CSE 572 - Data Mining

Assignment 1

Group 22:

ARAVIND THILLAI VILLALAN	1215121258	athillai@asu.edu
MOTTHILAL BASKARAN	1215168292	mbaskar2@asu.edu
SANDHYA CHANDRASEKARAN	1215159426	schand61@asu.edu
VAIDHEHI VASUDEVAN	1215127381	vvasude7@asu.edu

1. INTRODUCTION

Type 1 diabetes is a serious illness which requires complete monitoring of blood glucose levels. It can be treated with a proper nutrition plan, exercise and balanced medication and continuous monitoring using CGM or insulin basal and bolus infusions. CGM (Continuous Glucose Monitoring) sensors which when placed on a patient, monitors glucose levels every few minutes. We study the data from CGM sensors belonging to 5 patients.

Each CGM series data file has observations of the tissue glucose levels every 5 mins for 2.5 hrs during a lunch meal. The time series data begins at 30 minutes before the meal intake and goes upto 2 hrs after meal intake starts. The time stamps for each of the time series data is also recorded. Similarly, the time series data and the corresponding time stamps for the insulin basal and bolus infusions are recorded.

In the first phase of the project, we consider only the time series data obtained from CGM sensors for 5 subjects. Initially, the data is cleaned using pre-processing techniques. Using the pre-processed data, feature selection is performed and several features are analyzed. We create a feature matrix where each row is taken as a combination of features from each time series. This feature matrix is given as an input to PCA and the new feature matrix returns five principal components. The list of all the features obtained during feature selection, the performance of each of the features and principal components obtained from PCA for the CGM time series data and time stamp data are discussed in this first phase of the project.

1.1 Data Preprocessing

Before we extract the features we need to preprocess the data. Preprocessing the data means transforming the raw data into an understandable format. The given data in the first cell array which has tissue glucose levels every 5 mins for 2.5 hours during a lunch meal is inconsistent with missing or null values in between. So, we need to clean the data before we start extracting features from it. The data preprocessing technique that we have used here is that we calculate the mean for each row or each day's data and then we fill the missing/null values in that particular day's data with the mean value of that day.

2. FEATURE EXTRACTION

a) Extract 4 different types of time series features from only the CGM data cell array and CGM timestamp cell array.

The feature extraction techniques that we have used are,

1. Skewness
2. Entropy
3. Correlation
4. Velocity (Mean of Differences)

We have extracted these features for our dataset and the implementation is available in the Python code file.

b) For each time series explain why you chose such feature.

Skewness

Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The shape of the normal distribution of the values is depicted by the skewness. If the skewness value lies between -0.5 and +0.5, it is moderately skewed. If the value is 0 then the data is symmetric and if the value lies above +1 or below -1, then the data is said to be extremely skewed. For a normal distribution, skewness is zero. We use skewness as a feature to find how much the given distribution varies from the normal distribution.

We need to observe the glucose levels to find out when the person consumes food. By using skewness as a feature, we can find the deviation of this distribution. This deviation represents the increase in glucose levels which corresponds to the time at which food is consumed.

Entropy

Entropy is the rate of information production. Sample entropy is used for determining the complexity of physiological time series signal. It is more advantageous than approximate entropy as it is independent of the data length and exhibits relative consistency. The improved accuracy of Sample Entropy statistics makes them useful for biological time series. SampEn(m, r, N) is the negative natural logarithm of the conditional probability that two sequences similar for m points stay similar at the next point. Here, self-matches are not included in the calculation of probability. A lower value of Sample Entropy indicates more self-similarity in the time series.

Entropy is an important feature for biological time series data. Entropy is used to find the disorder in data. Encountering disorder in data is extremely critical for physiological analysis. So, we used Entropy as one of the primary features.

Correlation

Autocorrelation is the degree of correlation between the values of the same variables across different observations in the data. Autocorrelation is used in time series data where observations for same variable occur in different points of time. Correlation is the relationship between two quantities. Correlation is used for the prediction of one quantity from another. If the correlation coefficient is higher, both quantities are correlated and one of these can be removed. If the correlation coefficient is zero, the quantities are independent of each other and if the correlation coefficient is negative, both quantities are inversely proportional. We calculated the Autocorrelation values for each day and added it to the feature matrix.

In our data, there is a pattern of increase in glucose level whenever there is a food intake, which keeps repeating for each person. So, we used Autocorrelation feature to see if the variable is correlated.

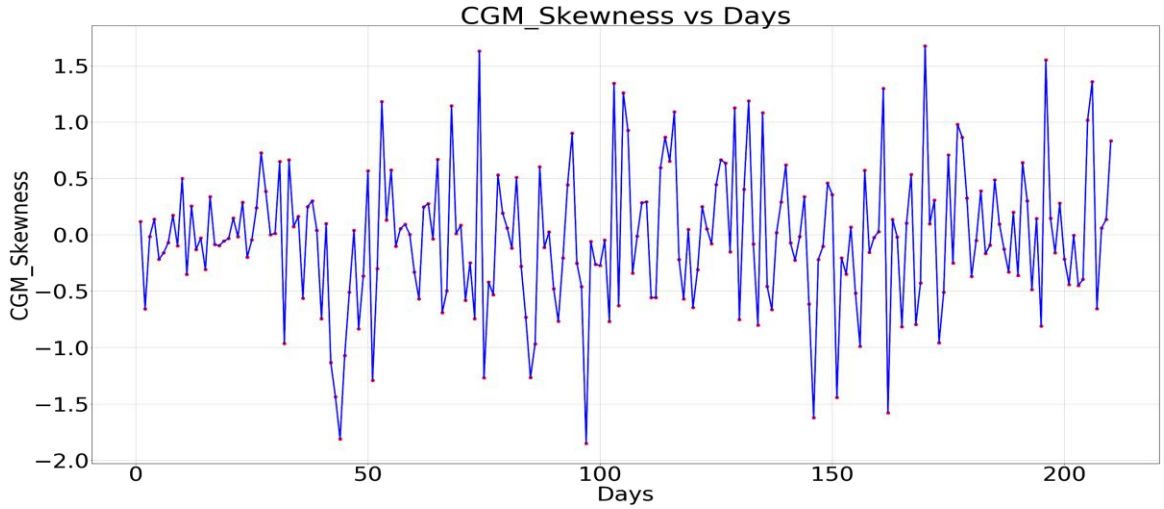
Velocity (Mean of Differences)

The velocity is required in cases where a large amount of data is involved. Here, the CGM sensor data is collected for an interval of every 5 mins for 5 subjects for a total of 210 days. The velocity is the mean of differences for a set of data. We take differences between consecutive data points in a set of data. We then perform mean of these difference values.

The mean of differences acts as an important metric since it shows where the eating pattern of the person occurs. The difference in consecutive points increases when there is a peak. This shows that when there is an increase in glucose level, the sum of differences increases which increases the mean of differences.

