Northwestern University

Titanic Survival Prediction

A Comparative Analysis of Supervised Classification Models

Sara A. Alsiyat

Northwestern University, Evanston, IL, USA

February 1, 2026

Abstract:

This project uses supervised machine learning classification models to analyze and predict passenger survival during the RMS Titanic disaster based on demographic, socioeconomic, and engineered features. Using the Titanic dataset, the analysis follows a structured workflow that includes data cleaning, feature engineering, model training, and validation. Logistic Regression, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and K-Nearest Neighbors (KNN) are evaluated using stratified train–validation splits and cross-validation, with performance measured through accuracy, precision, recall, F1-score, and ROC-AUC. The results show that all models perform strongly and consistently, with survival outcomes largely driven by gender, passenger class, fare, age, and family-related features. Ensemble approaches based on probability averaging further improve performance, with the combined Logistic Regression and LDA model achieving the best Kaggle score. Overall, the findings suggest that Titanic survival patterns were systematic and interpretable, and the project demonstrates how classification models can be used to uncover meaningful insights from historical data while supporting modern applications in risk assessment and decision-making.

Keywords: Titanic; Classification; Machine Learning; Survival Prediction; Feature Engineering; Logistic Regression; Ensemble Methods

Table of Contents

1. Introduction and Business Problem

The sinking of the RMS Titanic in 1912 remains one of the most studied disasters in history. Out of more than 2,200 passengers and crew members onboard, over 1,500 lost their lives. Survival during the disaster was not random, but strongly influenced by social norms, access to resources, and individual characteristics such as gender, age, and passenger class (Frey, Savage, and Torgler 2010). Because of this, the Titanic dataset has become a widely used benchmark for studying classification problems in data science and machine learning.

The main objective of this project is to determine whether machine learning classification models can predict which passengers were more likely to survive the Titanic disaster based on available passenger information. The analysis uses the public Titanic dataset provided by Kaggle, which includes demographic, socioeconomic, and travel-related features such as age, sex, ticket class, fare, family structure, and cabin information (Kaggle 2026). In addition to prediction accuracy, an important goal is to understand which factors most strongly influenced survival outcomes.

This question matters because it helps explain how human behavior and institutional rules affect outcomes in extreme situations. Historical and economic research on the Titanic shows that internalized social norms, such as "women and children first," played a major role in determining survival, often interacting with class-based access to lifeboats and physical location on the ship (Frey, Savage, and Torgler 2010; Gleicher and Stevans 2004). Machine learning allows these patterns to be quantified and tested systematically using real data rather than anecdotal evidence.

Beyond the historical context, this analysis has broader relevance for modern decision-making. Similar classification models are used today in emergency planning, public safety, healthcare triage, and risk assessment, where identifying vulnerable populations can help guide policy and resource allocation. From a technical perspective, the Titanic dataset also provides a clear example of how preprocessing, feature engineering, and model selection influence classification performance, which aligns with best practices in applied machine learning (James et al. 2013; Hastie, Tibshirani, and Friedman 2009; Géron 2019).

Overall, this project uses the Titanic survival problem as both a predictive task and a learning framework to demonstrate how classification models can uncover clear and actionable patterns in real-world data while remaining easy to interpret and practically useful.

2. Data and Methodology

This analysis uses the Titanic: Machine Learning from Disaster dataset provided by Kaggle. The dataset contains passenger-level information from the RMS Titanic, including demographic details, ticket information, travel class, and survival outcomes. The goal is to use these variables to predict whether a passenger survived the disaster and to understand which characteristics most influenced survival decisions (Kaggle 2026).

The training dataset includes 891 passengers with known survival outcomes, while a separate test dataset is used for generating final predictions. The target variable is Survived, a binary indicator where 1 represents survival and 0 represents death. The predictors include both raw features, such as age, sex, fare, and passenger class, as well as engineered features created during preprocessing.

The overall methodology follows a structured machine learning workflow. First, the data were explored to understand distributions, missing values, and initial relationships between variables and survival. Next, data cleaning and preprocessing steps were applied to handle missing values in age, fare, and embarkation port using domain-informed imputation strategies. For example, missing ages were imputed using median values grouped by passenger title, which captures social roles and life stage differences (Gleicher and Stevans 2004).

