Data Mining CSE572 Project1

Submitted to:

Professor Ayan Banerjee

Ira A. Fulton School of Engineering

Arizona State University

Team Members:

Pushparajsinh Zala(1217568222) - pzala@asu.edu

Vishwarajsinh Sodha(1217485776) - vmsodha@asu.edu

Samip Thakkar (1217014967) - sthakka2@asu.edu

Monil Nisar (1217111805) - mnisar2@asu.edu

1. Feature Selection

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

For feature extraction, we have used these techniques:

1.1.1. Fast Fourier transform: Fast Fourier Transform converts a signal from its original domain to representation in frequency domain.

1.1.2. Root Mean Square (RMS): Root mean square is the square root of the mean of the squares of all the values.

1.1.3. Binning (hyperglycemia, hypoglycemia): Divide the data into three parts: above, below and on the threshold.

1.1.4. Moving Average: It slides a window across the data and computes the mean of the windows' contents.

1.1.5. Time difference between meal time and maximum of glucose level: It denotes the time it took for maximum glucose ingression. So (max glucose time - meal time)

1.1.6. Is range larger than standard deviation: Boolean variable denoting if the standard deviation of row is higher than 'r' times the range (where range = difference between max and min of row). We are taking values of r = 0.2, 0.25, 0.3, 0.35 and 0.4
Value = std(x) > r * (max(x)-min(x))

1.1.7. Is the person taking Meal on regular time(Regularity in Meals): Here we evaluate the distribution of the meal times and estimate the parameters of the distribution from the training dataset. This will give us an overall understanding of the Regularity in Meals taken.

1.2 For each time series explain why you chose such feature.

1.2.1. Fast Fourier transform:

Reason: Fast Fourier Transform is specifically used for time series data. It converts the signal representation to the frequency domain. Thus, it is used for filtration purpose. As the representation is done in frequency domain, we can have a better analysis of the signal.**FFT** helps in converting the time domain in frequency domain which makes the calculations easier as we always deal with various frequency bands in communication system. Another very big **advantage** is that it can convert the discrete data into a continuous data type available at various frequencies. FFT takes N * log(N) operations and is faster as compared to the Discrete Fourier Transform which takes N ^ 2 operations.

1.2.2. Root Mean Square (RMS)

Reason: RMS, also known as quadratic mean as it takes roots and squares along with mean, will give the data sense regarding the magnitude of the data. We have calculated the rms value using python code. The numpy library performs basic mean, square and square root of the rows. RMS is basically calculated to determine the noise and variance of the data. If we consider normal mean error, it might be zero due to sign difference.

1.2.3. Binning (hyperglycemia, hypoglycemia)

Reason: The binning is the process of dividing the plot into three parts: Hypoglycemia, glycemia level, and hyperglycemia. A threshold is selected which can be considered as the glycemia value. Binning the data will divide the data in the three parts and we can determine whether the data is above the threshold (hyperglycemia), on the line (glycemia level) or lower than threshold decided (hypoglycemia). This can give insights to extreme events which may have occurred from training set. Binning can also reduces the effect of errors occurring during minor observations because it merges all data in one range.

1.2.4. Moving Average:

Reason: It is another measure to calculate the error in the data. It calculates the mean of the contents of a moving window. It is used to smooth the data series and helps in spotting the trends of the plot. It is specifically useful in use cases where the data is very volatile. It helps in calculating the support and errors in the data.

1.2.5. Time difference between meal time and maximum of glucose ingression:

Reason: It helps to get the time taken by glucose level to reach the peak value (maximum point) after the meal is taken. It will help in determining the time it takes to reach the highest glucose level after the insulin is supplied before taking the meal.

1.2.6. Is range larger than standard deviation:

Reason: Generally the standard deviation should be a fourth of the range of the values. Here we are taking r = 0.2, 0.25, 0.3, 0.35 and 0.4 in following formula for feature value.

value = std(x) > r * (max(x)-min(x))

This feature may capture variation on vector data, which can also help to capture irregularities.

1.2.7. Is the person taking Meal on regular time(Regularity in Meals):

Reason: We compute the meal time from InsulinBolus data and check the distribution of the lunch times. We assume normal distribution or gaussian distribution and compute its parameters like Mean and Variance for the lunch times. Once we get these parameters then we know the model distribution of person's lunch times in the form of probability.

