Instituto Tecnológico y de Estudios Superiores de Monterrey

Inteligencia artificial avanzada para la ciencia de datos I (Gpo 101)

Equipo 4

Momento de Retroalimentación: Reto Selección, configuración y entrenamiento del modelo

Integrantes:

Eliezer Cavazos Rochin A00835194

Facundo Colasurdo Caldironi A01198015

Saul Francisco Vázquez del Río A01198261

José Carlos Sänchez Gómez A01174050

Campus Monterrey **Fecha: 26/8/2024**

El desafío "Titanic - Machine Learning from Disaster" es un problema el cual estamos tratando de superar, esta, busca predecir la supervivencia de los pasajeros a bordo del Titanic. Este desafío implica construir un modelo de clasificación binaria capaz de predecir si un pasajero específico sobrevivió al hundimiento del Titanic, utilizando un conjunto de datos que incluye diversas características de los pasajeros, como su edad, sexo, clase en el barco, y otras.

El problema se adentra en el aprendizaje supervisado, donde la meta es, entrenar un modelo a partir de datos etiquetados para que pueda hacer predicciones sobre datos no vistos previamente. Dado que el objetivo es predecir una de dos posibles categorías de supervivencia de los pasajeros, se requiere de un modelo de clasificación binaria.

Para abordar este problema, es crucial comprender la naturaleza de los datos disponibles, los cuales se dividen en variables categóricas y numéricas.

Estas variables incluyen información crítica como el sexo del pasajero, su edad, la clase en la que viajaba, el número de familiares a bordo, entre otras, tal como se puede ver en la siguiente tabla:

Variable	Definición	Tipo de Datos	Modelos Compatibles
survival	Supervivencia	Categórico (Binario)	Regresión logística, árboles de decisión, SVM, redes neuronales, KNN, Naive Bayes, Random Forest
pclass	Clase de ticket	Categórico (Ordinal)	Regresión logística, árboles de decisión, Random Forest, SVM, KNN
sex	Sexo	Categórico (Nominal)	Regresión logística, árboles de decisión, SVM, redes neuronales, KNN, Naive Bayes, Random Forest
age	Edad en años	Numérico (Continua)	Regresión lineal, regresión logística, árboles de decisión, SVM, redes neuronales, KNN, Random Forest
subsp	Número de hermanos/esposos a bordo	Numérico (Discreta)	Regresión lineal, regresión logística, árboles de decisión, SVM, redes neuronales, KNN, Random Forest

parch	Número de padres/hijos a bordo	Numérico (Discreta)	Regresión lineal, regresión logística, árboles de decisión, SVM, redes neuronales, KNN, Random Forest
ticket	Número de ticket	Categórico (Nominal)	Generalmente no se utiliza como predictor directo en modelos, se puede transformar o excluir
fare	Tarifa del pasajero	Numérico (Continua)	Regresión lineal, regresión logística, árboles de decisión, SVM, redes neuronales, KNN, Random Forest
cabin	Número de cabina	Categórico (Nominal)	Generalmente no se utiliza como predictor directo en modelos, se puede transformar o excluir
embarked	Puerto de embarque	Categórico (Nominal)	Regresión logística, árboles de decisión, SVM, redes neuronales, KNN, Naive Bayes, Random Forest

Después de una investigación exhaustiva, identificamos el modelo de Random Forest como una de las opciones más robustas y efectivas para esta tarea. Random Forest es un modelo de ensamble que combina múltiples árboles de decisión, lo que permite manejar la complejidad de los datos y capturar las interacciones entre las características de manera más precisa.

Sin embargo, al ser conscientes de la necesidad de explorar distintos enfoques para la solución de un problema, decidimos implementar un modelo de clasificación. Los modelos de clasificación, por su naturaleza, son especialmente adecuados para problemas de predicción binaria, como el que se presenta en este reto, donde el objetivo es determinar si un pasajero sobrevivió.

Aunque el modelo de Random Forest es altamente prometedor, la implementación del modelo de clasificación nos proporciona una perspectiva adicional, asegurando que nuestras decisiones de modelado estén bien fundamentadas y que obtengamos el máximo rendimiento posible en la predicción de la supervivencia de los pasajeros del Titanic.

Las decisiones que tomamos durante la selección y el entrenamiento del modelo fueron:

a. Primero empezamos decidiendo el modelo de Clasificación ya que ocupamos predecir la supervivencia en base a SI o NO.
b. Para generar una predicción sencilla empezamos llenando los valores de edad vacíos en 0 para poder al menos generar una predicción de la

supervivencia de los pasajeros en base a las columnas de Age, Class y Sex.

c. Empezamos a buscar opciones para llenar los datos del campo de edad faltantes, para esto implementamos 3 métodos para intentar predecir la edad de las personas. Implementamos varias regresiones lineales para identificar con cuáles campos podríamos generar una mejor predicción de las edades, el segundo método fue la implementación de parámetros para intentar identificar las edades de los pasajeros en base a ciertos valores en base a varios campos como PClass, Sibsp y Parch, por último también intentamos llenar los datos en base a promedio de edades. De estos 3 métodos en la primera versión que teníamos no pudimos identificar un buen método para predecir las edades así que al final implementamos el último método de llenar las edades en base a promedio.
d. Volvimos a generar nuestra predicción ahora llenando los datos de edad por promedio y volvimos a mandarlo a Kaggle donde mejoró un poco el resultado
e. Volvimos a mejorar nuestros parámetros para intentar llenar las edades con nuestro Segundo Método y estamos comprobando que si mejoro mucho la predicción de edades con nuestras pruebas, así que vamos a utilizar este método para llenar los campos faltantes de edad
f. Estamos investigando la implementación del modelo de Random Forest ya que después de platicar con varios equipos pudimos identificar, que para el caso de este proyecto es el mejor modelo para generar la solución.

Link de los script en GitHub:
https://github.com/AinzWatch/RetoIA