Classification of Titanic Survivor Data using Supervised Machine Learning Models

Avalvir Kaur Sekhon, Zihuan Liu & Jordan Kaseram

February 6, 2026

Abstract

This project explores the application of superviesd machine learning techniques to predict passenger survival on the Titanic. Using an augmented Titanic dataset, we implemented and compared several classification models, including Linear Discriminant Analysis (LDA), Logistic Regression and K-Nearest Neighbors (KNN). The data preprocessing phase involved median imputation for missing age values stratified by passenger class, mode imputation for categorical data, and one-hot encoding for feature engineering. Models were evaluated using Repeated Stratified K-Fold cross-validation to ensure robustness against class imbalance. Experimental results demonstrate that feature selection, particularly the inclusion of passenger class and gender, significantly impacts predictive accuracy. The final comparison highlights the trade-offs between model interpretability and predictive performance.

Introduction

Supervised machine learning has become a cornerstone of predictive analytics, offering powerful tools for classification tasks across various domains. In classification, a model learns to map input features to discrete labels based on a labeled training set. This project focuses on the classic Titanic survival prediction task, which serves as a benchmark for evaluating classification algorithms.

The importance of supervised learning in this context lies in its ability to uncover non-linear relationships and interactions between features that may not be apparent through manual analysis. For instance, the ”women and children first” protocol suggests a strong interaction between age, sex, and survival. By utilizing models such as Logistic Regression and Random Forests, we can quantify the impact of these socioeconomic and demographic factors. Understanding these models' performance and the validity of our experimental methodology is crucial for ensuring that the predictions are both accurate and generalizable to unseen data.

Methodology

Titanic Data Set

The Titanic dataset consists of passenger-level information from the sinking of the RMS *Titanic*. Each observation corresponds to an individual passenger, with the response variable indicating survival status (1 = survived, 0 = did not survive). The dataset contains a mixture of demographic, socioeconomic, and voyage-related features, including age, sex, passenger class, fare paid, family composition, and cabin information.

In total, the dataset contains 891 observations with a diverse set of categorical and numerical variables. As is typical of real-world data, the raw dataset presents several challenges, including missing values, highly sparse columns, and correlated features derived from similar underlying information.

These characteristics make the Titanic dataset well suited for evaluating classification methods while also requiring careful preprocessing and feature selection prior to model fitting.

Data Preprocessing

Data preprocessing was performed using Python and the pandas library, which provides flexible and efficient tools for data cleaning, transformation, and preparation prior to machine learning analysis. Identifier variables such as passenger identifiers, ticket numbers, and booking references were removed, as they do not contain predictive information and may introduce unnecessary noise into the models.

Missing data were handled using simple, distribution-preserving imputation techniques. Missing values in passenger age, a numerical variable, were imputed using the median age within each passenger class. This approach allows age estimates to reflect socioeconomic differences across classes while remaining robust to outliers. Missing values in the port of embarkation, a categorical variable, were imputed using the mode, preserving the empirical distribution of this feature.

Categorical predictors were converted into numerical form using one-hot encoding creating dummy variables. To reduce redundancy and limit multicollinearity, several highly correlated or derived features were evaluated and pruned. For example, the variables `SibSp`, `Parch`, and `is_alone` were removed since `family_size` contains this information.

The resulting feature set provides a stable and interpretable foundation for training supervised machine learning models for binary classification.

Resampling and Validation Strategy

To ensure reliable model evaluation and reduce overfitting, the dataset was partitioned into training (75%) and testing (25%) sets using a stratified split to preserve the original survival class proportions. Stratification is particularly important for the Titanic dataset due to its class imbalance between survivors and non-survivors.

All model training, hyperparameter tuning, and feature selection were performed exclusively on the training set using stratified k-fold cross-validation with $k = 5$. This resampling strategy maintains class balance within each fold and provides stable performance estimates while preventing information leakage.

Final model performance was evaluated on the held-out test set, which remained untouched during model selection. This approach yields an unbiased estimate of generalization performance and ensures fair comparison across all classification methods.

Model Selection

Three classification methods were selected to evaluate the predictive performance of the features present in the Titanic data set. These methods include: Logistic Regression, Linear Discriminant Analysis (LDA) and K-Nearest Neighbours (KNN). Logistic Regression and LDA were considered for their interpretability due to linearity [2], and the Titanic prediction task is a binary response with a mix of numeric and categorical variables which these models perform well on [1].

