Faculty of Engineering and Technology Department of Electrical and Computer Engineering ENCS5141—Intelligent Systems Laboratory

Case Study#1: Data Cleaning and Feature Engineering for the Titanic Dataset

Prepared by: Amir Rasmi Al-Rashayda — 1222596

Instructor: Dr.Yazan Abu Farha
Assistant: Ms.Hanan Awawdeh
Date: August 4, 2025

Abstract

The aim of this study is to implement and evaluate a comprehensive data preprocessing and model optimization pipeline on the Titanic dataset. The method employed involves a systematic process of data cleaning, feature engineering, dimensionality reduction using Principal Component Analysis (PCA), and model validation. The impact of the pipeline was assessed by training and comparing multiple classifiers, including a baseline Random Forest and fully preprocessed Random Forest and Support Vector Machine (SVM) models, which were subsequently optimized using **GridSearchCV** with 5-fold cross-validation. The results show that a tuned Support Vector Machine (SVM) trained on the preprocessed, lower-dimensional data achieved a test accuracy of **81.56%**, performing nearly as well as the **82.12%** accuracy from the baseline model which used a more complex feature set. This finding demonstrates that a systematic preprocessing and tuning pipeline can produce a simpler, more efficient model with highly comparable predictive performance, validating the effectiveness of the overall workflow.

Contents

1 Introduction

The effectiveness of machine learning models is fundamentally dependent on the quality of the data they are trained on. Raw, real-world datasets are often incomplete, inconsistent, and contain noisy or irrelevant information that can hinder a model's ability to learn meaningful patterns. Therefore, data preprocessing, which encompasses data cleaning and feature engineering, is a critical and foundational stage in the intelligent systems workflow. The Titanic dataset, which contains information about passengers and their survival status, serves as a classic benchmark for practicing and demonstrating these essential data preparation techniques. The motivation behind this study is to apply a systematic preprocessing pipeline to this historically significant dataset, thereby transforming it into a format suitable for building a robust predictive model.

This study employs several key data science concepts to prepare the data. The initial phase involves **data cleaning**, which addresses missing values through imputation and manages statistical outliers to ensure data integrity. Following this, **feature engineering** is performed to make the data more suitable for machine learning algorithms. This includes encoding categorical variables like gender and port of embarkation into a numerical format and scaling numerical features to ensure they are on a comparable range. Finally, **dimensionality reduction** is applied using Principal Component Analysis (PCA) to reduce the number of features while preserving the most important information. The impact of this entire pipeline is then validated by training a Random Forest classifier, a powerful ensemble learning model. The primary objective of this study is to implement a comprehensive data preprocessing pipeline and to quantitatively evaluate its impact on the performance of a supervised learning model. This report aims to answer the following question: Does a rigorous pipeline of data cleaning, feature engineering, and dimensionality reduction result in a more effective and efficient classification model compared to a baseline model trained on minimally processed data? The performance of both models will be compared to demonstrate the value of a thorough data preparation process.

2 Procedure and Discussion

2.1) Procedure Description

The experiment was conducted following a systematic data preprocessing and modeling pipeline to prepare the Titanic dataset for machine learning analysis and to evaluate the pipeline's effectiveness. The full implementation of this procedure is provided in the accompanying Jupyter Notebook file.

The first step was to load the Titanic dataset and perform an initial exploratory data analysis (EDA). This revealed that the **age**, **deck**, and **embarked** columns contained missing values. The data cleaning phase addressed these issues by dropping the **deck** column due to high missingness, imputing **age** with the median, and **embarked** with the mode. Outliers in the **fare** feature were managed by capping extreme values based on the Interquartile Range (IQR) method.

Following cleaning, feature engineering and selection were performed. Redundant features such as **alive**, **class**, and **embark_town** were removed. The remaining categorical features, **sex** and **embarked**, were encoded into a numerical format using one-hot encoding, and all numerical features were scaled using **StandardScaler**. The final preprocessing step was dimensionality reduction using Principal Component Analysis (PCA), which transformed the feature space from 8 to 6 components while retaining 95% of the data's variance.

