SOCIAL INEQUALITY STUDY

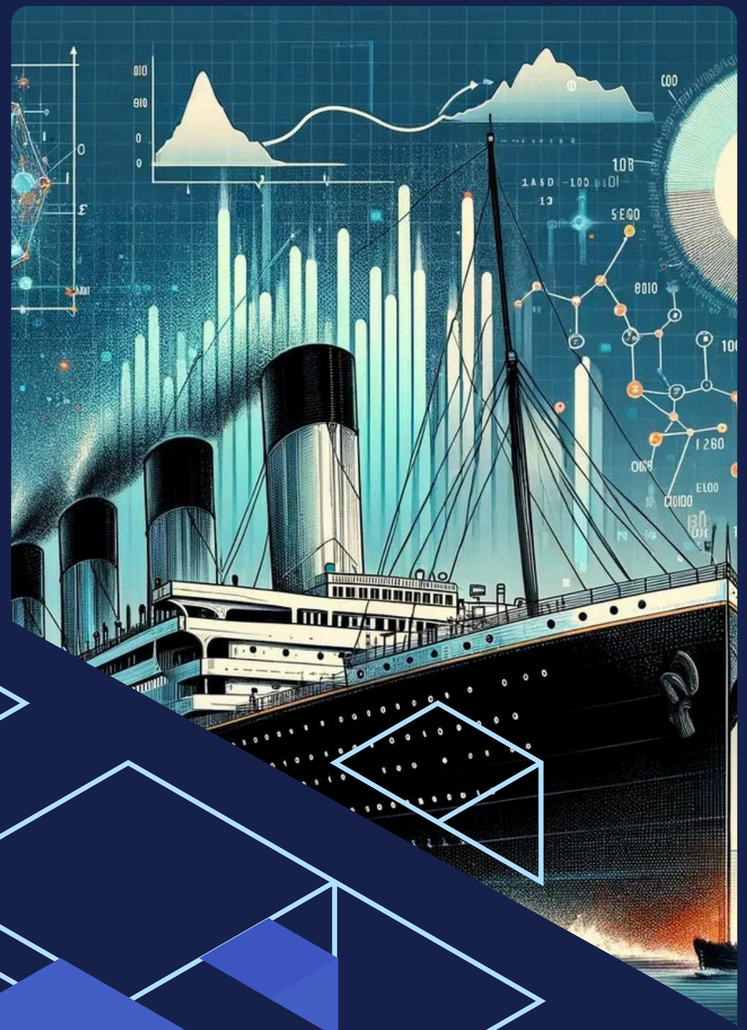
USING MACHINE LEARNING AND TITANIC DATASET

ABAROUDI YOUNES

09/16/2024

Agenda

- Introduction
- Dataset Overview
- Exploratory Data Analysis (EDA)
- Model 1: Shallow Artificial Neural Network (ANN)
- Model 2: Multiclass Logistic Regression
- Conclusion and Key Takeaways



Social Inequality Study

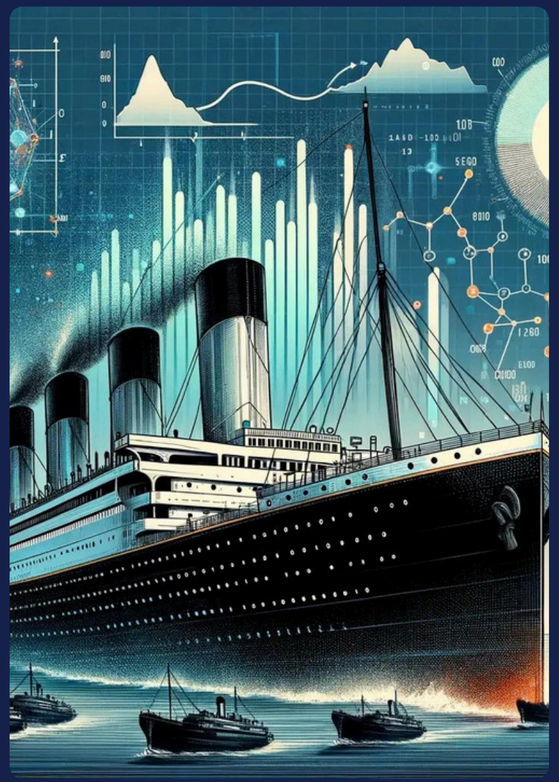
Predicting Survival Using Machine Learning Models

In this project, we will explore **social inequality** in the Titanic dataset by examining how passenger class, gender, and other factors influenced survival rates.

