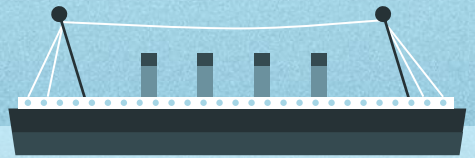
— Titanic data set analysis—

lart proj2

Up202108689 – António Azevedo
Up202108794 – José Martins
Up202108776 – Tomás Martins



— The data set features —

1. Passenger class;
2. Survived;
3. Home destination;
4. Name;
5. Sex;
6. Age;
7. Sibling/Spouse;
8. Parent/Children;
9. Ticket;
10. Fare;
11. Cabin;
12. Embarked;
13. Boat;
14. Body;

— The analysis—

Goal

Predict the survival of a Titanic passenger using the features mentioned before. This fall into a binary classification task, where we check whether a passenger survived (1) or not (0).

NaN Value

Some features contain a substantial number of NaN values, which can decrease their usefulness in predicting passenger survival. Given their limited contribution, these attributes will be removed from consideration.

— Algorithms—

Logistic Regression

Estimates the probability of an event occurring

Neural Networks

Uses strategies that replicate how real neurons collaborate to identify phenomena, weight options, and reach conclusions.

Random Forest

An algorithm that uses different decision trees in order to improve the overall accuracy

Decision Tress

A representation of the features that may lead to our goal in a tree-like structure

— Tools—

Pandas

“Pandas is a Python library used for (…) analyzing, cleaning, exploring, and manipulating data.”

Seaborn

“Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.”

Scikit-Learn

“Scikit-learn is a library in Python that provides many unsupervised and supervised learning algorithms.”

Matplot

“Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.”

TensorFlow

“TensorFlow makes it easy to create ML models that can run in any environment.”

— Work Done and Sources—

LINKS

- Titanic Dataset – Kaggle;
- What is a Linear Regression – IBM;
- What is a Decision Tree – IBM;
- Random Forest – GeeksForGeeks;
- What is a Neural Network – IBM;
- Pandas – W3School;
- Seaborn – Seaborn;
- Scikit-learn – Codecademy;
- Matplot – Matplotlib;
- Tensorflow – Tensorflow;

WORK ALREADY DONE

- NaN values cleaned;
- Necessary features chosen;
- Simple analys of the data provided;