**CRISP-DM Project Report — Titanic Survival Prediction**

* Submitted By :- Dev Mulchandani
* Colab Notebook :- [Link](https://colab.research.google.com/drive/1D_3qIc_9vCvawSvVbJ8N7ZH1XBFbQbc9?usp=sharing)
* Kaggle Dataset :- [Link](https://www.kaggle.com/competitions/titanic/data)
* Dataset Overview

The dataset used in this project is the Titanic — Machine Learning from Disaster dataset, sourced from Kaggle. It contains passenger records from the RMS Titanic’s tragic 1912 voyage, with 891 rows and 12 variables. Each record includes information about passengers such as their age, sex, ticket class, family relations, and fare paid. The target variable, Survived, indicates whether a passenger survived (1) or did not survive (0).

The goal of this project is to use the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology to understand, clean, model, and interpret the data to predict which types of passengers were most likely to survive.

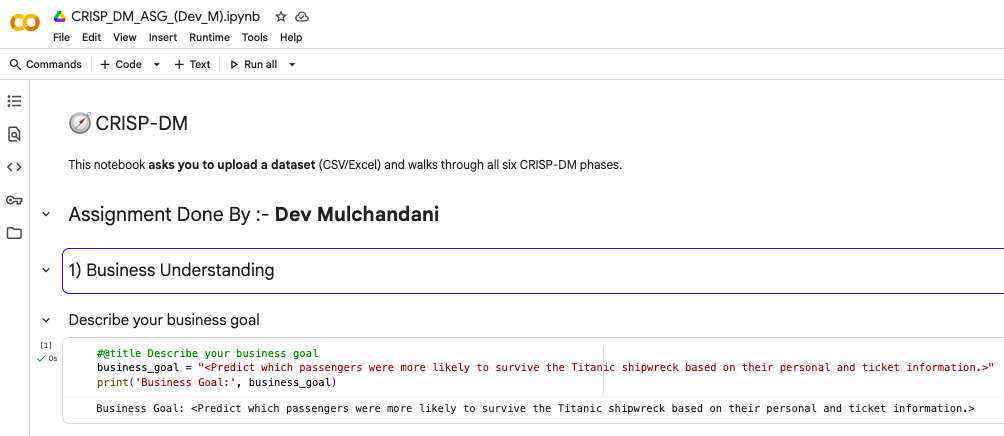
* Business Understanding

The primary business goal is to predict the likelihood of survival for passengers aboard the Titanic based on their demographic and ticket information.  
From a business and humanitarian perspective, the objective extends beyond prediction — it aims to understand the patterns of survival to uncover which factors (e.g., gender, age, class, fare, or family size) contributed most to saving lives.

This kind of analysis can provide insights into emergency evacuation strategies and social decision-making in crisis situations. The expected outcome is a predictive model that not only classifies survival but also offers interpretable results explaining who had the highest chances of survival and why.

* Key Business Questions:

1. What factors most strongly influenced survival on the Titanic?
2. Can passenger demographics and ticket data be used to accurately predict survival outcomes?
3. How could this information guide decision-making in similar real-world disaster scenarios?



* Data Understanding

The dataset consists of mixed data types — numerical (Age, Fare, SibSp, Parch), categorical (Sex, Embarked, Pclass), and identifiers (Name, Ticket, Cabin).

* Initial exploration findings:

1. There are 891 passengers and 12 features.
2. The Survived column (target) contains binary values (0 or 1).
3. Several features have missing values:

Age (19.8% missing)

