datos. El proceso de reparación de datos consiste en evaluar cada una de la columnas con sus datos faltantes correspondientes (Datos nulos) y reemplazarlos por el mejor tipo de dato que se adapte dado el caso.

Para el dataset que se tenía, se realizaron los siguientes reemplazos de los datos en cada una de las columnas:

- **Name** Se decide eliminar los datos faltantes ya que no representan una cantidad significativa en comparación con los datos totales, además de que al tratarse de nombre personales, no se recomienda reemplazar con datos existentes porque cada persona es única, ni crear nombres aleatorios.
- **Destination** Se decide reemplazar con el dato mas recurrente, pues es razonable que los datos faltantes se reemplacen con el destino mas concurrido por los usuarios.
- **Cabin, CryoSleep, Age** Se decide reemplazar con el valor del dato anterior.

- `RoomService`, `Foodcourt`, `ShoppingMall`, `Spa`, `VRDeck` Se decide reemplazar con el valor medio.
- `VIP` Se decide reemplazar con el dato que mas se repite.

## 2.1 Preprocesado de los datos

Una vez que el dataset esta reparado se procede con su preparación ya que para el proyecto asignado debe cumplir con los siguientes requisitos:

- Al menos 5000 instancias (filas, imágenes, etc.).
- Al menos 30 columnas.  
Al menos el 10% de las columnas han de ser categóricas.
- Al menos ha de tener un 5% de datos faltantes en al menos 3 columnas.

Para esto, se desarrollaron las siguientes seis funciones que fueron aplicadas sobre el dataset para generar más columnas de datos que permitieran categorizar de una forma más específica.

1. `split Passenger` Esta función se aplica sobre la columna `passenger` para dividirla en dos columnas nuevas que permitieran diferenciar entre ID y el grupo en el que se encuentran los pasajeros.
2. `one hot` Esta función se aplica sobre las columnas `HomePlanet` y `Destination` para dividirlas en en columnas discretas con cada uno de sus elementos.
3. `split Name` Esta función se aplica sobre la columna `Name` para dividirla en dos columnas diferentes llamadas `Name` y `Last Name`.
4. `TF` Esta función se aplica sobre las columnas `Cryosleep`, `VIP`, `Transported` para crear nuevas columnas booleanas.
5. `taken services` Esta función se aplica para todas las columnas de los servicios que ofrece el spaceship, para crear columnas categóricas con cada uno de ellos.
6. `age range` Esta funcion se aplica sobre la columna `Age`, para subdividirla en tres columnas con tres diferentes rangos de edades `Youth`, `Adult`, `Older`

Finalmente, una vez se confirma que se cumplen con los requisitos se inicia el proceso de análisis de datos para saber como proceder con el modelo predictivo.

## 2.2 Modelo Supervisado

Una vez se tiene el dataset totalmente reparado para el inicio del proceso de creación del modelo, se debe hacer un óptimo análisis de datos para comprender la correlación entre las variables y así definir las variables mas significativas para el modelo predictivo.

### 2.2.1 Análisis de Datos - Matriz de Correlación

Este proceso se inicia con la matriz de correlación ya que esta nos permite visualizar mediante un rango de colores la relación entre cada una de las variables. Como se visualiza en la siguiente figura, en esta matriz se tiene un rango entre  $[-1, 1]$  en donde cada numero es equivalente corresponde un color y refleja la relación entre variables, siendo el blanco = 1, la correlación más alta que se puede obtener, y el negro = 0 , la más baja.

