Instituto Tecnológico y de Estudios Superiores de Monterrey
Escuela de Ingeniería y Ciencias

Inteligencia artificial avanzada para la ciencia de datos I (TC3006C.102)

Momento de Retroalimentación: Reto Limpieza del Conjunto de Datos

Equipo 4:

Karla Andrea Palma Villanueva (A01754270)

Viviana Alanis Fraige (A01236316)

David Fernando Armendariz Torres (A01570813)

Alan Alberto Mota Yescas (A01753924)

Adrián Chávez Morales (A01568679)

Jose Manuel Armendáriz Mena (A01197583)

Docentes:

Alfredo Esquivel Jaramillo

Mauricio Gonzalez Soto

Frumencio Olivas Alvarez

Antonio Carlos Bento

Hugo Terashima Marín

Monterrey, Nuevo León, México. 20 de agosto de 2024

Índice

1. Introducción

La inteligencia Artifical se ha convertido en una herramienta fundamental en la industria, debido a su alto impacto en la toma de decisiones, principalmente en el ánalisis de datos. Entre sus ramas, el Machine Learning destaca por su capacidad para mejorar continuamente su desempeño al aprender de los datos, en lugar de depender únicamente de la programación explícita, esta disciplina se basa en la aplicación de modelos estadísticos que buscan hacer predicciones precisas a partir de datos históricos.

Machine Learning, o también conocido como Aprendizaje Automático, puede abordarse de diferentes maneras, y uno de sus enfoques más comunes es el Aprendizaje Supervisado. Se sabe que para realizar las predicciones se requieren de datos de entrada y de salida, por lo que para este tipo de modelos se requieren de datos etiquetados que sirven para el entrenamiento del mismo y así cumplir con su función de ejecutar las predicciones sobre los nuevos datos.

En este sentido, uno de los desafíos más comunes en la comunidad de las ciencias de Datos es el "Titanic - Machine Learning from Disaster". Dicho desafío además de aplicar las técnicas de aprendizaje automático, posee información realista mediante un análisis de datos del Titanic y permite a los algoritmos aprender de datos históricos para hacer predicciones confiables sobre eventos futuros.

2. Objetivo

El objetivo planteado corresponde a la creación de un modelo de aprendizaje automático con la capacidad de predecir la supervivencia de un individuo ante la tragedia del hundimiento del Titanic. Se busca generar un modelo de buen desempeño dada una entrada de caracterísitcas de un pasajero hipotetico.

Para cumplir este objetivo, se entrenarán multiples modelos de aprendizaje supervisado para clasificar el estatus del pasajero (sobrevivió, pereció). Los modelos serán entrenados mediante un set de datos de características de pasajeros pasados como su clase social, edad, sexo y tarifa, entre otras cosas.

Ante este objetivo, primeramente se deben preparar los datos, este es el objetivo puntual de esta primera etapa. Algo sumamente importante para poder aplicar cualquier tipo de

modelo son las actividades de limpieza y organización de datos, en las cuales, se trata de limpiar y suavizar cualquier inconsistencia que impida que el conjunto de datos sea coherente. El proceso de limpieza que se llevará a cabo se compone de manejo de datos nulos, caracteres problemáticos, inconsitencias en los datos, etcétera.

Este proceso es muy importante porque establecerá las bases para un análisis más extenso, permitiendo que los modelos de aprendizaje automático trabajen con un dataset ordenado y correctamente estructurado.de tal forma que el modelo sea más preciso y su rendimiento de resultados sea eficiente y viable.

3. Estructura de Datos

Los datos de los pasajeros del Titanic están divididos en dos conjuntos distintos: un conjunto de entrenamiento y uno de prueba. El conjunto de entrenamiento se utiliza para ajustar el modelo de aprendizaje, mientras que el conjunto de prueba se emplea para evaluar la precisión del modelo con datos desconocidos. El conjunto de entrenamiento contiene 891 instancias, mientras que el de prueba tiene 418. A continuación se presenta un diccionario de datos que describe las variables o características de ambos conjuntos, provenientes de la misma fuente en Kaggle (Cukierski, 2012).

