Title: Human Activity Recognition Using Machine Learning

Team: Jay Upadhyay (jxu9414), Sahana Murthy (sam2738), Sanchitha Seshadri (ss9886)

Problem Definition: The objective of this project is to use the data from the accelerometer in a person's smartphone to determine their physical activity. The set of activities to be determined includes walking upwards, sit-to-stand, stand-to-sit, sitting still, etc.

Motivation: The motivation behind this project is to gather useful data about the physical behavioral patterns of humans and can be used to provide insight about how much movement their lifestyle involves, which is very valuable data for monitoring health-related data for people such as number of steps walked, distance walked, etc per day.

The data: The dataset being used is called the HAPT Dataset, obtained from the UCI Machine Learning Repository webpage. The dataset contains 10929 instances of real-world data, and each instance has 561 attributes. The dataset and its description can be found here: http://archive.ics.uci.edu/ml/datasets/Smartphone-Based+Recognition+of+Human+Activities+and+Postural+Transitions

Intuition:

The problem is to determine the activity being performed by the person given the accelerometer data of their smartphones. There are several possible activities the person may be performing and the result is one of these options. Therefore, this is a classification problem where we need to decide what activity is being performed based on the features / attributes available in the data.

Now that we have determined that this is a classification problem, the next step is to choose the appropriate classification algorithms to solve this. As part of this project we examine three different classification algorithms and experiment with which of them works best for the given problem.

The first algorithm we use is the **decision tree algorithm**. The dataset has 561 attributes that need to be used to make a decision about the activity. Information gain is a metric that is used to quantify the amount of learning at each stage of the algorithm. The decision tree works by calculating the information gain for each of the attributes to choose the best feature at each step that enables to make the best decision about whether the activity is sitting, standing, etc. This helps select the features useful for the classification needed to perform, choosing the most effective ones first. This method ensures that the most relevant features are involved in the classification.

The second algorithm used for classification is the **logistic regression**. This algorithm is useful when the data to be classified is categorical in nature. The way it works is by finding a line to separate data to classify different instances correctly and identify that they are indeed different activities by examining the features. It works by estimating the probabilities of an instance pertaining to each activity to be predicted and the ultimate predicted event is the one with the largest probability attached to it. The logistic function is used to generate these. Thus, we see that this algorithm is a good choice to apply on the problem.

The third algorithm used for classification is the **multilayered perceptron (MLP)**. This model tries to optimize the log-loss function in order to perform the classification. In order to do so it uses the stochastic gradient descent. This is a neural network and is suitable for the problem since there are many attributes and the layers will work to determine the most useful ones.

Proposed method:

1. Clean and preprocess dataset

2. Partition dataset into training dataset and testing dataset

3. Select classification model and train model on training dataset

4. Once model is trained, predict events in the testing dataset

5. Measure mean square error and accuracy of model, get confusion matrix

6. Repeat for all three classification algorithms discussed above

7. Analyze and compare results by the metrics in step 5.

Experimental Results and Analysis:

We used a data set consisting of 3162 records for testing our models.