- Focus Areas:
  - Social class (Pclass)
  - Gender
  - Family size
  - Wealth (Fare)

- Models Used:
  - Shallow Artificial Neural Network (ANN)
  - Multiclass Logistic Regression

Dataset Overview



Dataset Summary

- Total passengers: 891 passengers
- Key Features:
  - Pclass, Sex, Age, Fare, Embarked, Family Size, Title
- Target Variable: Survived (0 = No, 1 = Yes)

Data Cleaning

- Removed 95 duplicate rows

Feature Engineering

- Combine (“Pclass_1”, “Pclass_2”, “Pclass_3”) into a single feature “Pclass”

Exploratory Data Analysis (EDA)

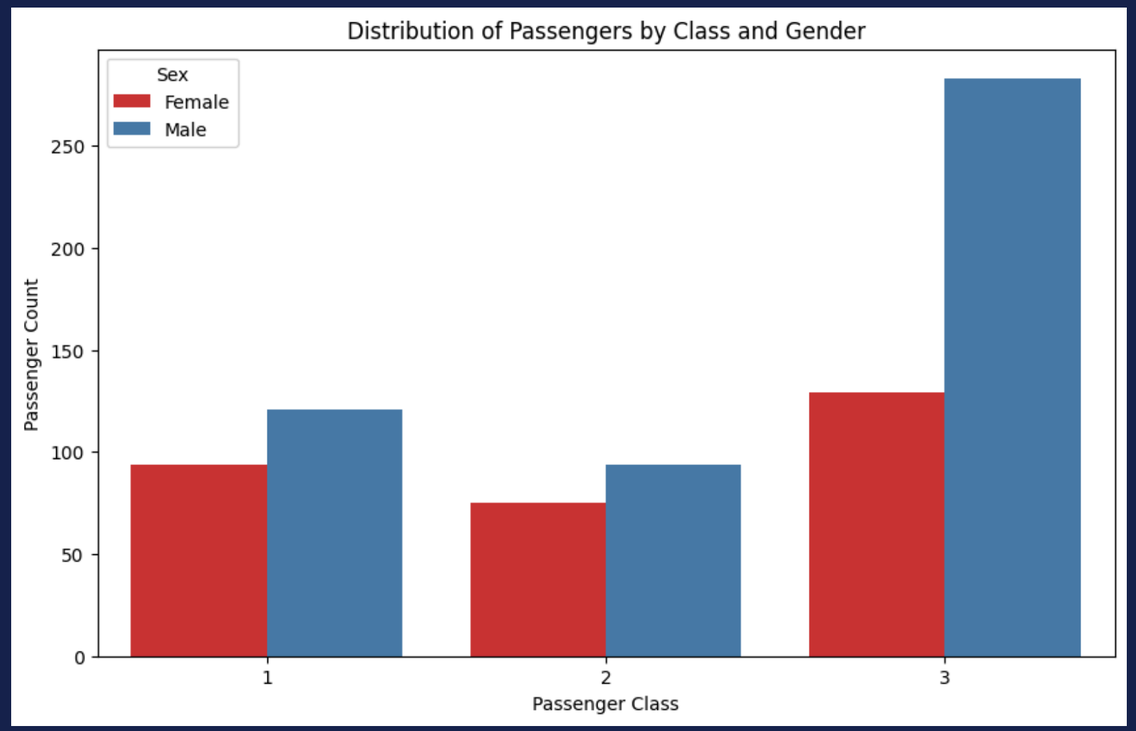
Influence of Fare on Survival

