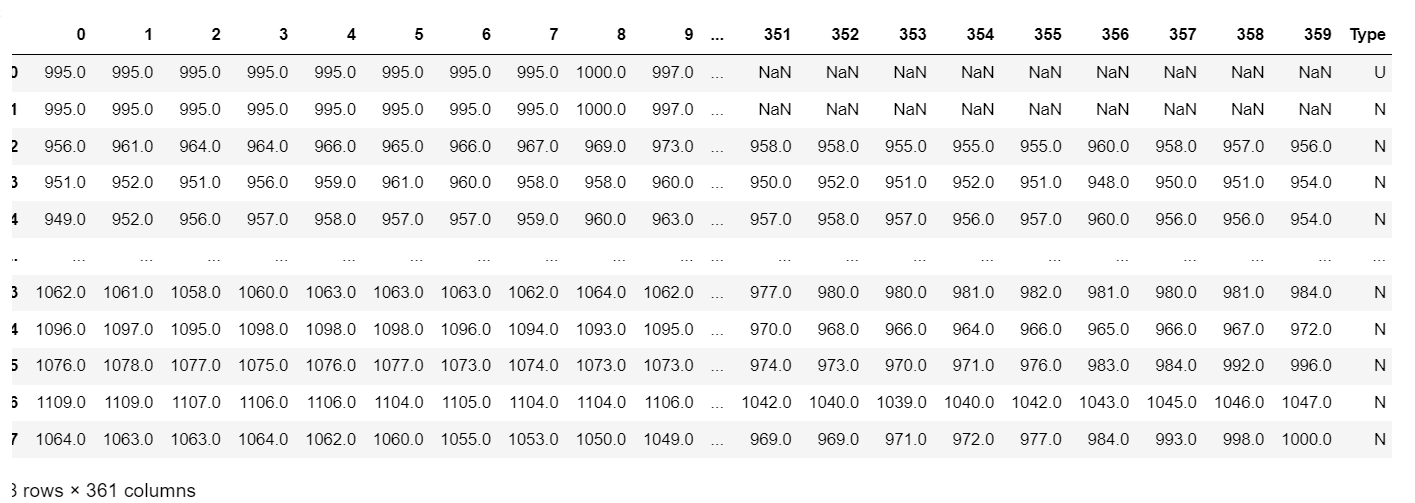
**Mini project 1 preliminary submission**

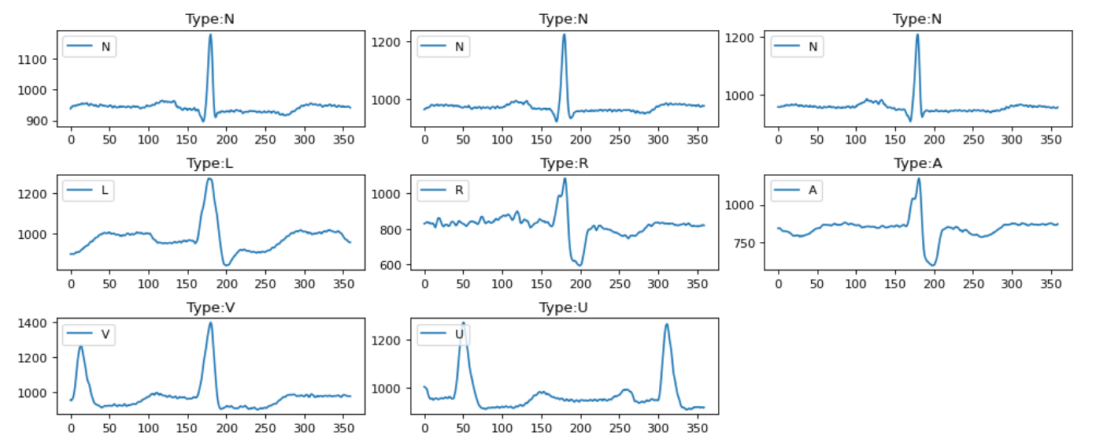
**1. preprocessing**

**1.1 preprocessor**

I transformed raw datasets from TXT files into CSV files, and then match them by sample numbers. Specifically, each type of sample has a range of numbers in CSV from minus 179 to plus 180. Then got 360 attributes with one ‘Type’.

**1.2 plotting**

Plotted 8 graphs with each of the types.