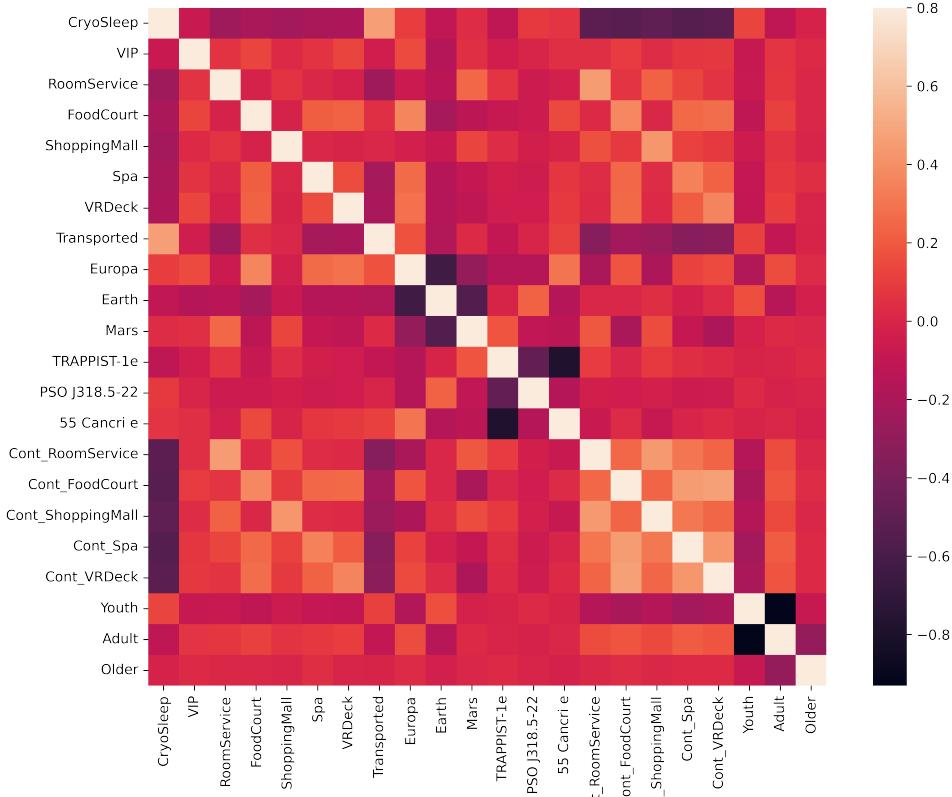


Figure 1: Matriz de Correlación

Es importante destacar que como la variable `Transported` es la que se desea predecir, esta es la que se debe analizar con mas detenimiento pues se necesita verificar cuales son las variables con una correlación mas significativa con respecto a ella. Por tanto, mediante esta matriz se puede visualizar que la variable `Cryosleep` es la variable con la mayor correlación. Sin embargo, es importante mencionar que según la matriz de correlación, la variable `youth` sugiere que los jóvenes tenían mas probabilidad en comparación con los viejos y los adultos de ser transportados. Asimismo, las personas que venían de Europa, como origen, tenían mayor probabilidad de ser transportados a otra dimensión.

Subsecuente a esto se procede a analizar la frecuencia de repetición de cada una de las variables en el dataset mediante histogramas para tener en cuenta otras posibles causas de que los pasajeros sean transportados y asimismo, para obtener una mejor visualización de la distribución de los datos.

### 2.2.2 Matriz de Confusión

Considerando la información obtenida mediante los histogramas de cada una de las columnas del dataset, ahora se analizara la correlación entre la variable `Cryosleep` y `Transported` mediante una matriz de confusión.

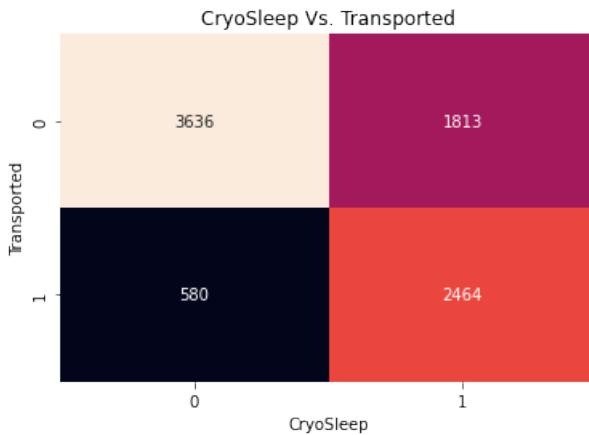


Figure 2: Matriz de Confusión

Como se visualiza, los pasajeros que estaban dentro de las cryosleep son los que tienen una mayor probabilidad de ser transportados a otras dimensiones durante el viaje en el spaceship a comparación de los pasajeros que no las utilizaron.