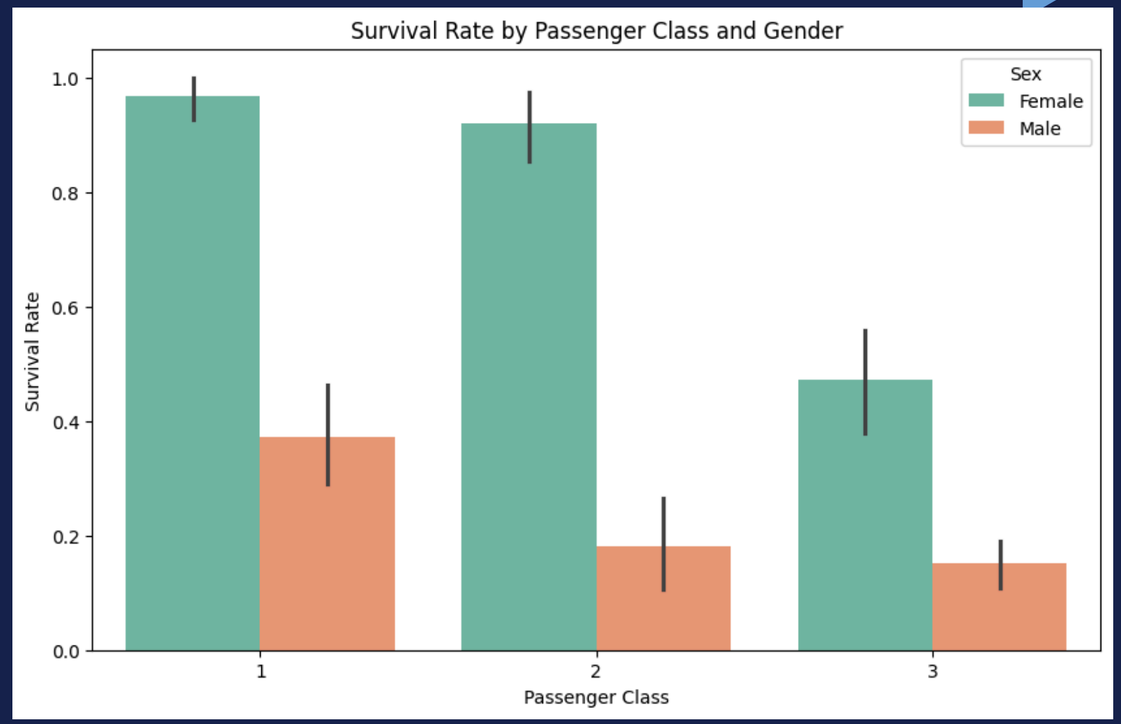
Influence of Family Size on Survival

Survival Rate Distribution Across Classes and Genders

Correlation Matrix for Social Inequality Features

Survival Rate Distribution Across Classes and Genders

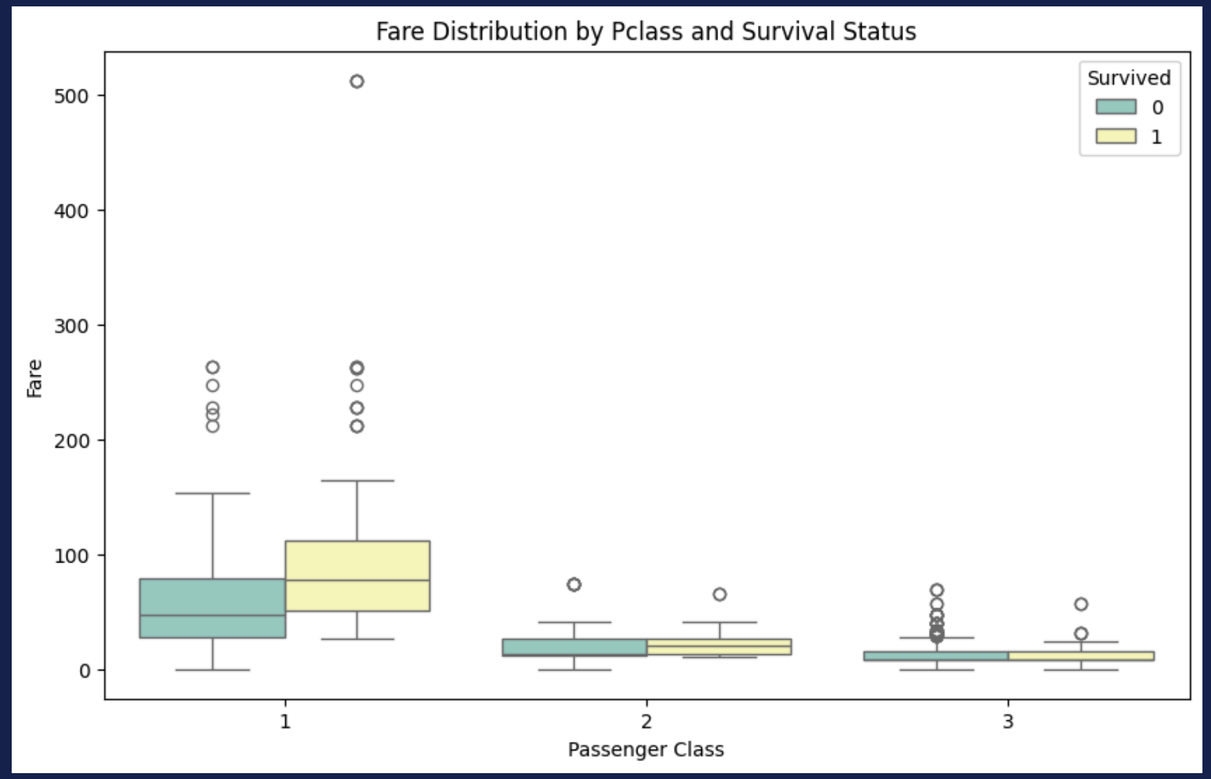




OBSERVATIONS

- **Wealthier passengers (first class)** had better access to lifeboats, while passengers from lower classes (third class) had a much lower chance of survival.
- **Gender** also played a significant role, as females were given priority during the rescue, leading to significantly higher survival rates among women.