Variable	Tipo	Especificaciones	Descripción
Survival	int	0 = No, 1 = Sí	Supervivencia del pasajero
pclass	int	1 = 1ª, 2 = 2ª, 3 = 3ª	Estatus socioeconomico
Name	string	-	Nombre y título del pasajero
sex	string	-	Sexo del pasajero
Age	number	La edad puede ser fraccionaria si es menor de 1 año; se usa la forma xx.5 cuando la edad se estima.	Edad del pasajero
sibsp	int	-	Número de hermanos o cónyuges a bordo
parch	int	Si los niños viajaron solo con una niñera el valor es cero en esos casos.	Número de padres o hijos a bordo
ticket	varchar	-	Número de ticket del pasajero
fare	float	-	Costo del ticket
cabin	number	-	Número de cabina
embarked	char	C=Cherbourg, Q=Queenstown, S=Southampton	Puerto de embarque

4. Limpieza de datos

Primeramente se llevó a cabo un analisis exploratorio de los datos, mediante el cual se realizó un descarte preeliminar de las variables o features que aparentan no ser significativas para la predicción del modelo (sibsp, ticket, fare, cabin, embarked). Mas adelante se detalle el motivo tras la decisión de trabajar sobre los otros features.

Una vez completado esto, se buscó lidíar con los valores ausentes del conjunto de datos de entrenamiento, en este caso, .^{A}ge.^{es} la unica varaible que presenta datos ausentes. Para lidiar con ellos, se decidió generar un feature adicional "Titles", correspondiente a los titulos

de cada pasajero extraidos de "Name". La principal intención de dicha generación es la de utilizar información generada de subgrupos de datos a base de los titulos para completar los valores faltantes. A continuación se muestran los títulos y el proceso de mamejo de nulos.

Titles	**Count**
Mr	517
Miss	182
Mrs	125
Master	40
Dr	7
Rev	6
Mlle	2
Major	2
Col	2
the Countess	1
Capt	1
Ms	1
Sir	1
Lady	1
Mme	1
Don	1
Jonkheer	1

Cuadro 1: Títulos y cuenta

Se identificaron los "Titlesçon una presencia de valores nulos en la variable de .ªge". Los títulos que presentaron un número significativo de datos faltantes en la edad fueron: "Master", "Mr", "Misszz "Mrs". De dichos títulos, se graficó la distribución de edades en gráficos de barras y caja y bigote.