The mean here gives the time at which person has the highest probability to take meal. We have computed this from the data given to us. Now if the person is too irregular than the variance would be high and becomes harder to predict the lunch time for the algorithm and in turn harder to regulate insulin flow on proper time. So, the attributes of this feature type are (1) Time until First Meal (2) Mean of Meal times (3) Variance of Meal times (4) Regularity in Meal (5) Deviation from Mean

So, now we have all the parameters from the gaussian distribution to determine the regularity of person in meals. This will help the algorithm to learn the meal habit pattern of the person. Given this parameters our goal is to learn the Probability of a Future Meal happening at a particular time. Once we know the probability of a Meal happening at a particular time the algorithm can decide accordingly whether to release insulin or not, and how much to be released.

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

1.3.1. Fast Fourier transform:

```
[98.5  0.5]
[338.   3.]
[133.  -1.]
[90.5 -1.5]
[120.5  -0.5]
[97.  1.]
[111.5   0.5]
[91.5  0.5]
[112.5  -1.5]
[138.5 -10.5]
[100.5  -1.5]
[76.  1.]
[113.   1.]
[104.  -2.]
[97.5 -3.5]
[99.5 -1.5]
[130.   0.]
[229.5   1.5]
[76. -1.]
```

Reason: The output of Fast Fourier Transformation is a 2-d matrix of shape 216 rows x 2 columns, where value and its frequency is given. FFT gives one of the maximum distinction among all techniques. As, the data is converted into frequency domain, the data can easily be analyzed. Hence, Fast Fourier Transform is a good feature extraction technique.

1.3.2. Root Mean Square (RMS)

```
199.42866310203922
304.8570156647211
219.55276510822327
175.77495081305906
149.84561499534556
184.99977477463767
154.42217565276476
177.3859962529925
159.38606170762444
161.19075138894704
195.87846997564586
178.09267250507529
170.66890460772285
168.00285215832895
107.69362407620363
166.5265244137802
189.96929576469282
145.3567450562007
141.34841704101254
```

Reason: The rms values of all the 216 rows is calculated as order: square, then mean and at last square root. The rms values signifies the errors and it should be as less as possible and here, we can see the values are not as expected. Hence, rms is not the best feature to extract here.

1.3.3. Binning (hyperglycemia, hypoglycemia)

```
[0.0, 0.2916666666666667, 0.7083333333333334]
[0.0, 0.0, 1.0]
[0.0, 0.20833333333333334, 0.7916666666666666]
[0.0, 0.3333333333333333, 0.6666666666666666]
[0.0, 0.2916666666666667, 0.7083333333333334]
[0.0, 0.25, 0.75]
[0.0, 0.3333333333333333, 0.6666666666666666]
[0.0, 0.375, 0.625]
[0.0, 0.2916666666666667, 0.7083333333333334]
[0.0, 0.20833333333333334, 0.7916666666666666]
[0.0, 0.20833333333333334, 0.7916666666666666]
[0.0, 0.4166666666666667, 0.5833333333333334]
[0.0, 0.25, 0.75]
[0.0, 0.2916666666666667, 0.7083333333333334]
[0.0, 1.0, 0.0]
[0.0, 0.375, 0.625]
[0.0, 0.25, 0.75]
[0.16666666666666666, 0.4166666666666667, 0.4166666666666667]
[0.0, 0.375, 0.625]
```

Reason: The blood glucose level is divided into three bins. Bins are for hyperglycemia(>130), normal(>70 & <130) and hypoglycemia(<70). There are 216 rows each with 3 bins.

1.3.4. Moving Average with 4 windows

```
[101.66666666666667, 170.0, 230.83333333333334, 256.3333333333333]
[333.6666666666667, 309.6666666666667, 282.6666666666667, 260.0]
[136.0, 200.5, 264.5, 251.0]
[92.83333333333333, 148.0, 208.0, 221.16666666666666]
[133.16666666666666, 163.5, 158.0, 142.16666666666666]
[103.33333333333333, 178.5, 219.5, 213.33333333333334]
[106.16666666666667, 145.33333333333334, 182.83333333333334, 170.5]
[93.66666666666667, 144.0, 208.5, 228.33333333333334]
[115.0, 152.33333333333334, 176.0, 184.33333333333334]
[149.5, 186.66666666666666, 166.16666666666666, 136.83333333333334]
[113.33333333333333, 197.66666666666666, 227.83333333333334, 221.33333333333334]
[79.83333333333333, 128.83333333333334, 217.0, 236.0]
[116.16666666666667, 163.16666666666666, 197.5, 192.66666666666666]
[110.33333333333333, 164.5, 202.16666666666666, 178.83333333333334]
[102.5, 101.16666666666667, 105.66666666666667, 120.0]
[102.66666666666667, 137.83333333333334, 189.0, 212.5]
[131.5, 180.83333333333334, 216.83333333333334, 216.83333333333334]
[216.16666666666666, 155.16666666666666, 88.5, 70.5]
[78.0, 133.33333333333334, 158.16666666666666, 174.66666666666666]
[122.16666666666667, 177.83333333333334, 200.16666666666666, 157.33333333333334]
```

Reason: The moving average is considering 4 windows and the output is thus a matrix of 216 x 3. The moving average will calculate the average of all the data that are present in the windows and from the data, it can be seen that covariance is very high among the data and have the maximum distinction.