These parametric tests assume a structured relationship between the predictors and the response, enabling stable parameter estimation and favourable bias-variance trade-offs. In contrast, KNN was included as a non-parametric test to capture potential non-linear patterns without assuming strong distribution assumptions. KNN provides a useful comparison model by classifying passengers based on identifying the observations that are nearest it rather than a global decision boundary [1].

Hyperparameter Tuning Strategy

Hyperparameter tuning was conducted using grid search, which is a simple memoryless method that can be used to explore predefined hyperparameter values [4]. This was combined with stratified cross-validation to identify the optimal model configuration.

For Linear Discriminant Analysis, hyperparameter tuning focused on the choice of solver (`svd` versus `lsqr`) and the use of covariance shrinkage, as shrinkage regularization is known to stabilize covariance estimation in the presence of correlated predictors and limited sample sizes [2, 4].

For Logistic Regression, tuning was performed over the regularization strength parameter C and the choice of ℓ_1 and ℓ_2 penalties, allowing control over model complexity and coefficient sparsity, which directly affects bias–variance trade-offs and interpretability in high-dimensional settings [1, 5].

For K-Nearest Neighbours, hyperparameters including the number of neighbours, distance metric, and weighting scheme were tuned to balance local versus global decision behavior and to regulate model sensitivity to noise and class overlap [1, 5].

Experiment

Experimental Design

The experimental design aimed to compare the predictive performance and interpretability of multiple classification models under a consistent evaluation framework. All models were trained using the same preprocessed feature set and assessed with identical performance metrics to isolate the impact of model choice and tuning decisions.

Each method was first evaluated in a baseline configuration to establish a performance reference. Subsequent experiments introduced hyperparameter tuning and feature selection to examine whether these refinements led to improvements in generalization performance.

Model comparisons were based on cross-validated accuracy estimates computed on the training data, followed by final evaluation on a held-out test set. This design enabled fair comparison between linear and non-linear models, as well as between parametric and non-parametric approaches, while controlling for differences in preprocessing and evaluation procedures.

Results and Analysis

Baseline Model Performance This section summarizes baseline predictive performance and examines both error patterns and feature contributions across models.

Table 1: Performance of baseline classification models on the Titanic dataset.

Baseline Model	CV Accuracy (Mean)	CV Accuracy (Std)	Test Accuracy	Test MSE
LDA	0.8249	0.0439	0.8386	0.1614
Logistic Regression	0.8368	0.0452	0.8251	0.1749
KNN	0.8139	0.0483	0.8027	0.1973

Overall, Linear Discriminant Analysis achieved the strongest test performance, while Logistic Regression performed comparably. K-Nearest Neighbours exhibited lower predictive performance, which is consistent with its sensitivity to class overlap and noise in feature space.

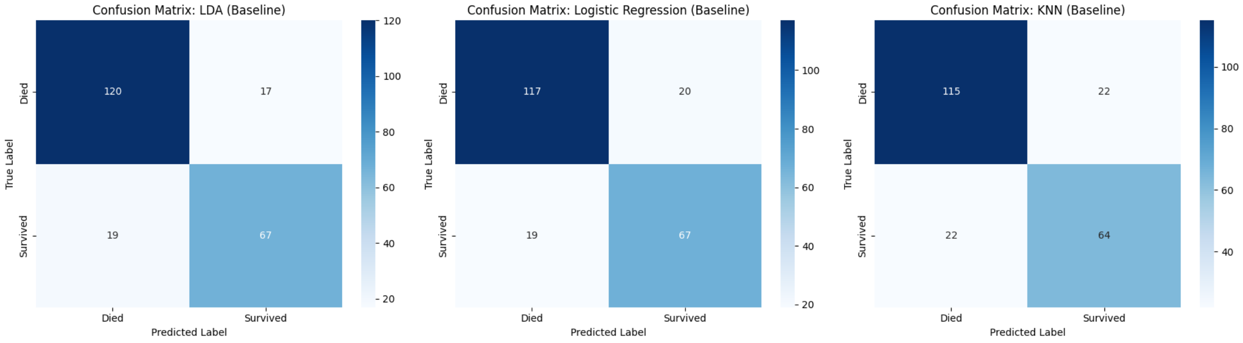

Figure 1: Confusion matrices for baseline classification models.

Both LDA and Logistic Regression maintain a balanced trade-off between precision and recall across survival classes. In contrast, KNN shows reduced recall for the survivor class, indicating greater difficulty separating survivors from non-survivors when observations overlap.