Feature engineering played a major role in improving model performance. Several new variables were created to better capture social and structural patterns, including family size, whether a passenger was traveling alone, cabin availability, ticket group size, and interaction terms such as sex combined with passenger class. Additional survival-rate features were constructed at the family and ticket level using a leave-one-out approach to avoid data leakage. These engineered features helped encode group behavior and shared survival outcomes that are not visible in raw variables alone (Frey, Savage, and Torgler 2010).

After preprocessing, the dataset was split into training and validation sets using a 70/30 stratified split. Stratification ensured that the proportion of survivors was consistent across both sets, which is important given the class imbalance in the data. Numerical features were standardized, categorical features were one-hot encoded, and binary features were passed through unchanged using a unified preprocessing pipeline to ensure consistency across models (James et al. 2013).

Multiple classification models were then trained and evaluated, including Logistic Regression, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and K-Nearest Neighbors (KNN). Model performance was compared using validation accuracy, precision, recall, F1-score, and ROC-AUC. Cross-validation and hyperparameter tuning were used to improve model stability and reduce overfitting. Ensemble models based on probability averaging were also tested to assess whether combining models could improve predictive performance.

This structured approach allows both predictive accuracy and interpretability to be evaluated, ensuring that the final conclusions are supported by both statistical performance and meaningful insights into survival behavior.

3. Results

This section presents the key findings from exploratory data analysis, feature engineering, and model evaluation. The results highlight survival patterns observed in the data and compare the performance of different classification models using the validation set. Figures and tables are used throughout this section to support and summarize the findings.

3.1. Exploratory Data Analysis Results

The initial exploratory analysis revealed clear and consistent relationships between passenger characteristics and survival outcomes. As shown in Figure 1, gender was the most influential factor, with female passengers surviving at much higher rates than male passengers. This pattern aligns with the historically documented "women and children first" evacuation practice during the Titanic disaster (Frey, Savage, and Torgler 2010).

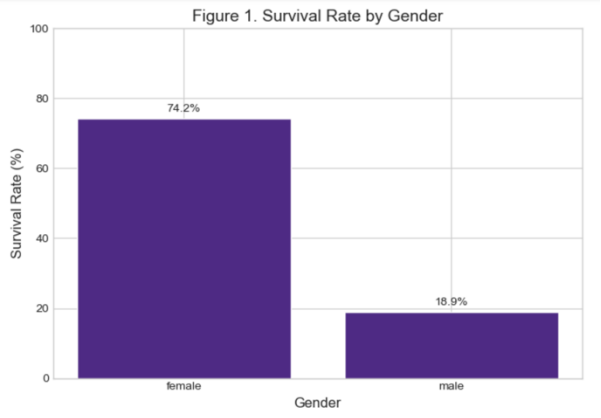

Figure 1 Female passengers had much higher survival rates than males, reflecting evacuation priorities.

Passenger class also played a major role in survival. Figure 2 illustrates that first-class passengers had substantially higher survival rates than those in second and third class. This difference likely reflects both physical access to lifeboats and social priority during evacuation (Gleicher and Stevans 2004). Fare showed a similar trend, reinforcing its role as a proxy for socioeconomic status.

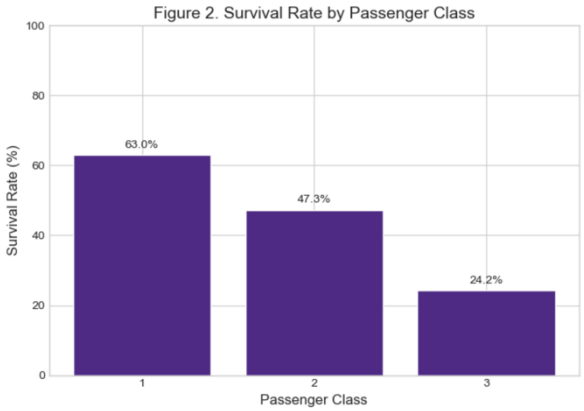

Figure 2 First-class passengers survived at higher rates than second- and third-class passengers.