1.3.5. Time Difference between meal and the highest value of glucose

```
0.07291666709352285
0.0
0.05902777798473835
0.0763888891087845
0.03125
0.05208333407063037
0.04861111205536872
0.07638888899236917
0.07986111100763083
0.034756944980472326
0.055555555038154125
0.06597222201526165
0.0486111108912155
0.05207175912801176
0.0763888891087845
0.06944444414693862
0.04861111112404615
0.0
0.07638888899236917
0.05208333407063037
```

Reason: The time difference from meal to highest glucose level denoted the time for maximum glucose ingression. It is a 216 x 1 matrix. There is less deviation in the data. Also at some point the value is zero, because the patient takes/reports insulin shot after the meal.

1.3.6. Is range larger than standard deviation:

```
[1, 1, 1, 1, 0]
[1, 1, 1, 0, 0]
[1, 1, 1, 1, 0]
[1, 1, 1, 1, 0]
[1, 1, 0, 0, 0]
[1, 1, 1, 1, 0]
[1, 1, 1, 1, 0]
[1, 1, 1, 1, 0]
[1, 1, 1, 1, 0]
[1, 1, 1, 0, 0]
[1, 1, 1, 1, 0]
[1, 1, 1, 1, 0]
[1, 1, 1, 1, 0]
[1, 1, 1, 1, 0]
[1, 1, 1, 0, 0]
[1, 1, 1, 1, 0]
[1, 1, 1, 1, 0]
[1, 1, 1, 1, 0]
[1, 1, 1, 1, 0]
[1, 1, 1, 0, 0]
```

Reason: As we increase the r value from 0.2 to 0.4 it increases the sparsity of the data. So at lower values of r we can use this feature to capture variation on cgm data which can help to find irregularities to reduce errors in prediction.

But checking with PCA values it's not performing well. Maybe we can argue that since is binary feature it seems it is not performing well compared to other features available.

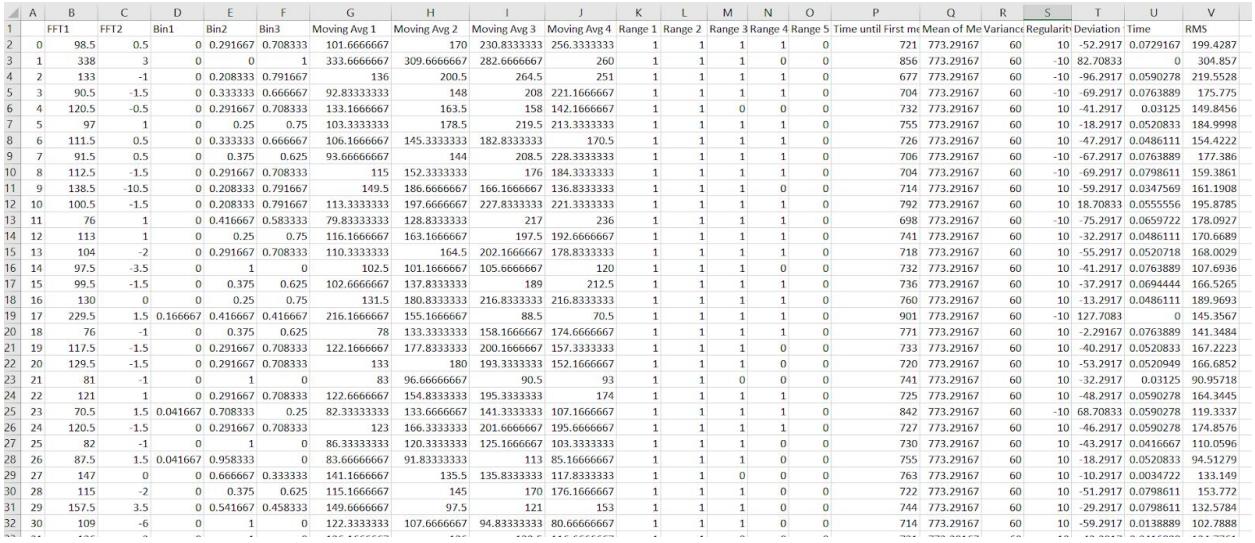
1.3.7. Is the person taking Meal on regular time(Regularity in Meals):

```
[721.0, 773.2916667000001, 60.0, 10.0, -52.29166667]
[856.0, 773.2916667000001, 60.0, -10.0, 82.70833333]
[677.0, 773.2916667000001, 60.0, -10.0, -96.29166667]
[704.0, 773.2916667000001, 60.0, -10.0, -69.29166667]
[732.0, 773.2916667000001, 60.0, 10.0, -41.29166667]
[755.0, 773.2916667000001, 60.0, 10.0, -18.29166667]
[726.0, 773.2916667000001, 60.0, 10.0, -47.29166667]
[706.0, 773.2916667000001, 60.0, -10.0, -67.29166667]
[704.0, 773.2916667000001, 60.0, -10.0, -69.29166667]
[714.0, 773.2916667000001, 60.0, 10.0, -59.29166667]
[792.0, 773.2916667000001, 60.0, 10.0, 18.70833333]
[698.0, 773.2916667000001, 60.0, -10.0, -75.29166667]
[741.0, 773.2916667000001, 60.0, 10.0, -32.29166667]
[718.0, 773.2916667000001, 60.0, 10.0, -55.29166667]
[732.0, 773.2916667000001, 60.0, 10.0, -41.29166667]
[736.0, 773.2916667000001, 60.0, 10.0, -37.29166667]
[760.0, 773.2916667000001, 60.0, 10.0, -13.29166667]
[901.0, 773.2916667000001, 60.0, -10.0, 127.70833329999999]
[771.0, 773.2916667000001, 60.0, 10.0, -2.291666667]
```