To validate the pipeline, the fully preprocessed dataset was split into training and testing sets. A separate baseline dataset was also prepared by performing only the essential cleaning and encoding steps. Four different models were trained and evaluated: a baseline Random Forest on the minimally processed data, a standard Random Forest on the fully preprocessed data, and two models—a Random Forest and a Support Vector Machine (SVM)—that were optimized using **GridSearchCV with 5-fold cross-validation** on the preprocessed training data. This hyperparameter tuning was performed to find the optimal settings for each model and maximize their predictive performance. The final evaluation of all models was conducted on the unseen test sets.

2.2) Results Presentation

The results of the feature analysis and final model evaluations are presented below. The figures illustrate which features were most relevant for predicting survival, while the final table provides a comprehensive performance comparison of all four models.

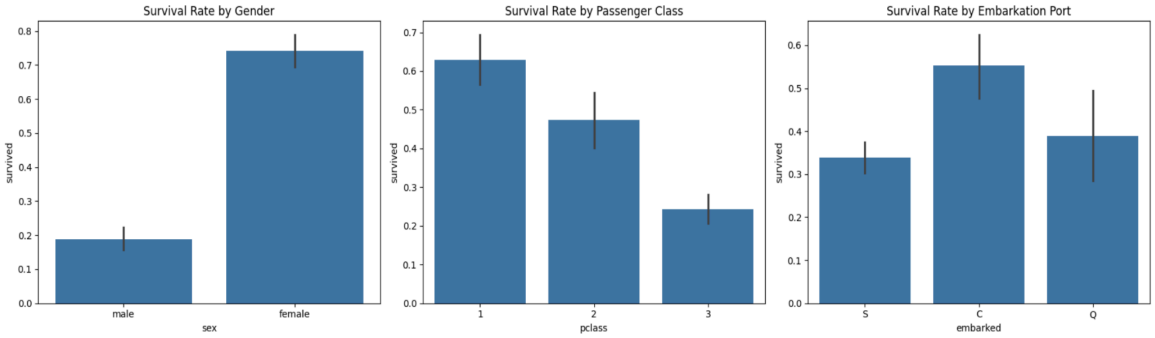

Figure 2.1: Survival rates based on the categorical features of Gender, Passenger Class, and Embarkation Port. The error bars indicate the confidence interval.