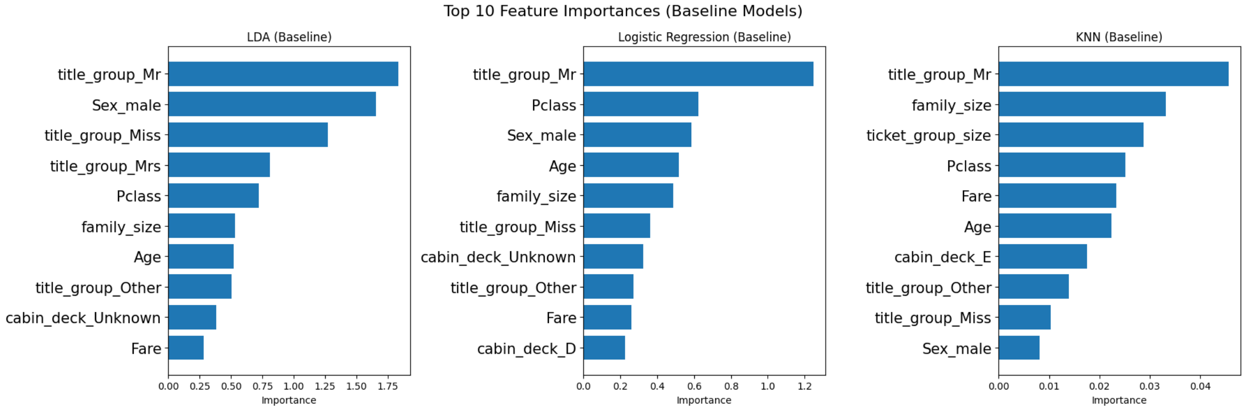

Figure 2: Top 10 features from baseline models for LDA, Logistic Regression, and KNN.

For the linear models, passenger title, sex, passenger class, and age were consistently identified as influential predictors. In contrast, KNN relied more heavily on proximity-based features such as family size and ticket group size. These findings align with historical accounts of the Titanic disaster, where it adopted a *"women and children first"* first policy for getting on the lifeboats.

Effect of Hyperparameter Tuning The optimal hyperparameter configurations selected via cross-validated grid search differed across models. For Linear Discriminant Analysis, the best-performing configuration used the `lsqr` solver with a shrinkage parameter of 0.05, indicating mild regularization of the covariance estimate. Logistic Regression achieved its best performance with an ℓ_1 penalty and regularization strength $C = 1$, encouraging sparse and interpretable coefficient estimates. For K-Nearest Neighbours, the optimal configuration used $k = 5$ nearest neighbours, the L_1-norm, and uniform weighting.

Table 2: Performance of tuned classification models on the Titanic dataset.

Tuned Model	CV Accuracy (Mean)	CV Accuracy (Std)	Test Accuracy	Test MSE
LDA	0.8269	0.0477	0.8117	0.1883
Logistic Regression	0.8408	0.0441	0.8251	0.1749
KNN	0.8249	0.0479	0.7713	0.2287

Hyperparameter tuning produced small improvements in cross-validation accuracy for Logistic Regression, increasing from 0.8368 to 0.8408, while test accuracy remained unchanged at 0.8251. This suggests that tuning improved validation performance without improving generalization to unseen data. In contrast, Linear Discriminant Analysis exhibited a slight increase in cross-validation accuracy but a noticeable decrease in test accuracy, indicating that shrinkage regularization introduced additional bias that did not translate to improved out-of-sample performance. K-Nearest Neighbours experienced the largest decline after tuning, highlighting the sensitivity of distance-based classifiers to hyperparameter choices in datasets with overlapping class distributions.