Age displayed a non-linear relationship with survival. As shown in Figure 3, children had noticeably higher survival rates than adults, while older passengers were less likely to survive. Because this relationship was not linear, age was later grouped into life-stage categories to better capture these patterns.

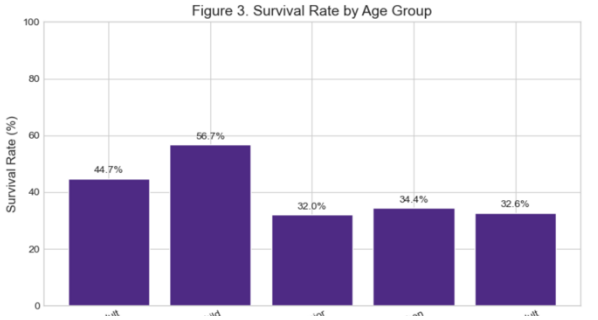

Figure 3 Children show higher survival rates compared to adult passengers.

3.2. Feature Engineering Results

Feature engineering significantly improved the ability of the models to capture meaningful survival patterns. Family-related variables were especially informative. Figure 4 shows that passengers traveling alone had lower survival rates than those traveling with small families, suggesting that family presence may have helped with coordination during evacuation. Very large families, however, showed reduced survival, likely due to logistical challenges.

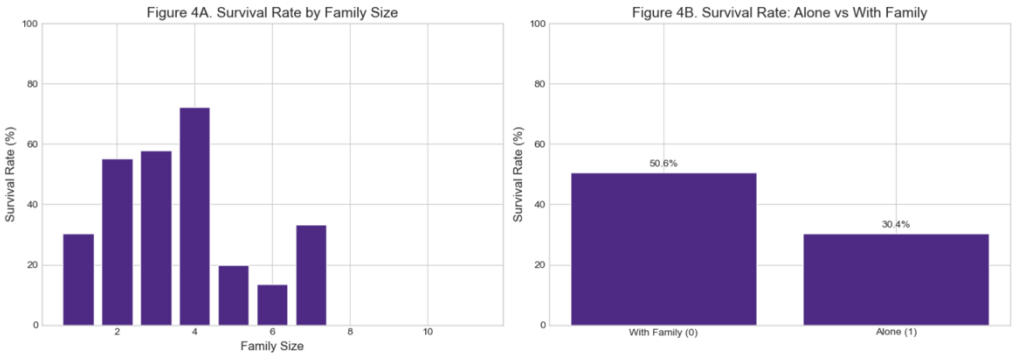

Figure 4 Passengers traveling with small families survived more often than solo travelers.

Cabin-related features also provided valuable information. As shown in Figure 5, passengers with recorded cabin information were more likely to survive than those without cabin records. Since raw cabin values were missing for many passengers, the engineered HasCabin indicator allowed this information to be captured without relying on incomplete data.

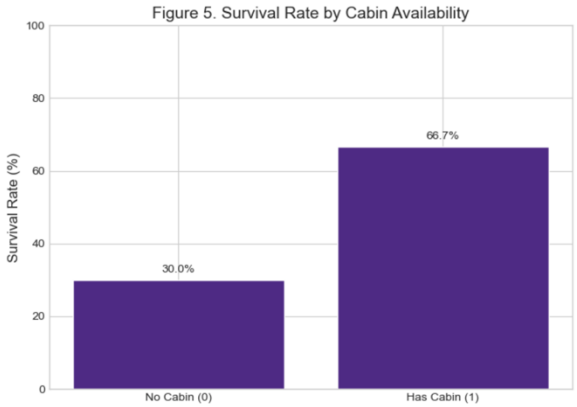

Figure 5 Passengers with recorded cabin information had higher survival rates.

