Human Activity Classification

First name	Last Name	IIT Email	Same project in ITMD 527 (Yes or No)
Bhumin	Shah	bshah40@hawk.iit.edu	No
Pratik	Tamhankar	ptamhankar@hawk.iit.edu	No

Introduction

Fitness has been an important activity of human cycle. From the start of the day till night, every activity a human performs counts. In the 20th century, world of machines and robot has overtaken human fitness by providing all the comforts. Starting from 2015 there was an exponential rise of fitness tracker which were introduced in the market. Over $2 billion of wearable fitness trackers were introduced in 2015 and ever since the need has been increasing. With over $3 billion in 2016 and $4 billion in 2017, technology and the application of these trackers have been in great demand. It's been predicted that by 2019 the growth of wearable devices is going to rise by 39%.

Humans using these wearable trackers have shown an increase of activity by 8.5 % from their daily routine. These tracking devices monitor the activity based on accelerometers, GPS, optical heart rate monitors, galvinic skin response sensors and other basic sensors. These numerous sensors incorporated in the single tracker ensures that correct reading of activity is tracked. We challenge the information received by these fitness devices due to lack of accuracy with the devices understanding of various moments by a human the rightful amount of activity made is inaccurately counted.

The dataset we are working on is been generated from Wireless Sensor Network (WSN), EvAAL competition technical annex. Through the readings accumulated by the WSN we would differentiate the activity each human has performed and be able to gauge the fitness.

1. Data Sets

Through our intensive reach, we came across this source of data the technique used for evaluation. This data was found at University of California, Irvine, machine learning repositories relevant to our area of interest:
https://archive.ics.uci.edu/ml/datasets/Activity+Recognition+system+based+on+Multisensor+data+fusion+%28AReM%29

The dataset broadly consists of activity data recorded by the user from time-series generated by a Wireless Sensor Network (WSN). This data is recorded by using the IRS nodes which basically comprises of Chest, Right ankle and Left ankle. Combinations of all the three nodes are considered while recording the values.

The dataset at the mentioned link is divided or arranged as per the different types of activities which were performed by individuals while recording the measurements. So, the dataset is logically divided into multiple activities of Bending, Cycling, Lying, Sitting, Standing and walking. Each of the activity data has a set of n datasets, the attribute information in each dataset is however same and at the same scale.

Below are the attributes in every dataset.

a. #Columns: time Indicates the Clock measure in millisecond when the recording
b. avg_rss12: Average of WSN recorded at Chest and Right ankle
c. var_rss12: Variance of WSN recorded at Chest and Right ankle
d. avg_rss13: Average of WSN recorded at Chest and Left ankle
e. var_rss13: Variance of WSN recorded at Chest and Left ankle
f. avg_rss23: Average of WSN recorded at Right ankle and Left ankle
g. var_rss23: Variance of WSN recorded at Right ankle and Left ankle

avg_rssn indicates the average rss value at nth measurement
var_rssn indicates the variance of rss value at nth measurement

Each folder will represent the classifier label and we need to treat each data accordingly in our preprocessing stage.

Citation Reference: F. Palumbo, C. Gallicchio, R. Pucci and A. Micheli, Human activity recognition using multisensor data fusion based on Reservoir Computing, Journal of Ambient Intelligence and Smart Environments, 2016, 8 (2), pp. 87-107

Dataset snapshot:

Sr No	Data Set Name	Total Number of files	Number of attributes
1	bending 1	7	7
2	bending 2	6	7
3	cycling	15	7
4	lying	15	7
5	sitting	15	7
6	standing	15	7
7	walking	15	7

2. Research Problems

In today's time there are innumerable activity tracking systems like fitness tracker and others which track fitness based on activity performed. These devices collect large sets of datasets at regular intervals. However, we need to ask questions, whether we are using this valuable information to predict and label activities into meaningful labels.

We believe that the datasets captured for identifying these results can have other potential uses too, hence we are interested in performing a classification task for identifying activities based on the factors of measurement. This meaningful segregation of information can be widely applicable in healthcare, sports and fitness domains.

This dataset represents activities of humans performing activities like bending, cycling, lying, sitting, standing and walking. Since there are many activities that might have a slight difference in the activity which correspondingly has the effect on the readings recorded. Our research problem would include finding the slightest difference and correspondingly differentiate the activity.