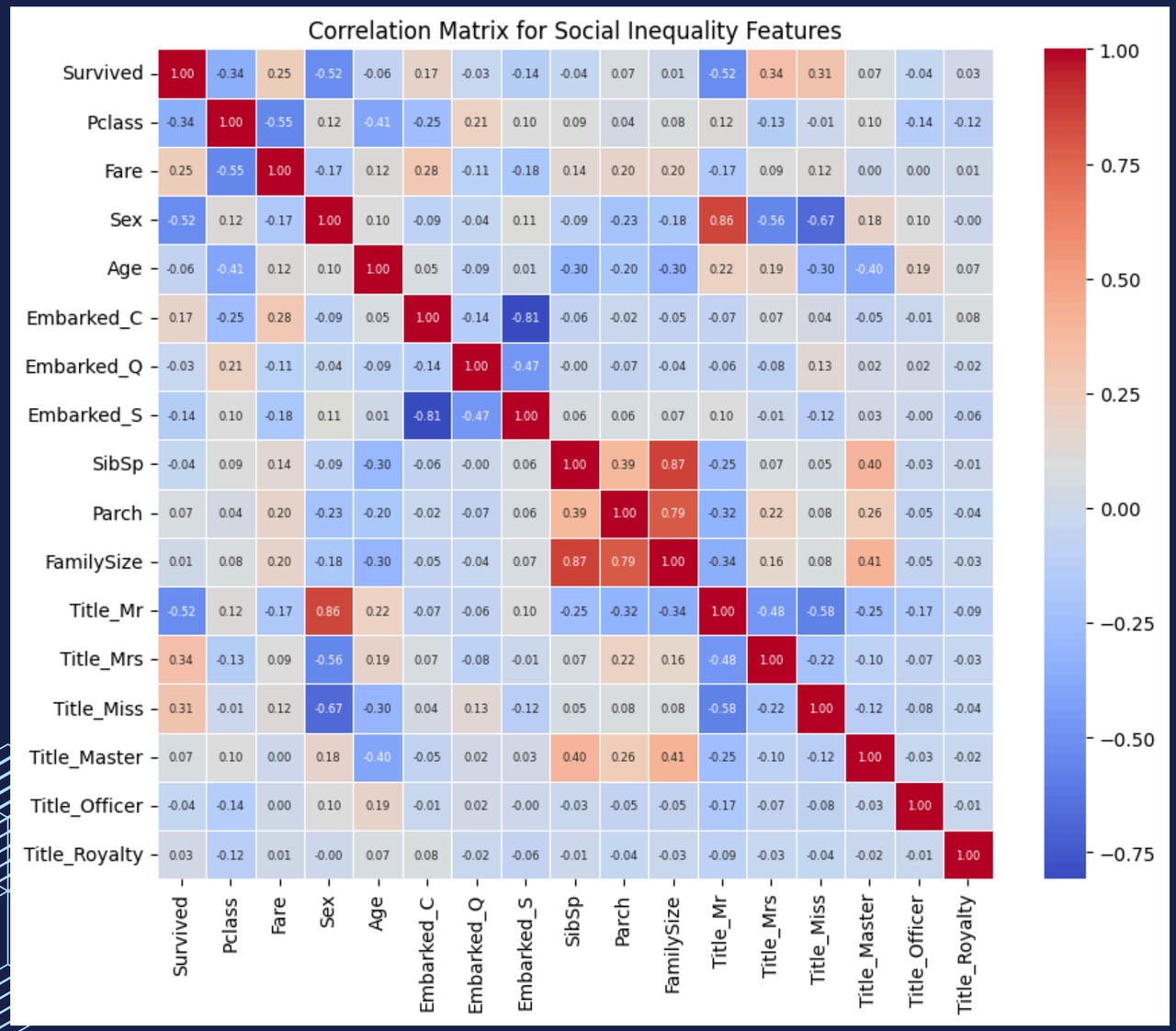
Influence of Fare on Survival



OBSERVATIONS

- **Fare is a proxy for wealth**, and wealthier passengers, who had better accommodations, were