**Data Wrangling**

A close up of text on a white background

Description automatically generated

*Figure 1.1*

The dataset provided for the analysis and machine learning has the directory structure as presented above in *Figure 1.1*. For each subject provided in the train set has N instances associated at different time stamps with each instance having a different predicted variable(label).

Source of the dataset - <https://bbdc.csl.uni-bremen.de/index.php/2019h/25-bbdc-2019>

The preprocessing step of the data involves –

1. Mapping the subjects from training set to each subject directory.
2. Interating through all the instances for each subject.
3. For each instance per subject from the training set, map to the respective labels(Ground Truth)
4. Create a new label with the ground truth and concatenate it with individual dataframe.
5. Append a list of dataframes and in the final iteration, concatenate all the dataframes into a main dataframe.
6. Added an extra dimension named Subject for exploratory data analysis purpose which will be dropped before training the model.
7. Added column names (taken from the dataset info file) to the main dataframe.

The shape of the main dataframe –

* Number of records/rows: 16367293
* Dimensions: 21

A screenshot of a cell phone

Description automatically generated

*Figure 1.2*

Histogram (*Figure 1.2*) represents an overview of the labels/ground truth for respective data records. The most common label which was recorded was found to be **Stair-up.**

A screenshot of a cell phone

Description automatically generated

*Figure 1.2*

The number of records grouped by subjects in the training set has been plotted in *Figure 1.2.*

Figure 1.3 shows the datatype of each predictor variable.

A close up of text on a white background

Description automatically generated

*Figure 1.3*

Kernel density estimation for one the variables EMG1 shows that the dominance of outliers in the dataset. Figure 1.4 represents this distribution. The same pattern can be observed for other predictor variables as well. The red line depicts the interquartile range.

A screenshot of a cell phone

Description automatically generated

*Figure 1.4*

Since the predicted variable was found to be a categorical value, it was converted to numerical value hence numbering the labels sequentially. This step was required at this stage because removing outliers(on categorical variables) wouldn’t have meant approprriate for further data analysis.

Data Value Space for the predicted variable, Label = {1,2,3,4,5,6,7……,20,21,22}

Next step involves filtering data 1 Standard Deviation away from the mean so that we can consider more meaningful data. After this our sample space included –

* Number of records/rows: 1921598
* Dimensions: 20

Hence, we can notice a reduction of approximately 88% of the less meaningful records which could have been a result of wrong reading by the sensors.

At this point, we can notice most of the outliers being removed from the dataset in

*Figure 1.5* and it mostly follows a Normal Distribution*.* This was noticed for all other predictor variables as well.

A screenshot of a cell phone

Description automatically generated

*Figure 1.5*

Now, the Label, ie, the ground truth variable was Encoded in such a way that one label has one column created and making sure to **avoid dummy variable trap** by removing one extra column.

As the dataset has continuous random variables with each having a varying range, it was considered necessary to scale them as it would reduce the running time of the Maximum Likelihood Estimation of Machine Learning algorithm by finding the optimal coefficients.

Using PCA to find the variables explaning the most variance caused the number of meaningful features to be reduced to 8 as seen in *Figure 1.6*

A screenshot of a cell phone

Description automatically generated

*Figure 1.6*