# Final Project Data Mining PPG_Signal

M. Imam Whidyarto
*Faculty of Electrical Engineering*
*Telkom University*
Bandung, Indonesia

## I. RESULT

### A. *Exploratory Data Analysis*

To begin this project, we will confidently utilize the PPG Signal dataset. Our first step is to conduct Exploratory Data Analysis (EDA) to gain a comprehensive understanding of the data. The dataset is comprised of 54 CSV files, which have been merged into one file for ease of analysis. Please refer to the image below to view the code used to merge the files..

**Fig 1**. Combining CSV file

```
df_list = []
for file in os.listdir(os.getcwd()):
    if file.endswith('.csv'):
        df_list.append(pd.read_csv(file))

master_df = pd.concat(df_list, ignore_index=True)
master_df.to_csv('PPG_Signal.csv', index=False)
```

Once the files have been merged, confidently proceed to read the merged data. As shown in the image below, confidently retrieve the top 5 entries in the dataset.

```
df = pd.read_csv('PPG_Signal.csv')
df.head()
```

| | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 0 | 1.689150 | 1.679374 | 1.667644 | 1.653959 | 1.634409 | 1.609971 |
| 1 | 1.221896 | 1.225806 | 1.228739 | 1.230694 | 1.230694 | 1.229717 |
| 2 | 0.466667 | 0.458824 | 0.447059 | 0.439216 | 0.427451 | 0.411765 |
| 3 | 0.301961 | 0.298039 | 0.294118 | 0.290196 | 0.282353 | 0.278431 |
| 4 | 2.449658 | 2.413490 | 2.383187 | 2.358749 | 2.338221 | 2.314761 |

5 rows × 4202 columns

**Fig 2**. Read dataset

```
df.describe()
```

| | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| count | 162.000000 | 162.000000 | 162.000000 | 162.000000 | 162.000000 |
| mean | 1.031910 | 1.028811 | 1.026817 | 1.026128 | 1.027141 |
| std | 0.767307 | 0.761986 | 0.757152 | 0.752781 | 0.749659 |
| min | 0.164223 | 0.161290 | 0.159335 | 0.157380 | 0.155425 |
| 25% | 0.396078 | 0.403922 | 0.403922 | 0.391984 | 0.403922 |
| 50% | 0.637295 | 0.629469 | 0.631373 | 0.633333 | 0.631373 |
| 75% | 1.529814 | 1.626588 | 1.667644 | 1.650049 | 1.595308 |
| max | 2.900293 | 2.920821 | 2.928641 | 2.923754 | 2.908113 |

8 rows × 4202 columns

**Fig 3**. Describe dataset

Apply the describe function to the dataset to obtain statistical summaries. Confirm that there are no empty values in the dataset, as shown in the figure below.

```
df.isnull().sum()
```

```
0                0
1                0
2                0
3                0
4                0
                ..
4197             0
4198             0
4199             0
Mean_NBP_Sys     0
Mean_NBP_Dias    0
Length: 4202, dtype: int64
```

**Fig 4**. Null data in dataset

Pada gambar dibawah ini menunjukkan plot histogram untuk data yang dipilih yaitu Systolic dan Diastolic.

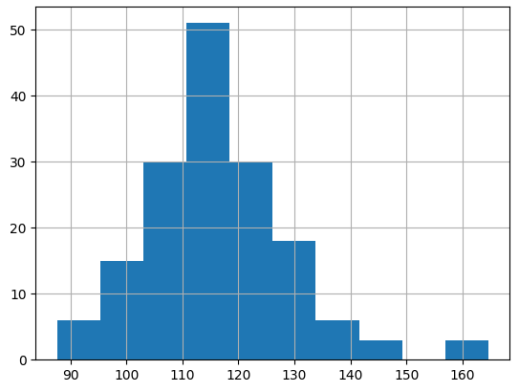


**Fig 5**. Plot histogram Systolic

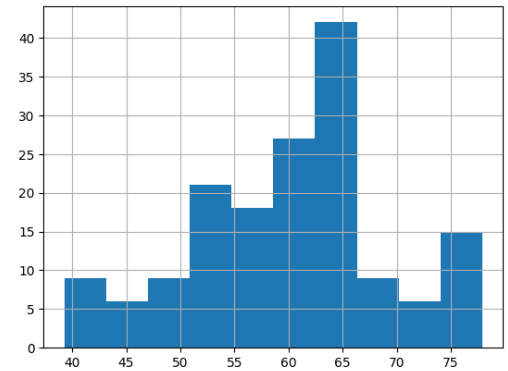


**Fig 6**. Plot histogram Diastolic

### B. *Prepocessing*

After performing exploratory data analysis (EDA), the next step is to preprocess the dataset by selecting the relevant features for predicting the values of Systolic and Diastolic. Any unnecessary columns should be deleted. Once the

preprocessing is complete, the X and Y values can be defined to create the training and testing datasets.

### *C. Training*

During the training stage, Linear Regression (LR), Random Forest (RF), and Support Vector Regression (SVR) models are utilized for Machine Learning (ML). This approach improves the accuracy of the prediction results, which are presented in the table below.

**Table 1**: Training Result

| Model | Prediction | Evaluation | Accuracy |
|---|---|---|---|
| Linear Regression | Systolic | RMSE | 5.4451 |
| | | R2 | 1.0 |
| | Diastolic | RMSE | 7.2603 |
| | | R2 | 1.0 |
| Random Forest | Systolic | RMSE | 1.0 |
| | | R2 | 0.9923 |
| | Diastolic | RMSE | 0.8369 |
| | | R2 | 0.9902 |
| SVR | Systolic | RMSE | 0.1000 |
| | | R2 | 0.9999 |
| | Diastolic | RMSE | 0.1000 |
| | | R2 | 0.9998 |

## II. DISCUSSION

The Random Forest model yielded the best results in predicting Systolic and Diastolic values among the three models tested (LR, RF, and SVR). The Random Forest model yielded the best results in predicting Systolic and Diastolic values among the three models tested (LR, RF, and SVR). The Random Forest model yielded the best results in predicting Systolic and Diastolic values among the three models tested (LR, RF, and SVR). It had high RMSE and R2 accuracy. In comparison, the LR and SVR models did not perform as well, especially the SVR model with a linear kernel. Although it had a small RMSE, the R2 value was good. However, when using the RBF and Sigmoid kernels, the RMSE value exceeded 1.0. Therefore, the best model for this dataset is the Random Forest model.