### 2.2.3 Evaluación del Modelo Supervisado

Para evaluar cuál es el mejor modelo para el dataset del problema se inició con la creación de una copia del dataset para evitar modificar los datos originales. Adicionalmente, se estableció el la columna de `Passenger` como el index del dataset para visualizar cuales de los pasajeros serian los posibles transportados, y se eliminan los datos discretos categóricos ya que estos no contribuyen al proceso de predicción.

Se divide el dataset de entrenamiento inicial en 70% de datos para entrenamiento y el 30% restante para datos de validación.

A continuación, se crea un ciclo *for* para evaluar los datos con 5 posibles estimadores y ver su porcentaje de acierto, con el cual, se decide cuál es el modelo mas óptimo para los datos.

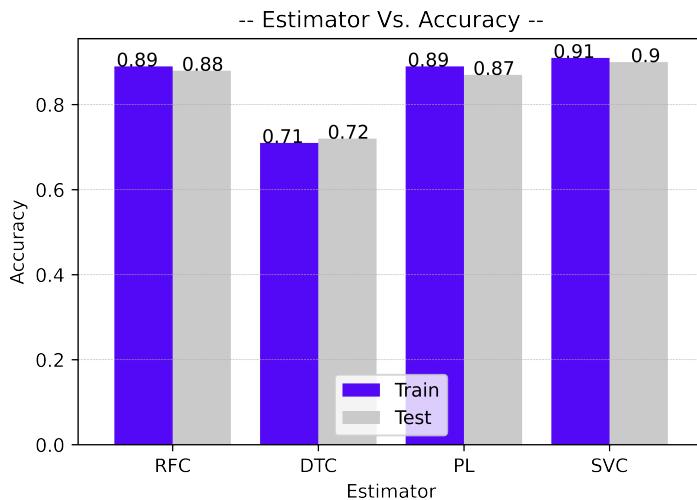


Figure 3: Comparación entre diferentes estimadores

Donde:

- *RFC* : Random Forest Classifier, con `max depth = 30` y `n estimators = 5`
- *DTC* : Decision Tree Classifier, con `max depth = 2`
- *PL* : Pipeline, con `n components = 2` y `SVC(gamma = 1)`
- *SVC* : Support Vector Classifier, con `gamma = 1`

Como se observa en la figura, el mejor porcentaje de acierto se obtuvo con el Support Vector Machine, lo que quiere decir que este es el estimador mas óptimo para el modelo de predicción.

Adicionalmente, una vez evaluados los diferentes estimadores, se decide re-evaluarlos para analizar su rendimiento y su variación debido al cambio en los hiper-parámetros n-estimators, max-depth, n-components, gamma.

Una vez realizadas las diferentes iteraciones de hiper-parámetros en los estimadores, se concluye que los mejores estimadores para el modelo supervisado es `Pipline` con n-components igual a 12 y un gamma igual a 1, y `SVC` con un gamma igual a 1, con ambos se llega a obtener un 91% de acierto para train y un 90% de acierto en test.

## 2.3 Modelo No Supervisado

Como nuestro objetivo específico es el de determinar la predicciones de cuantos y cuales pasajeros tienen mayor probabilidad de ser transportados a otras dimensiones durante su viaje interestelar debido a diferentes factores, se tomo la matriz de correlación obtenida con el análisis de datos para ver cuales variables tenían una mayor correlación respecto a la variable a predecir. En este caso, para la variable `transported`, se tiene las variables `Cryosleep`, `Europa` y `Youth`.

Con esta información se procede a crear un dataframe con dos de esas variables (`Cryosleep` y `Europa`) para definir los clusters de nuestro modelo no supervisado. Con esto en consideración,

se realiza un análisis de Silhouette para evaluar la calidad de los clusters que se podrán llegar a generar mediante los algoritmos de K-Means, esto en términos de qué tan bien se agrupan las muestras con otras muestras que son similares entre sí. La puntuación de Silhouette se calcula para cada muestra de diferentes clusters.