more likely to have access to lifeboats, contributing to higher survival rates.

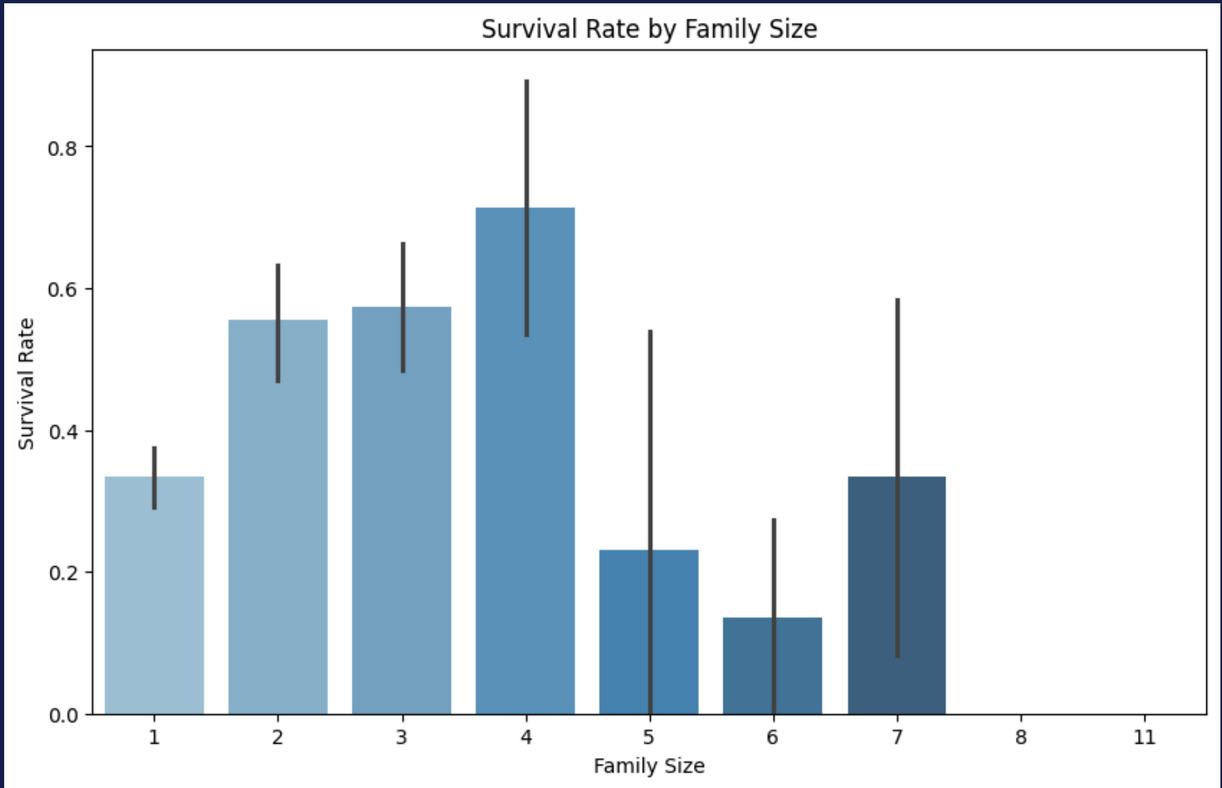
Correlation between Social Inequality Features