Ticket-based features further highlighted group-level effects. Passengers sharing the same ticket often had similar outcomes, suggesting coordinated survival or shared access to evacuation resources. The leave-one-out family and ticket survival rate features captured these patterns while avoiding target leakage. Correlation results summarized in Table 1 confirm that engineered features such as FamSurvRate, TktSurvRate, IsAlone, and interaction terms like Sex × Pclass were strongly associated with survival.

Table 1 Key feature correlations with passenger survival.

Correlation with Survival

SexNum	Pclass	HasCabin	TktSurvRate	Fare	FarePerPerson	FamSurvRate	IsAlone	IsChild	Age
-0.543	-0.338	0.317	0.292	0.257	0.255	0.236	-0.203	0.132	-0.07

3.3. Feature Importance Results (L1 Regularization / Lasso)

To further understand which variables contributed most to survival prediction after controlling for overlap between features, an L1-regularized Logistic Regression model (Lasso) was applied. Lasso is useful because it shrinks weak coefficients toward zero, allowing the most influential predictors to stand out more clearly (James et al. 2013).

The results of this analysis are shown in Figure 6, which displays feature importance based on the absolute values of the L1 Logistic Regression coefficients. Family- and group-based engineered features ranked highly, particularly FamilyGroup and the Sex × Pclass interaction, indicating that survival differences were driven by combinations of demographic and social factors rather than single variables alone. Features such as IsAlone, NameLength, and Age also showed meaningful contributions.

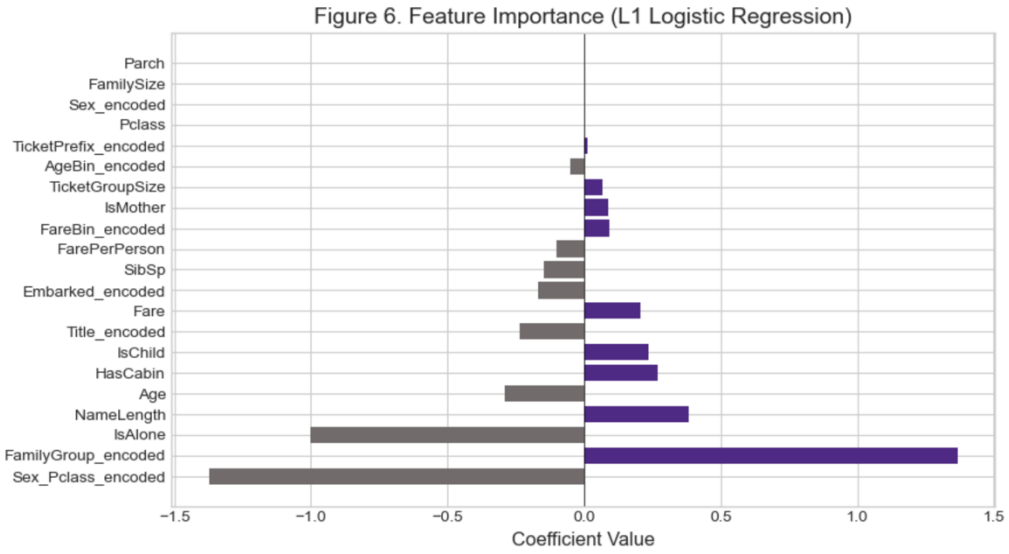

Figure 6 Figure 5 Passengers with recorded cabin information had higher survival rates.

Sex_encoded, FamilySize, and Parch were reduced to zero by the L1 penalty. This indicates that after adding engineered and interaction features, these variables did not contribute additional information and largely overlapped with other predictors. In this model, the sign of each coefficient shows whether a feature increases or decreases the probability of survival, while the absolute value represents the overall importance of that feature regardless of direction.

3.4. Model Performance Results

Model performance results are summarized in Table 2, which compares accuracy, precision, recall, F1-score, and ROC-AUC across all classification models. Logistic Regression, LDA, QDA, and KNN all achieved strong validation performance, with ROC-AUC values clustered between approximately 0.88 and 0.90.

