Group12 Project

Xiaoyu Liu[1]**,Hongyi Liu**[1]**, Tianhang Li**[1]**,Xiaofeng Wang**[1]**, Rui Huang**[1]

[1]Department of Statistics – University of Wisconsin-Madison

xliu969,hliu557,tli425,xwang2443,rhuang95

1. Introduction

The general purpose of our task is to predict each participant's age based on the brainwave data received. In this project, we perform an analysis on the data set of EEG signals which contains brainwave data and age of 1283 participants. We preprocess the raw data in parallel jobs by using the fast Fourier transform method to convert the time series data into frequency domain data, and then the power spectrum function is further used to obtain the power spectrum components, and the power density of the δ wave, θ wave, α wave, and β wave are obtained by integral calculation. Statistical models including KNN Regressor, Random Forest Regressor, SVR, and XGBoost Regressor are used for predicting age. The minimum MSE of the best XGBoost Regressor model reaches 48.85.

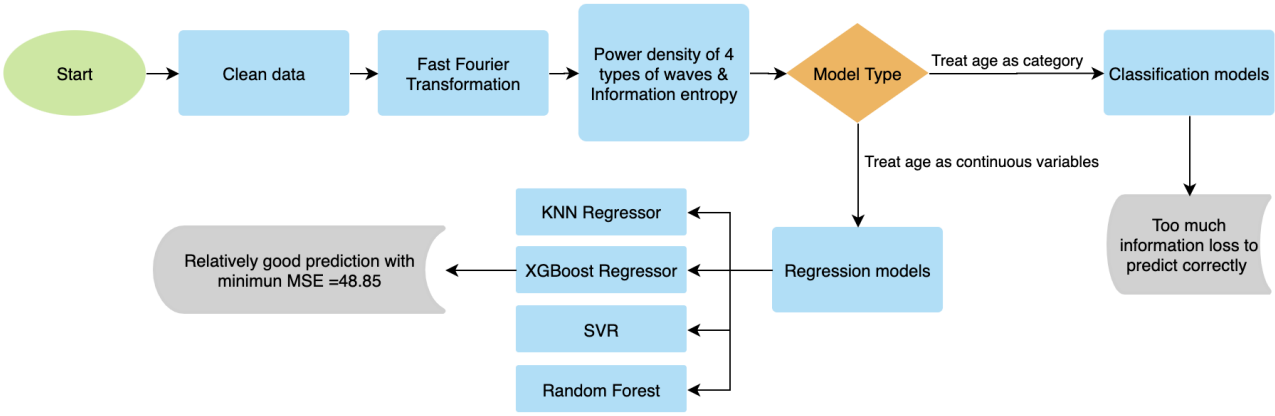

Figure 1. Project flowchart

2. Data

The data we use is EEG for Age Prediction from Kaggle. The data set contains brainwave data and age of 1297 participants which are all in CSV format. EEG is the recording of the electrical activity of a person's brain using a series of electrodes positioned strategically around a participant's head (Figure 3). In this data set, the signals are sampled with the frequency of 250HZ.

We use 1283 files from the data set because the size of other files is larger than 100MB which is beyond the expected size of CHTC input files. The size of each file we used is between 55MB and 100MB.

The first line of each CSV file is the age of the participant. The second line is the names of brainwave signals and tracked data for signals. There are 36 kinds of brainwave signals (some participants have less than 36 signals, but they have at least 24 signals in the CSV files), and each signal is followed by over 300,000 observations. In general, each column represents a different EEG channel, which can be divided into two types. The last four signals are PHOTIC-REF, IBI, BURSTS, and SUPPR. They represent photic sneeze reflex, interburst interval, burst, and suppression respectively. The other columns can be classified as a type that is linked with a certain zone in the brain (as the picture below shows). EEG features change as a function of age and health condition.

All the participants in this data set are assumed to be in similar health conditions. Therefore, the only cause of the variety we are considering is "age".

