**Laksh Budhrani & James Urbin**  
April 24, 2025  
CSCI B422: Data Mining  
Final Project Documentation

**Predicting Survival on the Titanic: A Machine Learning Approach**

**Introduction**

The sinking of the RMS Titanic on April 15, 1912, remains one of the most infamous shipwrecks in history. Of the estimated 2,224 passengers and crew aboard, only 722 survived. For our final project for "Data Mining B422," we were tasked with predicting whether an individual survives the sinking of the Titanic using the Kaggle Titanic dataset. We applied machine learning tools studied over the semester to identify key features influencing survival rates and to build predictive models for passenger survival outcomes. This project highlights practical applications of feature engineering, data preprocessing, and classification algorithms.

**Collaboration Approach**

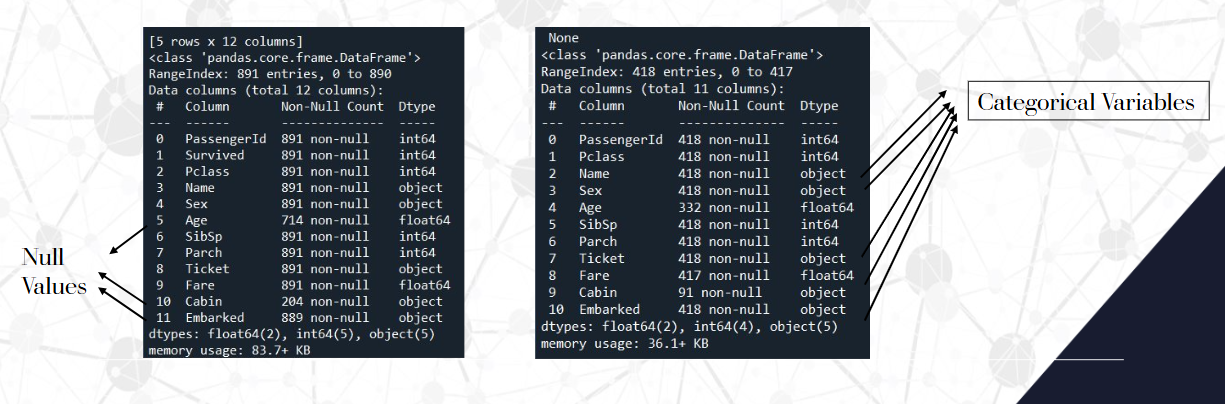
Throughout the last few weeks, we have spent a few hours discussing and working on the assignment. Initially, we had to do research on the 'Titanic' dataset. Our work focused on feature engineering, data preprocessing, model selection, and analyzing the eventual results. All decisions made were communicated to ensure both of us were happy with each new iteration.

**Timeline and Milestones**

* March 24- 28: Discussed dataset, looked over Kaggle website, discussed what we have learned in class thus far.
* March 31-April 3: Started working on the code collaboratively and looking at what is required and what features we should focus on
* April 4 - 8: Worked on feature engineering and seeing what average survival rates were on different features. Started preprocessing the data.
* April 9- 12: Model evaluation, submission preparation, started working on abstract
* April 13-15: Review all work completed along with going submitting abstract
* April 16-26: Optimizing the model, working on documentation, and preparing the video for the project.

**Data Exploration**

Our analysis began with thorough exploration of the dataset to understand its structure. The training dataset comprised 891 passengers with known survival outcomes, while the test set contained 418 passengers for whom we needed to predict survival.



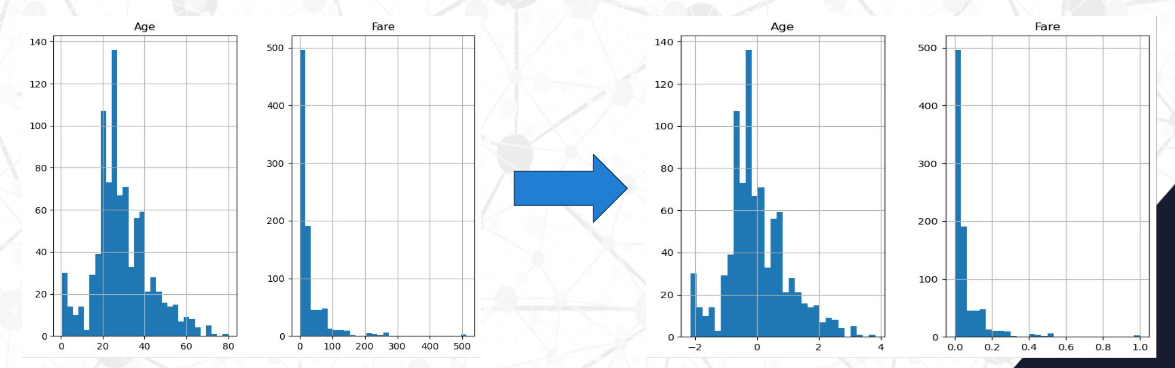
Initial exploration revealed several interesting insights:

* The dataset contained both numerical features (Age, Fare, SibSp, Parch) and categorical features (Sex, Pclass, Embarked)
* Several variables had missing values, particularly Age (19.9% missing), Cabin (77.5% missing), and Embarked (0.2% missing)
* Gender was encoded as a binary variable, allowing male and female passengers to be represented numerically for analysis.
* Cabin data was simplified into a "Has\_Cabin" feature to indicate whether a passenger had a recorded cabin, reducing complexity while preserving meaningful information.
* The port of embarkation was processed using one-hot encoding to effectively represent the categorical differences without imposing any order.
* Titles extracted from passengers' names provided additional context regarding their social status, making this feature potentially relevant for survival prediction.
* Ticket numbers and passenger IDs were excluded from the analysis as they lacked any meaningful predictive value.

**Data Preprocessing**

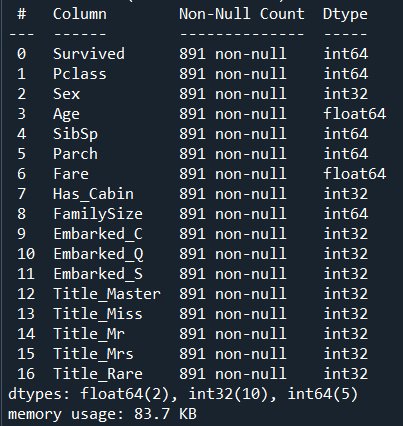
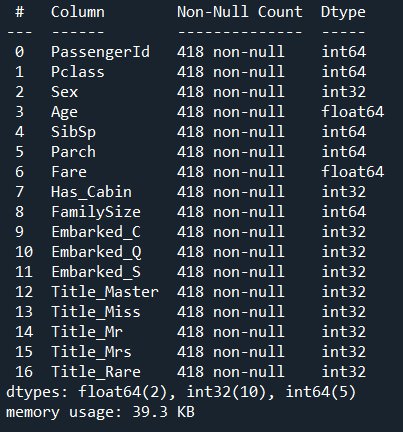
Our preprocessing pipeline addressed several challenges in the raw dataset:

1. **Missing Value Imputation**:
   * Age values were imputed with the median age grouped by Pclass and Sex, providing more accurate estimations based on passenger demographics
   * Missing Embarked values were filled with the mode (most common embarkation point)
   * Missing Fare values in the test set were imputed using the median fare
2. **Feature Transformation**:
   * Categorical variables like Sex were encoded using Label Encoding (Female = 0, Male = 1)
   * We employed Min-Max scaling for Fare to normalize values between 0 and 1
   * Age values were standardized using Z-score normalization
   * Categorical variables with multiple classes like Embarked and Title were transformed using one-hot encoding



1. **New Feature Creation**:
   * Created a Has\_Cabin binary feature (1 if Cabin exists, 0 otherwise) rather than using the original Cabin feature with numerous unique values
   * Generated a FamilySize feature by combining SibSp (siblings/spouse) and Parch (parents/children) counts and adding 1 for the passenger
   * Extracted Title from passenger names using regular expressions and mapped rare titles to broader categories

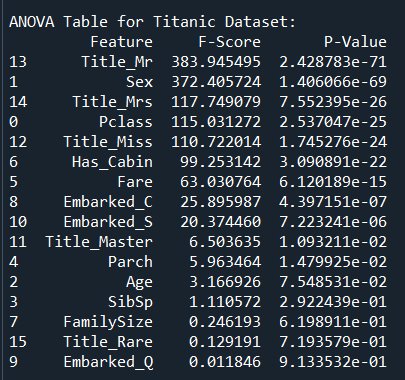
These preprocessing steps transformed raw, inconsistent data into a clean, normalized dataset suitable for machine learning algorithms.

**Feature Selection**

We employed multiple techniques to identify the most predictive features:

1. **ANOVA F-Test**: We used the f\_classif function to compute F-scores and p-values for each feature, which helped identify statistically significant predictors of survival. This analysis highlighted Title\_Mr, Sex, Title\_Mrs, Pclass and Title\_Miss as particularly important features.



