**Assignment 8: Technical Report**

**David D. Bui**

**Grand Canyon University**

**DSC-540**

**Professor Aiman Darwiche**

**December 22nd, 2021**

**GitHub:**

**https://github.com/DouglasBui/GCU/tree/main/DSC-540/Assignment5**

This report covers a statistical analysis of the Ensemble Methods for Classification of Physical Activities from Wrist Accelerometry article. The general purpose of this report is to follow a blueprint outline that article depicts step-by-step, and successfully come to the same results and conclusions. The ensemble method the researchers of the article used involved three methods of ensemble machine learning, which are Bagging, Boosting, and Random Forest. There is also a custom ensemble model that joins four algorithmic classifiers together. All these classifier models target on activity type prediction.

This requires an understanding of the dataset before diving into the methodology. The dataset is composed of 9 files that each represent different human individuals. The main features are ‘Timestamp’, ‘activityID’, and ‘3 wrist locations. These wrist locations are represented as ‘x-axis’, ‘y-axis’, and ‘z-axis’ since the data is being modeled to a 3-Dimensional scale. These wrist movements are then analyzed over 10 second time windows to denote feature extraction of 45 final features. These additional features reflect general statistics over the 10 second time window of movement. The goal is to create models of prediction with the target being activityID; this target indicates the type of activity the individual is performing, such as walking, sitting, resting. In this report we are going over a total of 8 activities. There are multiple models being used with 3 distinct datasets being analyzed, but for this report I am only focusing on one, which Is dataset 1. With this understanding its time to move onto the methodology.

Before the models can receive input, the data must be filtered to fit a specific understanding of interpretation. The Preprocessing step involves filling in missing NaN values via interpolation and reduction of the feature space to home in on each different dataset, mainly focusing on wrist movement. The next step is Feature extraction, where the now shrunken-down feature space is expanded with general statistical on the remaining features, such as mean, standard deviation, and variance, leading up to a total of 45 features. The last step before employing ensemble methods is the Normalization and feature selection. The datasets are scaled with a min-max function with the purpose of eliminating negative values, since our Naïve Bayes Classifier utilized requires this, and the features are selected based on correlation rates that reach up to 0.25 and above.

The time for ensemble methods has arrived, starting with the Bagging Classifier. This classifier, also known as bootstrapping, involves a random sampling method of the dataset or datasets above that allows for replacement of the selected sample. The boosting classifier also employs random sampling with replacement, but the replacement is done over weighted data. The Random Forest comes with the bagging method, the difference is that only a certain number of features are selected at random, and a series of decision trees are drafted and ensembled to approximate a prediction from a wide marginal base. The final fourth model is a Custom Weighted Majority Voting model. This model is composed from four different algorithms which are a Binary Decision Tree with a hierarchy, K-Nearest Neighbor, Support Vector Machine, and an Artificial Neural Network. These algorithms are ensembled via a Voting Classifier.

The evaluation of these models is evaluated within the article via f1-scoring and cross-validation with LeaveOneOut splitter. Here is where this report ran into an issue, I implemented each algorithm in with the bagging method. This resulted in my evaluation focusing on the precision of the WMV classifier. The results yielded a precision of 92.2% precision where in the article they produced a 92.7%. This gives reasonable evidence that my model was successfully able to replicate the article’s findings. This evaluation was also ran through a NB confusion matrix which is a similar process the article took.

References:

Chowdhury, K., ALOK, K., Tjondron, D., Chandran, V., Trost, S., (2017), Ensemble Methods for Classification of Physical Activities from Wrist Accelerometry, Medicine & Science in Sports & Exercise, doi: 10.1249/MSS.0000000000001291