Human Activity Recognition Using Smartphones: Feature Engineering Techniques And Stacking/ Ensemble Models for Accurate Classification

Name:	**Pratik Pal**
Registration No./Roll No.:	20213
Institute/University Name:	IISER Bhopal
Program/Stream:	Electrical Engineering and Computer Science
Problem Release date:	January 12, 2023
Date of Submission:	March 01, 2023

1 Introduction

The aim of this project is to develop a model that accurately recognizes human activity based on data collected from smartphones. The data set contains 563 different features and 8239 data points. In this project, we will be using different feature engineering techniques and classification models to accurately classify human activity into one of the six classes.

2 Methodologies to Solve the Problem

We plan to classify human activities using various machine learning models from the scikit-learn library, namely **k-Nearest Neighbors (kNN), Support Vector Machine (SVM), Random Forest, Decision Tree, and Multilayer Perceptron (MLP)**. For accurate predictions, we will be using the method of stacking/ensemble of classifiers. The stack will include a selection of the best-performing classifiers from the listed models. Initially, we stacked kNN, SVM, and Random Forest classifiers. The idea is to individually carry out hyperparameter tuning for the classifiers before stacking to get the most optimized parameters for the individual classifiers. We will carry out the optimization process till we get the best ensemble model. We are keen on using the evaluation metrics of **accuracy, Matthew's correlation coefficient (MCC), F1 score, and confusion matrix** for analysing the performance of each of the classifiers and the designed ensemble model.

The feature engineering techniques we plan to employ include **low-variance thresholding** [1] and **minimum redundancy maximum relevance (mRMR)** [2]. Low-variance thresholding will eliminate features below a particular variance. To determine the optimum threshold value, we took a range of values for variance and finally selected the one for which we achieved maximum accuracy. The mRMR is a feature selection method that eliminates features by evaluating the mutual information to get the value of relevance and redundancy [3]. We passed a range of values representing the number of features to find the K best features for which the obtained accuracy is maximum and plotted a curve of accuracy vs K best features from the mRMR algorithm. We could successfully get the desired results for our classification model by following the process in paper [2] in this initial phase of the project. We will also try to explore further feature engineering techniques to get better results.

3 Individual Contribution

The simulation for the low-variance thresholding feature engineering technique, kNN and MLP classifiers was run on the dataset and the performance evaluation for the same was done from my side. The design of the ensemble/stacking classifier model was mutually discussed and carried out.

References

[1] "sklearn.feature_selection.VarianceThreshold." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.VarianceThreshold.html

[2] A. Doewes, S. E. Swasono, and B. Harjito, "Feature selection on human activity recognition dataset using minimum redundancy maximum relevance," in *2017 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW)*. IEEE, jun 2017.

[3] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, aug 2005.