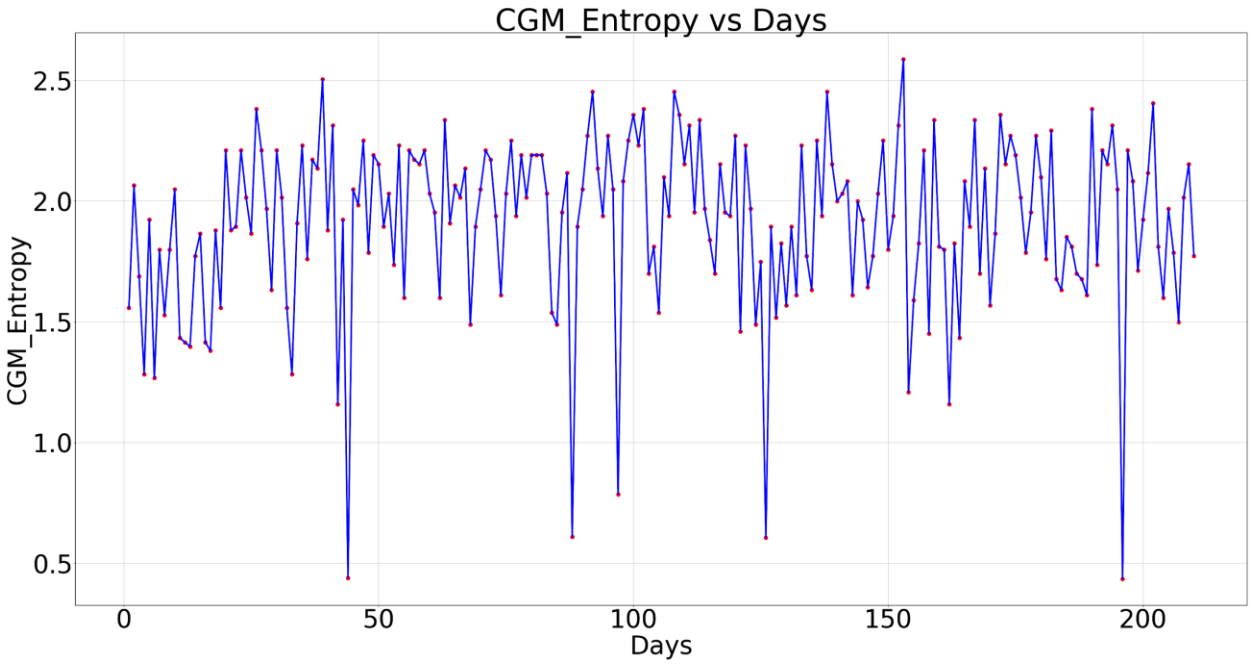
c) Show values of each of the features and argue that your intuition in step b is validated or disproved?

Skewness



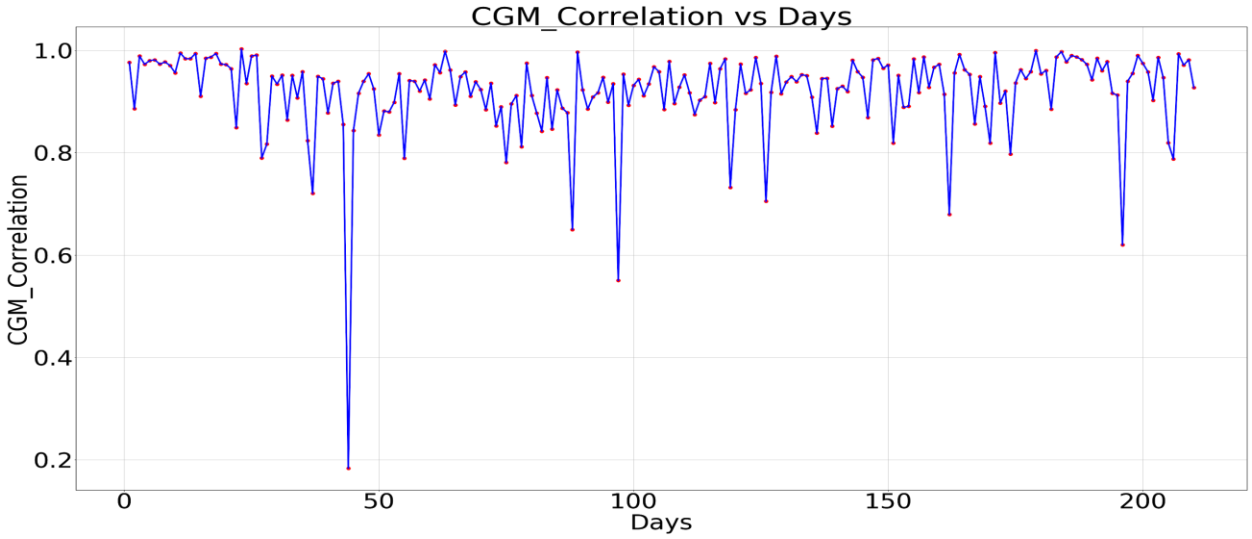
From the data matrix(first cell array), for each row which represents each day, we calculate the skewness for that particular day. Then we have plotted a graph that represents the skewness for all the 5 patients where the x axis represents the ‘days’ and the y axis represents the ‘skewness’ value. From the above plot, we find that our data is skewed since the plot has sudden rise and fall. We added skewness to PCA to validate its performance.

