Modeling Mathematics

Project 1 - Statistical Learning Methods

This project applies statistical learning techniques to predict Titanic passenger survival.

Dataset Selection and Documentation

Chosen Dataset: Titanic - Machine Learning from Disaster

- **Source:** https://www.kaggle.com/datasets/shuofxz/titanic-machine-learning-from-disaster
- **Reason for Selection:** The dataset is well-documented, publicly available, and widely used for classification problems, making it suitable for evaluating multiple statistical learning methods.
- **Description:** The dataset includes information about Titanic passengers, such as age, gender, ticket class, fare, and whether they survived the disaster.

Titanic Dataset because it covers classification, missing data handling, and feature engineering while being well-documented.

Key Questions:

1. What factors were most important in determining passenger survival?
2. Can we predict passenger survival using classification models?
3. Can clustering techniques reveal natural groupings of passengers based on characteristics like ticket class and fare?
4. How do different statistical learning models compare in terms of accuracy and interpretability?
5. **Deep Learning:** Can a neural network outperform traditional ML models

Hypotheses:

- **H1:** Female passengers had a significantly higher survival rate than males.
- **H2:** First-class passengers had a higher chance of survival than lower-class passengers.
- **H3:** Random Forest will outperform Logistic Regression in classification accuracy.
- **H4:** K-Means clustering will reveal meaningful passenger groupings related to survival.

Light Research on Existing Studies

- **Historical records confirm that first-class passengers and women had higher survival rates** due to evacuation protocols.
- **Previous studies using machine learning confirm that Random Forest models often perform best** on this dataset.
- **Clustering can help analyze passenger demographics but is not a predictive tool for survival.**

Preprocessing & Feature Engineering

- Handle **missing values** (`Age`, `Cabin`, `Embarked`).
- Convert categorical features (`Sex`, `Embarked`) into numerical format.
- Feature scaling (optional).

Data Preprocessing (Enhancements & Visualizations)

Handling Missing Data

We must discuss why we used **median/mode imputation** for missing values. Example:

- **Age:** Median is used to avoid bias.
- **Cabin:** Too many missing values → We **dropped it** instead

Statistical learning methods provide structured ways to **make predictions from data**. The Titanic dataset is a **classification problem** (survived or not), making it ideal for machine learning techniques.

Data Preprocessing

1. **Handled Missing Values:**
   a. Replaced missing age values with the median age.
   b. Filled missing embarkation points with the most frequent value.
2. **Feature Engineering:**
   a. Converted categorical variables (Sex, Embarked) into numerical format.
   b. Created a new feature: FamilySize = SibSp + Parch + 1.
3. **Data Normalization:**
   a. Scaled continuous variables (Age, Fare) to standardize distributions.

Model Selection & Justification

We need **three methods** from "An Introduction to Statistical Learning":

1. **Logistic Regression** (for classification).
2. **Random Forest Classifier** (for feature importance).
3. **K-Means Clustering** (for grouping passengers).

LogisticRegression	Good for binary classification (Survived = 1, Not Survived = 0).
Random Forest	Handles non-linearity & ranks feature importance.
KMeansClustering	Groups passengers into categories based on behavior (unsupervised).

Methodology

- Logistic Regression for binary classification
- Random Forest for feature importance & better prediction
- K-Means Clustering for passenger segmentation

Key Findings

- Women had a 3x higher survival rate than men
- First-class passengers had the best survival chances
- Random Forest outperformed Logistic Regression (Accuracy: 81% vs 78%)
- Clustering revealed 3 unique passenger groups.

Improvements

- Use Neural Networks for more accuracy
- Tune hyperparameters further
- Use external data (e.g., weather conditions on the Titanic).

Model Overview

Mode	Type	Use Case	Strengths	Weaknesses
Logistic Regression	Supervised Classification	Binary classification (Survived/Not Survived)	Simple, interpretable, and fast	Assumes linear relationships
Random Forest	Supervised (Ensemble Learning)	Predict survival and feature importance	Handles complex data, avoids overfitting, good accuracy	Slower, harder to interpret
K-Means Clustering	Unsupervised (Clustering)	Finds hidden passenger groups	Uncovers natural patterns in data	No survival labels, results depend on cluster choice

Metrics to Compare (for classification models)

- **Accuracy:** % of correct predictions.
- **Precision:** % of positive predictions that were correct.
- **Recall:** % of actual positives correctly identified.

- **F1-score:** Balance between precision & recall.

Error Analysis & Improvements

- Computing **confusion matrix** for classification models.
- Tuning hyperparameters (GridSearchCV).
- Comparing results with deep learning (using TensorFlow or PyTorch).
- Discussing **limitations & improvements** in the report.

Key Findings:

Random Forest outperforms Logistic Regression in all metrics.
Logistic Regression is weaker at recall, meaning it misclassifies some survivors.

- **Random Forest was the best classification model**, achieving 81% accuracy.
- **Logistic Regression, though less accurate, provided better interpretability.**
- **K-Means Clustering revealed meaningful passenger groupings** but was not predictive.
- **Gender, class, and fare were the strongest predictors of survival.**