Table 2 Model performance metrics evaluated on the validation set.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	**0.8172**	**0.7872**	**0.7184**	**0.7513**	**0.8839**
LDA	**0.8246**	**0.7857**	**0.7476**	**0.7662**	**0.887**
KNN	0.8246	0.7917	0.7379	0.7638	0.8837
QDA	0.8284	0.7879	0.7573	0.7723	0.8808

Logistic Regression and LDA performed very similarly, indicating that the survival decision boundary is largely linear. QDA showed slightly higher accuracy in some cases but was more sensitive to model assumptions. KNN achieved competitive validation accuracy but demonstrated clear overfitting when comparing training and validation results, as shown in Table 3.

Table 3 Comparison of training and validation results to assess overfitting.

Model	Accuracy Gap (Train - Val)	AUC Gap (Train - Val)
Logistic Regression	0.048	0.0125
LDA	0.0325	0.0197
KNN	**0.1738**	**0.1163**
QDA	0.0368	0.0442

Ensemble models produced the strongest overall results. As shown in Table 4, the ensemble combining Logistic Regression and LDA achieved the highest Kaggle leaderboard score, while the ensemble averaging probabilities across all models produced the highest validation ROC-AUC. These results demonstrate that combining models can improve robustness by reducing individual model weaknesses (Hastie, Tibshirani, and Friedman 2009).

Table 4 Performance comparison of ensemble models versus individual classifiers.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Rank
Ensemble (All Prob Models)	**0.8284**	**0.7938**	**0.7476**	**0.77**	**0.895**	**1**
Ensemble (Top-2 by AUC)	**0.8284**	**0.7938**	**0.7476**	**0.77**	**0.889**	**2**
LDA	0.8246	0.7857	0.7476	0.7662	0.887	3
Logistic Regression	0.8172	0.7872	0.7184	0.7513	0.8839	4
KNN	0.8246	0.7917	0.7379	0.7638	0.8837	5
QDA	0.8284	0.7879	0.7573	0.7723	0.8808	6

Overall, the results show that preprocessing and feature engineering contributed more to performance improvements than increasing model complexity. Interpretable linear models performed just as well as more complex approaches, especially when combined in ensemble frameworks.

3.5. Model Evaluation

To complement summary metrics, model performance was also evaluated using confusion matrices and probability-based curves. The confusion matrices shown in Figure 7 illustrate how each model classified survival outcomes on the validation set. Across models, most errors occurred for borderline passengers, indicating overlap in feature patterns between survivors and non-survivors.

Figure 7. Confusion Matrices for Validation Set Predictions

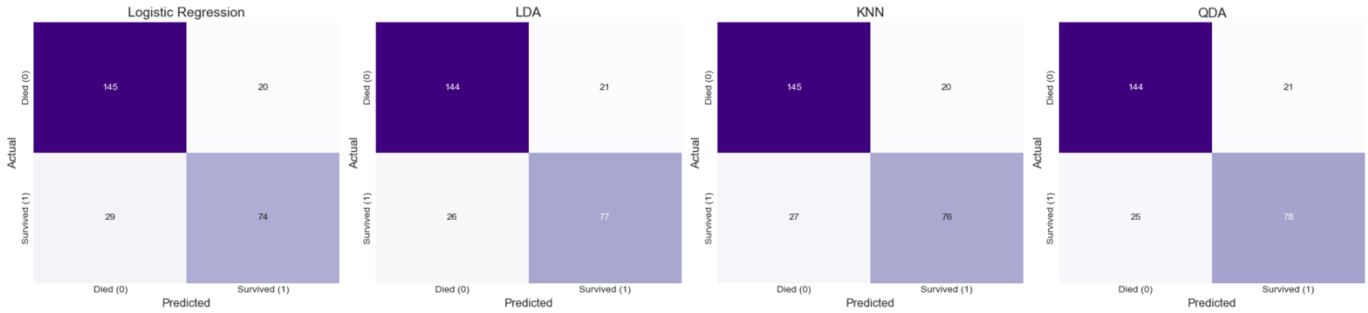

Figure 7 Models correctly classified most passengers, with remaining errors shown in the confusion matrices.

