Analyzing the Titanic Dataset with Logistic Regression in Rust

The dataset used for this final project was Titanic: Machine Learning from Disaster. This dataset contained information on the passengers aboard the Titanic when it sank, including their demographics, ticket class, cabin, and whether or not they survived. This is interesting to me since this dataset provides an opportunity to explore the relationships between these different factors and survival rates and to build predictive models to determine which passengers were most likely to survive.

The problem I am trying to solve is to predict the survival of passengers aboard the Titanic based on various factors such as their demographics, ticket class, cabin, and fare paid. This could help in understanding the factors that played a critical role in the survival of passengers during the disaster.

The results showed that certain factors significantly determined the passengers' survival rate. For instance, the passenger class (Pclass) showed a strong correlation with the survival rate, as passengers in higher classes (1st class) had a higher survival rate. Additionally, it was observed that female passengers had a higher survival rate than male passengers, which is likely due to the "women and children first" protocol during the evacuation. Furthermore, age had a varying impact on survival, with children having a higher survival rate.

To predict passenger survival, I implemented a logistic regression algorithm in Rust. Logistic regression is a binary classification algorithm that is suitable for predicting the probability of an outcome based on a set of input features. The data was pre-processed by handling missing values and converting categorical variables into numerical ones. After preprocessing, the logistic regression algorithm was applied using gradient descent to learn the weights associated with each feature.

A scatter plot was created to visualize the correlation between passenger class, age, and survival rate. The scatter plot clearly shows the differences in survival rates among the different passenger classes and age groups.

While our current implementation provides valuable insights into the Titanic dataset, there are some challenges and potential improvements. Firstly, some data points have missing values, particularly in the Age column. We filtered out these data points, but more advanced techniques like data imputation could be employed to retain and utilize all available data.

Secondly, we only considered a few features for our analysis. However, other features like fare, cabin, and embarked could also be relevant in determining survival rate. More advanced feature engineering techniques could be used to extract additional information from the dataset and improve prediction accuracy.

In conclusion, our analysis of the Titanic dataset using logistic regression in Rust provided insights into the factors affecting passenger survival. The implementation showcased the capabilities of Rust for machine learning and data analysis tasks. Future work could focus on improving data preprocessing and feature engineering to enhance prediction accuracy and gain further insights into the dataset.