Team Project 1

CS 53744 Machine Learning Project

Due Date : 11:59 PM, September 30, 2025
Instructor: Jongmin Lee

September 16, 2025

Task:

- **Classification problem: Titanic Survival Prediction (10 points)**
- Weight: 10% of total grade

Dataset & Platform:

- Kaggle Competition: Titanic - Machine Learning from Disaster
- Files: `train.csv`, `test.csv`, and example submission file `gender_submission.csv`

Submission Guidelines:

- Deliverables:
  1. **Prediction CSV file** (final Kaggle submission).
  2. **Jupyter/Kaggle Notebook** with code, intermediate results, and explanations.
  3. **Short Report** (2–3 pages, PDF) including:
     - Step-by-step approach (in the below steps)
     - Key insights from EDA
     - Which features were most useful
     - Final model and Kaggle score
- Format: PDF report, 11pt font, single-spaced.
- File name: `Assignment1_StudentID_Lastname_Firstname.pdf`
- Submit via the course portal before the deadline.

Late Policy:

- 1 day late: Maximum score is 50% of the total.
- 2 days or more late: Score is 0%.

Instructions:

1. **Step 0. Get Familiar with Kaggle**
   - Create a Kaggle account and join the *Titanic - Machine Learning from Disaster* competition.
   - Read the competition description, evaluation metric, and submission format.
   - Download the datasets (`train.csv`, `test.csv`, and example `gender_submission.csv`).
   - Set up your working environment (Kaggle Notebook or Google Colab).
2. **Step 1. Very Simple Prediction (Baseline 1)**
   - Predict all passengers as deceased (`Survived = 0`).
   - Create a submission file in the correct format (`PassengerId, Survived`).
   - Submit to Kaggle and check your score (∼61% accuracy).
3. **Step 2. Simple Rule-Based Prediction (Baseline 2: Gender)**
   - Predict all females as survived and all males as deceased.
   - Submit to Kaggle and check your score (∼78% accuracy).
4. **Step 3. Exploratory Data Analysis (EDA)**
   - Explore survival rates by gender, age, passenger class (Pclass), and family status (SibSp/ParCh).
   - Include simple visualizations such as bar charts, histograms, or boxplots.
5. **Step 4. Feature Engineering**
   - Create new features that may improve predictions, such as:
     - Family size (`SibSp + ParCh + 1`).
     - Titles extracted from names (`Mr`, `Mrs`, `Miss`, etc.).
     - Age groups (Child / Adult / Elderly).
6. **Step 5. Apply a Machine Learning Model**
   - Train a Logistic Regression model (or equivalent) using scikit-learn.
   - Split the training data into training and validation sets (e.g., 80/20 split) and report validation accuracy.
   - Generate predictions for `test.csv`.

- Submit to Kaggle and record your leaderboard score.
- Compare your result against the baselines from Steps 1 and 2.

7. **Step 6. Extend and Compare Models (Optional, Extra Credit)**

   - Try other algorithms such as Decision Tree, Random Forest, or SVM.
   - Conduct simple hyperparameter tuning (e.g., tree depth, number of estimators).
   - Compare models in terms of both validation accuracy and Kaggle leaderboard score.
   - Summarize the trade-offs you observe (e.g., complexity vs. performance).

Evaluation Criteria:

- (10%) Proper formatting of deliverables (CSV, Notebook, Report).
- (30%) Quality of data analysis and visualizations.
- (30%) Completeness of step-by-step workflow.
- (30%) Clarity of final results and report writing.

Extra Credit:

- Up to +20% for extending models (Step 6), hyperparameter tuning, or insightful analysis.

Connection to Course:

- This course emphasizes not only building machine learning models, but also understanding their behavior and limitations.
- The Titanic assignment introduces the **full ML workflow**: from baseline rules, to data exploration, feature engineering, and predictive modeling.
- By completing this assignment, students will gain practical experience with Kaggle, reproducible notebooks, and leaderboard evaluation — all of which are valuable skills for later course projects and research.