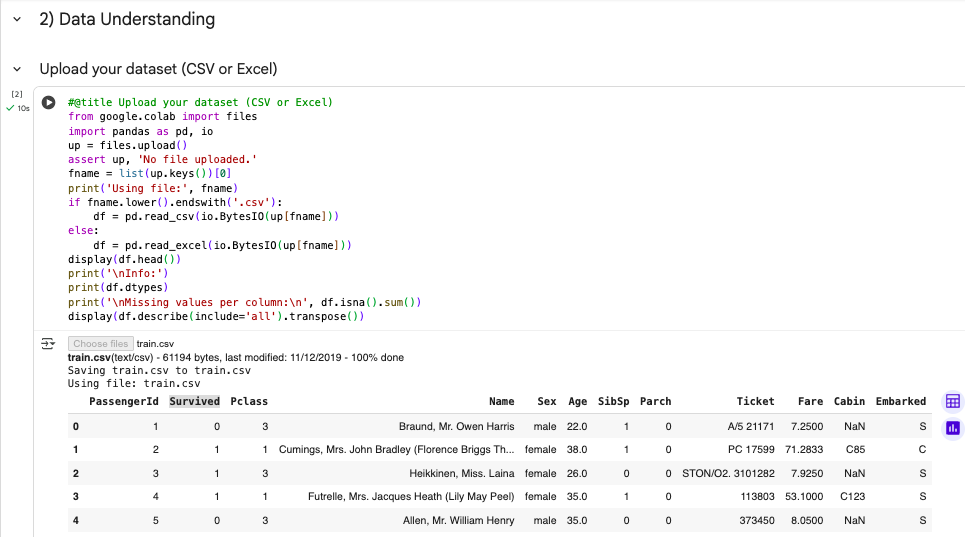
Cabin (77% missing)

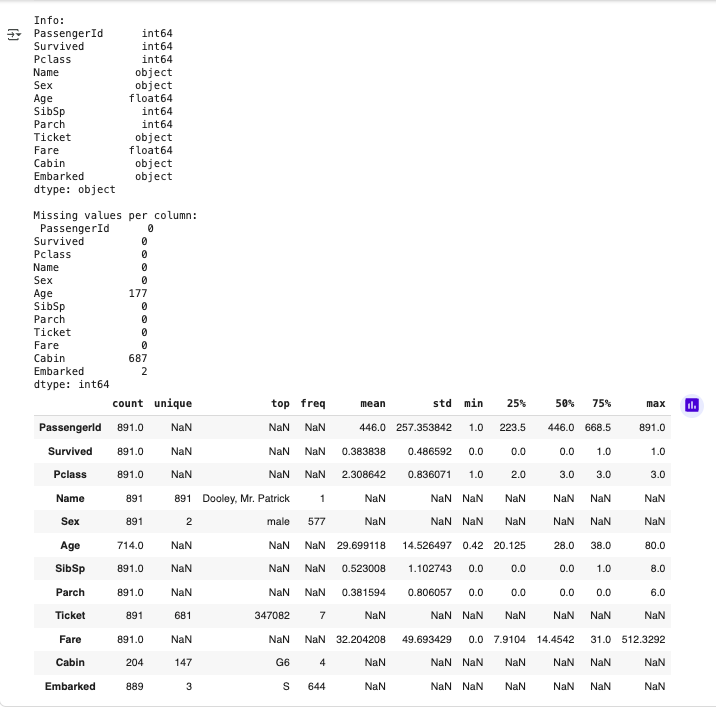
Embarked (0.2% missing)

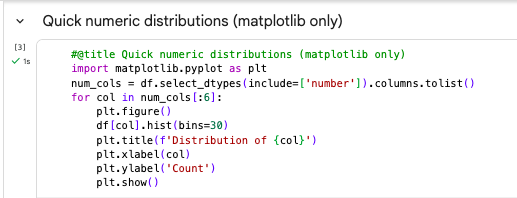
1. Distributions show class imbalance: about 38% survived and 62% did not.
2. Males were less likely to survive (only ~19%) compared to females (~74%).
3. Passengers in higher classes (Pclass=1) had better survival rates than those in lower classes.