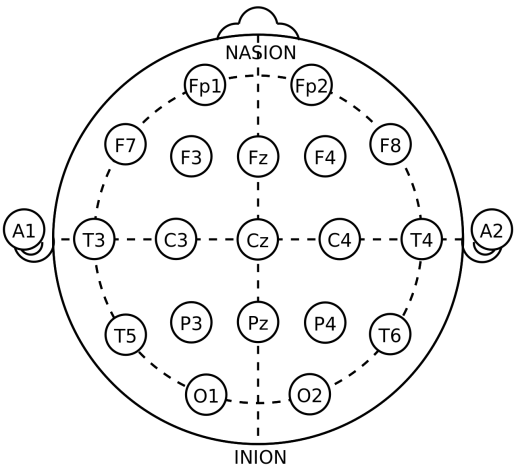

Figure 2. Electrode locations of EEG signals

3. Extracting feature

We firstly use CHTC to do data pre-processing. The single job will use a CSV file so there are 1283 jobs. Each job requires 1.5 GB as the requested memory. Besides, we use 1 CPU and require 10 GB of disk space. Each job will read a single CSV file and return 6 CSV files as outputs. In this processing, each job will also return the log files and the error files as the track of the transaction log and the error if there exist errors in the processing. Each job will take about 20 seconds and the submission takes about 19 minutes and 31 seconds.

What we want to do is to catch features in each signal channel, so for each sample, we have 24-36 "sub-features" to extract, we want to use the channels every sample has, which means we want to find the 24 common"sub-features" in every sample, so that data set will not include missing values. After this, we have 1283 samples(start with new_), each sample is a 24-36 dimensional vector in the sample space. Then we do dimensionality reduction that we only retain the coordinates which have the valid feature in all samples, in other words, not be implemented by us. The result should be a 24-dimensional space, which should be our feature space.

To find the feature of each participant, we read several papers about EEG signals. According to Al Zoubi et al., 2018, EEG signals could be transformed into four types of waves (Table 1).

δ(Hz)	θ(Hz)	α(Hz)	β(Hz)
1-3	4-7	8-13	14-30

Table 1. Four types of waves

Those waves can efficiently represent the activity and features of the brain, which indicates that the percentage of certain types of waves consisted of all types of waves might reflects age.

δ brainwaves are slow, loud brainwaves. They are generated in deepest meditation and dreamless sleep. δ waves suspend external awareness and are the source of empathy. θ brainwaves occur most often in sleep but are also dominant in deep meditation. In θ, our senses are withdrawn from the external world and focused on signals originating from within. It is that twilight state which we normally only experience fleetingly as we wake or drift off to sleep. α brainwaves are dominant during quietly flowing thoughts and in some meditative states. α is 'the power of now', being here, in the present. α is the resting state of the brain. α waves aid overall mental coordination, calmness, alertness, mind/body integration, and learning. β brainwaves dominate our normal waking state of consciousness when attention is directed towards cognitive tasks and the outside world. β is a 'fast' activity, present when we are alert, attentive, engaged in problem-solving, judgment, decision making, or focused mental activity. Besides, we also introduce the concept of "entropy"into the data.

To transform the original data into the four types of waves we mentioned above, we did the Fast Fourier Transformation, the formula is:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-i\pi k \frac{n}{N}} \quad k = 0, 1, 2, \ldots, N-1$$

Here X_k is the Fourier component, $X_n$represents the signal data in the data set, and N is the length of the signals, of course, we also need to change the index k to frequency based on frequency resolution.

Then we obtain the frequency domain spectrum (FDS) with a maximum of 250Hz, which is sample frequency. After this, for a better view of the signal features, we use the frequency domain to get the power spectrum, the formula is:

$$P_k = \frac{1}{N} |X_k|^2$$

The EEG of people mainly have 4 types of waves, they are δ(1-3Hz), θ(4-7Hz),α(8-13Hz),β(14-30Hz) wave, the power spectrum in these frequency bands varies with age. So based on the power spectrum, we can extract some features to help us predict the age. The power density calculated by the following formula can be a reasonable feature:

Suppose sets $\mathcal{A}_\sigma$,$\mathcal{A}_\theta, \mathcal{A}_\alpha, \mathcal{A}_\beta, \mathcal{A}_\gamma$ denote the frequency in the certain wave bandsfor example, the power density of the wave band σ is

$$p_\delta = \frac{\sum_{k \in \mathcal{A}_\sigma,} P_k}{\sum_{k \in \mathcal{A}_\sigma, \mathcal{A}_\theta, \mathcal{A}_\alpha, \mathcal{A}_\beta} P_k}$$

The power density for other waves can be calculated by the same formula, their sum is 1.