Further, we assume that data would have noise created due to combination of activities performed. These activities like standing and walking can often be confused with one another creating large amount of noise. All these measurements are being made with an electronic device. Chances of equipment failure or incorrect reading is quite a possible chance. Need to ensure that data is accurate, consistent and can be reliable.

3. Knowledge Discovery in Database (KDD)

3.1 Data Processing:

Data was organized as per the activities such as lying, bending, walking, cycling and sitting. Within each activity there were datasets that contributed to individual reading of actors performing those task.

NaN's:

In the dataset we found many NaN's and that could be fairly due the reasons of faulty instruments or node not attached properly to the actor or could be reason that the record has no value which in our case could be variance.

NaN's were dealt by removing those set of records and moving forward with the next part of the pre-processing.

Scalability:

In the description of the raw data it denotes that the values of the records range is large. This is due to difference in activities that records captured have a huge difference. To ensure that it does not impact the model we used min-max scaling to scale the range from 1-10 for all the records.

Irrelevant data:

In the captured dataset the variance in many occasions have turned out to the zero and especially in the case of node 23(both the ankles). This data did not seem relevant since it contained all zeros and could affect the accuracy.

Data after pre-processing has been quite normalized and useable. This can be justified with the scatter plot as denoted below.

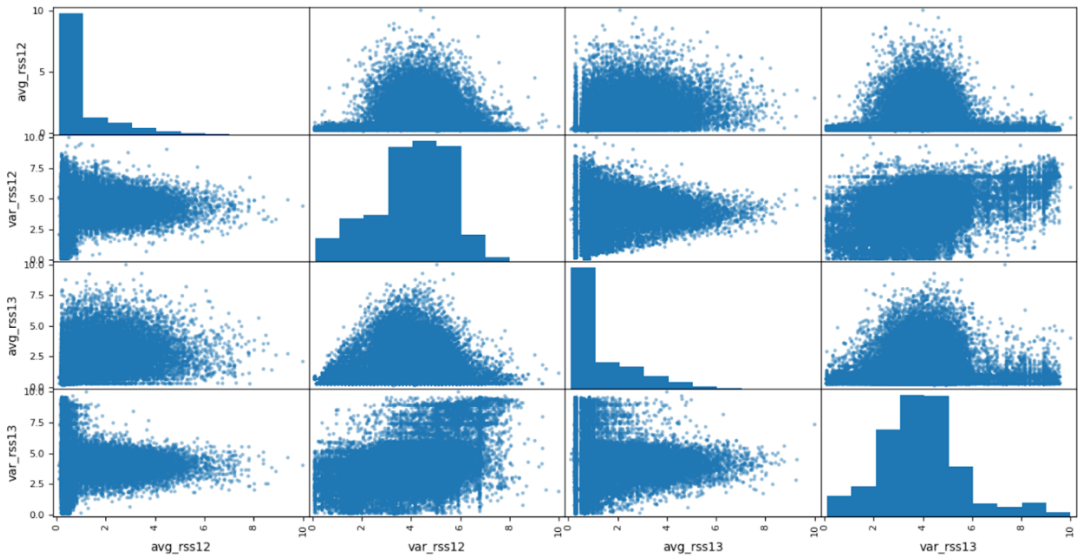

Figure 1: Scatter plot of Raw Data

Box Plots:

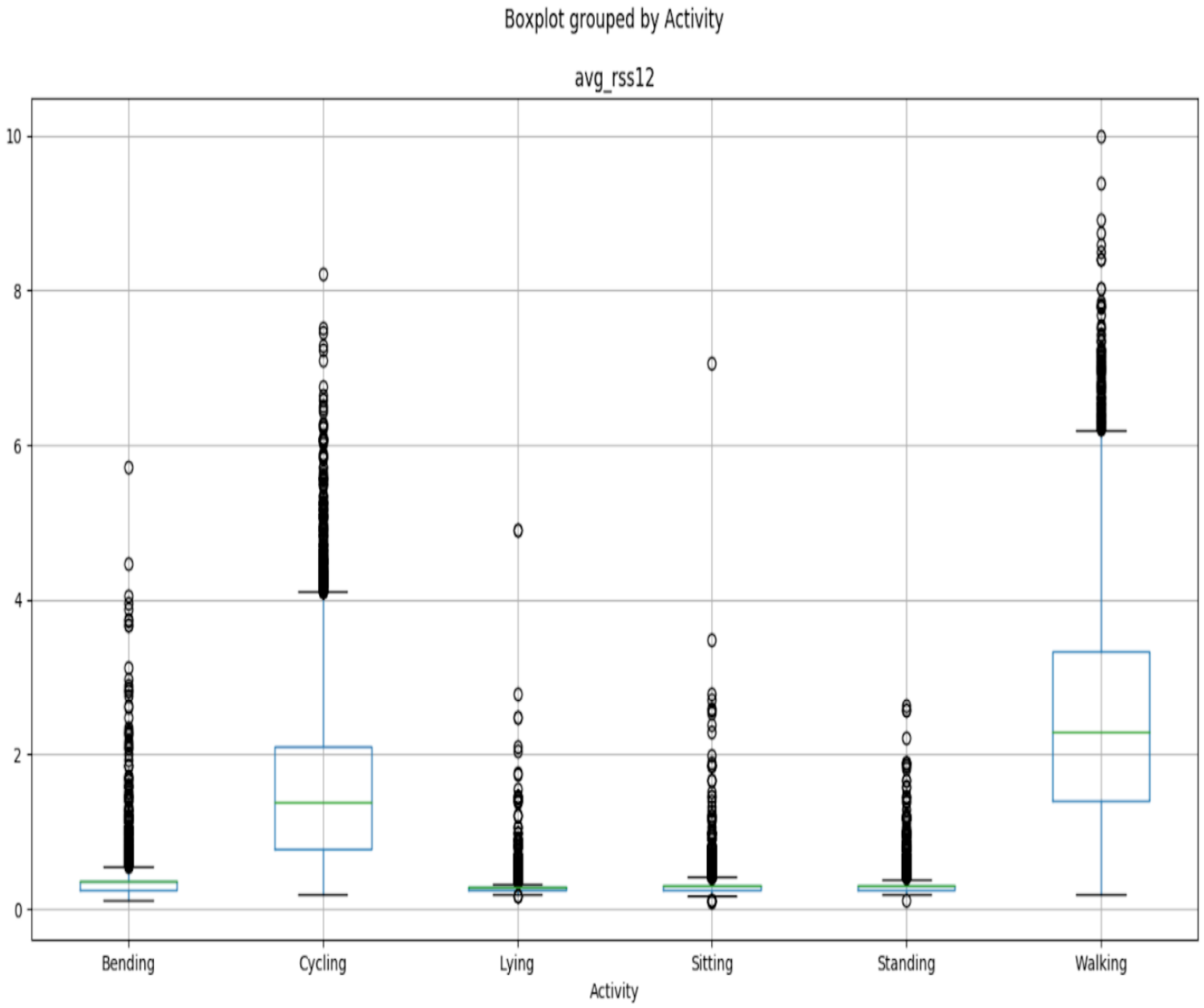

Figure 2: Box plot Avg_rss12

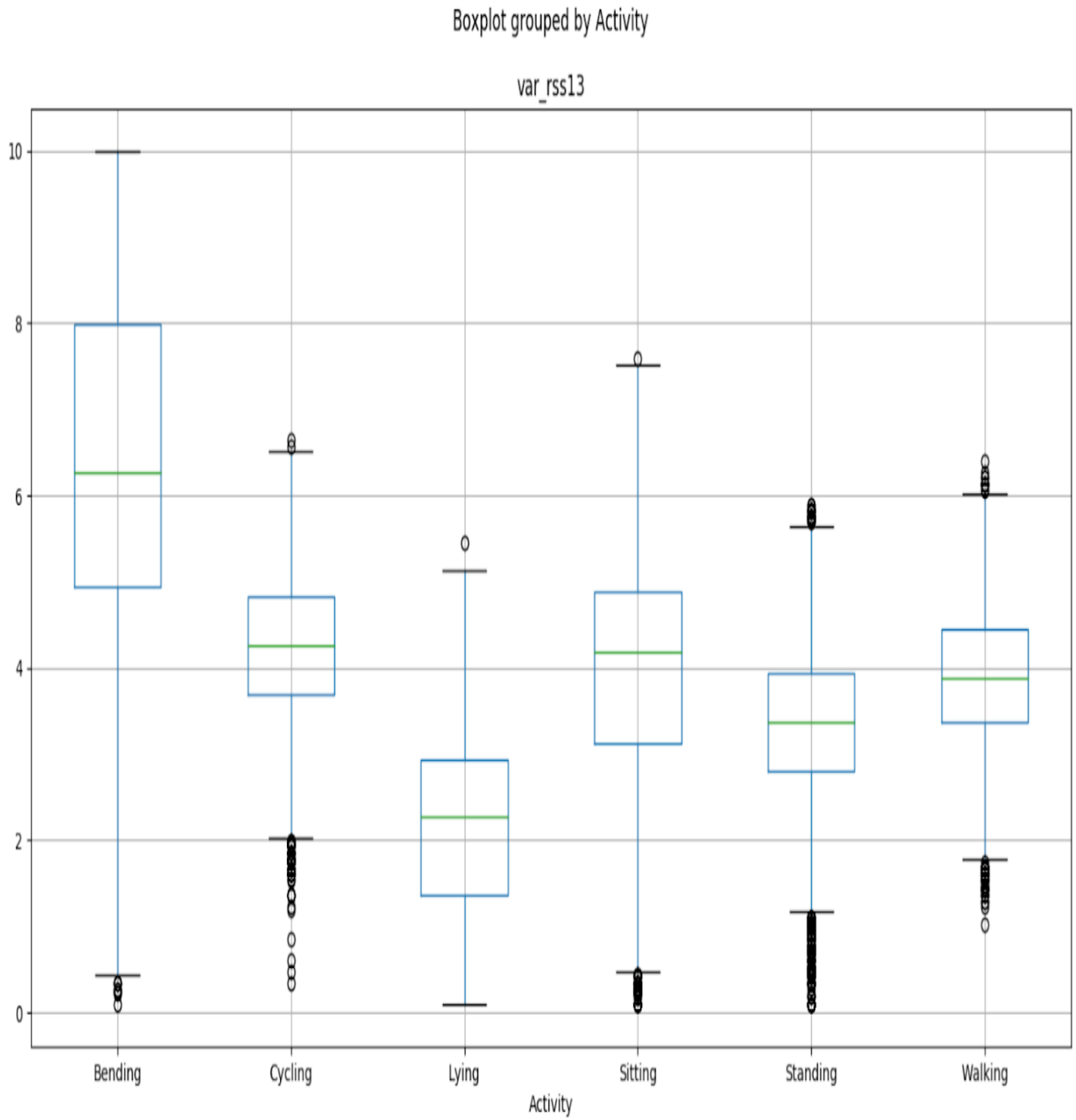

Figure 3: Box plot var_rss13

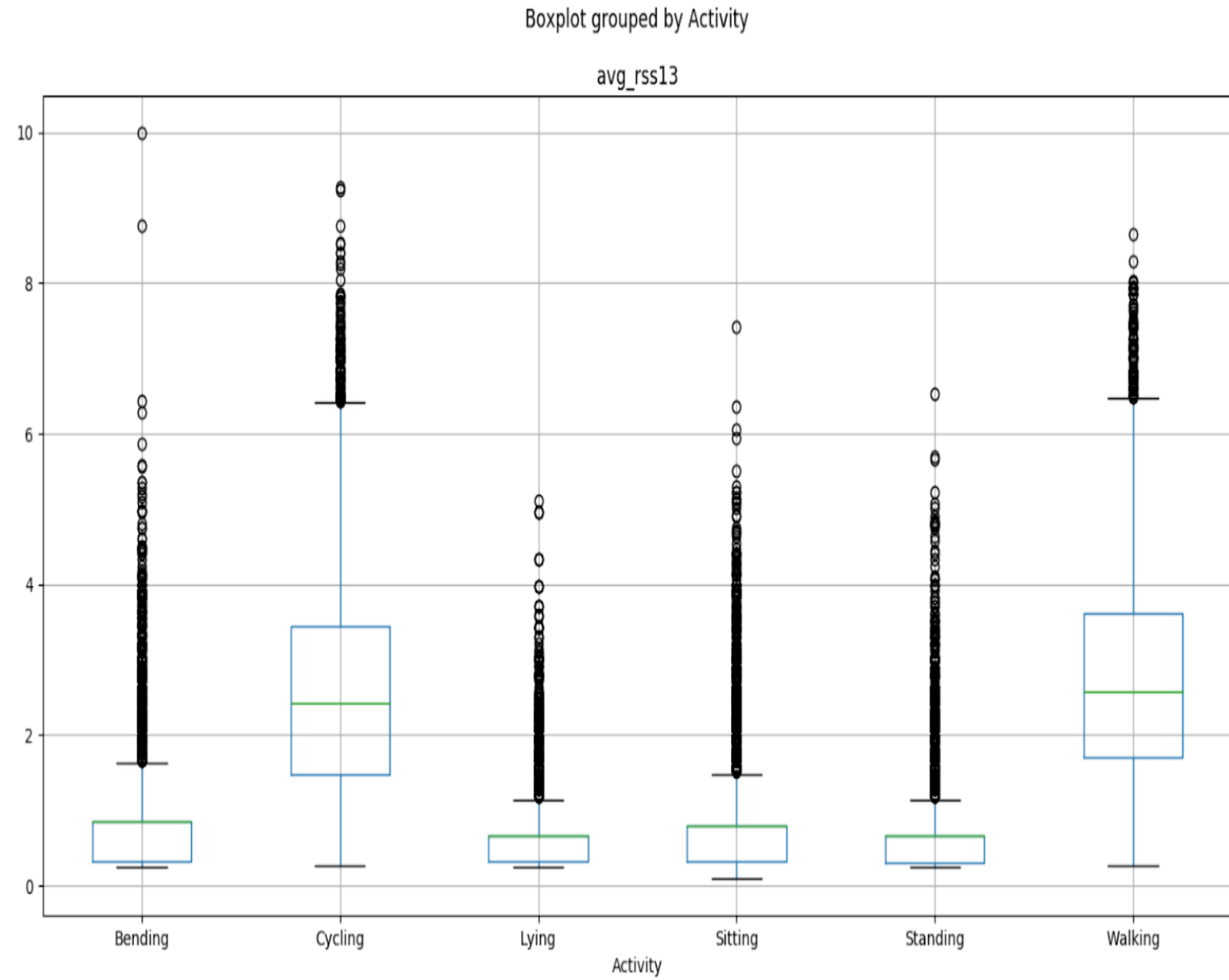

Figure 4: Box plot avg_rss13