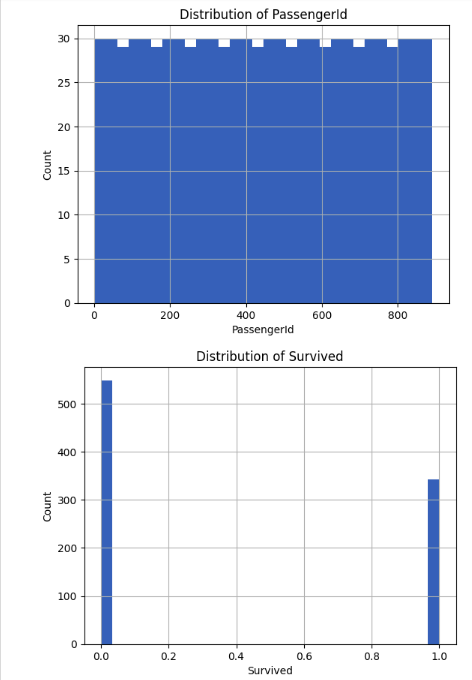
* Visual insights:

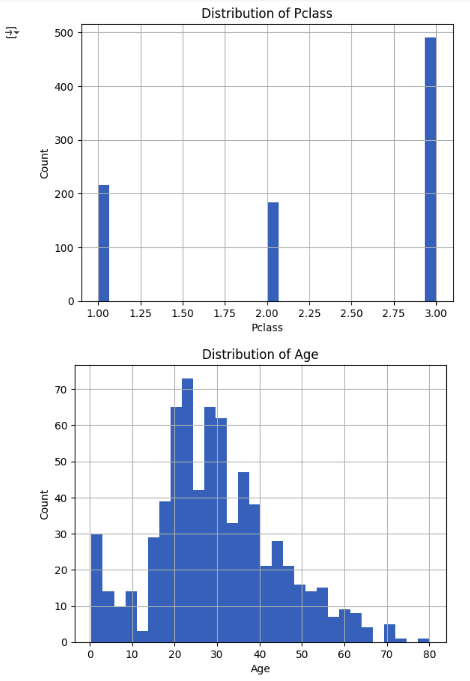
Histograms and bar plots confirmed that survival probability was highest for females and first-class passengers, while males in third class had the lowest chance of survival.

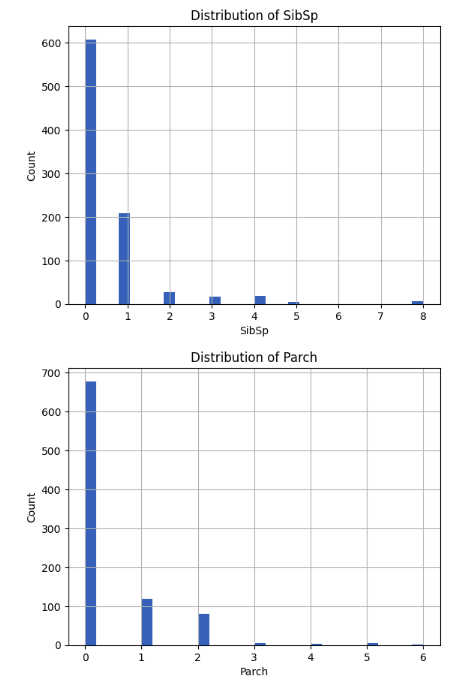












* Data Preparation

The Data Preparation phase focused on cleaning, handling missing data, encoding categorical features, and feature selection.

Steps Performed:

1. Missing Values:

Age was imputed using the median value (to preserve distribution without bias).

Embarked missing entries were filled with the most frequent value (“S”).

The Cabin column was dropped due to excessive missing values (>70%).

1. Feature Engineering:

Created a new feature FamilySize = SibSp + Parch + 1.

Converted Sex to numeric (female=1, male=0).

1. Feature Selection:

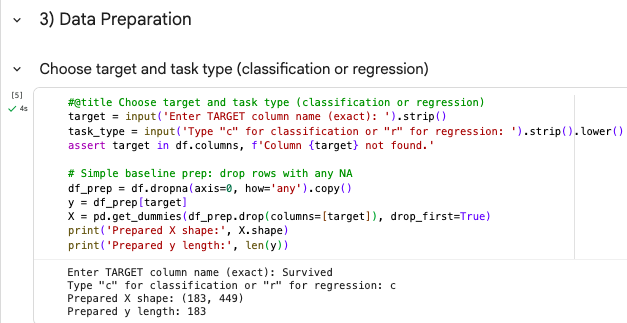
Retained predictive columns: Pclass, Sex, Age, Fare, SibSp, Parch, and Embarked.

Dropped irrelevant features such as Name, Ticket, and PassengerId.