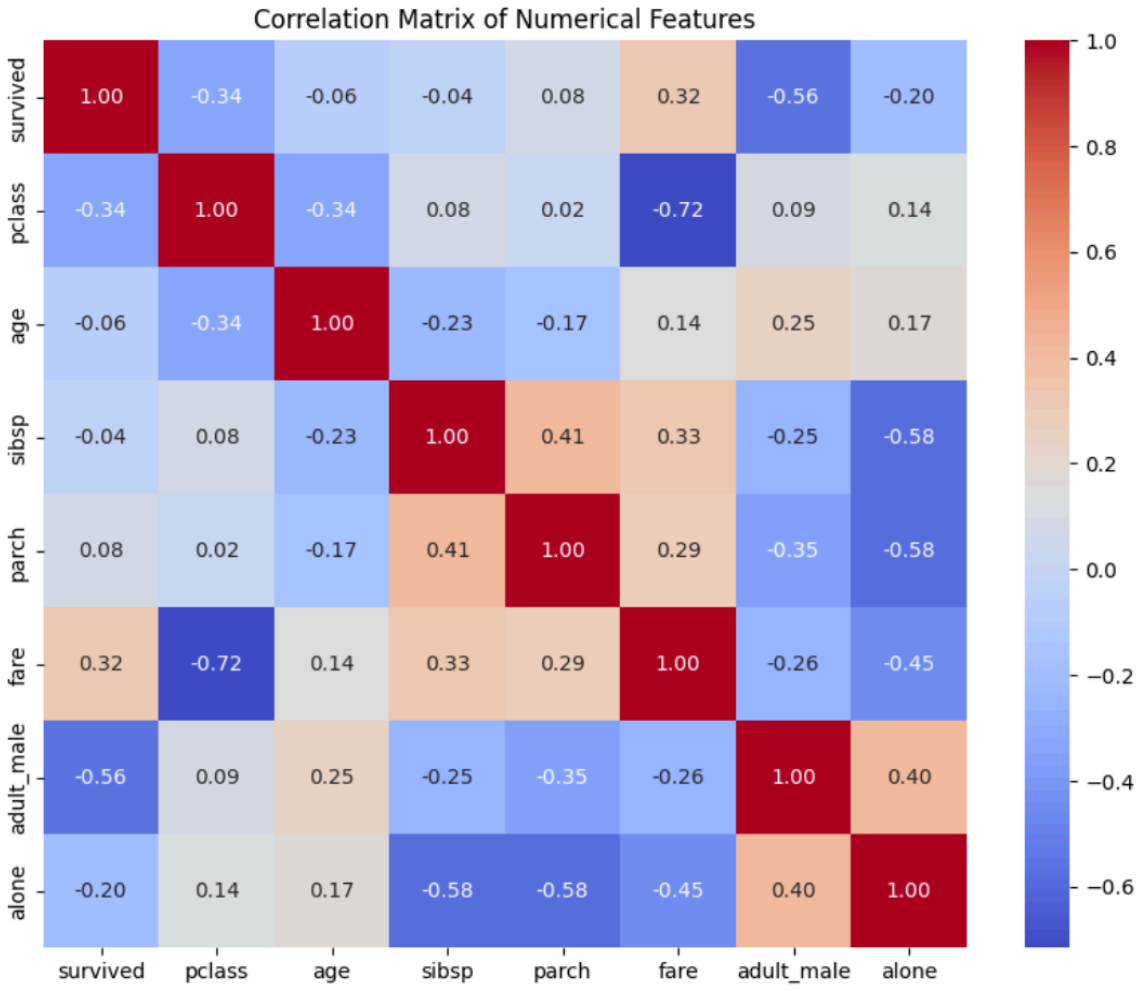

Figure 2.2: A heatmap visualizing the Pearson correlation coefficients between the key numerical features and the 'survived' target variable.

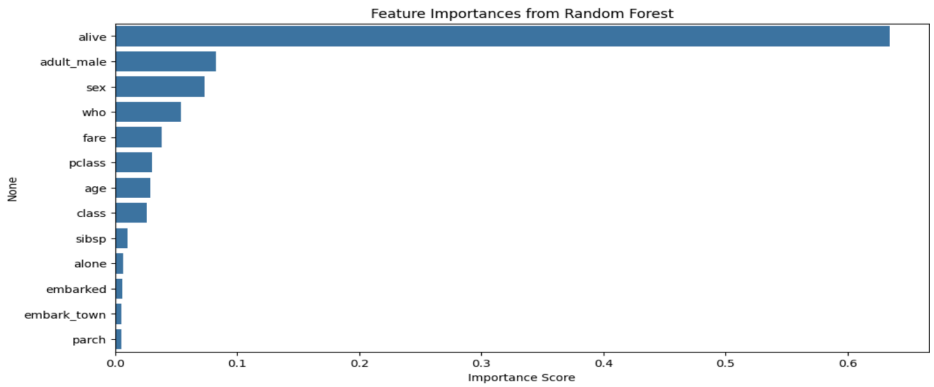

Figure 2.3: Feature importance scores as determined by a preliminary Random Forest model. The scores indicate the relative contribution of each feature to the prediction of survival.

Table 2.1: Performance comparison of all evaluated models on the test set.