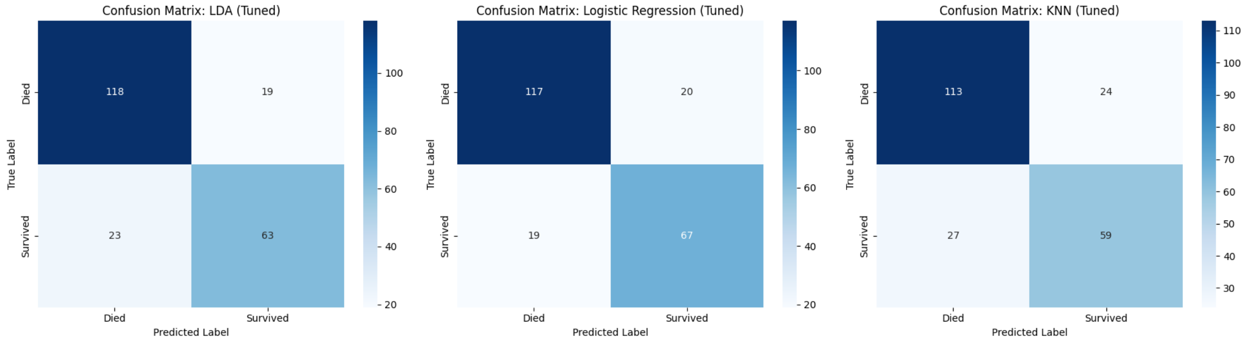

Figure 3: Confusion matrices for tuned classification models.

Compared to the baseline setting, tuned models generally show a similar error profile, with misclassifications concentrated among passengers whose covariates provide ambiguous survival signals. The tuned KNN model exhibits a larger reduction in recall for the survivor class, consistent with its lower test accuracy and increased susceptibility to overlapping neighborhoods.

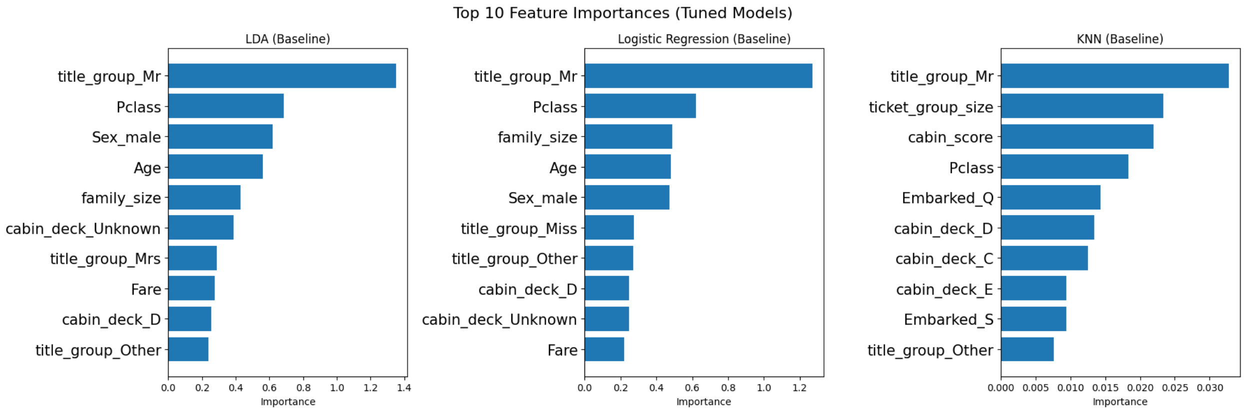

Figure 4: Top 10 features from tuned models for LDA, Logistic Regression, and KNN.

The tuned models largely preserved the same dominant predictors identified in the baseline analysis, including passenger title group, sex, passenger class, and age. This consistency indicates that the primary survival signal in the dataset is robust to moderate changes in model configuration, and that tuning primarily affects coefficient regularization rather than altering the overall feature ranking.

Discussion

Overall, linear models outperformed the non-parametric KNN approach on the Titanic dataset, suggesting that survival outcomes are primarily governed by global, structured relationships among predictors rather than highly localized decision boundaries. Logistic Regression and LDA benefited from their ability to model global trends and remained robust under both baseline and tuned configurations.

Hyperparameter tuning improved cross-validation performance but did not consistently improve test accuracy. In particular, tuning introduced additional bias for LDA and increased sensitivity for KNN, resulting in reduced generalization performance.

These findings underscore the bias-variance trade-off in supervised classification and demonstrate that increased model flexibility does not guarantee better performance.

Conclusion

References

[1] James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: Python edition.*

[2] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning.*

[3] Kaseram, J., Liu, Z., Sekhon, A. (2026). *Titanic survival classification project.* GitHub repository. Available at: `https://github.com/RoxyLiu66/data_572_group_5`

[4] Franceschi, L., Donini, M., Perrone, V., Klein, A., Archambeau, C., Seeger, M., ... & Frasconi, P. (2025). *Hyperparameter optimization in machine learning. Foundations and Trends in Machine Learning*, 18(6), 975-1109.

[5] Kuhn, M., & Johnson, K. (2013). *Applied predictive modelling (Vol. 26, p. 13).* New York: Springer.