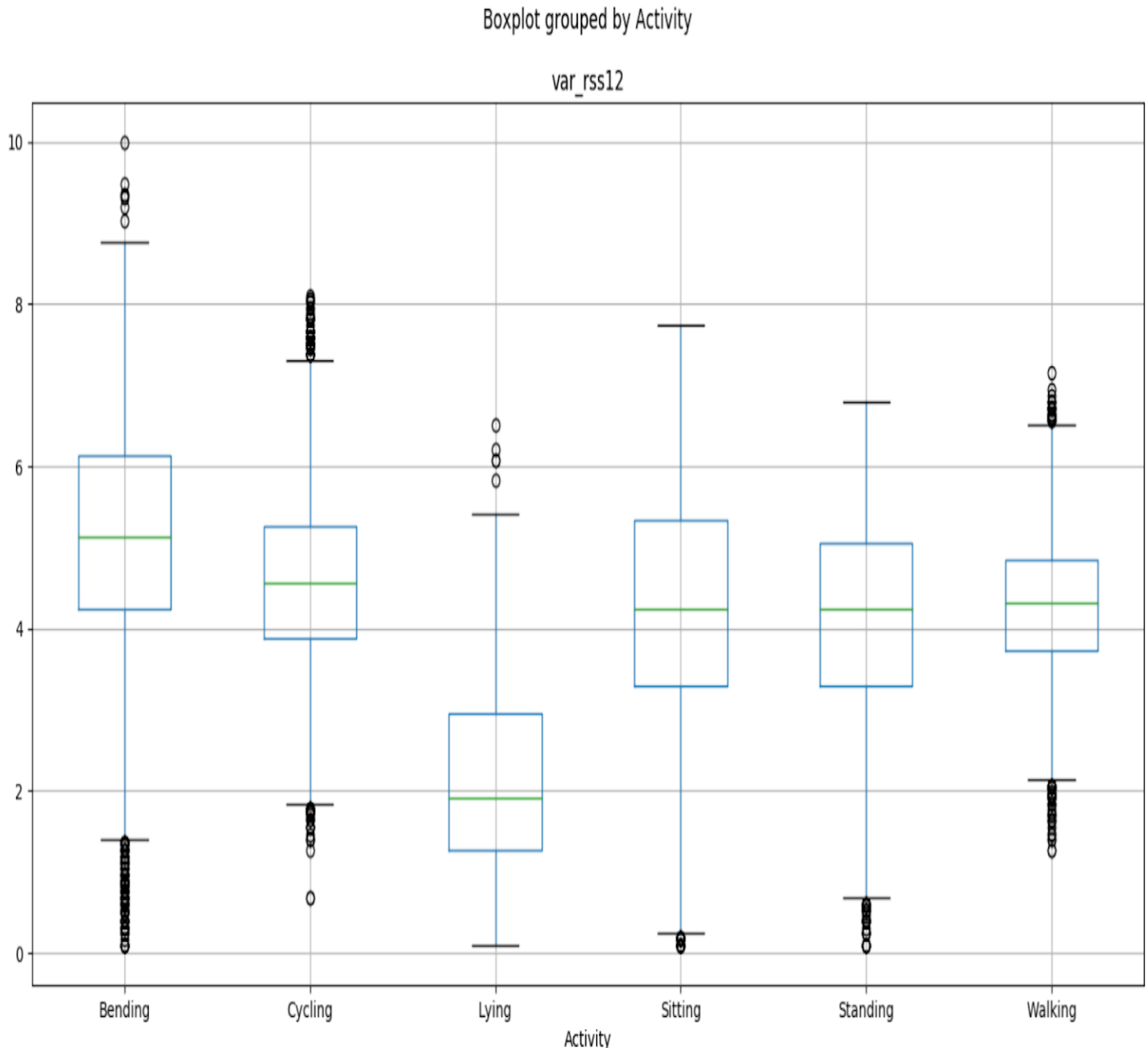

Figure 5: Box plot var_rss12

3.2 Data Mining Task:

During pre-processing we have given labels to each activity associated with their records and then merged them to build a model. Model building was