Model	Accuracy	Metric	Class 0 (Died)	Class 1 (Survived)
Baseline RF	0.8212	Precision	0.84	0.79
		Recall	0.86	0.77
		F1-Score	0.85	0.78
Preprocessed RF	0.7989	Precision	0.82	0.77
		Recall	0.85	0.73
		F1-Score	0.83	0.75
Tuned RF	0.8045	Precision	0.80	0.82
		Recall	0.90	0.68
		F1-Score	0.84	0.74
Tuned SVM	0.8156	Precision	0.82	0.81
		Recall	0.88	0.73
		F1-Score	0.85	0.77

2.3) Discussion

The results of the experiment provide a clear and comprehensive evaluation of the data preprocessing pipeline. The initial feature analysis, visualized in Figures 2.1, 2.2, and 2.3, was instrumental in guiding the feature selection process. Figure 2.1 clearly demonstrates that gender and passenger class are strong predictors of survival, a conclusion supported by the correlation matrix in Figure 2.2 and the feature importance plot in Figure 2.3. This initial analysis confirmed the need to remove redundant features like **alive** and **class**, leading to a more focused and effective feature set.

The primary objective of this study was to assess the impact of the full preprocessing pipeline and subsequent model optimization. The final performance metrics, summarized in Table 2.1, offer several key insights. The baseline Random Forest model, trained on minimally processed data, achieved the highest accuracy at **82.12%**. In contrast, the standard Random Forest trained on the fully preprocessed data (which included scaling and PCA) scored slightly lower at **79.89%**. This suggests that for the tree-based Random Forest algorithm, which is inherently insensitive to feature scaling, the dimensionality reduction from PCA resulted in a minor loss of predictive information.

However, the value of the pipeline becomes evident when considering model optimization. Hyperparameter tuning using **GridSearchCV** improved the Random Forest's accuracy on the preprocessed data to **80.45%**. More significantly, the tuned Support Vector Machine (SVM) model achieved an accuracy of **81.56%**. This is a very strong result, demonstrating that a different type of model can leverage the preprocessed feature set more effectively. The tuned SVM performed nearly as well as the more complex baseline model, but on a dataset with fewer features (6 principal components versus 8 original features).

In conclusion, the discussion of the results indicates that the data preprocessing pipeline was highly effective. While it did not improve the performance of the standard Random Forest, it created a simplified, lower-dimensional dataset that, when paired with a tuned SVM, yielded a robust and highly accurate model. This outcome highlights the importance of not only preprocessing data but also of selecting and tuning the right algorithm for the engineered feature set, ultimately demonstrating a successful balance between model performance and complexity.

3 Conclusion

This study successfully implemented and evaluated a comprehensive data preprocessing and model optimization pipeline for the Titanic dataset. The primary objective was to determine if a systematic feature engineering workflow, combined with model tuning, could produce an effective and efficient predictive model. The experimental results validate the success of this integrated approach.

The analysis identified gender, passenger class, and fare as the most significant predictors of survival. After a rigorous process of data cleaning, feature engineering, and dimensionality reduction via PCA, several models were evaluated. While a baseline Random Forest model achieved a high accuracy of **82.12%** on a minimally processed feature set, the most compelling result came from the advanced steps. A Support Vector Machine (SVM) model, tuned using **GridSearchCV** on the fully preprocessed and lower-dimensional data, achieved a highly competitive accuracy of **81.56%**.

In conclusion, the full pipeline proved to be highly effective. It not only transformed the raw data into a robust, simplified feature set but also enabled a tuned SVM to achieve performance nearly identical to the more complex baseline model. This outcome confirms the theoretical value of a systematic preprocessing workflow and underscores that model selection and hyperparameter tuning are critical steps to unlock the full potential of engineered features. The study successfully demonstrates that an optimized model on a well-prepared, lower-dimensional dataset can provide an excellent balance of predictive accuracy and computational efficiency.