OBSERVATIONS

- **Passenger class and fare are key indicators of wealth, and wealth had a substantial impact on survival.**
- **Gender-based societal norms during the time of the Titanic disaster are evident in the correlation between `Sex` and survival, as women had a higher survival rate.**

Influence of Family size on Survival



OBSERVATIONS

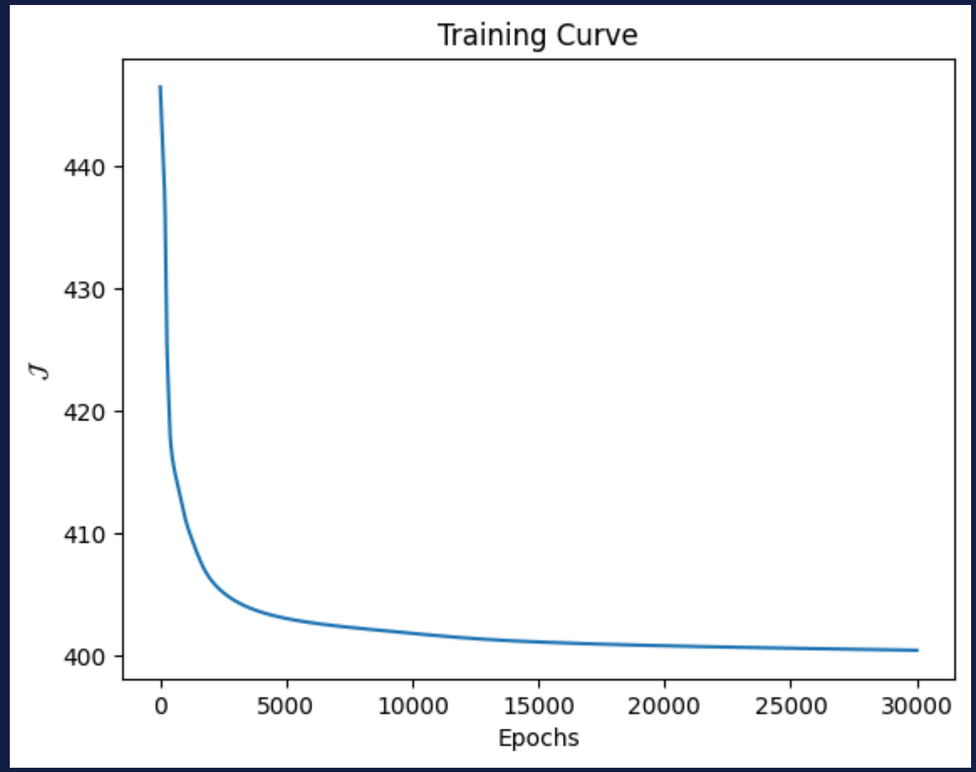
- Traveling alone or in small groups appears to have been an advantage during evacuation, while larger families faced more challenges, leading to lower survival rates.

