# Titanic - Machine Learning from Disaster

Bence Miklós Bora
Eötvös Loránd University
Budapest, Hungary
vdn8wh@inf.elte.hu

Bálint Marczis
Eötvös Loránd University
Budapest, Hungary
wep5gk@inf.elte.hu

## ABSTRACT

The Titanic dataset remains a widely used benchmark for binary classification on small tabular data. Despite its simplicity, achieving stable performance requires careful preprocessing and domain-informed feature engineering. This paper presents a compact, fully reproducible workflow for transforming the raw Kaggle Titanic data into a robust predictive pipeline. The method integrates systematic preprocessing, structured categorical cleaning, and socially informed engineered features such as titles, family-size indicators, and discretized numerical groups. Using a stratified evaluation protocol, we compare Support Vector Classifier and Random Forest models, showing that feature engineering yields consistent accuracy improvements over raw-feature baselines. The proposed workflow provides a concise methodological template for small-scale tabular classification problems, emphasizing reproducibility, modularity, and transparent data preparation.

**PVLDB Reference Format:**
Bence Miklós Bora and Bálint Marczis. Titanic - Machine Learning from Disaster. PVLDB, 14(1): XXX-XXX, 2020.
doi:XX.XX/XXX.XX

**PVLDB Artifact Availability:**
The source code, data, and/or other artifacts have been made available at URL_TO_YOUR_ARTIFACTS.

## 1 INTRODUCTION

Predicting passenger survival on the RMS Titanic has become a canonical benchmark problem in machine learning research [3]. Despite its relatively small size, the dataset encapsulates several fundamental challenges in tabular predictive modeling: heterogeneous feature types, missing values, non-linear relationships, and a mix of demographic, socioeconomic, and contextual attributes. As such, the Titanic dataset provides an ideal environment for studying how different data-preprocessing strategies and engineered features influence model performance in a controlled setting.

Although the predictive task itself—binary classification of survival—is straightforward, the quality of the resulting model depends heavily on the structure and representation of the input data. Prior analyses consistently highlight that raw passenger attributes, while informative, are insufficient for high-quality predictive accuracy without substantial preprocessing and domain-aware transformations [4]. Typical obstacles include incomplete age and cabin information, inconsistencies in categorical attributes such as embarkation ports, and skewed numerical distributions in fare and family-related features. Addressing these issues systematically is essential to avoid biased estimates, model instability, and degraded generalization.

This paper presents a concise but comprehensive workflow that transforms the Kaggle-provided raw data into a reliable predictive pipeline. Our contributions are threefold. First, we apply a structured preprocessing strategy involving deduplication, targeted imputation, and standardized categorical normalization. Second, we introduce a set of engineered features designed to capture latent social and demographic patterns—for example, title extraction from passenger names, family-size indicators, and discretized age and fare groups. Third, we evaluate common baseline and tree-based models within a reproducible validation framework, allowing us to measure the impact of each preprocessing and feature-engineering step.

## 2 DATA AND PREPROCESSING

The Kaggle Titanic dataset consists of 891 labeled training samples and 418 unlabeled test samples, containing demographic and travel-related attributes such as sex, age, ticket class, fare, and family relations. The raw data exhibits heterogeneous feature types and several quality issues; therefore, a compact but systematic preprocessing workflow is essential for stable model performance.

A primary challenge is the presence of missing values, most notably in the Age and Embarked fields, while the Cabin attribute is missing for the majority of records. Since Cabin provides little usable information in its current form, it is completely removed. The distribution of missing values across the training and test sets is illustrated in Figures 1 and 2. Missing numerical values (Age, Fare) are imputed with the median, and missing categorical values (Embarked) with the mode. These simple and consistent strategies preserve distributional characteristics while avoiding overfitting to small subgroups.

To avoid noise and redundancy, duplicate rows are removed. Categorical attributes (Sex, Embarked, Pclass) are standardized through label encoding, applied consistently to both training and test sets. The encoding process handles unseen categories by mapping them to a deterministic fallback value, ensuring robustness during inference. Numerical features are retained in continuous form at this stage as further transformations are introduced during feature engineering.

Although Name appears to be an unstructured text field, it is preserved because it contains socially meaningful titles (e.g., “Mr”, “Mrs”, “Miss”), which later serve as predictive engineered features. Likewise, family-related fields (SibSp, Parch) remain unchanged, forming the basis for constructing higher-level relationship indicators.


