Contents

[Data Collection And Preprocessing 1](#_Toc176772975)

[Abou the Titanic dataset 1](#_Toc176772976)

[Interpreting the dataset 2](#_Toc176772977)

[Removing columns that don’t add value 2](#_Toc176772978)

[Imputing missing values 2](#_Toc176772979)

[Scaling 3](#_Toc176772980)

[Feature Engineering 3](#_Toc176772981)

[Auto EDA using AutoViz 3](#_Toc176772982)

[Summary of the dataframe 4](#_Toc176772983)

[Auto EDA using SweetViz 5](#_Toc176772984)

[Summary of the dataframe 5](#_Toc176772985)

[Summary of each feature 6](#_Toc176772986)

[Auto EDA using missing no 7](#_Toc176772987)

# Data Collection And Preprocessing

## Abou the Titanic dataset

The Titanic dataset is a well-known dataset used in machine learning for classification tasks. The goal is to predict whether a passenger survived or not based on various features. It comes from the passenger manifest of the Titanic ship and includes several features (columns) with labels corresponding to survival status.

Features:

1. PassengerId: Unique identifier for each passenger (not particularly useful for modeling).

2. Survived: Target label (0 = No, 1 = Yes) indicating whether the passenger survived.

3. Pclass: Ticket class (1st = 1, 2nd = 2, 3rd = 3), representing the socioeconomic status of the passengers.

4. Name: The full name of the passenger, which can sometimes be parsed for additional information (e.g., titles like Mr., Mrs.).

5. Sex: Gender of the passenger (male, female).

6. Age: Age of the passenger in years. Some values are missing and may require imputation.

7. SibSp: Number of siblings or spouses aboard the Titanic.

8. Parch: Number of parents or children aboard the Titanic.

9. Ticket: Ticket number (can be useful in some advanced feature engineering, but often not directly valuable).

10. Fare: The fare paid by the passenger.

11. Cabin: Cabin number (many missing values, but can potentially indicate passenger class or location on the ship).

12. Embarked: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton), representing the departure point.

Label:

Survived: This is the target variable, indicating whether the passenger survived or not (1 = Survived, 0 = Did not survive).

# Interpreting the dataset

1. Pclass: There is a strong correlation between the ticket class and survival, with higher class passengers having better survival rates.

2. Sex: Gender plays a significant role; women were more likely to survive due to the "women and children first" policy.

3. Age: Younger passengers were often given preference during evacuation, so age is a crucial factor.

4. Fare: Higher farepaying passengers (typically in higher classes) were more likely to survive.

5. Embarked: The port of embarkation can offer insight, as passengers from different ports might belong to different socioeconomic classes.

6. SibSp and Parch: Having family members aboard can affect survival probabilities; passengers with more family members may have been more likely to survive.

These key features often contribute to building a successful prediction model for the Titanic dataset.

**url =** [**https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv**](https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv)

# Removing columns that don’t add value

'PassengerId', 'Name', 'Ticket', 'Cabin' don’t add value to the prediction and these can be removed from the dataframe.

# Imputing missing values

Missing Numerical values were replaced by median

Missing categorical values were replaced with most\_frequent

# Scaling

**StandardScaler** is a feature scaling technique from the **scikit-learn** library that standardizes data by removing the mean and scaling to unit variance. It ensures that each feature has a **mean of 0** and a **standard deviation of 1**.

A screenshot of a computer code

Description automatically generated

# Feature Engineering

SibSp: Number of siblings or spouses aboard the Titanic.

Parch: Number of parents or children aboard the Titanic.

Creating a new column with values as sum of SibSp and Parch as the family size

# Auto EDA using AutoViz

**AutoViz** is a Python library that automatically visualizes data with minimal code. It's designed for quick exploratory data analysis (EDA) by generating visual insights for any dataset. It handles both small and large datasets and can display charts for categorical, numerical, and time-series data. Here are key points:

## Summary of the dataframe

A screenshot of a computer

Description automatically generated

It has several plots , taken the scatter plot of Age vs Fare . As we see highest fares were purchased by people in the late thirtiess

A graph of age and age

Description automatically generated

Heatmap shows the correlation between features

A screenshot of a computer screen

Description automatically generated

For example there is negative correlation between Fare and Pclass

Implies that 1st class had greater fare compared to 2nd class and third class

# Auto EDA using SweetViz

**Sweetviz** is a Python library that automates the exploration and visualization of datasets. It generates high-density visual reports that help in understanding data distribution, missing values, correlations, and potential issues in the dataset. Sweetviz is useful for both quick exploratory data analysis (EDA) and detailed comparisons between datasets.

## Summary of the dataframe

A screenshot of a computer

Description automatically generated

Dataframe has 891 rows with 7 features out of which 5 are categorical and 2 are numerical

## Summary of each feature

A white rectangular object with a black border

Description automatically generated

Pclass feature (Passenger class) has 3 unique values 1,2 and 3

It has no missing values

Summary of Age

A screenshot of a computer

Description automatically generated

Numerical features like age also show maximum age, minimum age , average , median

# Auto EDA using missing no

**missingno** is a Python library used to visualize and analyze missing data in datasets. It helps you quickly identify patterns of missingness, which can inform decisions on how to handle the missing values.

From the chart it is seen Age and Embarked have missing values

A graph with a number of bars

Description automatically generated with medium confidence