**ÜSKÜDAR UNIVERSITY INSTITUTE OF SCIENCE AND TECHNOLOGY**

**MACHINE LEARNING – Final Project**

**Machine Learning Pipeline: Feature Selection and Hyperparameter Optimization**

**Project Objective**

In this project, you will work on the HIGGS Dataset, a large and high-dimensional dataset, to apply two critical components of the machine learning workflow: feature selection and hyperparameter tuning.

In this context, you will:

* Analyze outliers,
* Properly scale the data,
* Apply filter-based methods for feature selection,
* Train models using nested cross-validation with hyperparameter optimization,
* Use KNN, SVM, MLP, and XGBoost algorithms to compare performance,
* Evaluate model performance using ROC curves and other metrics.

**Deliverables**

1. GitHub Repository
   * Must include all code, graphical outputs, and comments.
2. Short Report (.pdf or .docx)
   * Should include all code, graphical outputs, and your interpretations.
   * Metric results for each model (in table format).
   * ROC curves.
   * Interpretation of the best-performing model and feature representation combination.

**Dataset Used**

* HIGGS Dataset (11 million samples, 28 features).
* A random sample of 100,000 rows will be used for this project.
* <https://archive.ics.uci.edu/ml/datasets/HIGGS>

**Section 1: Data Preprocessing**

**Outlier Analysis:**

* Outliers will be checked using the IQR method.
* If necessary, outliers will be removed or replaced with threshold values.

Feature Scaling:

* All numerical variables will be scaled to the [0, 1] range using MinMaxScaler.

**Section 2: Feature Selection**

**Filter-Based Feature Selection:**

* ANOVA F-score or Mutual Information will be used to select the top 15 features.

**Section 3: Modeling and Evaluation**

**Nested Cross-Validation:**

* Outer Loop: 5-fold CV
* Inner Loop: 3-fold CV
* **Flowchart A:** In the inner loop, different feature selection combinations are evaluated to determine the best model and feature set. (*Figure 1*)
* **Flowchart B:** In the inner loop, different hyperparameter combinations are evaluated to select the best model and hyperparameters. (*Figure 2*)
* **The outer loop is used to evaluate test performance.**

**Models to be Used:**

* K-Nearest Neighbors (KNN)
* Support Vector Machine (SVM)
* Multi-Layer Perceptron (MLP)
* XGBoost

**Hyperparameter Ranges (for Inner CV):**

* KNN: n\_neighbors =between 3and 11
* SVM: C = [0.1, 1, 10], kernel = ['linear', 'rbf']
* MLP: hidden\_layer\_sizes = [(50,), (100,)], activation = ['relu', 'tanh']

**Performance Metrics:**

* Accuracy
* Precision
* Recall
* F1 Score
* ROC-AUC

**ROC Curves (OVA method):**

* **ROC curves will be plotted for each class.**
* **AUC scores will be visualized and interpreted.**



Figure 1: Flowchart A



Figure 2: Flowchart B