National University

Of Computer & Emerging Sciences Faisalabad-Chiniot Campus

# Data Science for Software Engineering
# Assignment # 4

**Objectives:**

Design and implement a comprehensive machine learning pipeline applying Support Vector Machines (SVM), Logistic Regression, Perceptron, and Deep Neural Networks (DNN) for classification tasks on diverse datasets. Your work will ensure rigorous data preprocessing, feature extraction, model training, and critical evaluation, forming a solid foundation for comparative analysis and future advanced modeling efforts.

**Note: Carefully read the following instructions (*Each instruction contains a weightage*)**

1. First, think about statement problems and then write your program.
2. Write the Program in Python/IDE and save a notebook **for each program.**
3. Complete your assignment **within the given deadline**.
4. Please submit your IPYNB **files and GitHub repo link of the streamlit application.**
5. Submit your Assignment on Google Classroom.

## Dataset:

**Deepfake audio detection task:**

**from datasets import load_dataset**

**ds = load_dataset("CSALT/deepfake_detection_dataset_urdu")**

**Part 2: The Dataset file is attached with the assignment in CSV format.**

## TASKS

## Part 1: Urdu Deepfake Audio Detection (Binary Classification)

1. **Preprocessing:**
   - Extract relevant features from the audio files (e.g., MFCCs, Spectrograms).
   - Ensure uniform input size (e.g., fixed-length feature vectors).
2. **Model Building:**
   - **Train three classifiers:**
     - Support Vector Machine (SVM)
     - Logistic Regression
     - Single-Layer Perceptron (not multi-layer)

- ○ Also, design a Deep Neural Network (DNN) with at least 2 hidden layers.

3. **Training:**
   - ○ Train all models to classify audio as Bonafide vs Deepfake.
4. **Evaluation:**
   - ○ Use metrics: Accuracy, Precision, Recall, F1-Score, AUC-ROC.
   - ○ Compare the performance of all models on a common test set.

## Part 2: Multi-Label Defect Prediction (Multi-Label Classification)

**Tasks:**

1. **Preprocessing:**
   - ○ Analyze the dataset (uploaded CSV) for missing values, feature selection, scaling.
   - ○ Understand label distribution (is it highly imbalanced?).
2. **Model Building:**
   - ○ Train the following models separately for multi-label classification:
     - ■ Logistic Regression (one-vs-rest or other strategy)
     - ■ SVM (multi-label version)
     - ■ Perceptron
     - ■ Deep Neural Network (DNN)
3. **Training:**
   - ○ Split into train/validation/test sets.
   - ○ Hyperparameter tuning (e.g., C for SVM, learning rate for Perceptron).
4. **Evaluation:**
   - ○ Use multi-label metrics:
     - ■ Hamming Loss
     - ■ Micro-F1, Macro-F1
     - ■ Precision@k
5. **Challenge Element:**
   Train the Perceptron in **online learning mode** (update after each sample).

## Part 3: Interactive Streamlit App

**Objective: Build a real-time interactive UI for prediction.**

**Tasks:**

- Build a **Streamlit App** to allow:
  - ○ Uploading audio file → Predict if it’s Deepfake or Bonafide.
  - ○ Input feature vector for software defect data → Predict multiple labels.

- UI Requirements:
  - Clean layout (upload buttons, prediction outputs clearly shown).
  - Confidence scores for predictions.
  - Allow the user to **select which model** (SVM / Logistic Regression / DNN) to use for prediction at runtime.

**LinkedIn post and Medium blog are compulsory**.