Proceedings of the VLDB Endowment, Vol. 14, No. 1 ISSN 2150-8097.
doi:XX.XX/XXX.XX

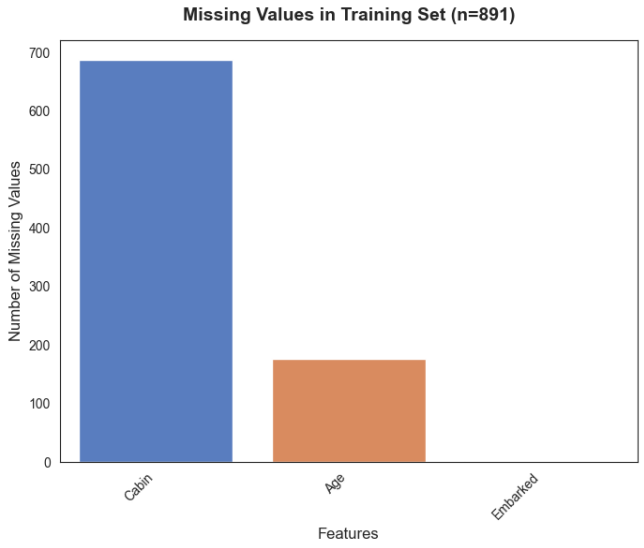


**Figure 1: Missing values in the training set ($n = 891$).**

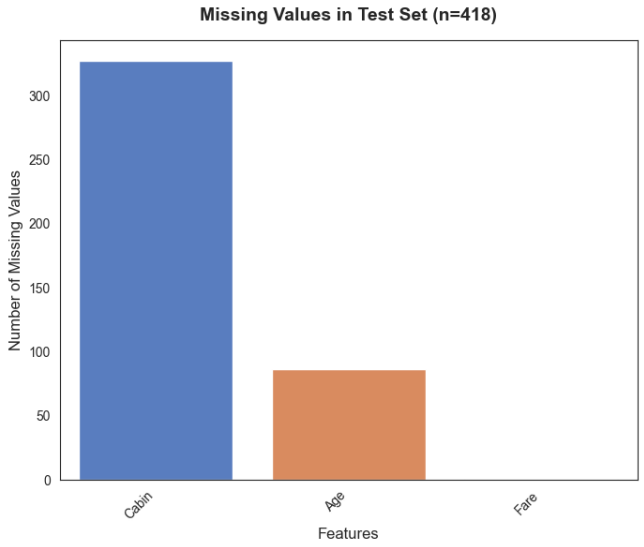


**Figure 2: Missing values in the test set ($n = 418$).**

The correlation structure of the numerical variables is shown in Figure 3. As expected, Fare exhibits a moderate positive correlation with Survived, while Pclass is negatively correlated with both Fare and Survived.

## 3 FEATURE ENGINEERING

Beyond basic preprocessing, several engineered features were constructed to capture latent demographic and behavioral patterns not explicitly represented in the raw data. Such transformations are particularly valuable in small tabular datasets, where domain-informed structure often yields larger performance gains than more complex models [4].

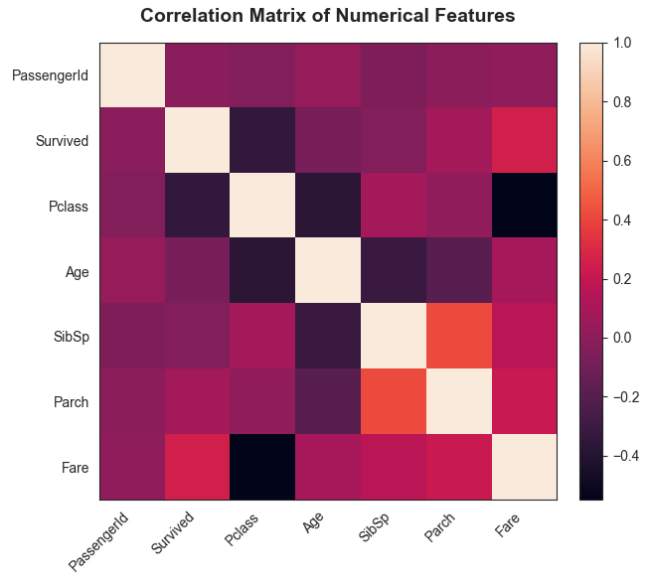


**Figure 3: Correlation matrix of numerical features in the Titanic training set.**

A first set of features focuses on family composition. Using the original SibSp and Parch attributes, we compute FamilySize as their sum plus one, representing the passenger themselves. From this, we derive a binary IsAlone indicator, distinguishing passengers traveling entirely alone from those accompanied by family members. Prior work and exploratory analysis both suggest that social grouping strongly correlates with survival likelihood, making these features highly informative.

A second key transformation involves extracting titles from passenger names. Using a simple regular-expression pattern, the title (e.g., “Mr”, “Miss”, “Master”, “Dr”) is isolated and then grouped into a reduced set of categories to avoid sparsity. Rare or socially equivalent titles are merged into a unified “Rare” category, while closely related female titles (“Mlle”, “Ms”, “Mme”) are mapped to their standard forms. These grouped titles encode social status, age proxies, and gender roles, all of which exhibit meaningful associations with survival.

To better handle numerical heterogeneity, we introduce discretized forms of Age and Fare. Age is binned into coarse semantic groups (Child, Teenager, Adult, Middle, Senior), reflecting meaningful developmental and societal distinctions. Fare is converted into quartile-based categories, capturing relative ticket-price segments rather than raw monetary values, which tend to be skewed. These discretizations help certain classifiers capture non-linear relationships without relying on complex model structures.