PassengerId	CryoSleep	Europa
0001_01	0	1
0002_01	0	0
0003_01	0	1
0003_02	0	1
0004_01	0	0
0005_01	0	0
0006_01	0	0
0006_02	1	0
0007_01	0	0
0008_01	1	1
0008_02	1	1
0008_03	0	1
0009_01	0	0
0010_01	0	0
0011_01	0	0
0012_01	0	0
0014_01	0	0
0015_01	0	0
0016_01	1	0
0017_01	0	0
0017_02	0	0
0020_01	1	0
0020_02	1	0
0020_03	1	0
0020_04	0	0

Figure 4: DataFrame para Clustering

Para el Dataset dado, se puede evidenciar que las variables, para determinar que un usuario sea teletransportado o no, son binarias, no continuas. Esto quiere decir que en términos de distribución y varianza, no obtenemos diferentes valores continuos, si no, únicamente ceros y unos, y esto lo podemos de igual manera observar en las siguientes gráficas donde se visualiza 4 puntos totalmente demarcados, donde no existe ni la mas mínima dificultad de diferenciación.

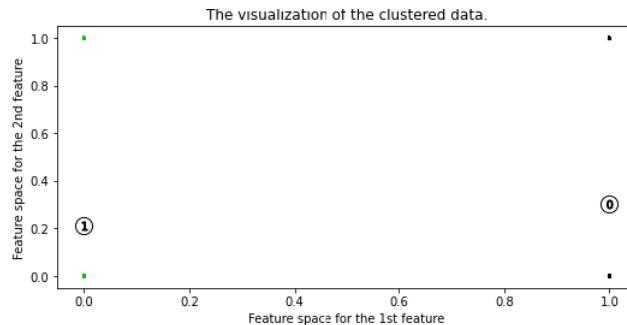


Figure 5: Clustering

Este hecho es de gran relevancia, pues conociendo este comportamiento y entendiendo la distribución de los datos y sus implicaciones en las predicciones, se concluye que para realizar un modelo de predicción no supervisado, es totalmente necesario tener datos continuos para los clusterings y saber las tendencias de separación de los datos por su variabilidad. En conclusión, debido a los datos que se tienen para este problema, no es posible conducir modelos no supervisados en el modelo que se ha probado.

### 3 Resultados, Métricas y Curvas de Aprendizaje

Al realizar el análisis con las curvas de aprendizaje correspondientes a cada uno de los mejores estimadores del modelo supervisado bajo la consideración del score de accuracy, evaluado previamente con sus mejores hiperparametros, y utilizando la librería Sklearn junto con el método Bootstrapping, se obtienen las siguientes gráficas.

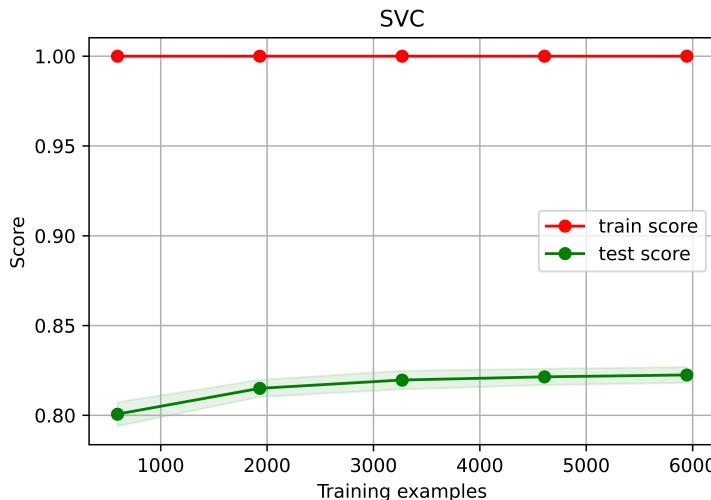


Figure 6: Support Vector Machine

Como se puede observar aquí en esta primera gráfica, al agregar mas datos, la gráfica no converge. Aunque se tiene un buen rendimiento en el entrenamiento del modelo, en el test sucede todo lo contrario, y se puede evidenciar el gran espaciamiento entre ambos, que nos habla de la gran diferencia entre desempeños. En otras palabras, se puede decir que el algoritmo memorizó los datos, hay un overfitting, y esto se puede deber a varias razones tales como:

- Se necesitan muchos mas datos.
- Requiere otro tipo de algoritmo.
- La complejidad del modelo es muy alta y se debe disminuir.