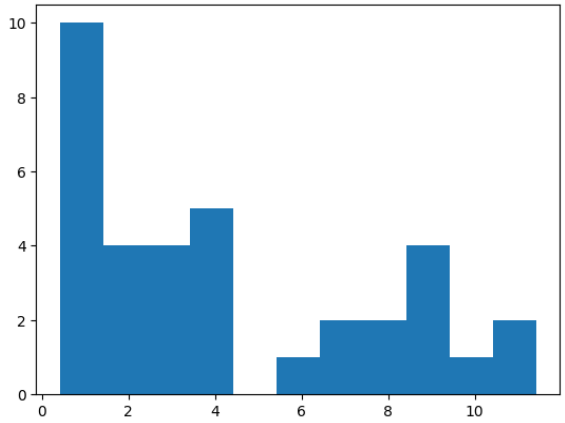

Figura 1: Distribución edades, título: Master

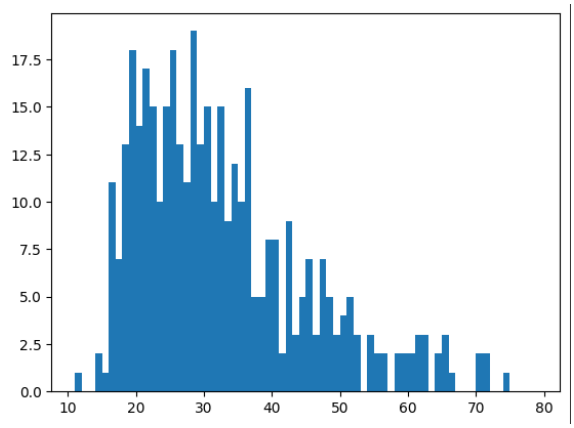

Figura 2: Distribución edades, título: Mr

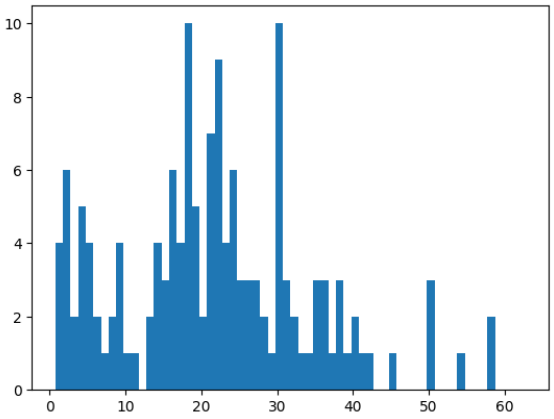

Figura 3: Distribución edades, título: Miss

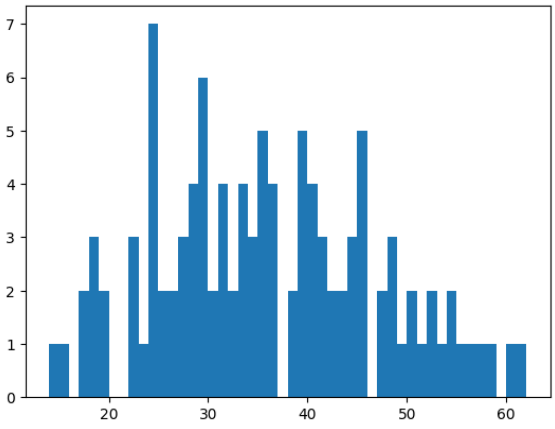

Figura 4: Distribución edades, título: Mrs

Tras realizar la visualización de los datos correspondientes a cada título, se decidió rellenar los valores faltantes en la variable de edad utilizando el una medida de tencencia central de la misma asociado a cada título respectivo. Los registros nulos en edad de los titulos con presencia de edades atípicas (Mr, Miss) se remplazaron con la media de los datos del respectivo titulo. En cambio para los que carecen de datos atípicos (Master, Mrs), se utiló la media. Esto con la intención de afectar lo mínimo posible la distribución inata de las edades

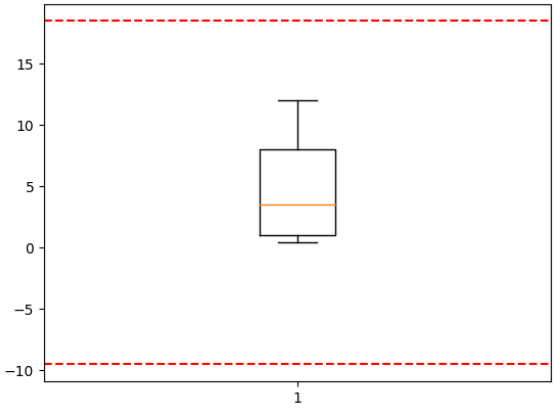

Figura 5: Boxplot edades, título: Master

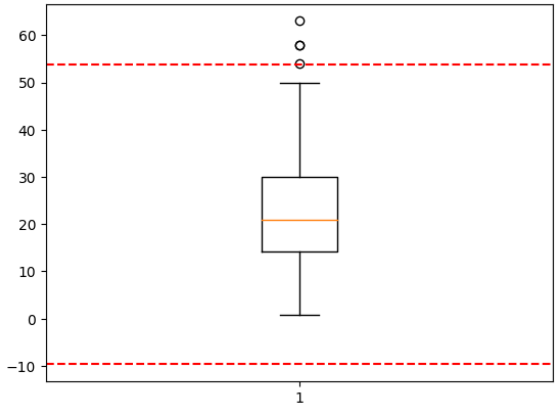

Figura 6: Boxplot edades, título: Mr

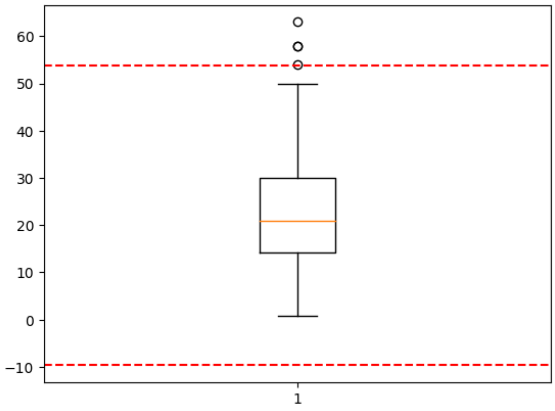

Figura 7: Boxplot edades, título: Miss

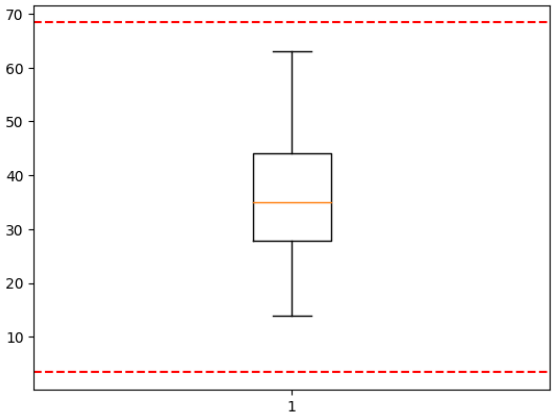

Figura 8: Boxplot edades, título: Mrs

por título.