Entropy



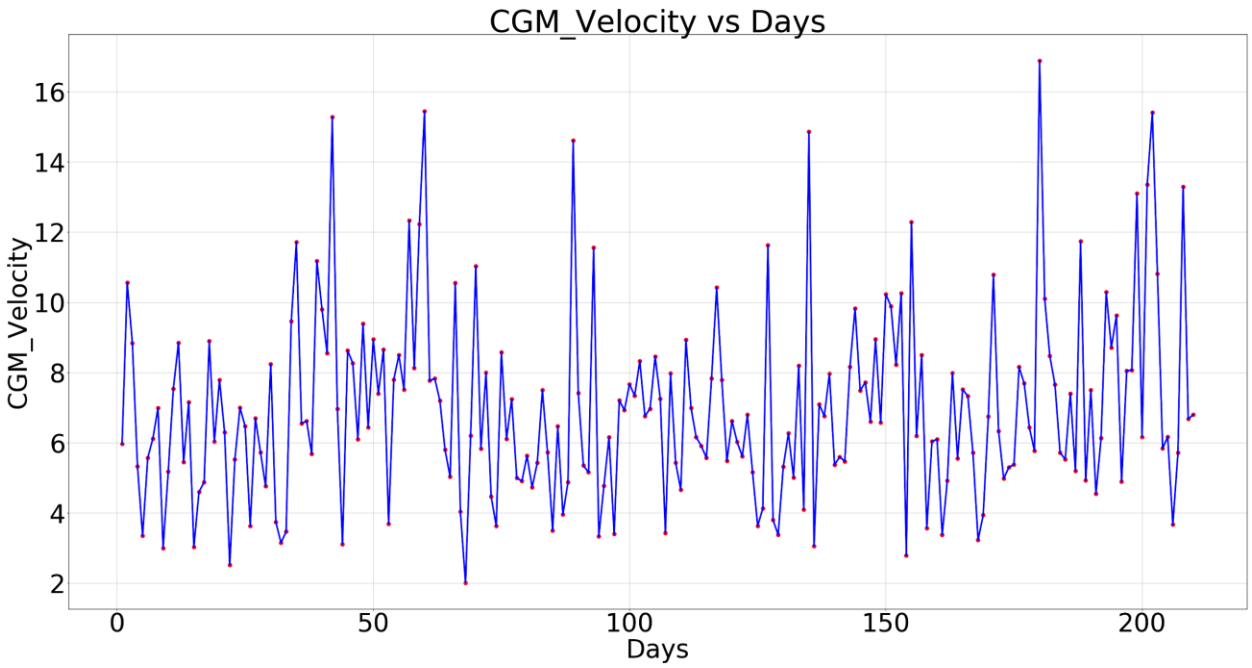
We have plotted a graph for representing Entropy with respect to the number of days for five patients. We assumed that our data has high disorderliness and so we chose to have entropy as a feature. The medical data should not be disordered. From the plot, we can observe that the entropy has pointed out all the disorderliness in the data.

Correlation



We have plotted a graph to represent the correlation with respect to the number of days for five patients. From the graph, we can infer that the data is correlated since there is a repetition in the pattern. Our assumption that the data is correlated is correct and is evident from the graph.

Velocity (Mean of Differences)



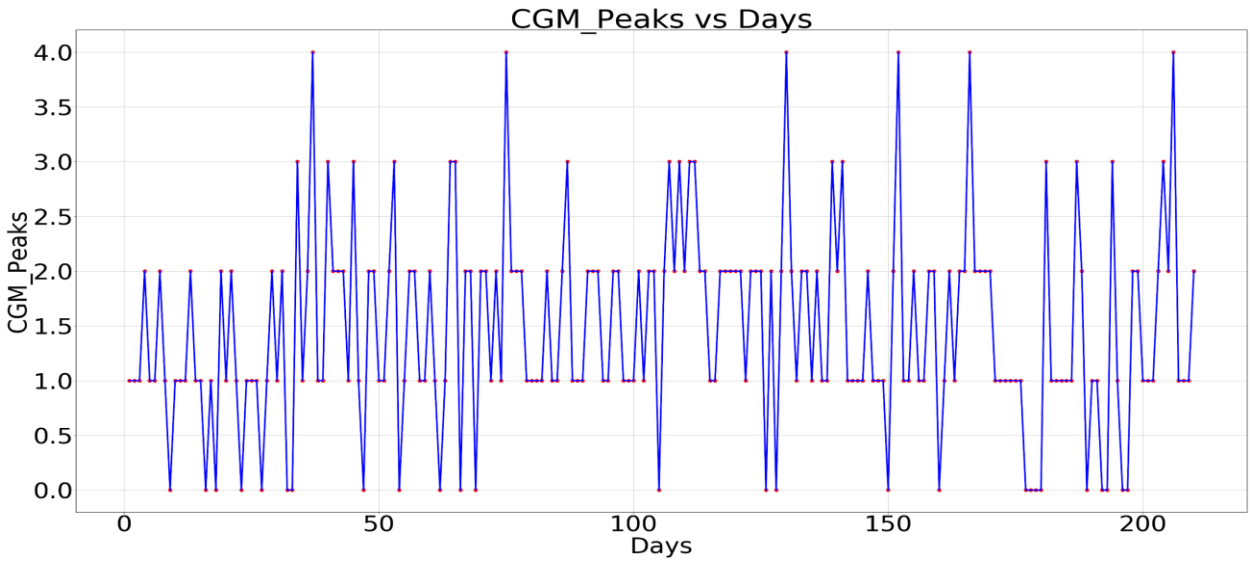
We plotted a graph representing the Velocity or Mean of Differences with respect to the number of days for all the patients. We thought that velocity is a good feature since it directly corresponds to the sudden increase in glucose levels. The graph shows sudden increase in glucose levels with increase in the mean of differences.

OTHER FEATURES

We have considered additional features for the feature selection process using PCA.

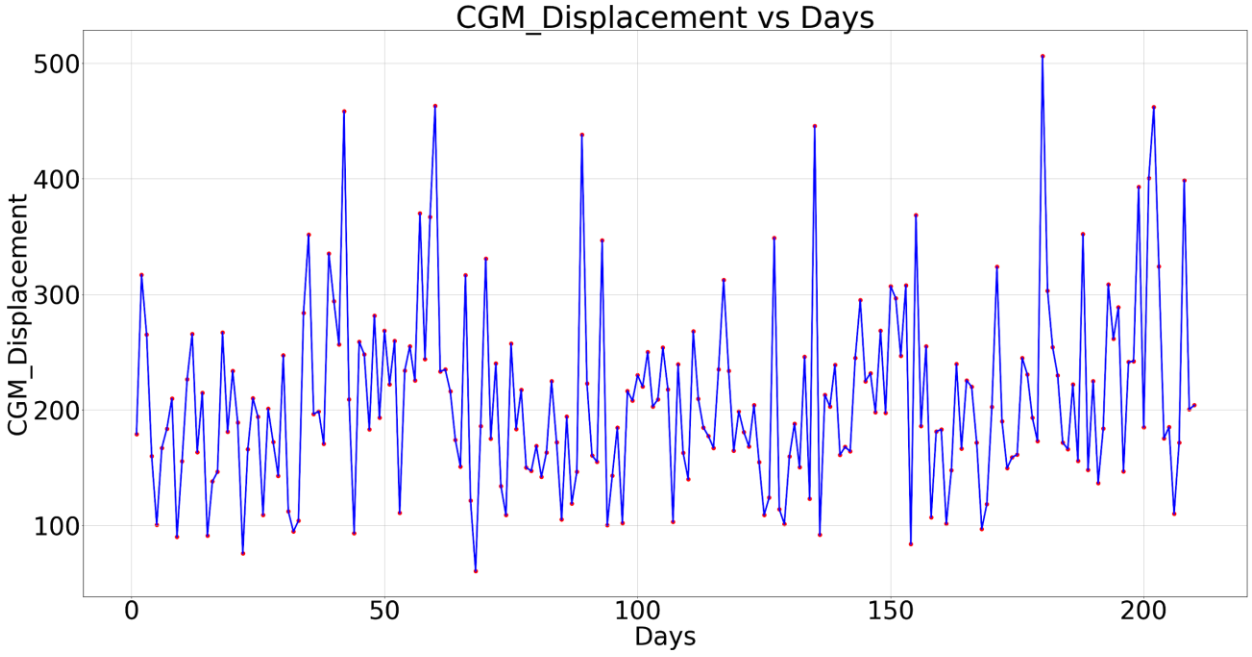
Number of Peaks

Peaks indicate significant events such as food intake in our example. We calculated the number of peaks in each day which indicates the presence of important events like the intake of food in data. We added this feature to the feature matrix.



