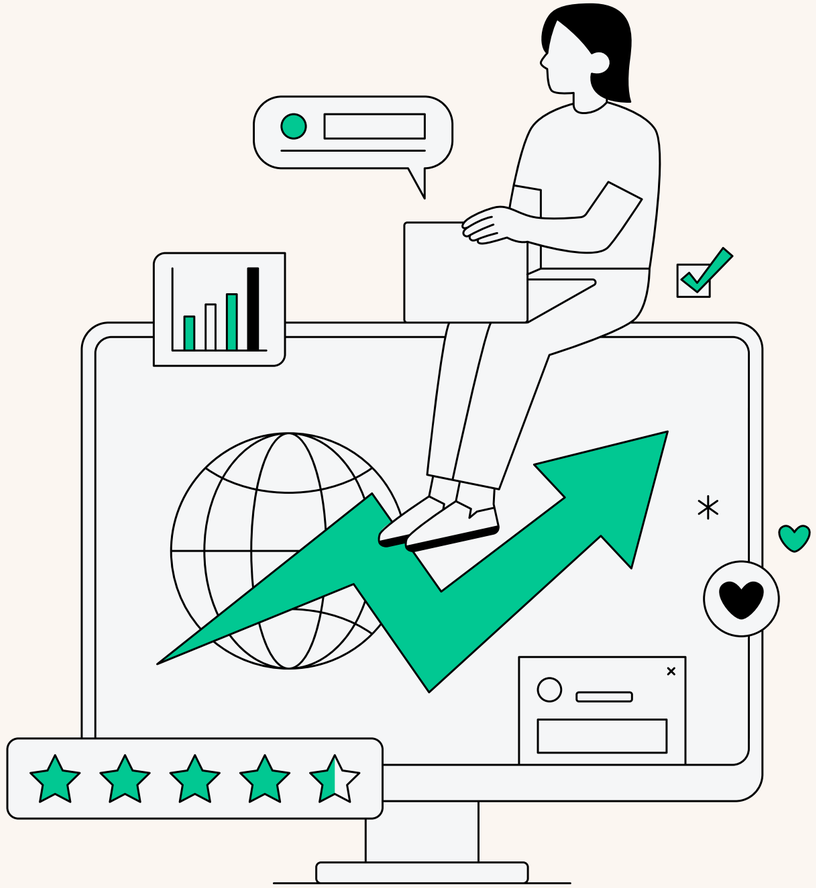
by Leonardo-Gabriel Marcu

# Titanic Survival Prediction

## using Logistic Regression



www.github.com/isntgabi

# Introduction

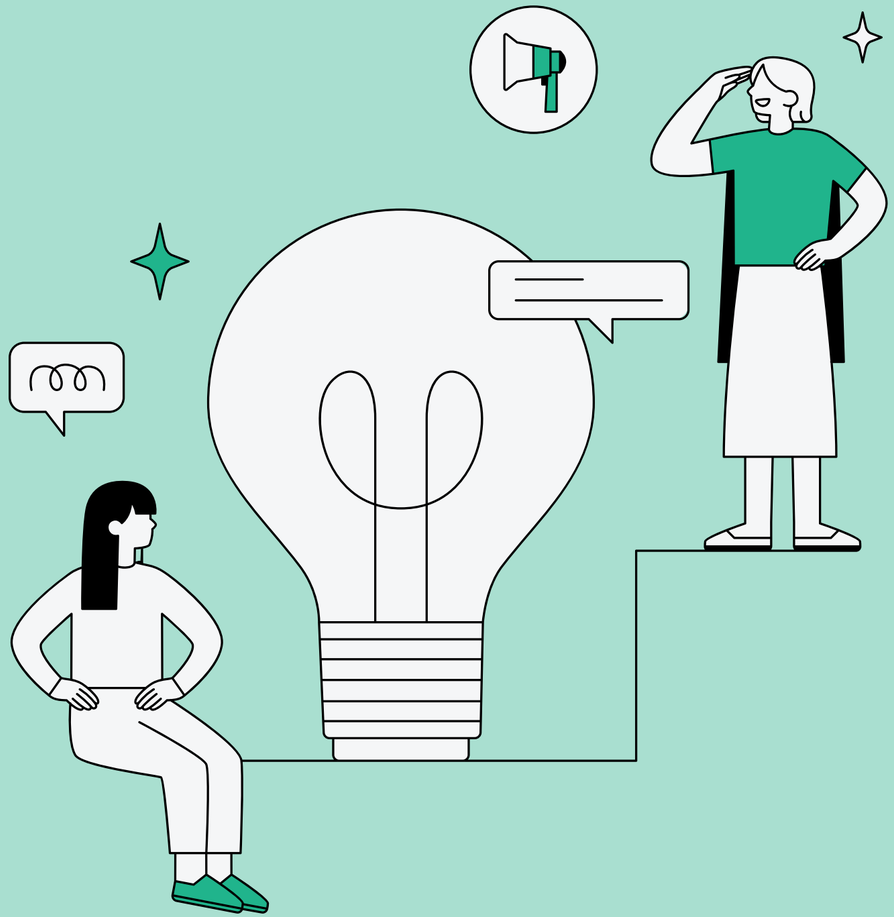
This project explores the use of **logistic regression** to predict the survival of passengers on the Titanic.

Logistic regression is a statistical model used for binary classification problems — situations where the output variable can take one of two possible outcomes (e.g., survived or not).

It models the probability that a given input point belongs to a particular class.
The dataset used is the well-known 'Titanic - Machine Learning from Disaster' dataset from Kaggle.

It provides information about the passengers such as age, sex, class, family aboard, fare paid, etc.

Using these features, we aim to build a predictive model that estimates the chances of survival.