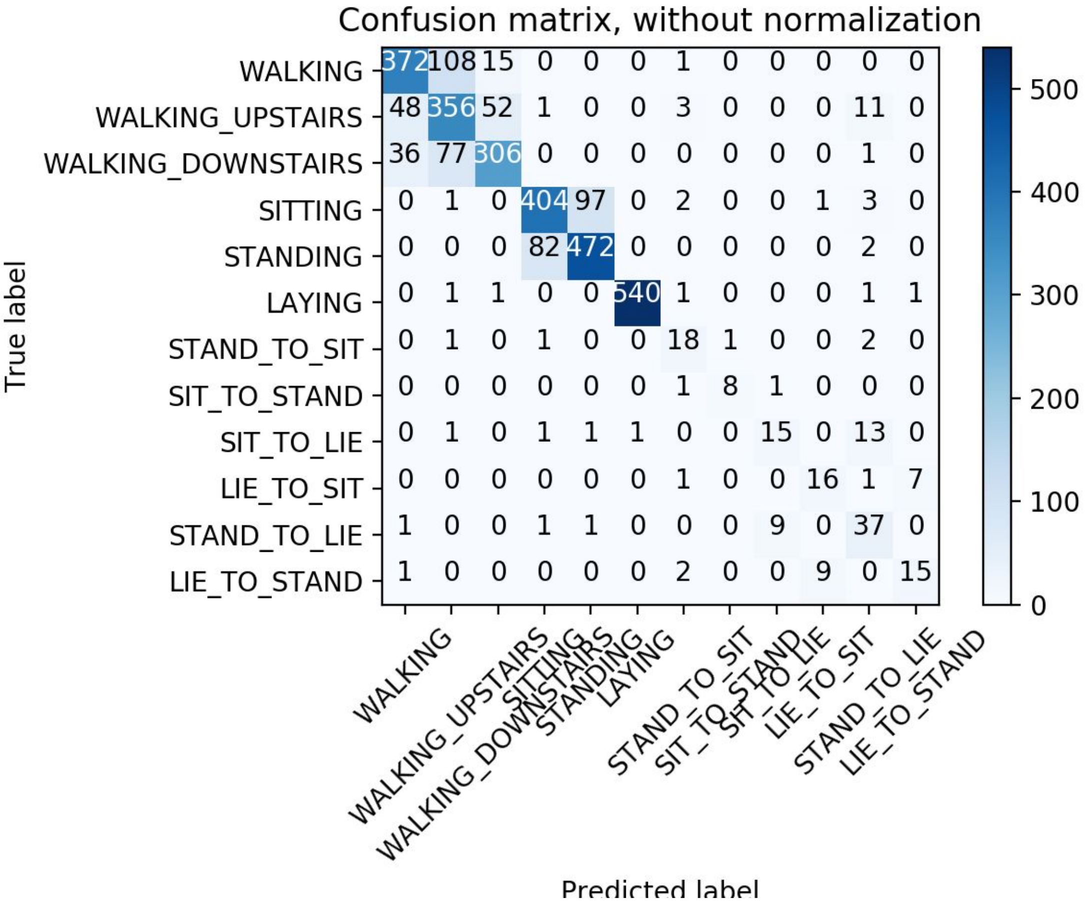
Model 1: Decision Tree

Number of correctly classified records: 2562

Accuracy: 81.025%

Mean Squared Error: 0.577

The confusion matrix obtained for this model is shown below:



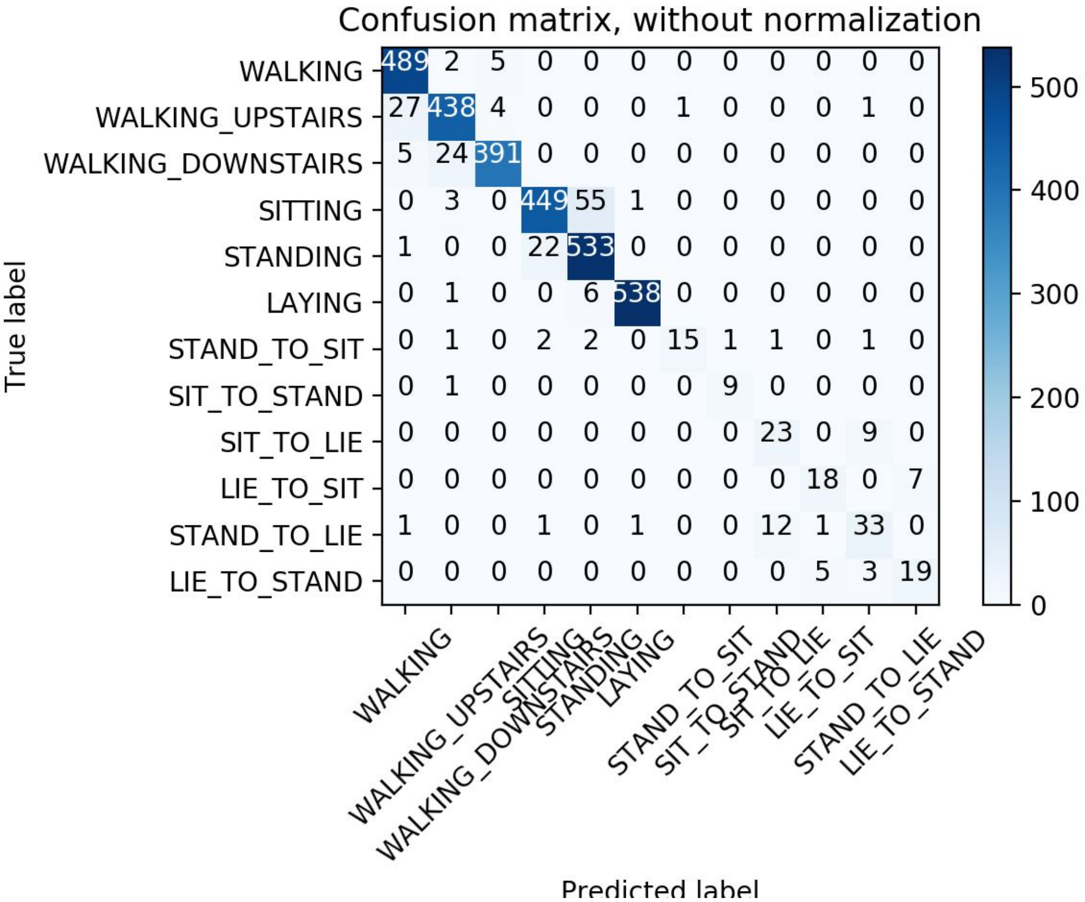
Model 2: Multi-Layer Perceptron

Number of correctly classified records: 2914

Accuracy: 92.16%

Mean Squared Error: 0.285

The confusion matrix obtained for this model is shown below:



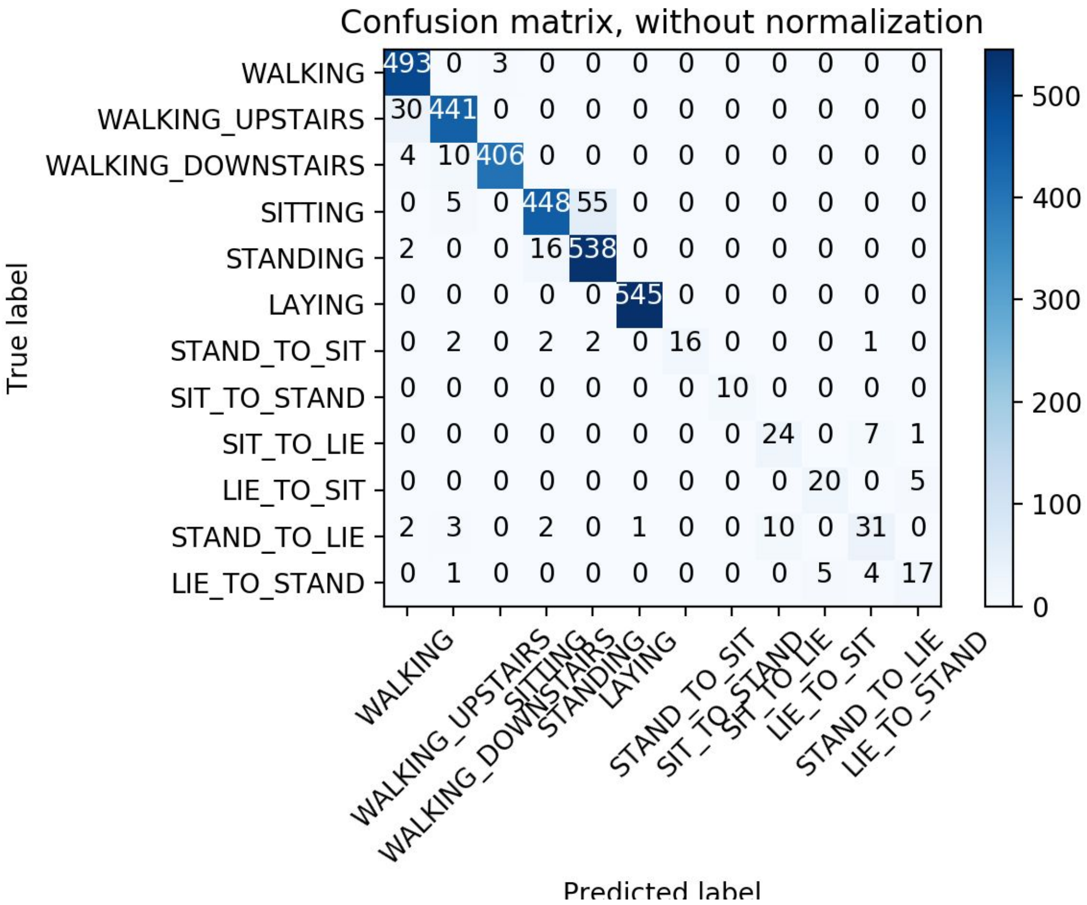
Model 3: Logistic Regression

Number of correctly classified records: 2989

Accuracy: 94.53%

Mean Squared Error: 0.338

The confusion matrix obtained for this model is shown below:



We started with the process of creating a Decision Tree for which we calculated Mean Squared Error, Confusion Matrix and Accuracy. Since, the accuracy was low we went ahead and implemented Multi-Layer Perceptron Classifier and calculated the metrics for the Neural Network. There was a boost in the accuracy by around more than 10% and the Mean Squared Error was decreased a lot. We got a model which was around ~93% accurate and both models took around similar amount of time to train the model. After that we implemented Logistic regression wherein we got an even more accurate model with accuracy more than 94% but surprisingly the Mean Squared error rose and the time for training this model was more than twice the time taken to train the other two models.

Conclusion:

By analyzing the above data, we concluded that for the given data we can use either MLP Classifier or Logistic Regression depending on the requirement i.e. if we need more accuracy, we can use Logistic Regression and when we have less time to train the model, we can use MLP Classifier.

Another conclusion we made is that the Mean Squared Error is not suitable for classification models as we can see that it is not a good parameter to determine the accuracy of our models.