Displacement

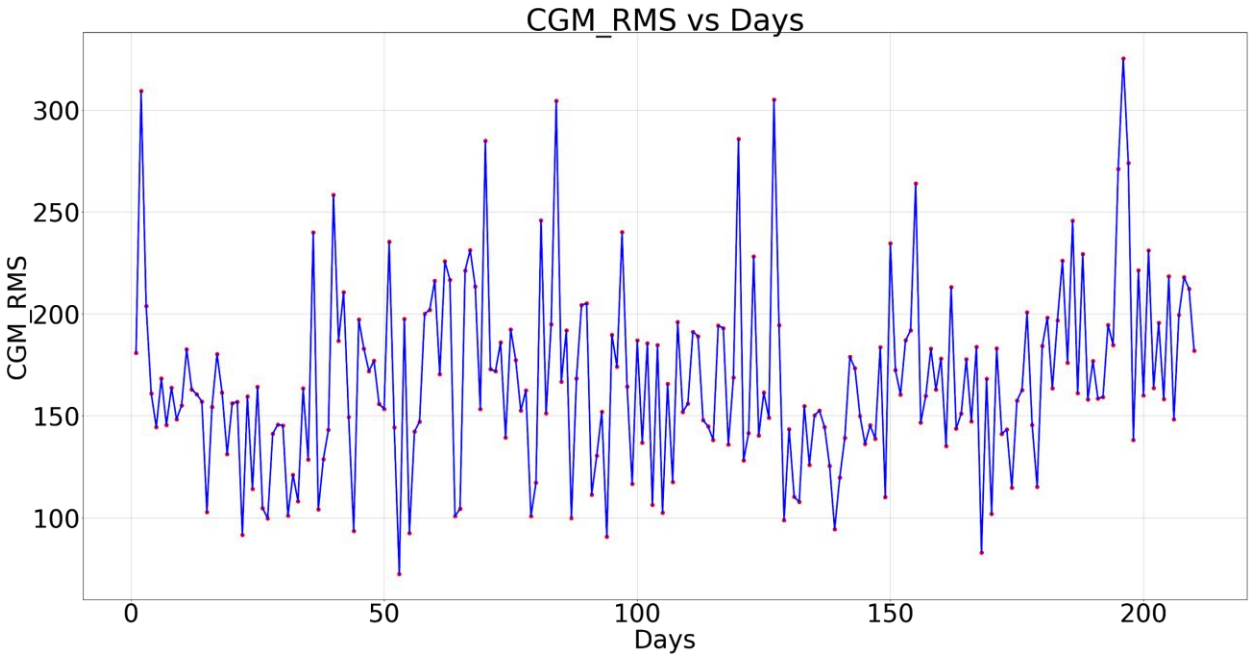
The Displacement refers to the sum of differences in data. The difference between every two consecutive points is performed on a set of data. The sum of these differences gives the displacement in data.



RMS

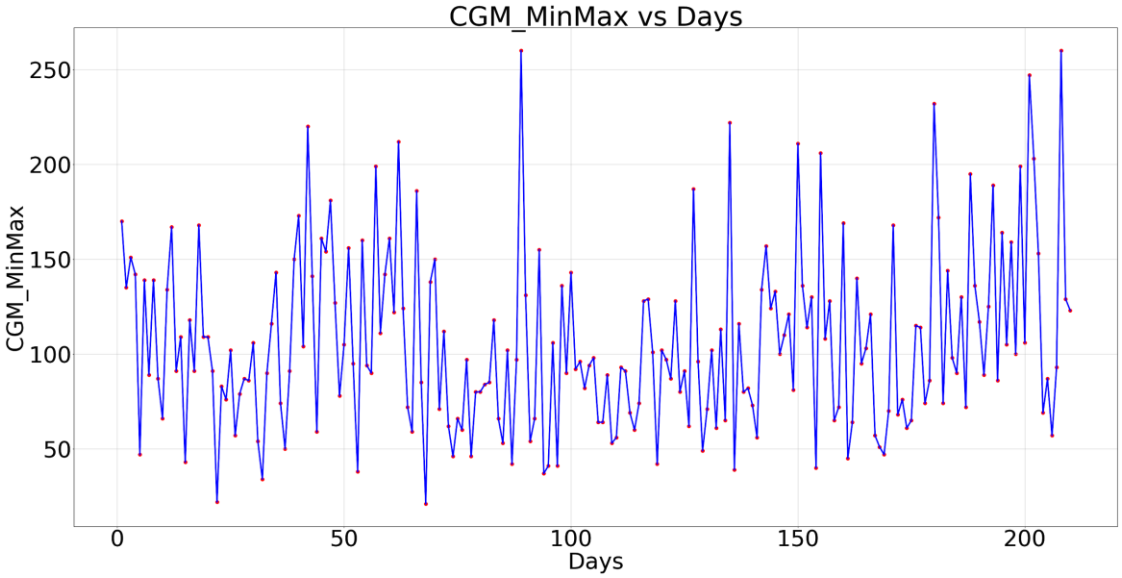
The Root Mean Square or the quadratic mean is the square root of the mean square of the data points. Mean square is the arithmetic mean of the squares of the data points. The RMS value is used to represent the magnitude of data. The average of a set of data points does not

give us the magnitude of that set. So, RMS is a method of finding the magnitude by taking the average of unsigned values. RMS is proportional to the amount of energy produced over a period of time. We found the RMS values for each day and added it to the feature matrix.



