

Titanic Dataset — Data Cleaning & Preprocessing

Internship Task 1 Report

Abstract

Real-world datasets are often incomplete, noisy, and inconsistent, making data preprocessing a critical step in any machine learning pipeline. This project focuses on performing structured data cleaning and preprocessing on the Titanic dataset to prepare it for machine learning modeling. Techniques such as missing value imputation, categorical feature encoding, numerical feature scaling, outlier detection, and feature engineering were applied. The final output is a clean, high-quality dataset suitable for further exploratory data analysis and predictive modeling.

Objective

The primary objectives of this task were to handle missing values, encode categorical features, normalize numerical data, detect and remove outliers, and prepare a machine-learning-ready dataset using real-world preprocessing techniques.

Dataset Description

The Titanic dataset contains passenger information such as age, gender, ticket class, family relationships, fare, and survival status. It presents common real-world data challenges such as missing values, inconsistent data formats, and categorical features.

Methodology

1. Data Exploration: Understanding dataset structure, data types, and missing values.
2. Missing Value Treatment: Median and KNN imputation for Age, Mode for Embarked, and removal of Cabin.
3. Feature Encoding: Binary encoding for Sex and one-hot encoding for Embarked.
4. Feature Scaling: Standardization of Age and Fare using StandardScaler.
5. Outlier Detection: Boxplots and IQR method used to remove extreme values.
6. Feature Engineering: Creation of FamilySize and IsAlone features.

Results & Observations

The dataset was successfully cleaned and transformed into a machine-learning-ready format. Outlier removal improved data consistency. Feature engineering introduced meaningful new attributes. Correlation analysis and visualization revealed important survival patterns based on gender and passenger class.

Conclusion

This task successfully demonstrated an end-to-end real-world data preprocessing pipeline. The cleaned dataset is now fully prepared for exploratory data analysis and machine learning modeling. This project significantly strengthened practical understanding of applied data science techniques.