carried in Python using a library called "sklearn".

We considered classifying activities using the following algorithm:

- Logistic Regression
- Linear Discriminant Analysis
- K – Neighbors
- Decision Tree
- Random Forest
- Support Vector Machines

Following are the result outcomes

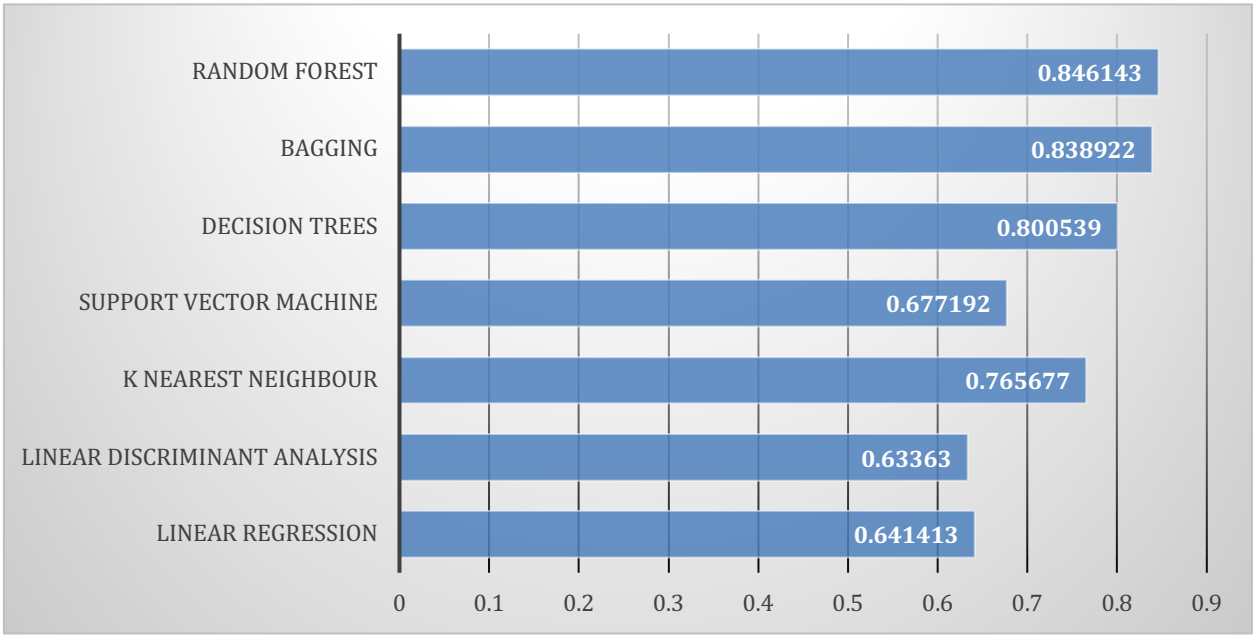

Figure 6: Model evaluation

Based on the above obtained result set we see the best precision outcome was with Random Forest at 0.86.

