

Project Writeup: Personalized Activity/Workout Recommendation System

Team: ML Mavericks

Team Members:

- Shubham Kothiya
- Yugm Patel
- Soumya Bharathi Vetukuri

Homework Assignment: Classification, Amalgamation, and Performance Evaluation

Objective

This homework focuses on integrating multiple datasets to enhance classification performance for our personalized activity/workout recommendation system. We will examine how combining datasets improves model accuracy and decision-making.

Task Overview

Each team member will work on different classification approaches using an incrementally amalgamated dataset. The key tasks include:

1. **Find and integrate additional datasets** (Dataset 2 and 3) with the existing Dataset 1.
2. **Run classification models** (each team member should choose a different classification algorithm).
3. **Evaluate how performance improves** with each dataset combination.
4. **Perform a Muller Loop** on the datasets to systematically compare performance.
5. **Visualize and summarize results** in tabular format.

Datasets

We will use three datasets:

- **Dataset 1 (ds1):** Daily activity tracking data (steps, distances, active minutes)
- **Dataset 2 (ds2):** Sleep monitoring data
- **Dataset 3 (ds3):** Heart rate monitoring data

Each dataset contains valuable insights that, when combined, can enhance the accuracy of our personalized activity recommendations.

Methodology

Step 1: Load and Preprocess Data

- Load the three datasets from CSV files.
- Handle missing values using imputation techniques.
- Normalize/scale numerical features to ensure consistent model input.
- Merge datasets incrementally to analyze classification performance at each stage.

Step 2: Classification Models

Each team member will implement and evaluate a different classification model:

- **Logistic Regression**
- **Random Forest Classifier**
- **Support Vector Machine (SVM)**

Step 3: Model Training and Evaluation

- Perform a train-test split.
- Train classification models on:
 - **Dataset 1 only (ds1)**
 - **Dataset 1 + Dataset 2 (ds1 + ds2)**
 - **Dataset 1 + Dataset 2 + Dataset 3 (ds1 + ds2 + ds3)**
- Evaluate model performance using:
 - **Accuracy Score**
 - **Precision, Recall, and F1 Score**
 - **Confusion Matrix Analysis**

Step 4: Running a Muller Loop

A Muller Loop is used to measure incremental performance improvements by adding datasets sequentially:

1. Run classification on **Dataset 1 (ds1)**
2. Run classification on **Dataset 1 + Dataset 2 (ds1 + ds2)**
3. Run classification on **Dataset 1 + Dataset 2 + Dataset 3 (ds1 + ds2 + ds3)**
4. Compare results across these iterations.

Step 5: Results Analysis

- **Visualize results** in tables and plots to demonstrate the impact of dataset amalgamation.
- Discuss insights on how dataset integration influences classification accuracy and model robustness.
- Assess the best-performing classification model based on key performance indicators (KPIs).

Expected Outcomes

- A structured approach to integrating multiple datasets.
- Insights into the effectiveness of different classification models.
- Evidence of performance improvements via dataset amalgamation.
- Well-documented results supporting personalized activity recommendations.