Teniendo en cuenta esto, se decide disminuir la complejidad del modelo con el estimador de SVC disminuyendo el valor del gamma a 0.2 y 0.01. De lo anterior, se obtienen las siguientes gráficas.

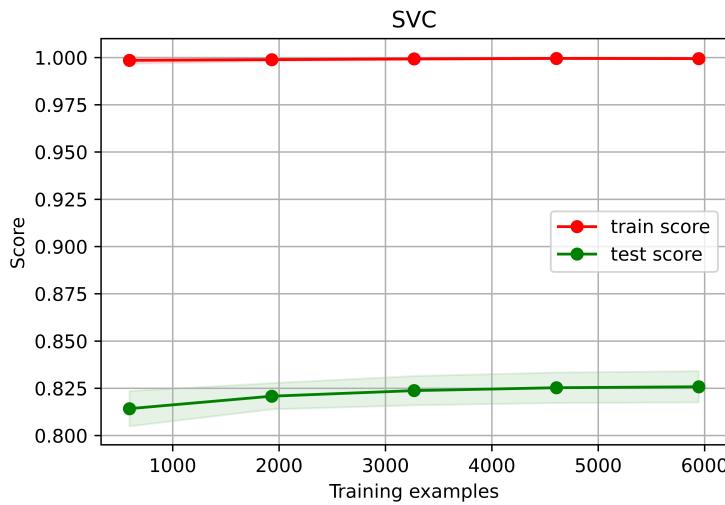


Figure 7: Support Vector Machine con gamma=0.2

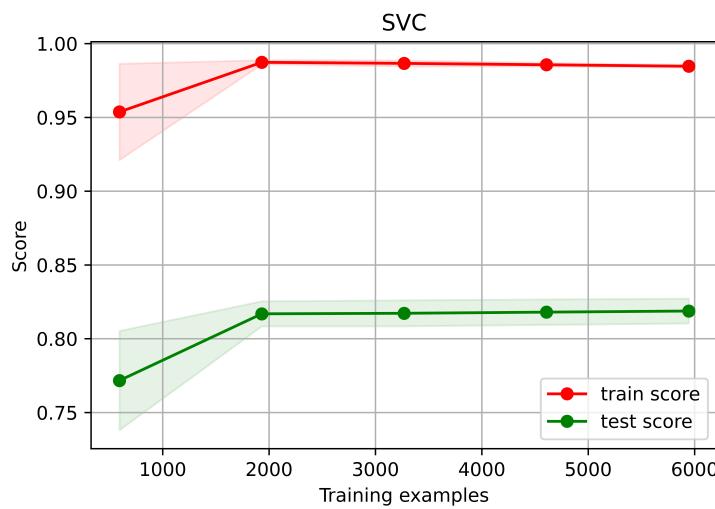


Figure 8: Support Vector Machine con gamma=0.01

Como se observa, el rendimiento y el comportamiento de los gráficos no cambia de forma considerable y esto significa que necesitamos mas datos.

Para el caso del estimador Random Forest Classifier se puede observar que el rendimiento en test aumentó según la cantidad de datos que se tienen. Sin embargo, el rendimiento en entrenamiento es deficiente, ilustrando que en este caso, el algoritmo esta memorizando los datos.