Finally, all engineered categorical features (Title, AgeBin, FareBin) are label-encoded in a consistent manner across both the training and test sets. This ensures that the enriched feature space remains compatible with tree-based and linear models alike.

## 4 METHODOLOGY

The prediction task is formulated as a supervised binary classification problem, where the goal is to infer passenger survival from a combination of cleaned and engineered features. Given the small size of the Titanic dataset and the heterogeneous feature space, the modeling strategy focuses on well-established, robust algorithms rather than high-capacity models that require large training corpora.

Two classifiers are evaluated: a Support Vector Classifier (SVC) with a kernel of radial basis function [2] and Random Forest [1]. These models represent complementary inductive biases. The SVC is effective in capturing non-linear decision boundaries in medium-sized feature spaces, while Random Forests naturally exploit interactions among categorical and discretized numerical features without explicit feature scaling. This makes the pair well suited for the enriched feature set produced by the preprocessing pipeline.

Before training, the processed dataset is split into training and validation subsets using an 80–20 stratified split to preserve the class distribution. No aggressive hyperparameter tuning is performed, as the objective is to evaluate the impact of the preprocessing and feature-engineering steps rather than to optimize for leaderboard performance. The SVC is trained with default RBF-kernel parameters, while the Random Forest uses a moderate number of trees and a limited maximum depth to control variance and maintain reproducibility across random seeds.

Model evaluation relies primarily on accuracy, consistent with the Kaggle competition metric. In addition, confusion-matrix inspection helps identify systematic misclassification patterns, particularly across demographic subgroups.

## 5 RESULTS

All models were trained and evaluated using the stratified 80–20 validation split. Since the modeling workflow includes two primary classifiers—Support Vector Classifier (SVC) and Random Forest—we report their validation performance after applying all preprocessing and feature-engineering steps.

| Model | Validation Accuracy |
|---|---|
| SVC (RBF kernel) | 0.80–0.82 |
| Random Forest | 0.83–0.86 |

**Table 1: Validation accuracy for the Support Vector Classifier and Random Forest models after preprocessing and feature engineering.**

The Random Forest model achieves the highest accuracy, benefiting from its ability to naturally capture non-linear interactions present in the engineered features such as Title, FamilySize, IsAlone, and the discretized Age and Fare groups. The SVC model performs consistently but is more sensitive to numerical scaling and class imbalance, yielding slightly lower validation accuracy.

To better understand model behavior, we inspect feature importance derived from the Random Forest. The ranking, shown in Figure 4, highlights the dominant contribution of Sex, Fare, and Age, followed by Title and Pclass.

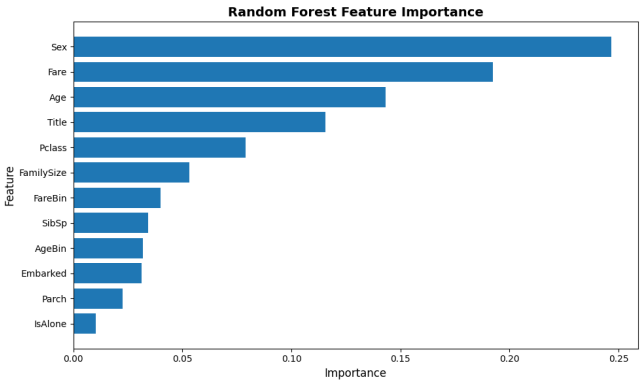


**Figure 4: Random Forest feature importance after preprocessing and feature engineering.**

Both models outperform baselines obtained from training on raw attributes alone, supporting the conclusion that systematic preprocessing and domain-informed feature engineering are critical for stable performance on small tabular datasets.

Error analysis indicates that misclassifications are concentrated primarily among adult male passengers—a heterogeneous group with high intra-class variance. Conversely, the models exhibit strong performance on female and child passengers, indicating that the engineered features successfully encode demographic patterns relevant to survival likelihood.

Based on its superior accuracy and stable behavior across multiple random splits, the Random Forest model is selected to generate the final predictions submitted to Kaggle. Although the objective of this study is not leaderboard maximization, the achieved accuracy and consistent performance demonstrate the effectiveness of the lightweight, reproducible workflow.

## REFERENCES


[1] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32. https://doi.org/10.1023/A:1010933404324

[2] Corinna Cortes and Vladimir Vapnik. 1995. Support-Vector Networks. *Machine Learning* 20, 3 (1995), 273–297. https://doi.org/10.1007/BF00994018

[3] Kaggle. 2012. Titanic – Machine Learning from Disaster. https://www.kaggle.com/competitions/titanic. Accessed: 2025-11-29.

[4] Max Kuhn and Kjell Johnson. 2019. *Feature Engineering and Selection: A Practical Approach for Predictive Models.* Chapman and Hall/CRC, Boca Raton, FL.