We also use entropy as one of the features to represent the change of waves. The information entropy reflects the degree of chaos, the entropy, representing the power proportions of 4 waves, changes greatly when the proportion of each wave changes.

Now we obtain 4 power density p in each channel, for this time, we decide to use information entropy H as the feature for each brainwave signal:

$$H = \sum_{i \in \sigma, \theta, \alpha, \beta} -p_i \ln p_i$$

Every sample has 1 entropy in each coordinate, if the coordinate has no valid data, the entropy should be denoted as NA. So our data is changed into 1283 samples, each sample records its entropy. Moreover, the proportion of a certain wave is also recorded for further analysis.

We save the entropy and 4 waves in a new CSV file for every data set, then we have 1283 new CSV file, after deleting the NA column in the files, we can easily get the name of every column, then we use functions to find the union set of names, the number should be 24, save all the data below these names to 1283 new files and these are our feature space.

After pre-processing the original data sets, we combine them into five merged set. The entropy set includes each observation's age and the entropy in each of the EEG signal channels. Wave sets of four types of waves include each observation's age and the power density of each type of wave in each of the EEG signal channels.

4. Statistical Models

Last time we grouped ages into age groups, which might lose most of the information. When treating the problem as a classification problem, we noticed that the accuracy of prediction invalidation might not prove the accuracy of the model itself. Because there are only a few groups and random prediction might get rather high accuracy. The other problem is that age of our samples is not uniformly distributed and a large part of them are under 40 years old, which might cause unbalance if we group them. However, if we do not group them, the problem of unbalance might be less serious because samples of a certain age are less centered compared to the case of groups. (Figure 3)

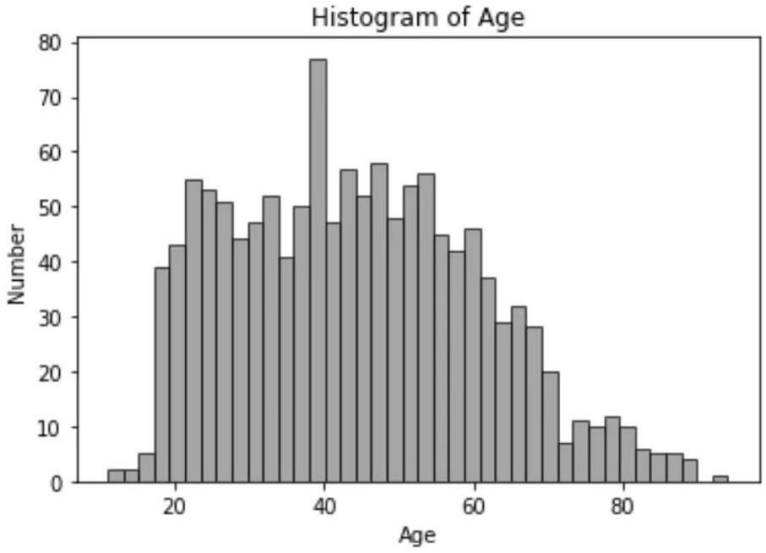

Figure 3. Distribution of age

Therefore, this time we treated age as continuous variables that could be fitted with regression models.

To choose proper features among the features we extracted as mentioned above, we calculated the Pearson correlation of Features and drew the scatter plots of the data set. From the Pearson correlation between alpha wave and age, we notice that more than 4 signal channels' correlation coefficients between age are larger than 0.1.(Figure 4) We also noticed that the wave index from signal channels are very likely to be linearly related, which indicates that we need to solve the linearity between variables.

Among those features, we chose the power density of θ wave, the power density of α wave, and entropy for further research according to the plots and research from Kanokwan.etl, 2019.

We applied PCA to eliminate the linearity between explanatory variables(namely the value in different signal channels). Then we tried five types of models to predict the age of a participant this time. To obtain optimal hyperparameter, we use the grid-search method to calculate the accuracy of a different combination of hyperparameters on split sets. (Table 2) The metric we chose is mean squared error(denoted as MSE), which can reflect the difference between our prediction and the true value.