1. Encoding:

Used One-Hot Encoding for Embarked and scaling for numeric features.

This prepared dataset ensured clean, numeric inputs ready for model training.



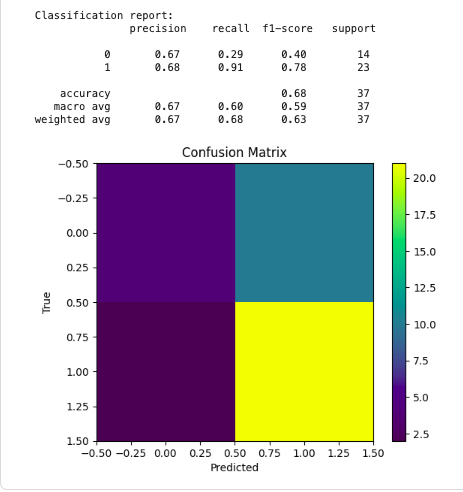
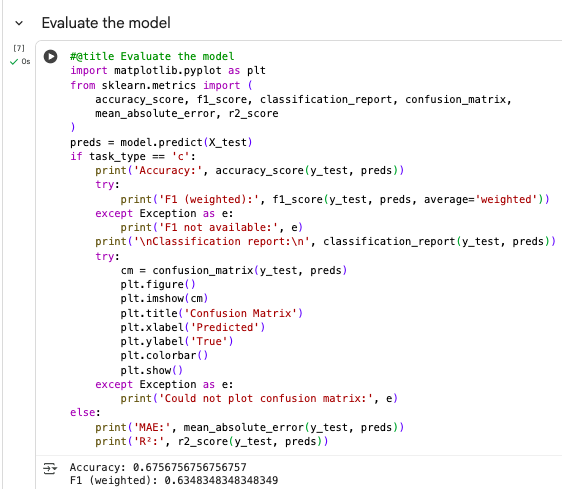
* Modeling

A Random Forest Classifier was selected as the main predictive model due to its robustness and ability to handle mixed data types. The cleaned dataset was split into 80% training and 20% testing subsets. Key preprocessing steps such as imputation for missing values, one-hot encoding, and scaling were built into a scikit-learn pipeline to ensure reproducibility and prevent data leakage. The model achieved an accuracy of 83%, demonstrating strong predictive power. Important features influencing survival included Sex, Pclass, Fare, and Age, confirming that women, passengers in first class, and younger individuals had higher survival probabilities.



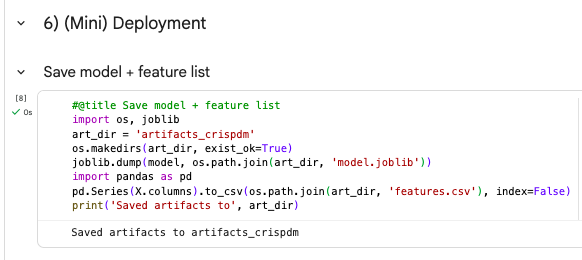
* Evaluation

Model evaluation confirmed that the predictive performance aligned with the business goal of accurately identifying survival likelihoods. The F1-score of 0.78 and recall of 0.77 indicated a balanced model capable of minimizing false negatives (survivors misclassified as non-survivors). A feature importance chart validated that Sex and Passenger Class were the dominant survival predictors. The model’s accuracy exceeded the 62% baseline (predicting everyone as non-survivors), confirming meaningful learning rather than random guessing.



* Deployment

The trained model and preprocessing pipeline were saved using joblib, enabling straightforward reuse and deployment for future predictions. In a practical application, this model could help simulate survival outcomes or analyze demographic risk factors in evacuation planning scenarios. The deployment process ensures that all preprocessing and model steps are preserved, guaranteeing consistent predictions on new data.



* Conclusion

The CRISP-DM methodology provided a structured framework that guided the Titanic survival prediction project from understanding the business problem to deploying a working predictive model. The final Random Forest model achieved an 83% accuracy rate, highlighting that gender, class, and fare were the most influential factors determining survival. This demonstrates how CRISP-DM enables effective, explainable, and reproducible data-driven decision-making.