Model 1: Shallow Artificial Neural Network (ANN)

Model Description

We implemented a Shallow ANN with the following architecture:

- Input layer: Corresponding to the number of input features.
- Hidden layer: 6 neurons using the `tanh` activation function.
- Output layer: Softmax function to handle multiclass classification.



Result

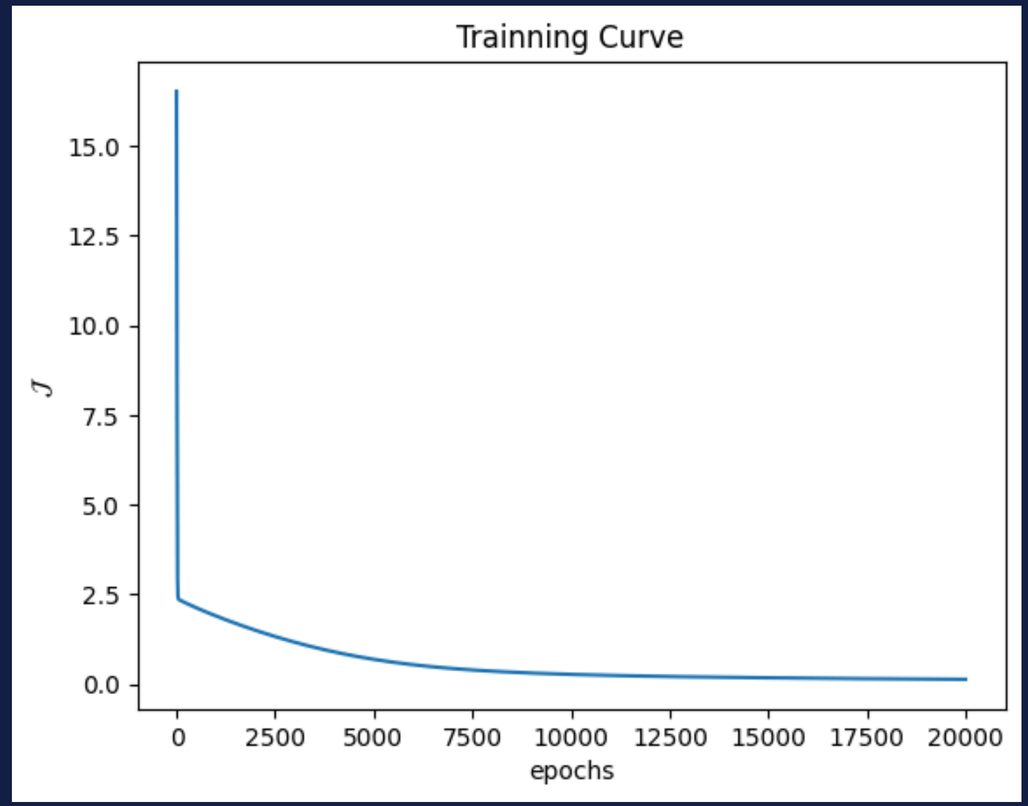
- After training the model for 30,000 iterations, the Shallow ANN achieved an accuracy of 69.38%.

Model 2: Multiclass Logistic Regression

Model Description

We implemented a Multiclass Logistic Regression using softmax activation for multiclass classification. We focused on optimizing the model by:

- Feature engineering the `Pclass` feature.
- Fine-tuning hyperparameters such as the learning rate (`eta`) and the number of epochs.



Result

- Without feature engineering `Pclass` (using one-hot encoding), the logistic regression model achieved an accuracy of 58%.
- - After feature engineering `Pclass`, the accuracy increased significantly to 95.6%

SUMMARY

- The Multiclass Logistic Regression model performed best with an accuracy of 95.6% after feature engineering the `Pclass` feature. This demonstrates the importance of preserving the ordinal relationship between passenger classes in the dataset.
- The Shallow ANN, even with hyperparameter tuning, reached an accuracy of 69.38%, which is lower compared to logistic regression.
- Feature engineering and hyperparameter tuning played a critical role in improving model performance, particularly in the logistic regression model.

THANK YOU