According to Figure 7, Figure 8, Figure 9, we chose the optimal hyperparameter for KNN model. The hyperparameter we focus on this time is the number of k. We only included the process of choosing the hyperparameter for KNN model as an example for there is only one hyperparameter to choose.

As shown in Figure 7, Figure 8 and Figure 9, the MSE of the training set is increasing when parameter k is increasing, which is reasonable because when k increases, the prediction took more data points into concern when finding nearest points in neighborhood and this would increase the

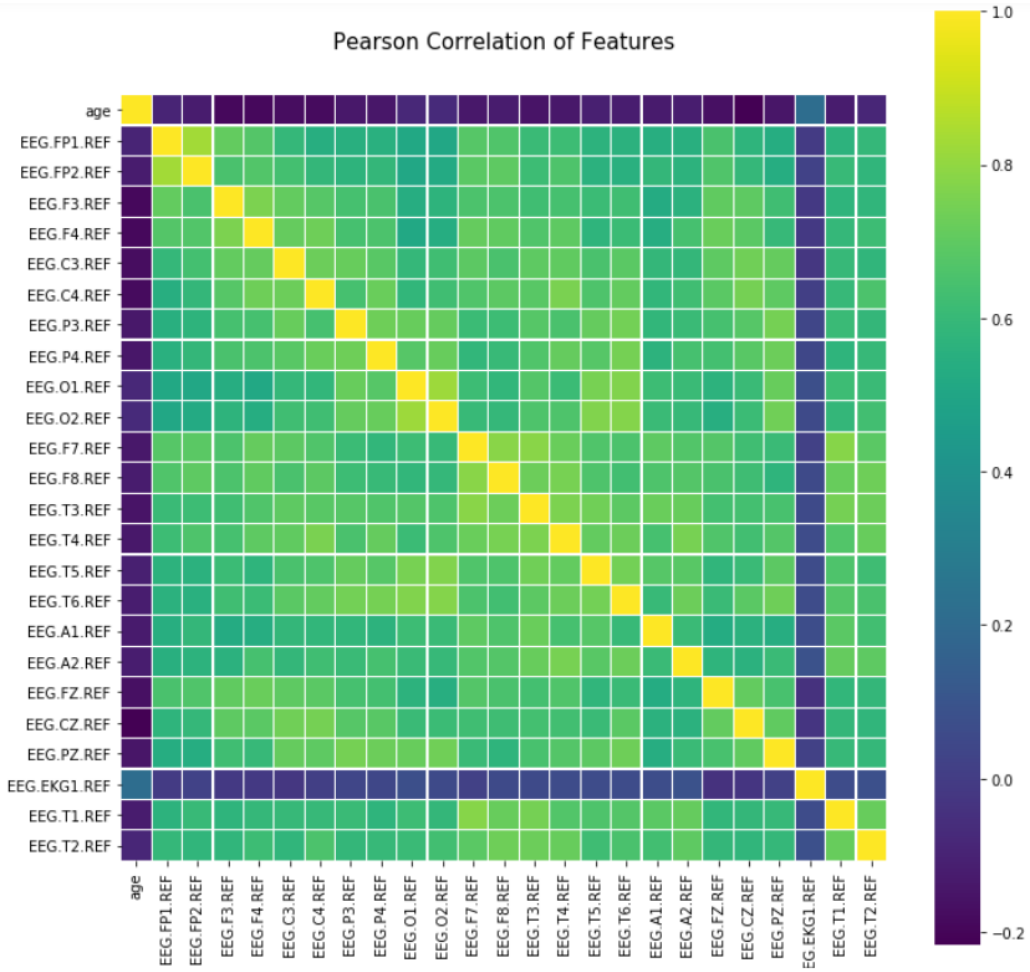

Figure 4. Pearson relationship of θ wave

generality at the cost of bias. At the same time, as k increases, MSE of testing set is decreasing before it reaches the best point, which indicates that the model is more accurate at the best point. However, the MSE of validation sets is not high enough especially for the model of α wave and entropy. On the other hand, the trend of accuracy as k increase is not very reasonable either because the trend is not increasing steadily but keeps changing the direction of variation.

Compared with the model of α wave and entropy, the model of θ wave performs much better on MSE. Thus, we think the power density of θ wave might be a more proper feature for the KNN regressor model. We also compared those features in the random forest model, XGB regressor model, and SVR.