MinMax

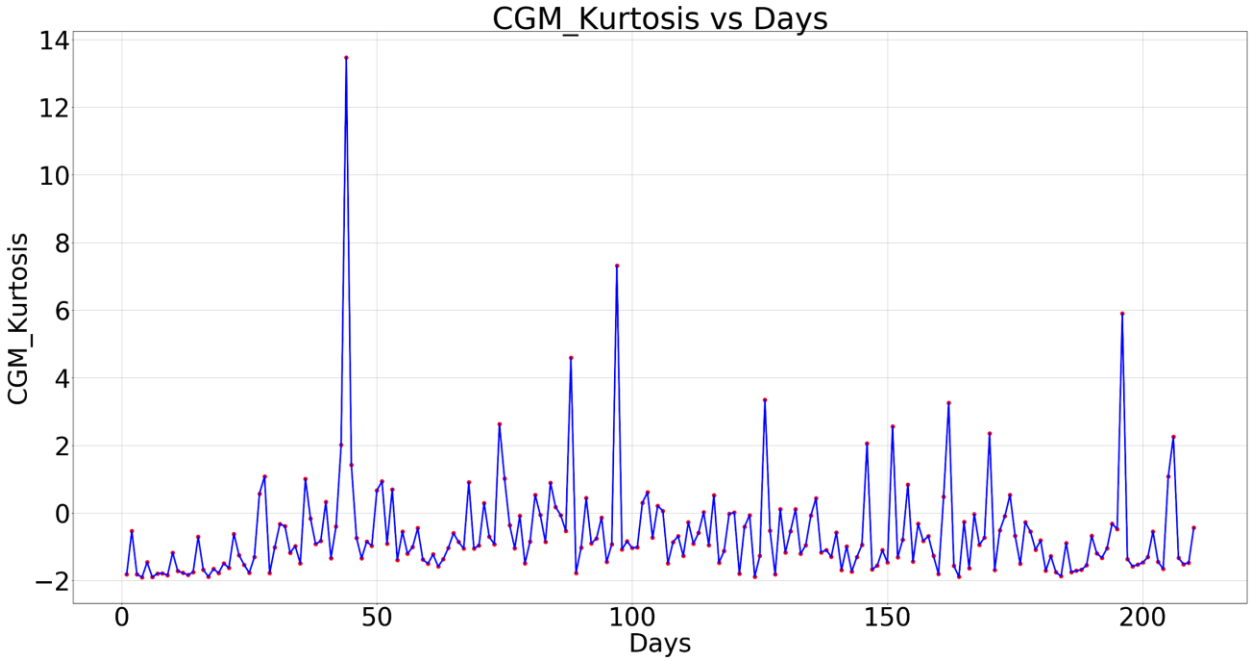
MinMax refers to the difference between maximum and minimum data points from the dataset. MinMax defines the range of dataset and is the simplest measure of spread. The range feature can be useful when we have a variable that should not cross a certain threshold.



Kurtosis

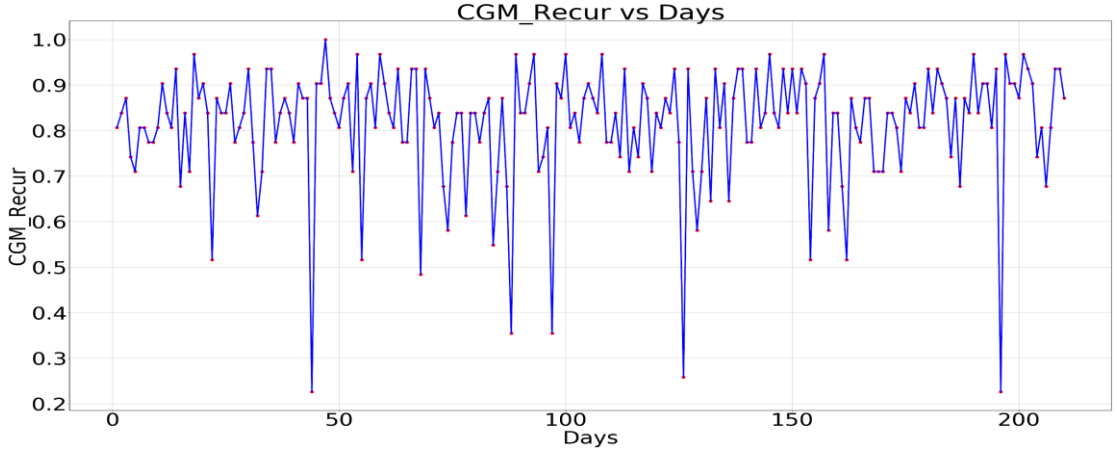
Higher the value of kurtosis means that the points lie away from the mean and actually has more points lie away from the mean. Kurtosis gives measure of distance of the points away from the mean by measuring the fourth central movement of the variable. Kurtosis value is generally referred to as the fourth central movement of the variable (X) to the square of the Standard deviation with a difference of 3.

$$\operatorname{Kurt}[X] = \operatorname{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mu_4}{\sigma^4} = \frac{\operatorname{E}[(X-\mu)^4]}{(\operatorname{E}[(X-\mu)^2])^2},$$



Value to Time Series Ratio (Recur)

Returns a factor which is 1 if all values in the time series occur only once, and below one if this is not the case. In principle, it just returns the ratio of the number of unique values to total number of values present.



d) Create a feature matrix where each row is a collection of features from each time series. So, if there are 75 time series and your feature length after concatenation of the 4 types of features is 17 then the feature matrix size will be 75 X 17.

PRINCIPAL COMPONENT ANALYSIS

Principal component analysis is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components [1]. It is a statistical procedure which by using a smaller set of "summary indices" allows you to summarize the content in large data tables. PCA works by projecting the data geometrically onto lower dimensions called principal components. Such dimensionality reduction is really useful for visualizing and processing high-dimensional datasets, while still maintaining as much of the variance in the data possible.

Creating a feature matrix:

During the feature extraction process, we collected upto 10 features to give it as an input to the PCA. Entropy, RMS, Correlation, Peaks, Velocity, MinMax, Skewness, Displacement, Kurtosis, Recurrence. The values of these features are calculated for the combined dataset of 210 records and stored in individual dataframes. These dataframes are merged as a single feature dataset matrix and this matrix is given as an input to PCA. Now, the feature matrix will be of the size 210x10. The 210 rows are obtained by combining the rows of all five subjects. We perform PCA on this feature matrix for feature selection. PCA returns a new feature matrix from which most useful features can be inferred.

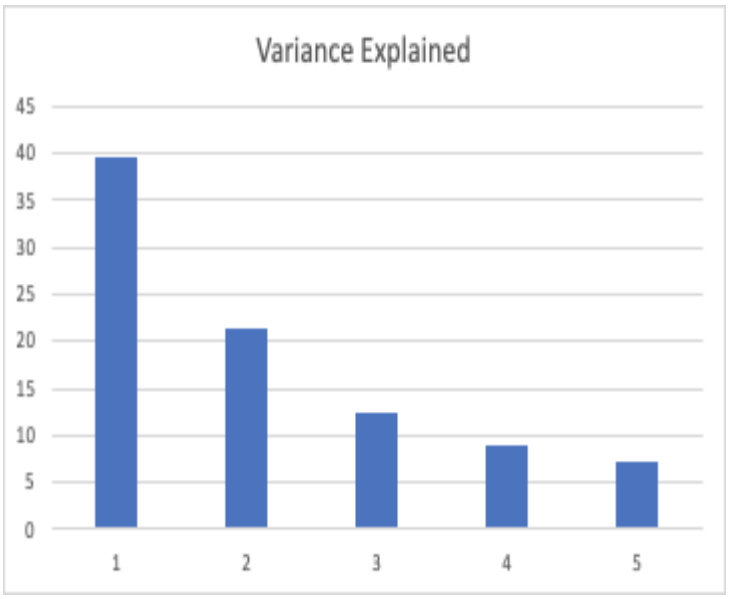
e) Provide this feature matrix to PCA and derive the new feature matrix. Chose the top 5 features and plot them for each time series.