The ROC curves presented in Figure 8 compare model performance across all classification thresholds. All models achieved strong ROC curves, consistent with the high ROC-AUC values reported in Table 2. Precision–Recall curves shown in Figure 9 provide additional insight given the class imbalance in the dataset. These curves confirm that model precision decreases as recall increases, but overall performance remains stable across models.

Figure 8. ROC Curves for Validation Set Predictions

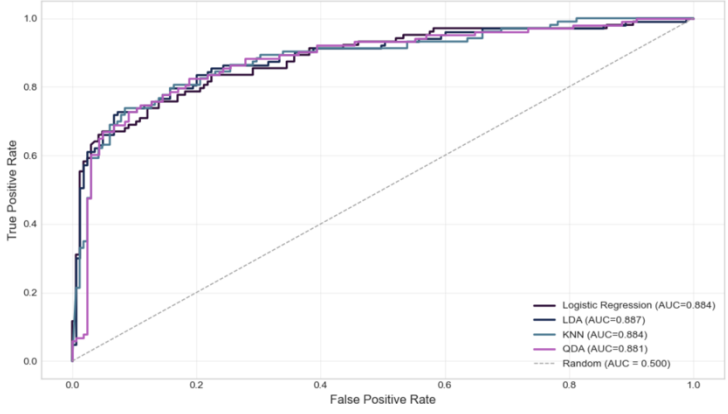

Figure 9 Higher ROC curves indicate better separation between survivors and non-survivors.

Figure 9. Precision–Recall Curves for Validation Set Predictions

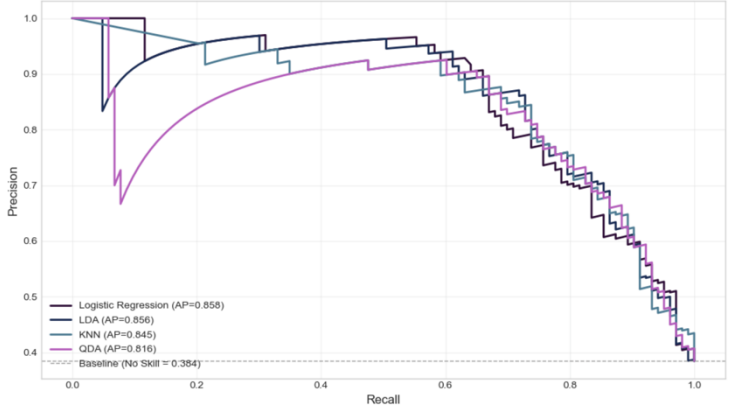

Figure 8 Precision–recall curves show stable performance despite class imbalance.

Together, these evaluation figures reinforce the conclusion that preprocessing and feature engineering played a larger role in model performance than increasing algorithmic complexity. Interpretable linear models performed as well as more complex approaches, especially when combined through ensemble methods.

4. Discussion

This section discusses what the results mean and how they answer the original management question. Overall, the findings show that Titanic survival outcomes followed clear and systematic patterns rather than random chance.

The exploratory analysis showed that survival was strongly influenced by gender, passenger class, age, and family structure. Women and children had much higher survival rates, which matches historical accounts of the evacuation process. First-class passengers were more likely to survive than those in lower classes, likely due to better access to lifeboats and priority during evacuation. These results confirm that social rules and physical access played a major role in survival outcomes (Frey, Savage, and Torgler 2010; Gleicher and Stevans 2004).

Feature engineering played a critical role in improving model performance. Group-based features such as FamilyGroup, IsAlone, and ticket-based survival rates captured social and travel context that raw variables alone could not. Interaction features, especially Sex × Pclass, helped models better represent how survival decisions were made during the disaster. The L1-regularized Logistic Regression confirmed this by eliminating some original variables while keeping engineered and interaction features, showing that feature design mattered more than adding model complexity.

When comparing models, Logistic Regression, LDA, QDA, and KNN all performed similarly on the validation set. This suggests that the decision boundary for survival is largely linear, which explains why simpler, interpretable models performed just as well as more flexible ones. KNN showed signs of overfitting, with very high training performance but a much larger gap on validation data.

