Titanic Dataset Machine Learning Report

Authors: Adrian-Nichita Zloteanu, Shafan Nazeer Ahmed

Institution: University of the Cumberlands

Course: Advanced Big Data and Data Mining

Date: 11.07.2025

Dataset Choice

For this project, we picked the Titanic dataset. It was an exciting dataset to work with, having a manageable amount of data and an interesting mix of numeric and categorical features. The survival outcome is easy to understand, and the data has enough quirks (missing values, skewed fares, social status) to make things interesting. Additionally, everyone is familiar with the story, making it easy to relate to the results.

Data preparation

To prepare the data, we decided first to drop the ‘Cabin’ column. It had too many missing values, though it would have been interesting to see which cabins had the highest survival rate. Then, we filled in the missing data for the Age and Embarked columns, using median and mode, respectively..

And finally, we removed duplicate rows and checked for some weird possible values, such as negative fares.

Exploratory Data Analysis

EDA revealed some interesting data: most passengers were in their 20s and 30s, but there were also infants and seniors present. “Fare” was very skewed; most paid under $50, but a few paid hundreds for 1st class accommodations.

Survival rates were much higher for women and first-class passengers. Age and fare, however, also played a role.

Feature Engineering

For feature engineering, we attempted to extract the title from the passenger names (e.g., Mr., Mrs.) to capture social status. This might lead to interesting results correlating higher-class titles with survival rates. We also created a FamilySize column to detect whether traveling alone or in a group affected survival rates.

Ultimately, we also log-transformed the Fare column to address the aforementioned skewness and stabilize performance, resulting in improved model performance.

Modeling Results

Regression

The goal of the Regression step was to predict a passenger's fare using features such as class, sex, age, family size, and title. Fare wasn't a critical target, but it worked well to demonstrate regression techniques.

For that purpose, we used two models: first we used a simpler linear regression model, and then a ridge regression model.

In this dataset, both models performed almost identically. Regularization did not add much value. The Ridge model performed almost identically to plain linear regression, which suggests the features weren't overfitting much to begin with.

Classification

The classification algorithm was the most interesting one. The goal here was to predict whether a passenger would survive the Titanic based on the above parameters. For that, we used two models, a decision tree and k-NN.

Both models confirmed the obvious: being female, traveling in first class, and being younger all boosted survival odds. Hyperparameter tuning (using GridSearchCV) helped optimize the Decision Tree, but the results remained essentially unchanged.

The Decision Tree achieved an accuracy of approximately 77%, while k-NN reached around 71%. The tree was also easier to interpret, so it ultimately proved to be the better option overall.

Clustering

The goal here was to identify natural groupings of passengers without directly using the survival column. I used K-Means and included `Sex` as one of the features. K-Means isn't ideal for categorical features, but it worked well enough for this case

The results were pretty interesting and in line with previous findings:

Cluster 0: All men, mostly in third class, traveling alone or in small groups. They had the lowest fares and the lowest survival rate.

Cluster 1: Mostly women in first class, older on average, with the highest fares. This group had the best survival rate.

Cluster 2: Mostly women and children in the third class, often traveling with bigger families. Their fares were mid-range, and their survival rate was decent.

Overall, gender, class, and group size had a significant influence on the outcome.

Association Rule Mining

The goal here was to find hidden patterns in the data using the Apriori algorithm. We

used a minimum support of 0.1 and a lift threshold of 1.2 to filter the rules. This gave us a focused set of high-interest patterns. We then looked at the top 10 rules with the highest lift.

The results matched what we've been seeing:

- Surviving was mostly tied to being female and in first class.
- Young adult men in third class, who had purchased cheap tickets, had the worst chances.
- Young women were way more likely to survive than young men.

These rules confirm the same patterns we saw earlier with class, gender, and age.

Practical Recommendations

This kind of analysis can be helpful beyond just the Titanic:

- Emergency planning should consider group-specific risk — for example, women and children had higher survival rates, but that was due to social norms, not safety protocols. Understanding who's at risk and why helps improve response strategies.
- Policies should be based on data like this to make safety procedures fairer and effective.
- Travel and insurance companies could use the same techniques to spot high-risk groups and offer more personalized services.
- It also demonstrates how data mining can corroborate historical accounts and aid in the study of social patterns in other contexts.

Ethical Considerations

Ethical Considerations

The Titanic dataset is publicly available and anonymized, so there are no real privacy concerns here. However, in real-world projects, you'd want to ensure that any personal or sensitive data is properly protected.

The models identified survival differences based on gender, socioeconomic class, and age. That lines up with how things worked in 1912, but in a modern context, blindly following patterns like that could cause problems. Any system like this should be thoroughly checked to ensure it’s not merely repeating historical biases.

This dataset reflects the social structure of the time — women and first-class passengers got priority. Our models learned that, which is fine for analysis, but in real-world use cases, you’d want to be careful how you apply that kind of logic.

What I did:

- Added features, such as title and family size, to provide a more detailed analysis.
- Looked at results in the context of history, not just what the model got right.
- Talked about how bias shows up and why it matters in real decision-making.

Conclusion

Ultimately, the Titanic dataset proved to be an effective means of exploring various machine learning techniques. Each method—regression, classification, clustering, and rule mining—helped reveal patterns that matched our expectations based on historical data. It was also a good reminder that models don't just predict; they reflect the data they're trained on, including biases. That's why it's just as important to interpret results as it is to build models.