

Shivam Patel – Machine Learning Project Report

Course: CSCI 5371 – Machine Learning

Instructor: Dr. Zachary Stine (Assistant Professor)

Project Title: Titanic Survival Prediction

1. Data Description

The dataset used for this project is the Titanic dataset from Kaggle, which contains information about passenger on the titanic, including demographics, ticket details, and survival outcome.

Key features:

- Pclass: Passenger class (1 = 1st, 2 = 2nd, 3 = 3rd)
- Sex: Gender of the passenger
- Age: Age of the passenger in years
- SibSp: Number of Siblings/spouses aboard
- Parch: Number of parents/children aboard
- Fare: Ticket fare
- Embarked: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

Target variable:

- Survived: (0 = did not survived, 1 = survived)

Dataset Size: 891 entries

2. Experiment design

The goal of the project is to predict whether a passenger survived using machine learning models.

Preprocessing steps:

- Fill missing numerical values (Age) with the mean.
- Fill missing categorical values (Embarked) with the mode.
- Drop unnecessary columns (Cabin, Ticket, name).
- Encode categorical features:
 - Sex -> 0 = female, 1 = male
- Scale features using StandardScaler for models that require it (Logistic Regression and Neural Network).
- Split dataset into 80% training and 20% testing sets.

Models and modelling decisions:

Model Type	Variation / Hyperparameter
Logistic Regression	C=0.1 1.0 10.0
KNN	n_neighbor=3 5 7
Neural Network	Hidden_layer_sizes=(8,) (16,8) (32,16,8)

3. Result and interpretation

Example:

I. Logistic Regression

- Best parameter: C=1.0
- Confusion Metrix

```
[[90 15]
 [19 55]]
```

- precision/recall/F1:

Class	Precision	Recall	F1	Support
0	0.83	0.86	0.84	105
1	0.79	0.74	0.76	74

II. Decision Tree

- Best parameter: n_neighbor=7
- Confusion Metrix

```
[[92 13]
 [17 57]]
```

- Precision/recall/F1:

Class	Precision	Recall	F1	Support
0	0.84	0.88	0.86	105
1	0.81	0.76	0.78	74

III. Neural Network

- Best parameter: Hidden layer = (8,), activation = relu
- Confusion Metrix

```
[[96 09]
 [21 53]]
```

- Precision/recall/F1:

Class	Precision	Recall	F1	Support
0	0.82	0.91	0.86	105
1	0.85	0.72	0.78	74

Example Prediction:

- Likely Survivor: 1st Class, female, age 28 – Model prediction = 1
- Likely non-survivor: 3rd Class, male, age 30 – Model prediction = 0

4. Reflection/Lesson Learned

- Preprocessing is crucial for model performance.
- Trying multiple models allows comparison and understanding.
- Logistic Regression is simple and interpretable; Decision Trees capture non-linear relationships; Neural Networks are powerful but require careful tuning.
- Predicting new passenger is straightforward after preprocessing and scaling.

Output screenshot:

```
Step 1.....Complete
Step 2.....Complete
Step 3.....Complete
Step 4.....Complete
Step 5.....Complete
Step 6.....Complete
-----Project 1 Complete-----
===== Detailed Model Reports =====

Model: Logistic Regression
Best Hyperparameter: C = 1.0
Confusion Matrix:
[[90 15]
 [19 55]]

Classification Report:
      precision    recall  f1-score   support

     0       0.83       0.86       0.84       185
     1       0.79       0.74       0.76        74

   accuracy       0.81       0.80       0.81       179
  macro avg       0.81       0.80       0.80       179
 weighted avg       0.81       0.81       0.81       179

Model: KNN
Best Hyperparameter: n_neighbors = 7
Confusion Matrix:
[[92 13]
 [18 56]]

Classification Report:
      precision    recall  f1-score   support

     0       0.84       0.88       0.86       185
     1       0.81       0.76       0.78        74

   accuracy       0.82       0.82       0.83       179
  macro avg       0.82       0.82       0.82       179
 weighted avg       0.83       0.83       0.83       179

Model: Neural Network
Best Hyperparameters: hidden_layer_sizes = (8,), activation = relu
Confusion Matrix:
[[96  9]
 [21 53]]

Classification Report:
      precision    recall  f1-score   support

     0       0.82       0.91       0.86       185
     1       0.85       0.72       0.78        74

   accuracy       0.84       0.82       0.83       179
  macro avg       0.84       0.82       0.82       179
 weighted avg       0.83       0.83       0.83       179

===== Comparing models =====
Logistic Regression: Accuracy = 81.01%
KNN: Accuracy = 82.68%
Neural Network: Accuracy = 83.24%
=====
Best performing model: Neural Network
=====
Example passenger 1 prediction (0=Not Survived, 1=Survived): 0
Example passenger 2 prediction (0=Not Survived, 1=Survived): 1
```