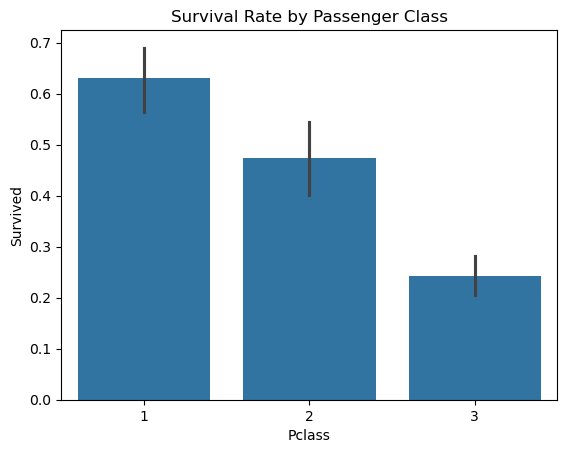
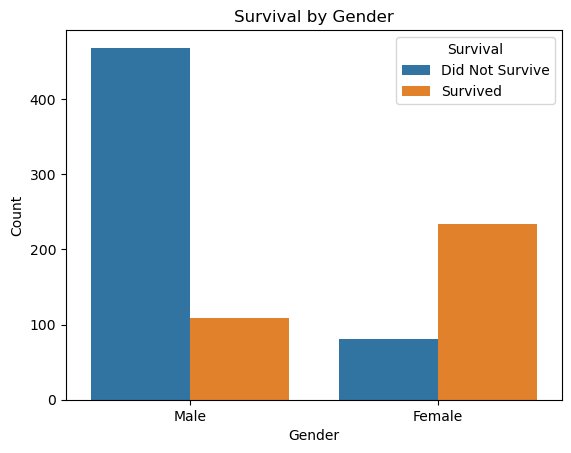
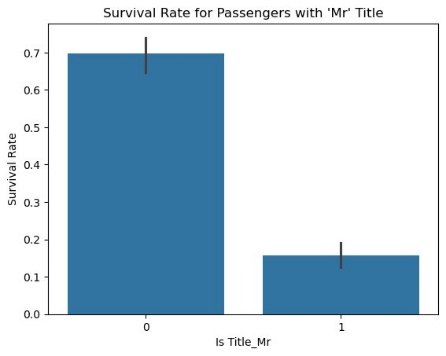
1. **ElasticNet Regularization**: We applied ElasticNet (combining L1 and L2 regularization) to identify relevant features while handling multicollinearity. Features with non-zero coefficients were selected for model training, effectively reducing dimensionality and reducing overfitting.

Through this process, we identified the following key predictors of survival:

* Sex (being female significantly increased survival chances)
* Title\_Mr (having the title "Mr." was negatively correlated with survival)
* Pclass (higher passenger class correlated with better survival rates)



Visualization of these relationships further confirmed our feature selection approach, showing clear patterns between these variables and survival outcomes.

**Technical Approach and Model Selection**

We've focused on making new features based on family size, title extraction from names, ticket information, and cabin information. For modeling, we implemented and compared six different classification algorithms:

1. Random Forests: An ensemble method that typically offers robust performance
2. Decision Trees: Provides clear decision rules and feature importance
3. K-Nearest Neighbors: A non-parametric approach that performs well on smaller datasets
4. Support Vector Machines (SVC): Effective for complex decision boundaries
5. Logistic Regression: A baseline model offering good interpretability
6. Gaussian Naive Bayes: A probabilistic classifier that works well with high-dimensional data
7. Neural Network using Keras with TensorFlow backend: A deep learning model capable of capturing complex non-linear relationships in data

**Results**

Our model evaluation showed varying performance across different algorithms:

* SVC and Decision Tree achieved the highest accuracy at 78.23%
* Random Forest, Logistic Regression, Gaussian Naive Bayes, and Neural Network all achieved 77.27% accuracy
* K-Nearest Neighbors showed the lowest performance at 76.56%

A screenshot of a computer

AI-generated content may be incorrect.

Our analysis of the Titanic dataset showcased how both simple and complex models can be used to predict survival outcomes, with most models achieving similar accuracy (around 77–78%). Interestingly, more advanced models like Random Forest and Neural Networks did not significantly outperform simpler methods such as Decision Trees and SVC. This may indicate that the dataset’s patterns are relatively straightforward or that the limited sample size (891 passengers) restricts complex models from capturing deeper insights. The findings emphasize that feature engineering and preprocessing often have a greater impact on performance than model complexity. Key takeaways from the project include:

* Demographic variables, including gender and the Title "Mr.", emerged as significant indicators of survival.
* Socioeconomic features, such as passenger class (Pclass), demonstrated considerable predictive power as well.
* Feature engineering and data preprocessing were critical to achieving higher accuracy
* Model selection was secondary to the quality and relevance of the input features

This project successfully applied the data mining techniques and methodologies learned throughout the semester to a real-world dataset, demonstrating both the power and limitations of predictive modeling in historical analysis.