We applied similar methods when fitting XGBRegressor, SVR, Random Forest model. Regression result could be found in Table 2.

After comparing regression result of models including feature of θ wave, α wave and entropy, we found out that MSE of regression models are smaller than others if we choose θ wave as the feature. On the other hand, XGB regressor seems to have the best performance on prediction. The best model should be the XGB regressor considering the power density of θ wave, with the MSE of 48, which is rather small compared with the mean of ages(44.24) or variance of ages(270.08). Therefore, we think the prediction is successful and our process of dealing with EEG data is reasonable.

5. Source

[1] EEG data, Kaggle, https://www.kaggle.com/ayurgo/data-eeg-age-v1

[2] O.Al Zoubi, C. Ki Wong, R. T. Kuplicki, H.-w. Yeh, A. Mayeli, H. Refai, M. Paulus, and J.

Bo-durka, “Predicting age from brain eeg signals—a machine learning approach,”Frontiers in Aging-Neuroscience, vol. 10, p. 184, 2018.

[3] S. Kanokwan, W. Pramkamol, K. Wipatcharee, W. Warissara, R. Siwarit, S. Sompiya, B. Onuma, and S. Mitra, “Age-related differences in brain activity during physical and imagined sit-to-stand inhealthy young and older adults,”Journal of Physical Therapy Science, vol. 31, no. 5, pp. 440–448,2019.