**1.3 cleaning**

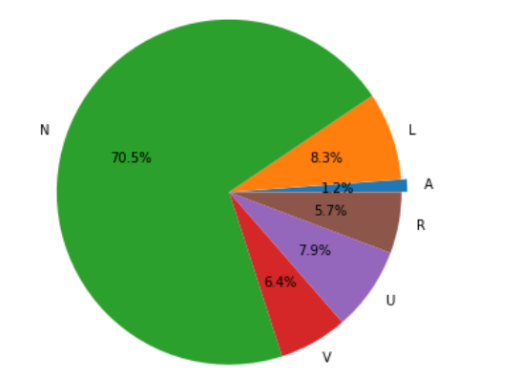
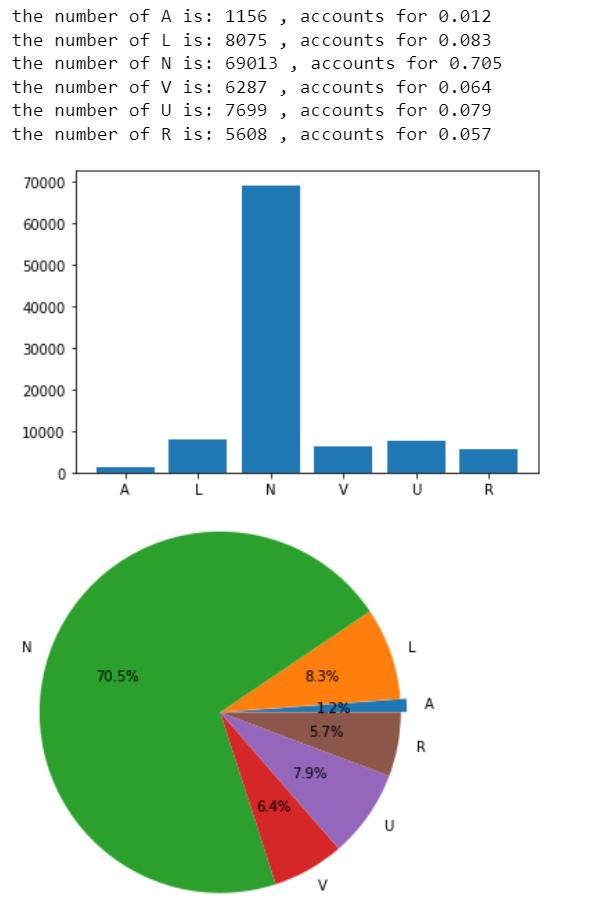
Cleaned infinity and outliers in the whole dataset.

**1.4 normalizer**

Then I added two data normalization methods: one is to normalize to the range [0,1], and another is to use StandardScaler().

**1.5 class\_imbalance\_checker**

Obviously, types are not balanced.