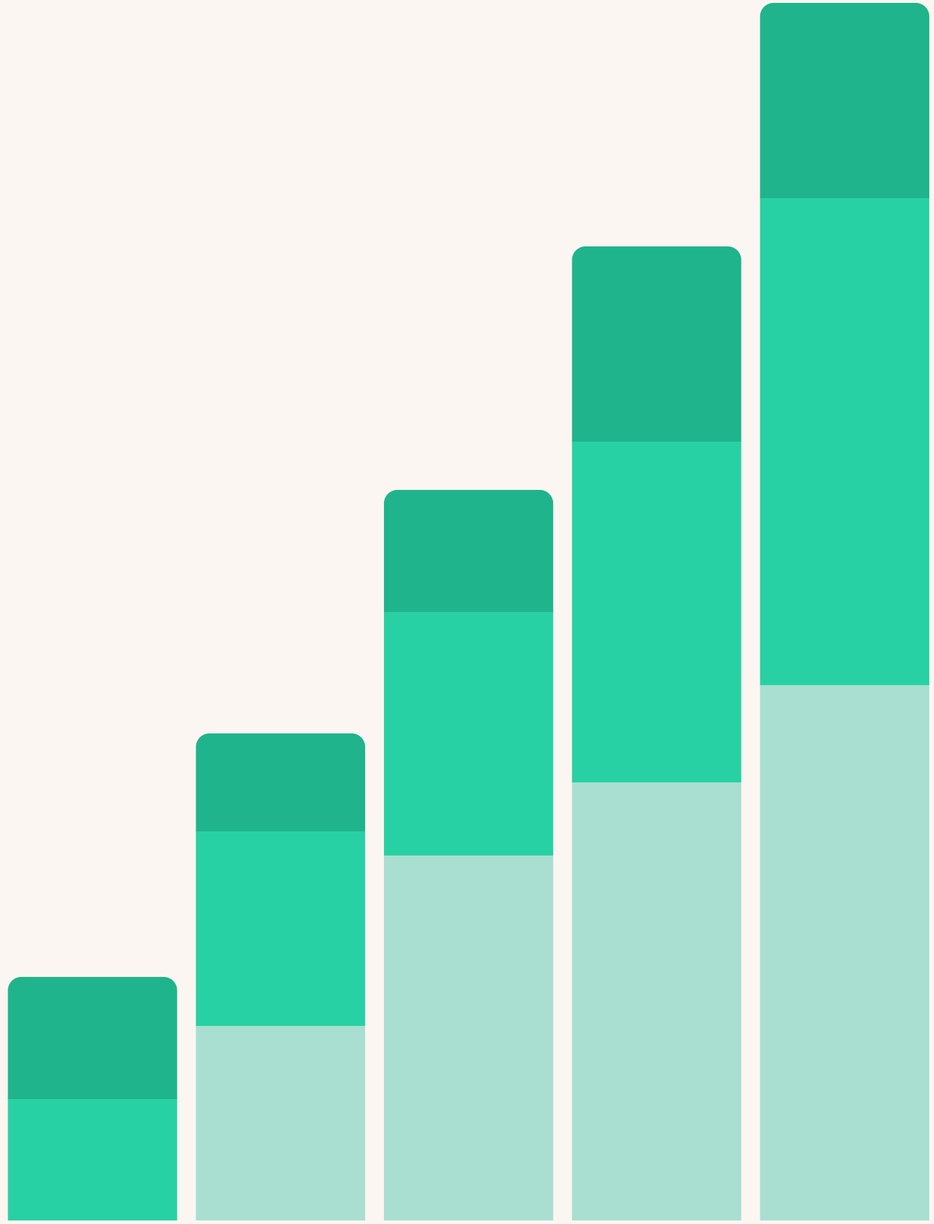
# Dataset description

The dataset consists of 891 entries (passengers) with the following relevant columns:

- Survived: 0 = No, 1 = Yes (target variable)
- Pclass: Ticket class (1 = 1st, 2 = 2nd, 3 = 3rd)
- Sex: Gender of the passager
- Age: Age in years
- SibSp: Number of siblings/spouses aboard
- Parch: Number of parents/children aboard
- Fare: Ticket fare
- Embarked: Port of embarkation (C = Cherbourg, Q=Queenstown, S = Southampton)

The dataset underwent preprocessing steps:

- Missing values were filled using mean for numerical and mode for categorical columns
- Categorical columns like 'Sex' and 'Embarked' were encoded using label encoding and one-hot encoding respectively.
- Features such as 'Cabin' and 'Ticket' were dropped due to sparsity or irrelevance.