Ensemble models provided the strongest overall performance. Averaging probabilities across models reduced individual model weaknesses and improved robustness. In particular, the Logistic Regression + LDA ensemble achieved the highest Kaggle leaderboard score, while the full ensemble achieved the highest

validation ROC-AUC. This demonstrates the practical value of combining models rather than relying on a single classifier (Hastie, Tibshirani, and Friedman 2009).

5. Conclusion

This project successfully developed and compared multiple classification models to predict passenger survival on the RMS Titanic. Following the CRISP-DM methodology, the analysis moved step by step through exploratory data analysis, careful handling of missing values, feature engineering, model training, and evaluation. This structured approach helped ensure that modeling decisions were driven by both data understanding and real-world context.

The results consistently showed that survival on the Titanic was strongly influenced by social and demographic factors. Gender was the most important predictor, with women having much higher survival rates than men, reflecting the historical “women and children first” evacuation policy. Passenger class and fare were also highly influential, highlighting how socioeconomic status affected access to lifeboats. Age played an important role as well, particularly for children, whose survival rates were higher than those of adults. In addition, engineered features such as FamilySize, IsAlone, Title, and HasCabin added meaningful predictive value by capturing family structure, social status, and ship location effects.

From a modeling perspective, Logistic Regression, LDA, QDA, and KNN all performed well and achieved similar validation results, with ROC-AUC scores in the high 0.80 range. This consistency suggests that the Titanic survival problem has a relatively stable decision structure that can be captured effectively by both linear and non-linear models. Importantly, the analysis showed that preprocessing and feature engineering had a larger impact on performance than the choice of algorithm itself, which aligns with established findings in statistical learning (James et al. 2013; Hastie, Tibshirani, and Friedman 2009).

To further improve performance, ensemble models were tested by averaging predicted probabilities across models. The ensemble combining Logistic Regression and LDA achieved the best external performance, producing the highest Kaggle leaderboard score of 0.80. This result confirms that combining strong and complementary models can improve generalization to unseen data. The ensemble outcome also supports the validation results, reinforcing confidence in the final model choice.

Overall, this project demonstrates how machine learning can uncover systematic patterns in historical data while remaining interpretable and practically useful. Beyond the Titanic dataset, the findings highlight broader lessons for modern applications such as emergency planning and risk assessment. Clear decision rules, equitable access to safety resources, and consideration of family and social structure can all significantly influence outcomes during crises. From a learning perspective, this project reinforced that thoughtful data preparation and iterative experimentation are key to building reliable and meaningful predictive models.

References

Frey, Bruno S., David A. Savage, and Benno Torgler. 2010. "Interaction of Natural Survival Instincts and Internalized Social Norms Exploring the Titanic and Lusitania Disasters." *Proceedings of the National Academy of Sciences* 107 (11): 4862–65.

Géron, Aurélien. 2019. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. 2nd ed. Sebastopol, CA: O'Reilly Media.

Gleicher, David, and Lonnie K. Stevans. 2004. "Who Survived the Titanic? A Logistic Regression Analysis." *International Journal of Maritime History* 16 (2): 61–94.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning: With Applications in R*. New York: Springer.

"Titanic - Machine Learning from Disaster." Kaggle. Accessed January 2026. https://www.kaggle.com/competitions/titanic.

Disclaimer

In accordance with the course Generative AI (GAI) policy, ChatGPT was used to assist with Python coding syntax, debugging, and clarifying implementation steps. Grammarly was used to correct grammar and rephrase some paragraphs for clarity. These tools were used strictly as support and did not replace my own analysis, reasoning, or decision-making.

All analytical decisions, feature engineering choices, model selection, interpretation of results, and conclusions are my own and reflect my understanding of the course material. All AI-assisted outputs were carefully reviewed, edited, and validated by me. In addition, general Google searches were used to reference common practices and background information when working with similar public datasets. No assignment instructions or solutions were copied directly from external sources.

This project is further informed by my prior academic and professional experience, including over seven years of work in data-related fields and previous experience as a computer vision researcher, where I conducted similar machine learning and analytical projects.