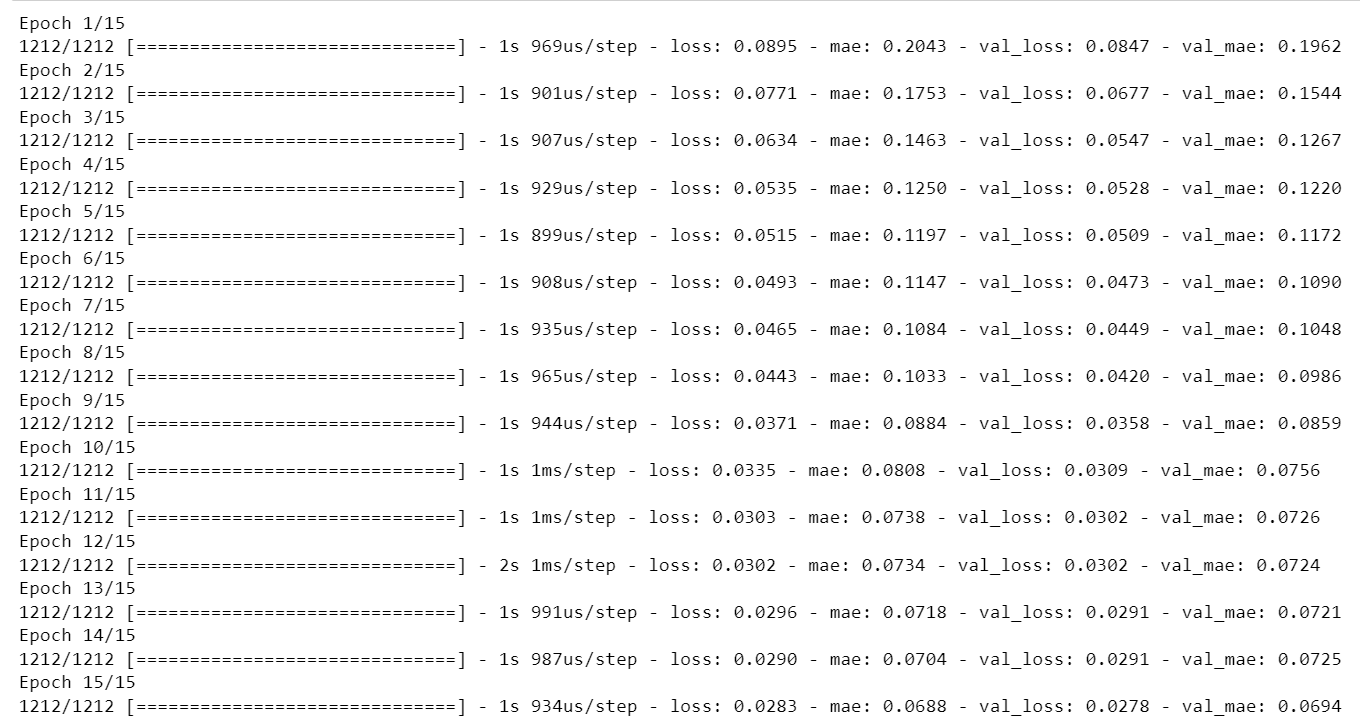


**1.6.1 imbalance\_remover**

I implemented two removers: one is the combination of undersampler and SMOTE, and the other is just SMOTE.

**1.6.2 autoencoder**

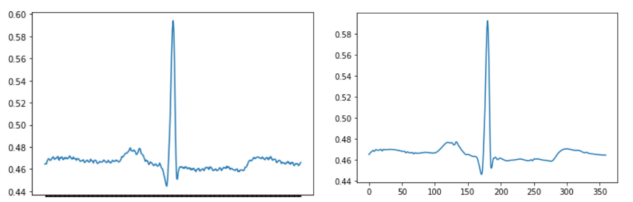
AE is used in imbalanced\_remover as the third way with 15 epochs.



After removing the imbalance, the shape of the data is as follows:



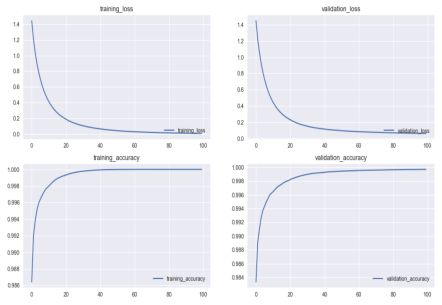
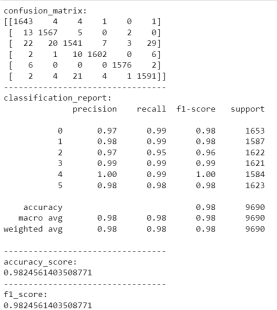
**1.7.1 noise\_remover**

Implemented two removers: Wavelet transforms using pywt library and the mean filter.

**2. training**

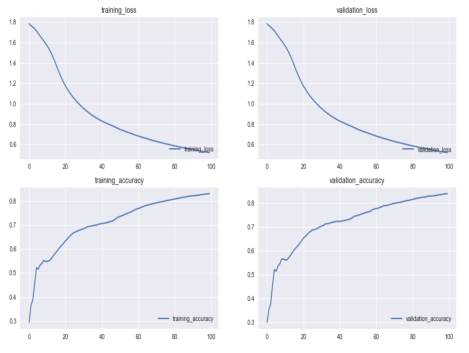
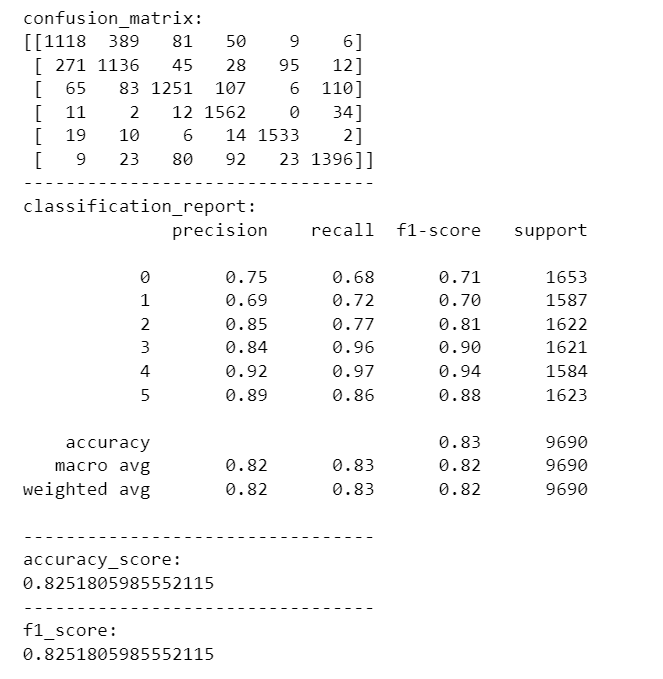
**2.1 model1**

The first model is lightGBM, which has 100 iterations, over 98% accuracy in the test set. Besides, here are some graphs and reports about it:



**2.2 model2**

MLP was used in this part, which is about 82% accuracy in the test set.



2.3 compare

|  |  |  |  |
| --- | --- | --- | --- |
|  | precision | recall | F1-score |
| lightGBM | 0.98 | 0.98 | 0.98 |
| MLPclassifier | 0.83 | 0.83 | 0.83 |