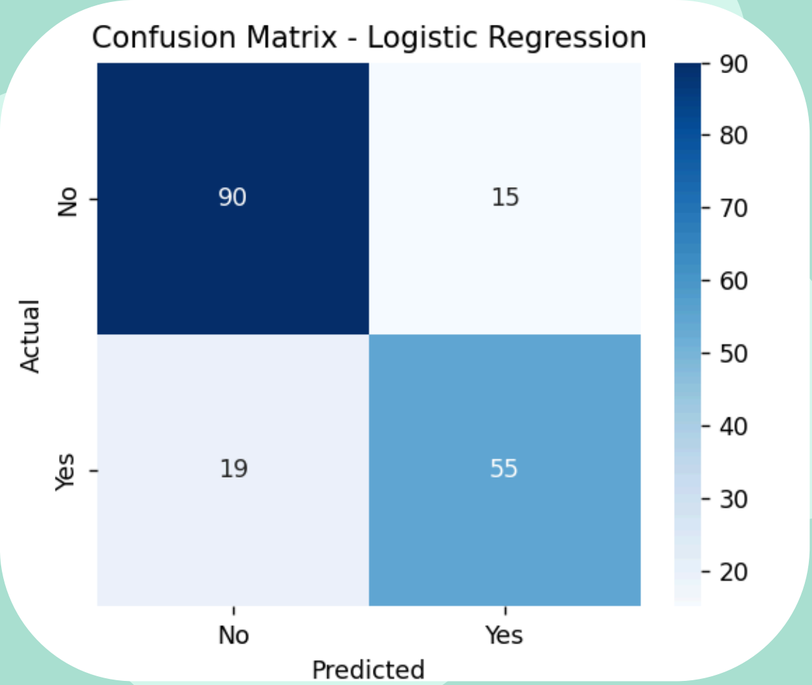


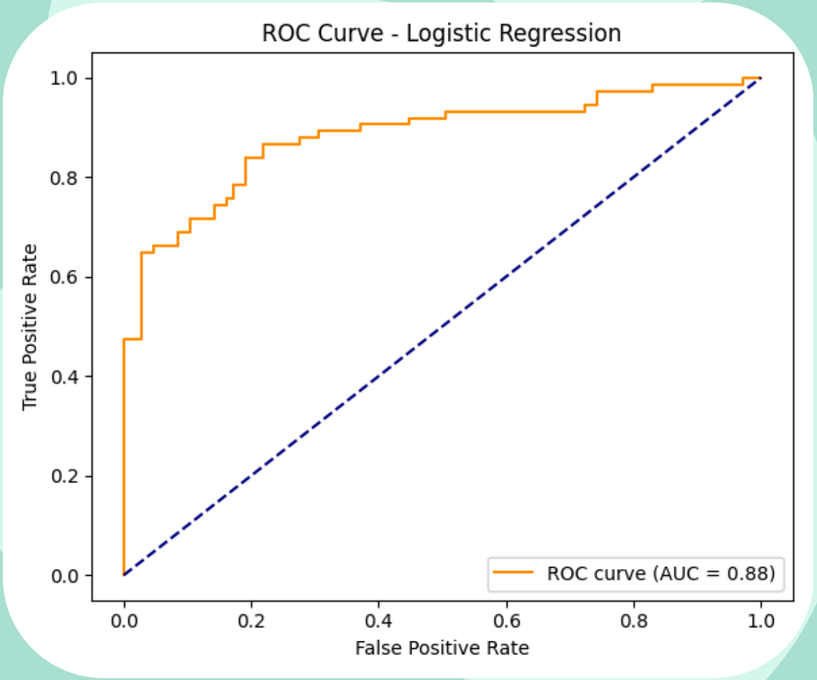
# Analysis and interpretations

I trained a logistic regression model with an 80/20 train-test split. The model was also validated using 5-fold cross-validation, yielding an average accuracy of 0.789. Below are the evaluation results on the test set:

Confusion Matrix: Shows that 90 true negatives and 55 true positives were correctly predicted, while 15 false positives and 19 false negatives occurred.

ROC Curve: The model achieved an AUC score of 0.8824, indicating high discriminative power. The closer the curve is to the top-left corner, the better the performance.



# Analysis and interpretations

## Precision, recall and F1-score by class

- Class 0 (did not survive): Precision = 82.57% , Recall = 85.71%, F1-score = 84.11%
- Class 1 (survived): Precision = 78.57%, Recall = 74.32%, F1-score = 76.38%

- *Overall accuracy: 81.01%*

- Average cross-validation accuracy: 78.90%

## Feature importance based on model coefficients (top influencing features):

- Sex: +2.59 → being female increases chances of survival
- Pclass: -0.94 → higher class (1st) increases chances
- SibSp, Parch: negative → large families had lower survival chances
- Fare: positive → more expensive ticket increased chances
- Embarked_C: slightly positive

# Conclusions

The logistic regression model performed well, achieving over 81% accuracy and an AUC of 0.88.

The most significant factor in predicting survival was gender — females had a higher survival rate. Other influential features include passenger class, fare paid, and port of embarkation.

Possible future improvements:

- Feature engineering (e.g., titles from names, family size)
- Hyperparameter tuning
- Comparing with more complex models (e.g., Random Forest, SVM)

Presented by Leonardo-Gabriel Marcu

# Thank you very much!

www.github.com/isntgabi