Execution of PCA:

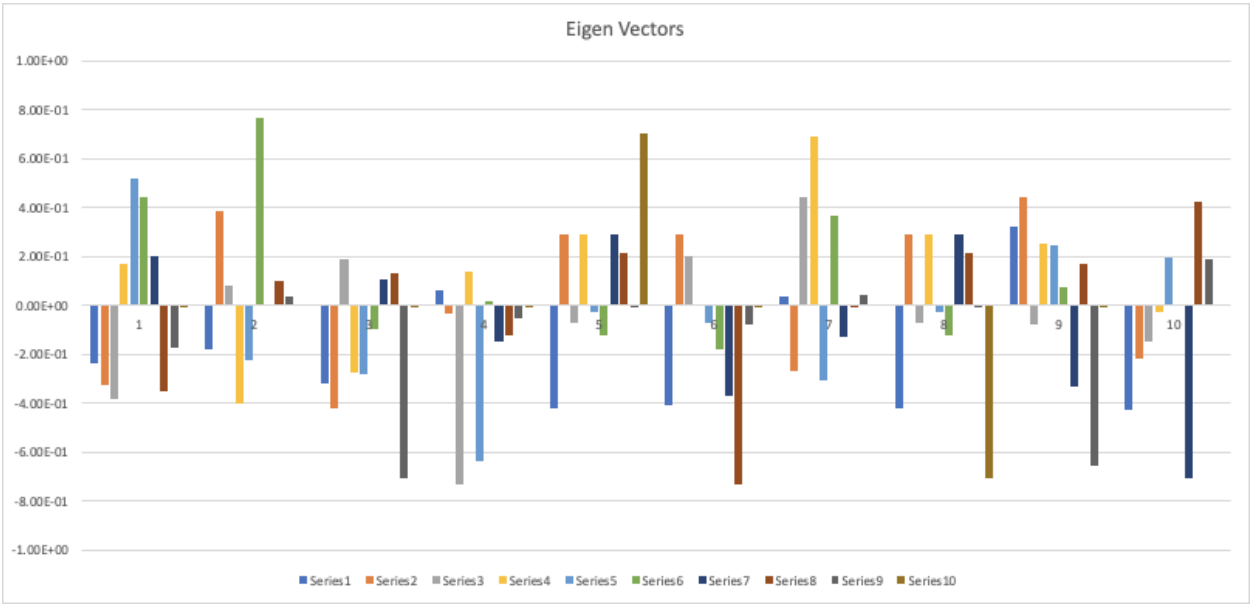
We give the feature matrix of size 210x10 as an input to PCA. PCA is implemented in Python using SciKit learn. Since the feature matrix consists of values of all the features, it is important to scale the data before performing PCA. We used StandardScaler to perform normalization of the data. PCA gives the principal components and variance. The new feature matrix is of size 210x10 and therefore we obtain 10 principal components. Now we have to select the top 5 features from the obtained 10 principal components. Finding eigenvectors and eigenvalues of the covariance matrix means fitting the principal components along the variance of the data. The 10x10 'coeff' matrix returned by the PCA represents the eigenvectors in the decreasing order of their variance from left to right which implies the principal components. We can drop those eigenvectors which have low variance since they are less representative of the data and we can chose those top 5 eigenvectors with higher weightage of the variance. We identify the weightage of the eigenvectors using SkLearn's PCA with the command 'explained_variance_ratio_'

In order to choose the top 5 principal components which represents the highest variance among the data we calculated the weighted ratio of all the top 10 components.

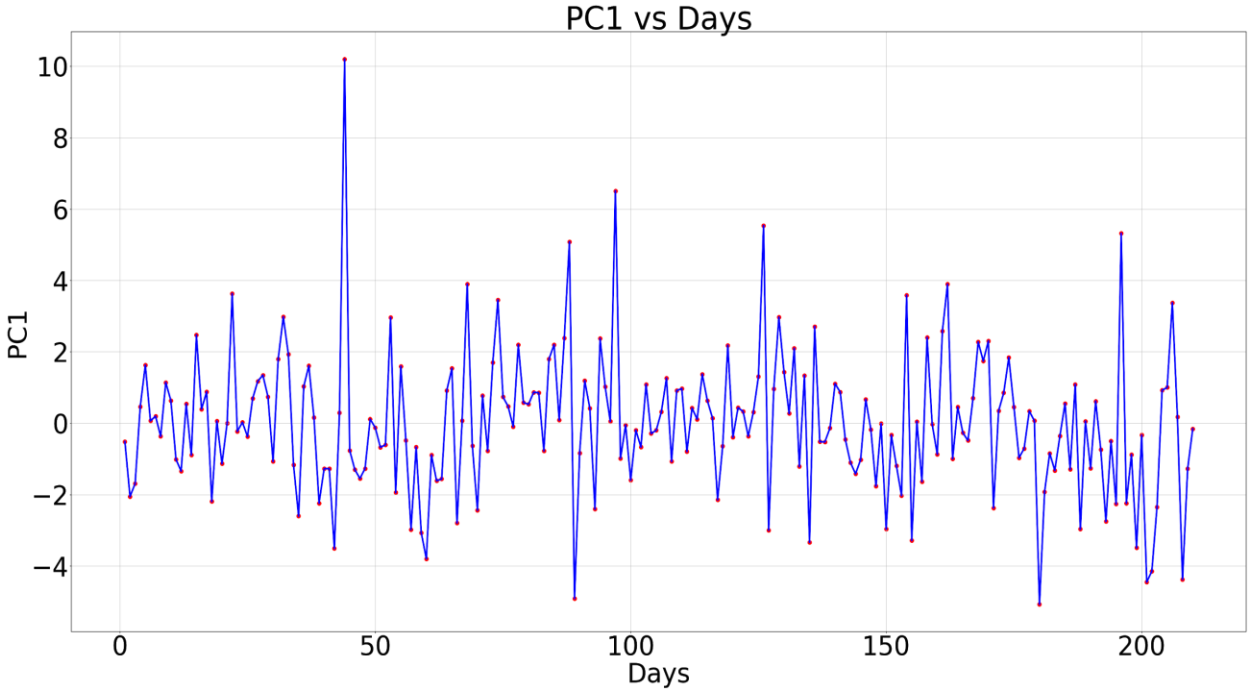
	Weighted_Ratio
0	0.395778
1	0.213109
2	0.123849
3	0.091528
4	0.073812
5	0.061441
6	0.021394
7	0.010702
8	0.008386
9	0.000000