[4] Brainwaves, https://brainworksneurotherapy.com/what-are-brainwaves

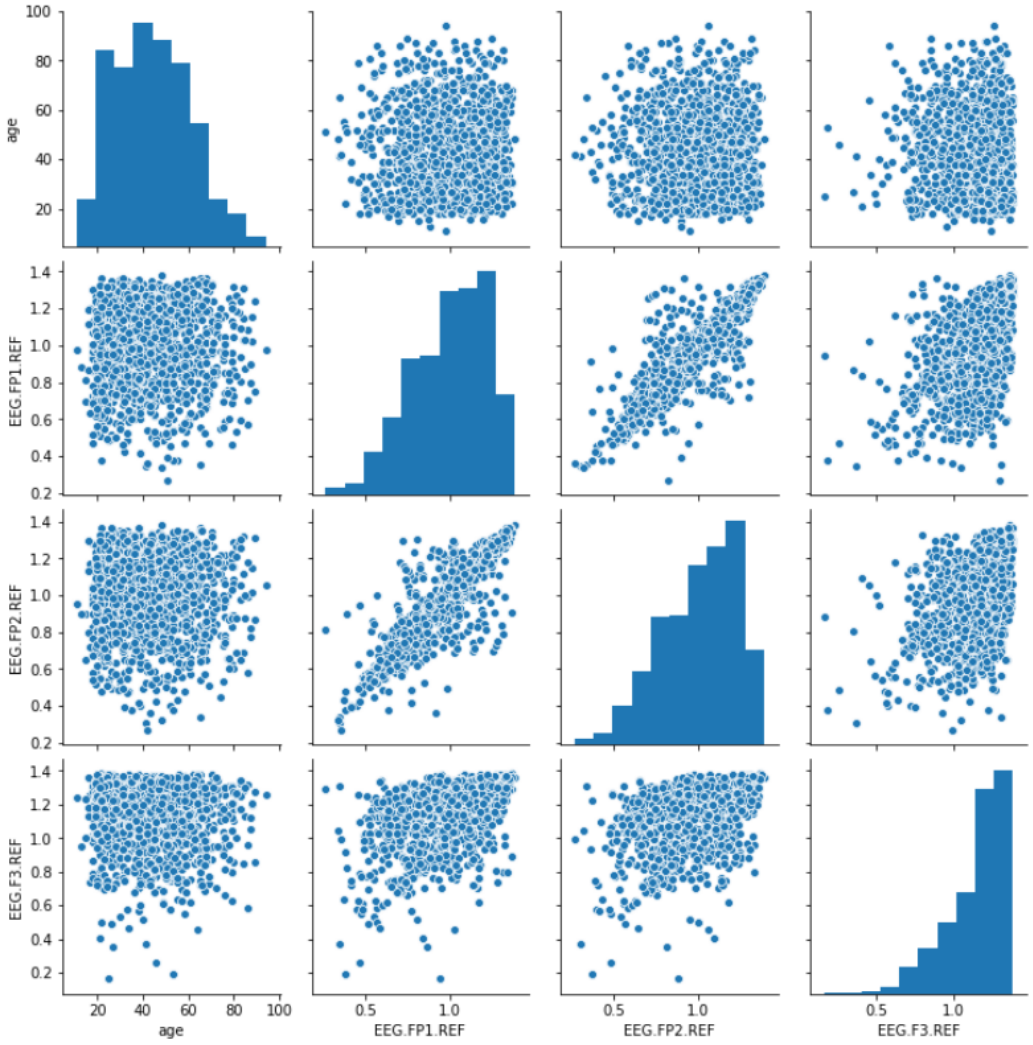

Figure 5. Scatter plot of α wave

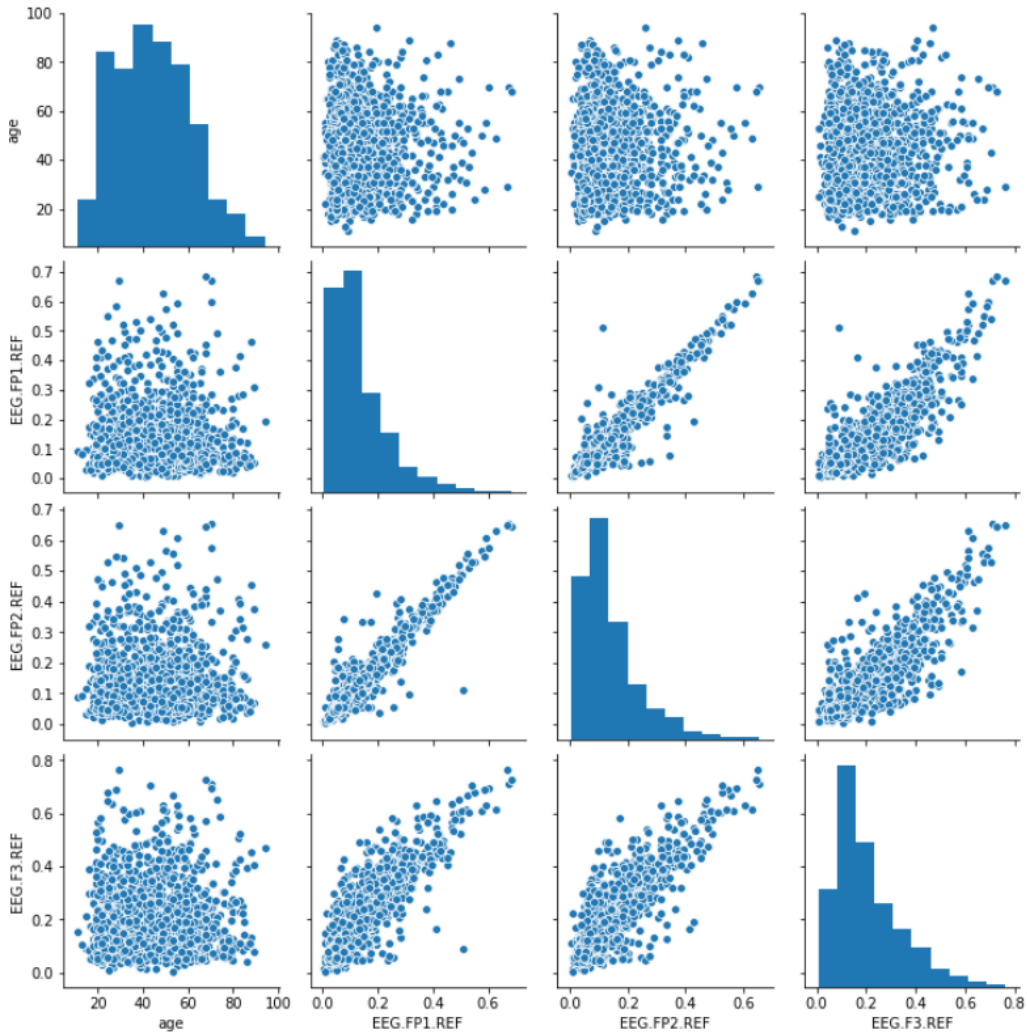

Figure 6. Scatter plot of entropy

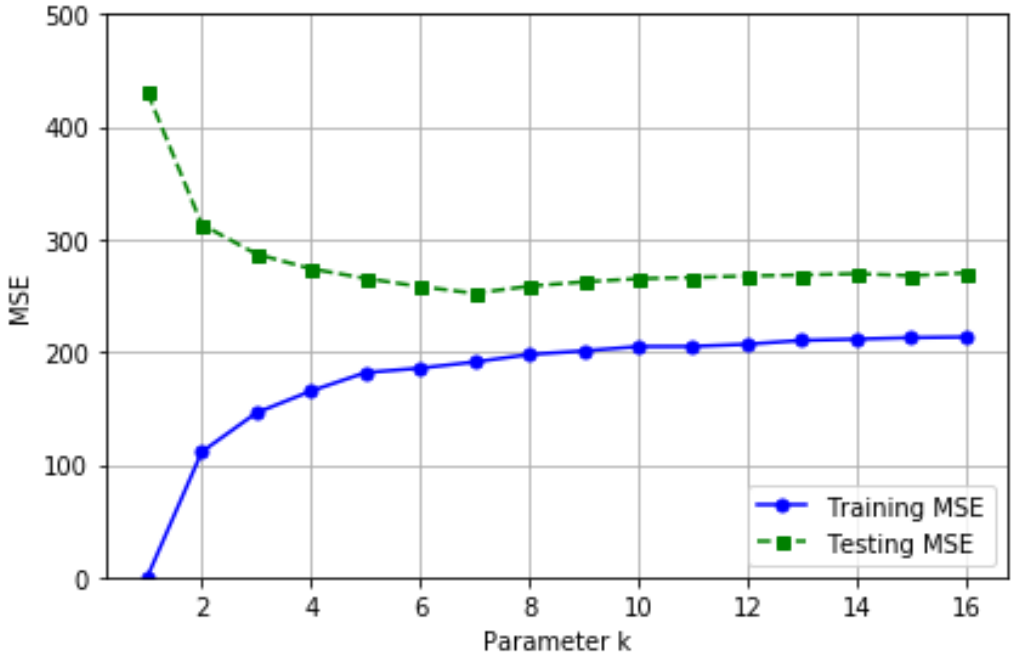

Figure 7. Tune parameter for KNNRegressor of θ wave

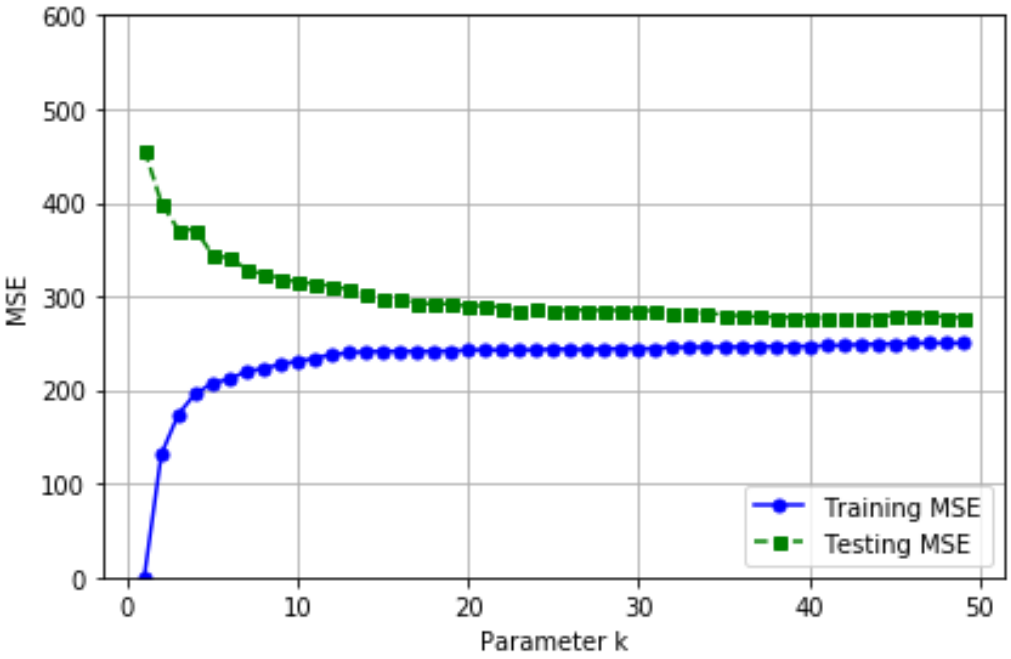

Figure 8. Tune parameter for KNNRegressor of α wave

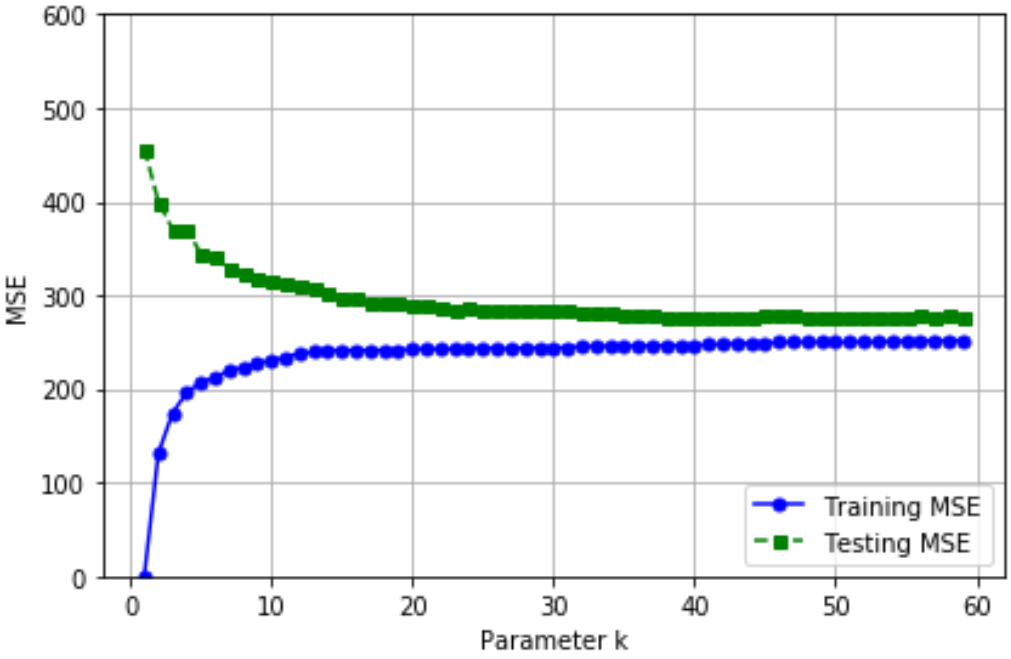

Figure 9. Tune parameter for KNNRegressor of entropy