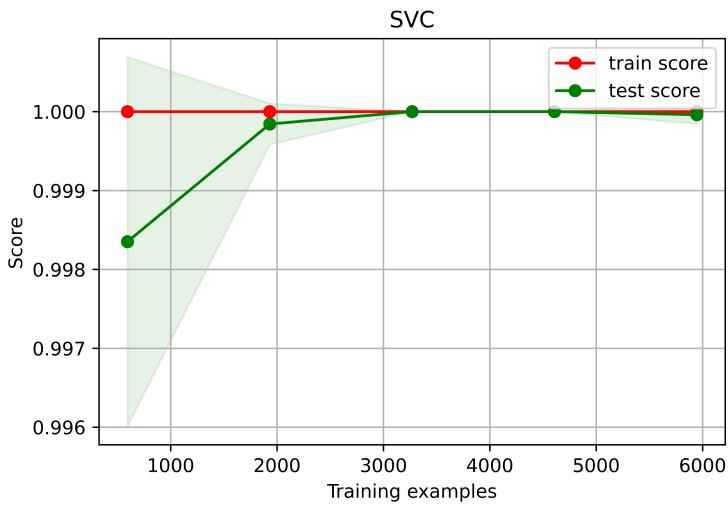


Figure 9: Random Forest Classifier

Evaluando el estimador al variar los hiperparámetros de `n_estimators` y `max-depth`.

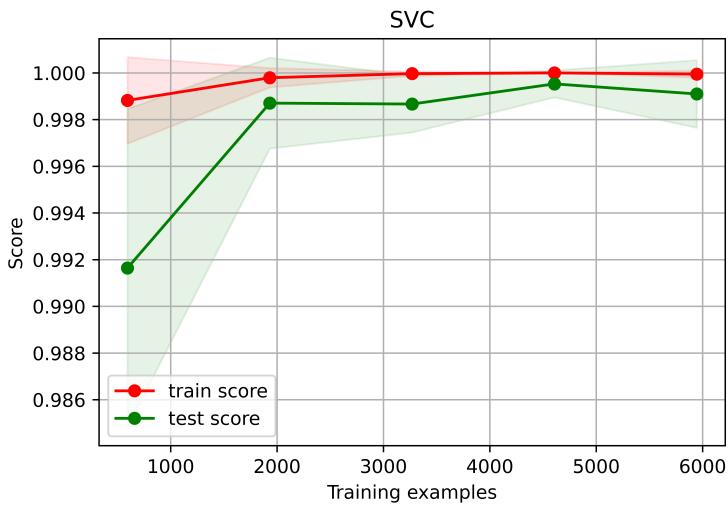


Figure 10: Random Forest Classifier

En conclusión, el mejor estimador para el modelo es el `Random Forest Classifier`

## 4 Retos y Consideraciones de Despliegue

En los viajes interestelares son muy comunes las anomalías espacio-temporales y debido a esto se desea implementar en las nuevas agencias de transporte interestelar un método de predicción que nos permita brindarle con seguridad a cada pasajero información acerca de su probabilidad de ser transportado a otra dimensión, y así, este pueda ser recuperado fácilmente.

Teniendo esto en cuenta se define el 90% como el nivel mínimo de desempeño para el despliegue en producción, ya que como en los viajes estelares es de suma importancia el bienestar y la seguri-

dad de nuestros usuarios, el factor de seguridad que se brinda debe ser el mas alto posible. Así mismo, esta certificación de seguridad en el desempeño productivo de los viajes interestelares se lograría en gran medida gracias al acompañamiento de otras empresas profesionales en seguridad y monitoreo de pasajeros en naves interestelares.

Junto a estas empresas se podrían establecer programas de seguimientos y monitoreo de la ubicación de cada uno de los pasajeros en tiempo real, para saber su ubicación constantemente y poder verificar minuto a minuto si hay alguna variabilidad espacio-temporal que ponga en peligro la vida de los pasajeros. Entre estos procesos pueden contemplarse el uso de manillas de ubicación, o ships temporales para rastrear las coordenadas de cada pasajero durante el transcurso de su viaje.

## 5 Referencias

- [1] *Spaceship Titanic | Kaggle*. (n.d.). Retrieved July 1, 2022, from <https://www.kaggle.com/competitions/spaceship-titanic/overview>