Problems

1. Survival Analysis

2. Exploratory Data Analysis (EDA)

- Demographic Analysis: Understanding the demographics of the passengers, such as age distribution, gender distribution, and class distribution.
- Social-Economic Status (SES) Analysis: Exploring how SES, represented by passenger class and fare, affected survival rates.
- Family Dynamics: Analyzing the impact of family size (sibsp: # of siblings/spouse aboard, parch: # of parents/children aboard) on survival chances.

3. Feature Engineering

- Creating New Features: Deriving new features like family size, is_alone (based on family size), title extracted from name, and deck extracted from the cabin number to see if they provide additional predictive power.
- Categorical Feature Encoding: Exploring different ways to encode categorical variables (like embarked port, sex) and their impacts on model performance.

4. Statistical Analysis

- Correlation Analysis: Investigating the correlation between different features and survival. For example, checking if fare and passenger class are correlated to survival rates.
- Hypothesis Testing: Testing hypotheses about survival rates across different groups (e.g., comparing survival rates between genders, passenger classes, or age groups).

5. Model Comparison and Evaluation

- Comparing Models: Building various machine learning models and comparing their performance in predicting survival.
- Evaluation Metrics: Exploring different evaluation metrics such as accuracy, precision, recall, F1 score, ROC AUC to understand model performance.

6. Clustering

- Passenger Segmentation: Using unsupervised learning techniques (like K-means or hierarchical clustering) to segment passengers into groups based on features like age, fare, or family size, and analyzing if these segments show different survival patterns.

7. Missing Value Imputation

- Handling Missing Data: Investigating strategies for handling missing data (e.g., age, cabin, embarked) such as imputation with mean/median/mode, predicting missing values, or using algorithms that can handle missing values directly.

8. Time Series Analysis

Although the Titanic dataset is not a time series dataset, if timestamped data on events leading up to the sinking were available, one could explore time series analysis to study patterns over time (e.g., the timeline of when people boarded lifeboats).

9. Text Analysis

- Analyzing Names or Ticket Information: Extracting and analyzing textual information, such as titles from names or patterns in ticket numbers, to see if they offer insights into social status, family relations, or survival likelihood.

10. Ethical Analysis

- Bias and Fairness: Discussing the ethical aspects of model predictions and the fairness of decision-making processes that could have been influenced by biases in features like class, sex, or age.

Methods

1. Exploratory Data Analysis (EDA)

- Visualization: Utilize graphs such as histograms, box plots, scatter plots, and pie charts to understand the distribution and relationship between variables.

- Summary Statistics: Generate summary statistics like mean, median, mode, standard deviation, and quartiles to get insights into the data's central tendency and dispersion.

2. Data Preprocessing

- Handling Missing Values: Techniques include imputation (using mean, median, or mode), deletion, or prediction models to fill in missing values.
- Feature Engineering: Creating new features from existing ones to improve model performance, such as extracting titles from names, or engineering family size from SibSp and Parch.
- Normalization and Standardization: Adjusting the scale of features to normalize the distribution or standardize the range.
- Encoding Categorical Variables: Converting categorical variables into numerical values using methods like one-hot encoding, label encoding, or binary encoding.

3. Statistical Analysis

- Correlation Analysis: Identifying relationships between features using Pearson correlation, Spearman's rank correlation, or Kendall's tau.
- Hypothesis Testing: Performing t-tests, chi-squared tests, or ANOVA to test hypotheses about the data.

4. Machine Learning Models for Classification

- Logistic Regression: A linear model for binary classification tasks.
- Decision Trees: A model that uses branching methods to represent decisions and decision making.
- Random Forests: An ensemble method using multiple decision trees to improve classification accuracy.
- Support Vector Machines (SVM): A powerful classifier that works well in high-dimensional spaces.
- K-Nearest Neighbors (KNN): A simple, instance-based learning algorithm where the class of a sample is determined by the majority class among its k nearest neighbors.
- Gradient Boosting Machines (GBM) and XGBoost: Ensemble techniques that build models sequentially to correct the errors of previous models.
- Neural Networks(potentially, depends on class material): Using basic feedforward neural networks or more complex architectures like convolutional neural networks (CNNs) for classification, depending on feature engineering and data representation.

5. Model Evaluation and Selection

- Cross-Validation: Techniques like k-fold cross-validation to assess how the model's results generalize to an independent dataset.
- Performance Metrics: Metrics such as accuracy, precision, recall, F1 score, ROC-AUC curve for classification tasks to evaluate and compare model performances.
- Confusion Matrix: A table used to describe the performance of a classification model on a set of test data for which the true values are known.

Datasets

[Titanic - Machine Learning from Disaster | Kaggle]

Work Distribution (Impromptu)

Week 1-2: Data Acquisition and Initial Exploration Both: Download the dataset from Kaggle or another source. Partner 1: Focus on understanding basic features like Age, Sex, Pclass. Partner 2: Explore more intricate features like Cabin, Embarked, Fare. Outcome: Gain a comprehensive understanding of the dataset's features.	**Week 3-4**: Data Cleaning and Preprocessing Both: Work on handling missing values, feature encoding, and data normalization. Partner 1: Specifically focus on handling missing values and outliers. Partner 2: Concentrate on feature engineering and categorical data encoding. Outcome: Clean and preprocessed dataset ready for analysis.	**Week 5-6**: In-Depth Data Analysis and Visualization Both: Perform detailed data analysis with a focus on different aspects. Partner 1: Analyze demographic data (Age, Sex, Pclass) and visualize their impact. Partner 2: Investigate socio-economic aspects (Fare, Cabin, Embarked). Outcome: Deeper insights into how different features influence survival rates.
Week 7-8: Model Building and Initial Testing Partner 1: Implement traditional classification methods (Logistic Regression, Decision Trees, Random Forest). Partner 2: Develop a basic neural network model for comparison. Both: Evaluate initial models using accuracy, precision, recall. Outcome: A set of baseline models for survival prediction.	**Week 9**: Model Refinement and Advanced Testing Partner 1: Fine-tune traditional models and experiment with ensemble methods. Partner 2: Optimize the neural network architecture and parameters. Outcome: Improved model performance with fine-tuned parameters.	**Week 10**: Final Evaluation and Documentation Both: Compare all models and compile findings. Partner 1: Lead the writing of the methodology and model evaluation sections. Partner 2: Focus on results, discussion, and conclusion. Both: Prepare a joint presentation or report summarizing the project. Outcome: A comprehensive report and presentation of the project findings.