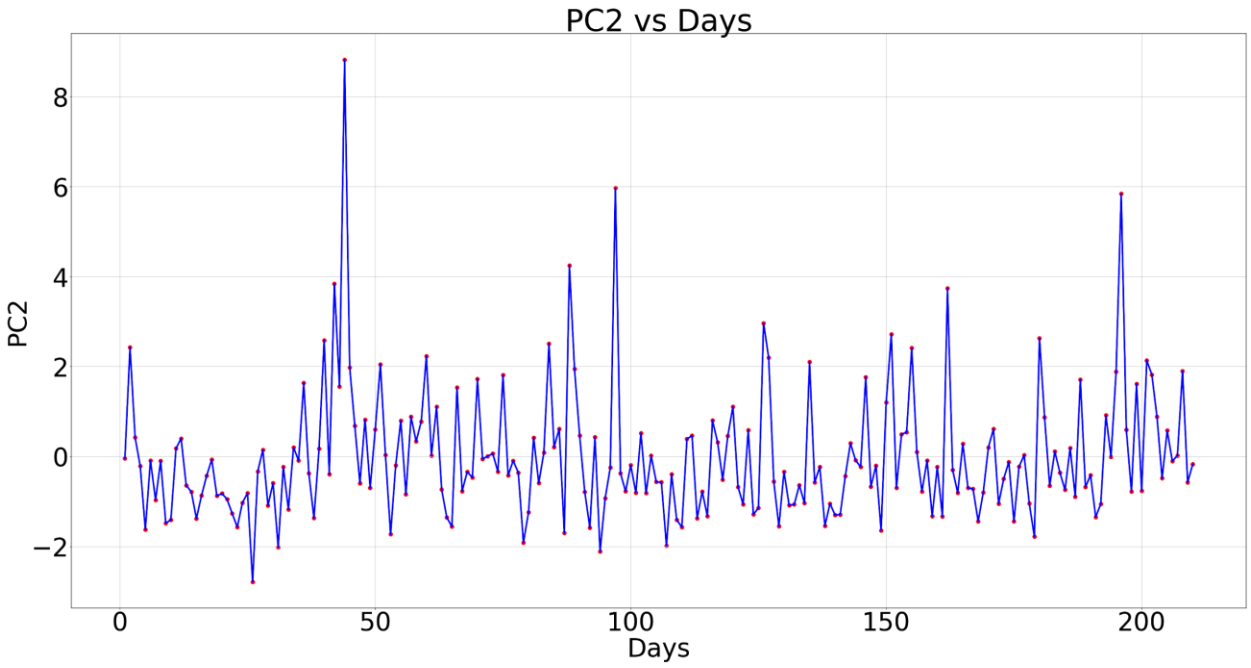


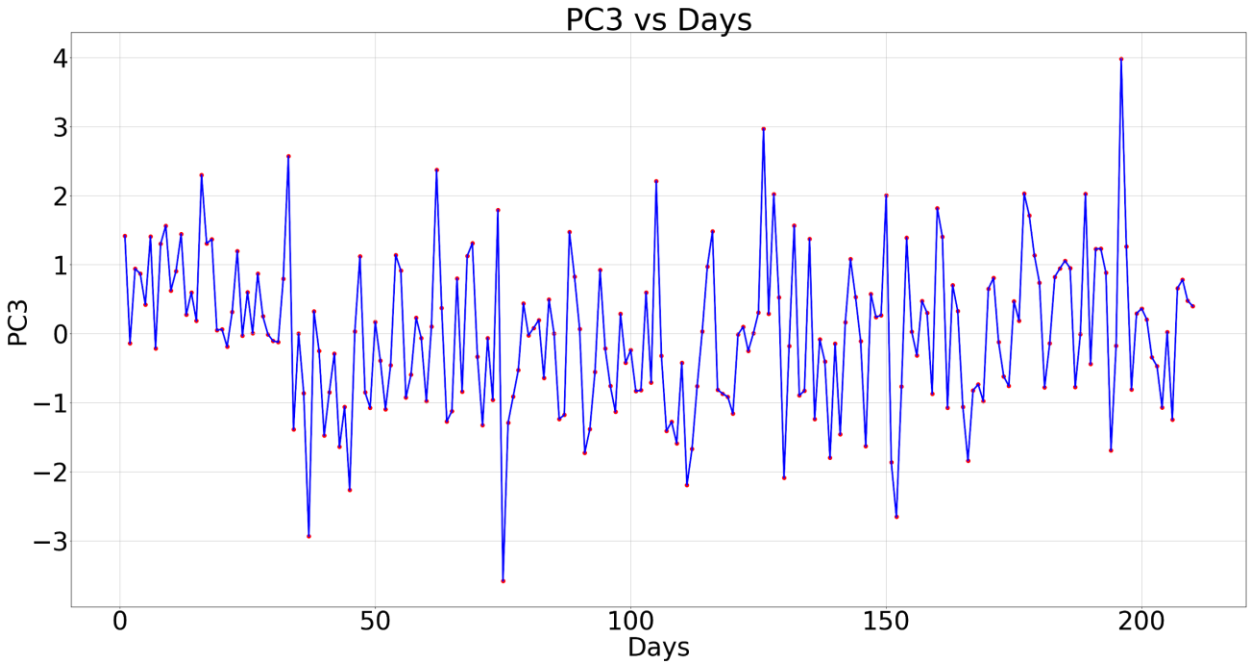
Now, after calculating the weighted ratio we choose the top 5 components that contribute the most weight. The chosen top 5 weighted components represents about 89% of the information. The First Principal component alone represents about 39% of the information in the above graph.

