Project Report: Analysis of the Titanic Dataset

Introduction

This report provides a comprehensive analysis of the Titanic dataset, a classic in the field of data science and machine learning. The project explores various data mining techniques to extract meaningful insights from this historical data. The analysis covers data preprocessing, exploratory data analysis, regression, classification, clustering, and association rule mining. The report concludes with practical recommendations based on the findings and a discussion of the ethical considerations involved in the project.

1. The Dataset and Its Significance

The project utilizes the well-known Titanic dataset, which contains information about the passengers aboard the RMS Titanic, which tragically sank in 1912. The dataset includes demographic and travel-related information for each passenger, such as age, sex, passenger class, fare paid, and whether they survived the disaster.

This dataset was chosen for its richness and complexity, making it an excellent case study for a wide range of data mining tasks. It contains a mix of numerical and categorical data, as well as missing values, which provides a practical opportunity to apply various data cleaning and preprocessing techniques. Furthermore, the clear and compelling objective of predicting passenger survival makes it an ideal dataset for demonstrating the power of classification models. The historical context of the data also allows for a nuanced discussion of social structures and their impact on survival, adding a layer of depth to the analysis.

2. Data Preprocessing, EDA, and Feature Engineering

The initial phase of the project focused on preparing the data for analysis. This involved several key steps:

- **Data Cleaning:** The dataset had missing values in the 'Age', 'Cabin', and 'Embarked' columns. The 'Age' was imputed with the median age of the passengers, a robust measure against outliers. The 'Cabin' column, having a significant number of missing values, was dropped from the dataset. The few missing 'Embarked' values were filled with the mode, which is the most frequent port of embarkation.
- **Exploratory Data Analysis (EDA):** EDA was conducted to uncover patterns and relationships within the data. Visualizations were key to this process. Bar charts of survival counts by passenger class and sex revealed stark differences. A significantly higher proportion of first-class passengers and female passengers survived. This initial analysis strongly suggested that socioeconomic status and gender were critical factors in determining survival.
- **Feature Engineering:** To enhance the predictive power of the models, new features were created from the existing data. 'Age' was binned into categorical groups ('Child', 'Teen', 'Young Adult', 'Adult', 'Senior') to better capture the non-linear relationship between age and survival. A 'FamilySize' feature was also created by combining the 'SibSp' (number of siblings/spouses aboard) and 'Parch' (number of parents/children aboard) columns. This allowed for an analysis of how family size influenced survival, rather than looking at the two variables in isolation.

3. Modeling and Results

The project employed a variety of data mining techniques, each providing a different lens through which to view the data.

- **Regression Analysis:** The goal of the regression analysis was to predict the 'Fare' paid by passengers based on features like 'Pclass', 'Sex', 'Age', 'Embarked', and 'FamilySize'. Three models were compared: Linear Regression, Ridge Regression, and Lasso Regression. The models were evaluated using R^2, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). The Ridge regression model performed the best, with a slightly higher R^2 value of 0.4249, indicating that it explained about 42.5% of the variance in fare prices. The regularization provided by the Ridge model likely helped to prevent overfitting and improve its predictive accuracy.
- **Classification Analysis:** For the classification task, the objective was to predict whether a passenger 'Survived'. Two models were used: a Decision Tree and a k-Nearest Neighbors (k-NN) classifier. The Decision Tree model was optimized using GridSearchCV to find the best hyperparameters, which were determined to be a max_depth of 5 and min_samples_split of 10. The k-NN model had a slightly higher F1-score (0.6667) than the Decision Tree (0.6435), suggesting it was better at balancing precision and recall. However, the Decision Tree had a higher accuracy (0.7709) than the k-NN model (0.7374). The ROC curves for both models were also plotted, showing that both performed significantly better than a random classifier.
- **Clustering Analysis:** K-Means clustering was used to segment the passengers into three distinct groups based on their characteristics, without using the 'Survived' label. The features used for clustering were 'Pclass', 'Sex', 'Age', 'Embarked', 'SibSp', 'Parch', and 'Fare'. The resulting clusters were analyzed to understand their defining characteristics.

    - **Cluster 0:** Characterized by younger passengers with large families and lower fares, predominantly in third class.
    - **Cluster 1:** Comprised of older, wealthier passengers, often traveling with smaller families in first class.
    - **Cluster 2:** Represented passengers with average age, smaller family sizes, and lower fares, mostly in third class. This clustering provides a useful segmentation of the passengers, which could be used for more targeted analysis or for understanding the different passenger profiles on the Titanic.
- **Association Rule Mining:** This technique was used to find interesting relationships in the data, with a focus on rules that had 'Survived' as the consequent. The results showed strong associations between survival and being female, particularly for those in first and second class. For example, the rule {female, Pclass=1} -> {Survived} had a very high confidence of 0.968, meaning that first-class female passengers had a 96.8% chance of survival. This reinforces the findings from the EDA and classification analysis, providing a clear and interpretable set of rules that highlight the key factors driving survival.

4. Practical Recommendations and Insights

The insights gleaned from this analysis can be translated into practical recommendations for similar emergency situations:

- **Prioritization in Evacuation:** The data overwhelmingly shows that women and children, especially those in higher-class accommodations, had a significantly higher chance of survival. In a modern-day crisis, this historical data could inform evacuation protocols, suggesting that prioritizing vulnerable groups can have a substantial impact on outcomes.

- **Resource Allocation:** The correlation between passenger class and survival suggests that access to lifeboats and other resources was not evenly distributed. This highlights the importance of equitable resource allocation in emergency planning to ensure that everyone has a fair chance of survival, regardless of their socioeconomic status.

5. Ethical Considerations

While the Titanic dataset is historical and publicly available, it is still important to consider the ethical implications of the analysis:

- **Data Privacy:** The dataset contains personal information, and although the individuals are no longer living, it serves as a reminder of the importance of data privacy and the responsible handling of personal data in any data science project.
- **Fairness and Bias:** The dataset reflects the societal biases of its time. The "women and children first" protocol, while seemingly chivalrous, also reflects a patriarchal society. The strong influence of passenger class on survival is a stark reminder of social inequality. It is crucial to be aware of these biases when building predictive models. A model trained on this data could, if not carefully handled, perpetuate these biases. For example, a model that simply learns that "first-class passengers survive" could be seen as unfair if used to make decisions in a different context.
- **Addressing Concerns:** The project implicitly addresses these concerns by not just building predictive models, but by using the analysis to highlight and understand these historical biases. The EDA and association rule mining, in particular, serve to make these patterns explicit. By understanding the "why" behind the predictions, we can be more critical of the models and their potential applications. The project does not shy away from

the uncomfortable truths in the data but rather uses them as a point of discussion and learning.

By being mindful of these ethical considerations, the project not only demonstrates technical proficiency but also a responsible and thoughtful approach to data analysis.