

# E0-270 (O): Assignment 1

## Binary Classification

**Due Date:** 24 March 2025

### Instructions

1. Answer all questions carefully, providing full justifications.
2. All the code should be done by you and not copied from any source.
3. You can use Python libraries such as sklearn.
4. Total marks for the assignment is 20.

## Problem 1: Binary classification

In this problem, you will explore and implement the following methods.

1. Naive Bayes
2. Logistic Regression
3. K-Nearest Neighbor(KNN)
4. Support Vector Machines(SVMs)

using the given data set. This data set is used for **binary classification**, where the target is **Survived**.

**Task:** Implement all mentioned algorithm along with hyper parameter search using

1. Grid Search
2. Bayesian Search

In Case of SVM, Hyper parameter search must include "C", "gamma" and the kernels(Linear and RBF).

### 3.1 Load and Preprocess the Dataset

- Use ONLY COMMA(,) as file separator while opening or loading data file.
- Handle **missing values**.
- Encode **categorical features** (One-Hot Encoding).
- Standardize **numerical features**.
- Split data into **70% training** , **10% validation** and **20% test** sets.

### 3.2 Test Set

- Train split should be used for model training.
- Validation split should be used to compare models for different hyperparameters.
- Based on the best hyperparameter obtained, use test split to report final model performance.

Column Name	Column Description
<b>PassengerId</b>	Unique ID for each passenger (just an identifier, no predictive value).
<b>Survived</b>	Target variable: 0 = No, 1 = Yes (whether the passenger survived).
<b>Pclass</b>	Passenger class (ticket class): 1 = 1st, 2 = 2nd, 3 = 3rd.
<b>Name</b>	Passenger's full name (can be parsed for titles like Mr, Mrs, etc.).
<b>Sex</b>	Gender of the passenger (male or female).
<b>Age</b>	Age of the passenger (in years). Some values are missing.
<b>SibSp</b>	Number of siblings/spouses aboard the Titanic.
<b>Parch</b>	Number of parents/children aboard the Titanic.
<b>Ticket</b>	Ticket number (can sometimes be used for feature engineering).
<b>Fare</b>	Passenger fare (in British pounds).
<b>Cabin</b>	Cabin number (many missing values, sometimes used for deck information).
<b>Embarked</b>	Port of Embarkation: C = Cherbourg, Q = Queenstown, S = Southampton.

Table 1: Dataset Column Descriptions

## Deliverables

Component	Details
<b>Code</b>	Implement all mentioned algorithms with hyperparameter optimization.
<b>Report</b>	Discuss <b>accuracy and hyperparameter tuning results</b> .
<b>Submission</b>	Submit a ZIP file: <code>FirstName_Last5DigitsOfSRNo.zip</code> , containing: <ul style="list-style-type: none"> <li>• <code>naive_bayes.py</code></li> <li>• <code>logistic.py</code></li> <li>• <code>knn.py</code></li> <li>• <code>svm_linear.py</code></li> <li>• <code>svm_rbf.py</code></li> <li>• <code>grid_search.py</code> (Grid Search for Hyperparameter Tuning)</li> <li>• <code>bayes_search.py</code> (Bayesian Search for Hyperparameter Tuning)</li> <li>• <code>report.pdf</code> (Observations &amp; Analysis)</li> </ul>