While configurations used were default however with parameters changed for K – Neighbor or Random Forest does not yield much difference.

4. Potential Solutions

Preprocessing:

Before beginning with model building we shall carry out the knowledge discovery steps of checking for any missing values, check for data quality, to merge these datasets, perform random sampling i.e. Data Integration, perform necessary transformations if required to bring the data on same level, eliminate and summarize the cleansed data

For addressing the above, we have identified that we will perform Binning by smoothening by bin medians for bringing the range of values at same level

In case of outliers, we shall perform Clustering so that the data points which are extremes are identified and are not part of the model building process

We have identified that since our datasets are scattered and requires a step of consolidation, we shall perform Data Integration as a step for carrying out this step. Since our datasets has numeric values, we will have to perform Correlation Analysis to identify positively correlated, negatively correlated and independent datasets. Pearson correlation will be used for this purpose.

If we find that the range of dataset values are having high variance then we shall decide to perform data transformation to bring the range of values at same level.

Model Building:

The dataset we have in hand needs certain steps of preprocessing, further we shall build various models of classification on this dataset, measure the accuracy and cost functions, test how the prediction accuracy works on various models and finally select the best model.

The algorithms which will be considered for model building will be as below:

a. Naïve Bayes Classifier
b. Random forest classifier
c. Support Vector Machine(SVM)
d. Logistic Regression

We shall perform steps of building a model on training datasets, run it on testing datasets, measure its accuracy on testing datasets, identify if there is an overfitting problem, take necessary steps for improving the models and re build the model.

5. Evaluations

We propose that this dataset be split into training and testing sets using Hold Out Evaluation technique. Based on the dataset we shall perform classification tasks; each activity will be considered as a Label for classification.

Initially we shall set the percentage as 80-20 i.e. 80% training and 20% testing data. Our evaluation matrix would be comparing the accuracy based on the different techniques used. First level of comparison would be between accuracy at training and testing data. Comparing the training and testing dataset can arise classification problems of overfitting.

To eliminate the overfitting problem, we would check and eliminate margin of error. This would give us an optimal result having the best classification technique for the given data set.

Our outcome from the modelling yielded us the best model being the Random forest on which the model was built and tested with the validation dataset.

Table 1: Result set of Model Evaluation

Sr No	Classification Model	Mean	Standard Deviation
1	Linear Regression	0.641413	0.009063
2	Linear Discriminant Analysis	0.63363	0.007763
3	K nearest neighbour	0.765677	0.007916
4	Support Vector Machine	0.677192	0.005641
5	Decision Trees	0.800539	0.006334
6	Bagging	0.838922	0.004553
7	Random Forest	0.846143	0.006429

```
Confusion Metrics
[[1186   22   13   41   12   10]
 [  10 1011    0   29   17  373]
 [   3    3 1377    2   14    3]
 [  38   32   11 1226  131    8]
 [   5   19   14  138 1231    8]
 [   5  316    2    6    6 1126]]
```

Figure 7: Confusion Matrix on Random Forest

Classification report

	precision	recall	f1-score	support
Bending	0.95	0.92	0.94	1284
Cycling	0.72	0.70	0.71	1440
Lying	0.97	0.98	0.98	1402
Sitting	0.85	0.85	0.85	1446
Standing	0.87	0.87	0.87	1415
Walking	0.74	0.77	0.75	1461
avg / total	0.85	0.85	0.85	8448

Figure 8: Precision of each activity

The screen grab of the results mentioned above is been showing us the precision for each activity. We can further understand the count in terms of precision from the confusion matrix.

The first row and column of the confusion matrix refers to the number of classified records as Bending activity which outcomes the precision of 0.95.

6. Conclusion

By performing the data mining techniques/algorithms on given dataset, we have observed and conclude following things:

- Activities such as Bending, Lying gave a high precision compared to other activities. Which could be assumed since the momentum of activities are less compared to others
- Total precision of 0.85 which can be considered in building a machine learning algorithm

7. Futuristic scope

With current algorithm of Random Forest, we can consider in machine learning task and try understanding the patterns and activities and classify data.