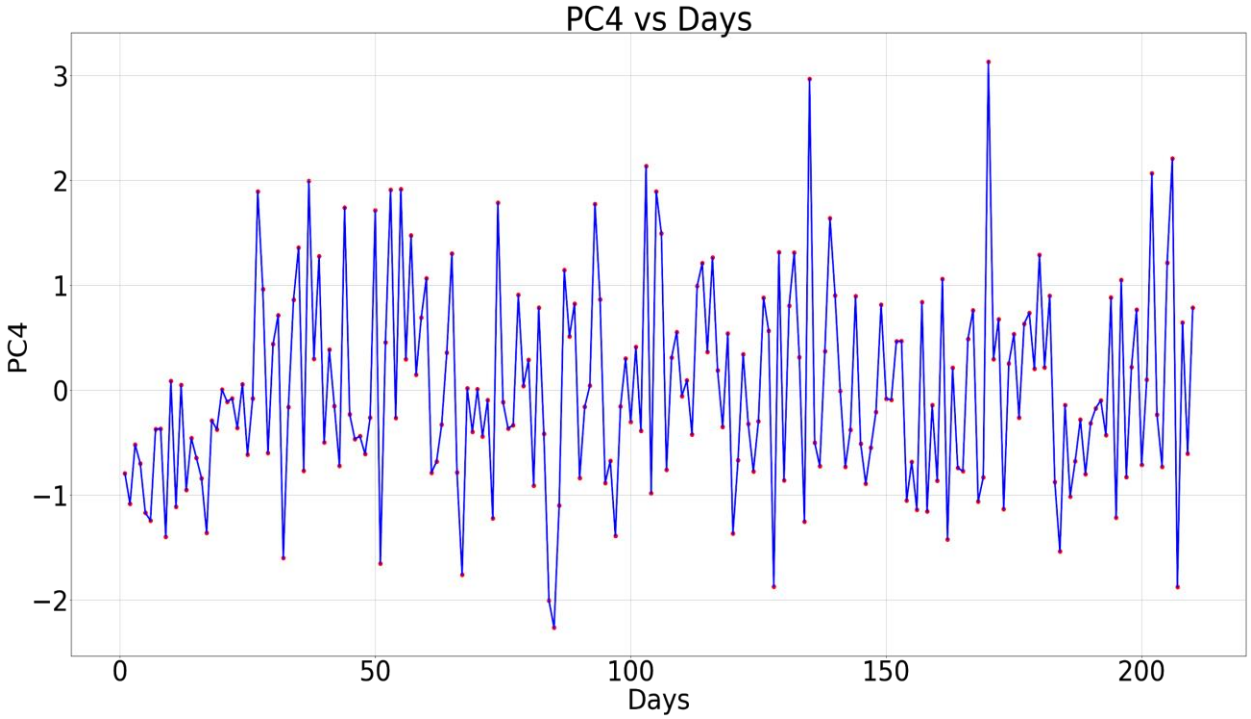


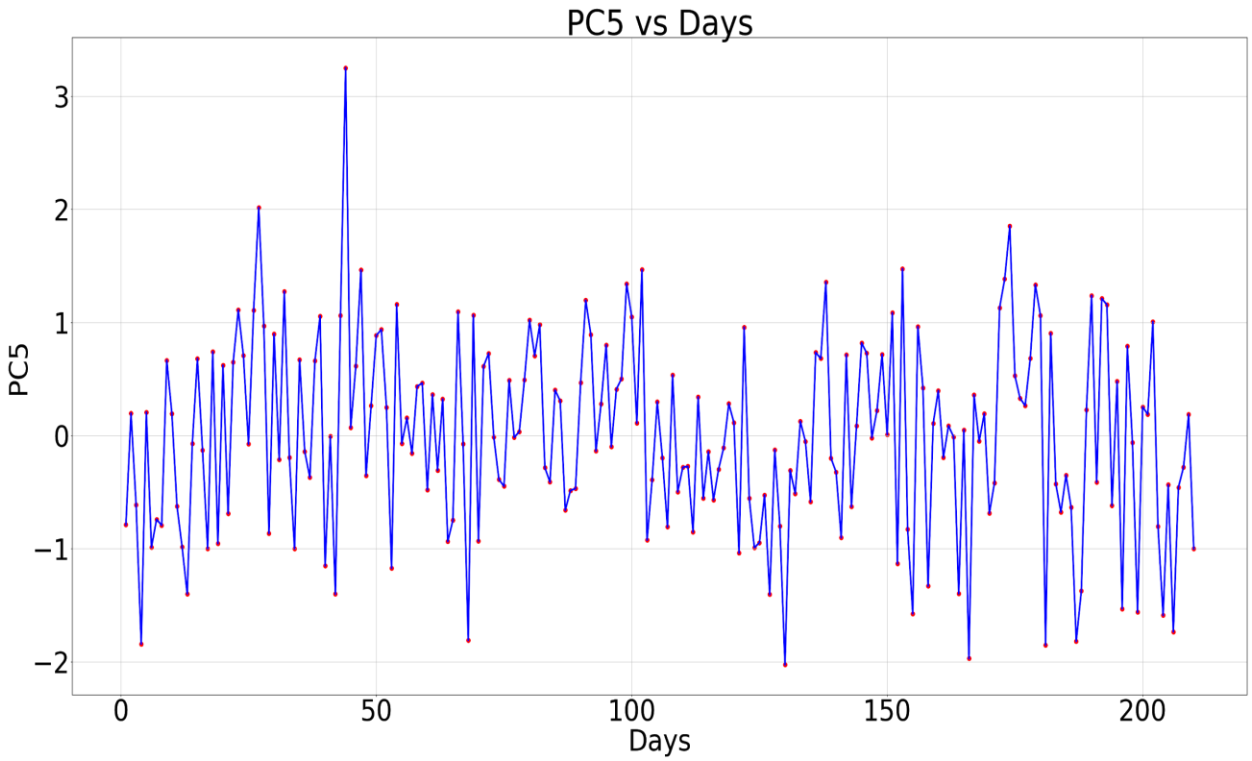
The plots given below represents the top 5 principal components against the time series (days).


PC1 vs Days
PC1
Days
10
8
6
4
2
0
−2
−4
0
50
100
150
200


PC2 vs Days
PC2
Days
8
6
4
2
0
−2
0
50
100
150
200


PC3 vs Days
PC3
Days
4
3
2
1
0
−1
−2
−3
0
50
100
150
200


PC4 vs Days
PC4
Days
3
2
1
0
−1
−2
0
50
100
150
200



f) For each feature in the top 5 argue why it is chosen as a top five feature in PCA?

In the 'EigenVectors' bar graph plot, each unique color of the bar represents the eigenvectors which represents the individual weights along the original features. The higher the values of the bars, the more is the significance that the data lies along that feature component of the eigenvector. The components with the highest bars or values are the top 5 principal components. From the above graph, we can deduce that "MinMax", "Correlation", "Displacement", "Value to Time Series Ratio" and "Skewness" are those top 5 principal components.

Displacement

The displacement in our data can be used as a feature since the sum of differences increases when there is a sudden change in the glucose levels which indicate that the person has eaten at that point.

Correlation

In our data, there is a pattern of increase in glucose level whenever there is a food intake, which keeps repeating for each person. So, we used Autocorrelation feature to see if the variable is correlated.

Skewness

We need to observe the glucose levels to find out when the person consumes food. By using skewness as a feature, we can find the deviation of this distribution. This deviation represents the increase in glucose levels which corresponds to the time at which food is consumed.

Value to Time Series Ratio

In this data, there won't be any unique values since the pattern repeats itself for every patient. This is also evident from the graph provided. This can be an important feature in determining the distinctiveness of the data.

MinMax

The range feature can be useful when we have a variable that should not cross a certain threshold. In our data, the sudden rise in glucose level during the food intake is beyond a threshold. So, this MinMax feature can be of use to predict the food intake.

References

[1] https://en.wikipedia.org/wiki/Principal_component_analysis