También, en el feature de 'sex', se cambió el tipo de dato a enteros, cambiando los valores 'male' por 0, y 'female' por 1. Permitiendonos manipular más facilmente estos valores.

5. Estructura final del dataset

Como resultado del análisis, se seleccionaron seis características que se considera son clave para realizar un estudio profundo de las personas que viajaron y sobrevivieron en el Titanic. Además, se centró el análisis en el objetivo principal: determinar si la persona sobrevivió o no. Las variables seleccionadas son las siguientes:

- **PassengerId**: Es el identificador único para cada pasajero.
- **Pclass**: Es relevante para saber la clase social o estatus económico del pasajero.
- **Sex**: El género del pasajero es relevante para futuros analisis sobre hombres y mujeres que abordaron.
- **Age**: La edad del pasajero, al igual que el sexo, es de importacia para analisis sobre los diferentes rangos de edades que abordaron y sobrevivieron.
- **Parch**: Indica el número de padres y/o hijos en el Titanic. De esta manera se asigna un valor y vemos si la persona viajo con familiares o, sin ellos.
- **Titles**: El título o distinción social del pasajero, esto es relevante en las clases sociales y los diferentes títulos de personas que abordaron en aquella época.
- **Survived**: Es el indicador de supervivencia.

Con estas variables, se buscó llevar a cabo un análisis mas detallado, para asi identificar posibles patrones y factores que influyeron en la supervivencia de los pasajeros, lo que permitirá generar conclusiones significativas sobre los sucesos del Titanic.

5.1. Pre-procedamiento de datos

Para la variable *'sex'*, se realizó una codificación binaria, asignando el valor 0 a 'male' y el valor 1 a 'female'.

En el análisis de los títulos, se decidió agrupar aquellos que representaban la misma clase social en una sola categoría. Esto se hizo con el propósito de reducir la cantidad de datos

únicos y facilitar el análisis, manteniendo asi la relevancia de la información sobre el estatus social de los pasajeros. La agrupación se realizó de la siguiente manera:

- **Mr**: Se mantuvo como una categoría individual.
- **Miss**: Agrupó títulos equivalentes como *Miss*, *Mlle* (Mademoiselle) y *Ms*.
- **Mrs**: Incluye únicamente a las pasajeras con el título *Mrs*.
- **Master**: Corresponde al título *Master*, utilizado para varones jóvenes.
- **Social Workers**: Agrupamos a pasajeros con títulos como *Dr* y *Rev* (Reverendo), aunque no sean trabajadores sociales en un sentido moderno.
- **Military**: Incluye a pasajeros con títulos militares como *Major*, *Col* (Coronel) y *Capt* (Capitán).
- **Nobility**: Los títulos de nobleza como *Sir*, *Lady*, *Mme* (Madame), *Don* y *Jonkheer* fueron agrupados en una sola categoría.

Después de realizar esta agrupación, se aplicó una codificación *one-hot*, lo cual generó columnas binarias para cada grupo de títulos. De esta forma, un valor de `True` indica que el pasajero pertenece a esa categoría, tal y como se ve en el cuadro 2. Finalmente, se eliminó la columna original de *Titles* para evitar redundancias y trabajar únicamente con las nuevas columnas generadas. Esto permitirá analizar los datos de manera más eficiente y estructurada.

PassengerId	Pclass	Sex	Age	Parch	Mr	Miss	Mrs	Master	Social Worker	Military	Nobility
1	3	0	22.0	0	True	False	False	False	False	False	False
2	1	1	38.0	0	False	False	True	False	False	False	False
3	3	1	26.0	0	False	True	False	False	False	False	False
4	1	1	35.0	0	False	False	True	False	False	False	False
5	3	0	35.0	0	True	False	False	False	False	False	False

Cuadro 2: Clasificación de títulos por clase social

Referencias

Cukierski, W. (2012). *Titanic - machine learning from disaster.* Kaggle. Descargado de `https://kaggle.com/competitions/titanic`