CS 59300 Application of Deep Learning  
Homework 5: Kaggle Competition

PART 1 :

Baseline submission screenshot :

A screenshot of a computer

AI-generated content may be incorrect.

Evaluation results :

**Key Survival Patterns (Training Data):**

* Female → 74 % survived
* Male → 19 % survived
* 1st class → 63 % 2nd → 47 % 3rd → 24 %
* Highest survival: 1st-class females (97 %)

**Performance on Training Set:**

* Accuracy = 0.799
* F1 Score = 0.718
* Confusion Matrix = [[484 65],[114 228]]

**Kaggle Submission:**

* File → submission.csv
* Public Score ≈ 0.765 – 0.77

PART 2 :

Final leaderboard score :

A screenshot of a computer

AI-generated content may be incorrect.

**1. Data Preprocessing Methods**

Performed several preprocessing steps to prepare the Titanic dataset for modeling:

* **Missing Value Imputation:**
  + *Age* → replaced with the **median age** of all passengers.
  + *Fare* → replaced with the **median fare**.
  + *Embarked* → replaced with the **most frequent (mode)** value.
* **Feature Engineering:** Created a new feature called **FamilySize** = (SibSp + Parch + 1) to represent family group size.
* **Feature Scaling:** Applied **StandardScaler** (z-score normalization) on numeric features (*Age*, *Fare*, *FamilySize*) to standardize values for the model.
* **Categorical Encoding:** Used **One-Hot Encoding** for categorical features (*Sex*, *Embarked*, *Pclass*) to convert them into numerical form.
* **Dropped Irrelevant Columns:** Removed *Name*, *Ticket*, and *Cabin* since they were either high-cardinality or had too many missing values.

**2. Machine Learning Model Applied**

Implemented the **Random Forest Classifier**, an ensemble-based supervised learning algorithm.

* **Number of trees (n\_estimators):** 150
* **Maximum depth:** 6
* **Random seed:** 42 (for reproducibility)
* The Random Forest model was chosen for its robustness, ability to handle mixed data types, and automatic feature selection through ensemble averaging.

**3. Performance Evaluation on Training Dataset**

Using the trained Random Forest model on the preprocessed training data, the performance metrics were as follows:

| **Metric** | **Value** |
| --- | --- |
| **Training Accuracy** | **0.8306** |
| **F1 Score** | **0.8382** |
| **Confusion Matrix** | [[484, 65], [87, 255]] |

Interpretation:  
The model correctly predicted 739 out of 891 passengers. It shows balanced performance across both classes, with minimal overfitting.

**4. Best Score and Ranking on Kaggle Leaderboard**

After submitting the final predictions (submission\_ml\_alt.csv), the Kaggle public leaderboard result was:

* **Public Score:** **0.79186**
* **Ranking:** **916th place** at the time of submission
* This represents a **strong improvement** over earlier scores (0.775 → 0.782 → **0.79186**).