IAC-31 Assignment 01: Titanic Survival Prediction Report

Student Name: Mahmud Galib
Student ID: 164549
Course: IAC-31
Institution: Bangladesh University of Engineering and Technology (BUET)
Instructor: Kowshic Roy
Date: July 30, 2025

INTRODUCTION

OBJECTIVE:
The primary goal is to build a Decision Tree classification model that can accurately predict whether a Titanic passenger survived or died based on available passenger information.

DATASET:
The analysis uses the famous Titanic dataset from Kaggle, containing:
• **Training data**: 891 passengers with known survival outcomes
• **Test data**: 418 passengers requiring survival predictions
• **Features**: Age, gender, ticket class, family relationships, and embarkation details

APPROACH:
This study implements a systematic machine learning pipeline including data preprocessing, feature engineering, Decision Tree model training, and performance evaluation on the Kaggle platform.

FEATURES USED IN THE MODEL

The following 8 features were carefully selected for the Decision Tree classifier:

Feature	Description
Pclass	Passenger class (1=First, 2=Second,

	3=Third)
Sex	Gender (converted to numeric: male=1, female=0)
Age	Passenger age (missing values filled with mean)
SibSp	Number of siblings/spouses aboard
Parch	Number of parents/children aboard
Fare	Ticket fare (missing values filled with mean)
Embarked	Port of embarkation (S=0, C=1, Q=2)
FamilySize	Created feature (SibSp + Parch)

FEATURE ENGINEERING DECISIONS:
- **Missing Age values**: Filled with mean age
- **Missing Embarked values**: Filled with most common port
- **Missing Fare values**: Filled with mean fare
- **Gender encoding**: Male=1, Female=0 for numerical processing
- **Family Size creation**: Combined siblings/spouses and parents/children counts
- **Removed irrelevant features**: Name, Ticket, Cabin (too many missing values)

METHODOLOGY

DATA PREPROCESSING:

1. **Missing Value Treatment**:
   - **Age**: Filled with mean (29.7 years)
   - **Embarked**: Filled with mode ('S')
   - **Fare**: Filled with mean fare

2. **Feature Encoding**:
   - **Sex**: Male=1, Female=0
   - **Embarked**: S=0, C=1, Q=2
   - **Family Size**: SibSp + Parch

3. **Data Splitting**:
   - **Training set**: 712 samples (80%)
   - **Validation set**: 179 samples (20%)
   - Stratified split to maintain class distribution

MODEL IMPLEMENTATION:
• **Algorithm**: Decision Tree Classifier (scikit-learn)
• **Parameters**: max_depth=5, random_state=42

EVALUATION APPROACH:
• Train-validation split for performance assessment
• Accuracy, precision, recall, and F1-score calculation
• Confusion matrix analysis for detailed performance insight
• Final evaluation on Kaggle test set

RESULTS AND ACCURACY ACHIEVED

MODEL PERFORMANCE SUMMARY:

PRIMARY ALGORITHM - DECISION TREE:
• **Training Accuracy**: 86.38%
• **Validation Accuracy**: 75.98%
• **Kaggle Test Accuracy**: 77.033%

DETAILED PERFORMANCE METRICS:

Class	Precision	Recall	F1-Score
Died (0)	0.77	0.87	0.82
Survived (1)	0.74	0.58	0.65
Overall	0.76	0.76	0.75

CONFUSION MATRIX ANALYSIS:
• **True Negatives (Correctly predicted deaths)**: 96
• **False Positives (Incorrectly predicted survival)**: 14
• **False Negatives (Incorrectly predicted deaths)**: 29
• **True Positives (Correctly predicted survival)**: 40

KAGGLE SUBMISSION RESULTS:
• **Final Score**: 77.033% (0.77033)
• **Public Leaderboard Rank**: #9,819
• **Test Predictions**: 418 passengers

• **Predicted Survival Rate**: 34.9%

OPTIONAL: COMPARISON WITH OTHER ALGORITHMS

To provide comprehensive analysis, additional machine learning algorithms were tested

ALGORITHM COMPARISON RESULTS:

Algorithm	Accuracy
Decision Tree	75.98%
Logistic Regression	80.45%
Neural Network	79.89%
KNN	69.83%

REFERENCES

REFERENCES:

1. **Kaggle Titanic Competition**: https://www.kaggle.com/competitions/titanic
2. **Scikit-learn Documentation**: https://scikit-learn.org/
3. **Pandas Documentation**: https://pandas.pydata.org/
4. **Matplotlib Documentation**: https://matplotlib.org/
5. **Other Extra Resources**: tiny.cc/aml31