Method	Feature	Hyperparameters	MSE
KNN Regressor	α wave	K(number of neighbors): 37	280.12
KNN Regressor	θ wave	K(number of neighbors): 7	256.08
KNN Regressor	Entropy	K(number of neighbors): 52	280.06
XGBRegressor	α wave	Learning Rate: 0.006, Max Depth: 5, Min Child_weight: 5	308.35
XGBRegressor	θ wave	Learning Rate: 0.01, Max Depth: 6, Min Child_weight: 3	48.85
XGBRegressor	Entropy	Learning Rate: 0.006, Max Depth: 5, Min Child_weight: 5	315.89
SVR	α wave	C: 5, Coef0: 0.01, Degree: 3, Gamma: scale, Kernel: rbf	283.65
SVR	θ wave	C: 10, Coef0: 0.01, Degree: 3, Gamma: scale, Kernel: rbf	281.56
SVR	Entropy	C: 1, Coef0: 0.01, Degree: 3, Gamma: scale, Kernel: rbf	276.68
Random Forest Regressor	α wave	Max Depth: 15, Max Features: sqrt, Max Leaf Nodes: 30, N_estimators: 80	276.17
Random Forest Regressor	θ wave	Max Depth:20, Max Features: auto, Max Leaf Nodes: 32, N_estimators: 100	248.93
Random Forest Regressor	Entropy	Max Depth:16, Max Features: sqrt, Max Leaf Nodes: 40, N_estimators: 55	272.94

Table 2. Regression results of all methods