Reason: The feature type works because the data points are having high variation in the direction of the attribute vectors of this feature type. The normal distribution of the data and the parameters estimated for it are highly indicative of the data patterns in our dataset.

1.4 Create a feature matrix where each row is a collection of features from each time series.

We have 7 features with 21 feature columns in total and a total of 216 rows.
216 X 21 feature matrix.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1		FFT1	FFT2	Bin1	Bin2	Bin3	Moving Avg 1	Moving Avg 2	Moving Avg 3	Moving Avg 4	Range 1	Range 2	Range 3	Range 4	Range 5	Time until First me	Mean of Me	Variance	Regularity	Deviation	Time	RMS
2	0	98.5	0.5	0	0.291667	0.708333	101.6666667	170	230.8333333	256.3333333	1	1	1	1	0	721	773.29167	60	10	-52.2917	0.0729167	199.4287
3	1	338	3	0	0	1	333.6666667	309.6666667	282.6666667	260	1	1	1	0	0	856	773.29167	60	-10	82.70833	0	304.857
4	2	133	-1	0	0.208333	0.791667	136	200.5	264.5	251	1	1	1	1	0	677	773.29167	60	-10	-96.2917	0.0590278	219.5528
5	3	90.5	-1.5	0	0.333333	0.666667	92.83333333	148	208	221.1666667	1	1	1	1	0	704	773.29167	60	-10	-69.2917	0.0763889	175.775
6	4	120.5	-0.5	0	0.291667	0.708333	133.1666667	163.5	158	142.1666667	1	1	0	0	0	732	773.29167	60	10	-41.2917	0.03125	149.8456
7	5	97	1	0	0.25	0.75	103.3333333	178.5	219.5	213.3333333	1	1	1	1	0	755	773.29167	60	10	-18.2917	0.0520833	184.9998
8	6	111.5	0.5	0	0.333333	0.666667	106.1666667	145.3333333	182.8333333	170.5	1	1	1	1	0	726	773.29167	60	10	-47.2917	0.0486111	154.4222
9	7	91.5	0.5	0	0.375	0.625	93.66666667	144	208.5	228.3333333	1	1	1	1	0	706	773.29167	60	-10	-67.2917	0.0763889	177.386
10	8	112.5	-1.5	0	0.291667	0.708333	115	152.3333333	176	184.3333333	1	1	1	1	0	704	773.29167	60	-10	-69.2917	0.0798611	159.3861
11	9	138.5	-10.5	0	0.208333	0.791667	149.5	186.6666667	166.1666667	136.8333333	1	1	1	0	0	714	773.29167	60	10	-59.2917	0.0347569	161.1908
12	10	100.5	-1.5	0	0.208333	0.791667	113.3333333	197.6666667	227.8333333	221.3333333	1	1	1	1	0	792	773.29167	60	10	18.70833	0.0555556	195.8785
13	11	76	1	0	0.416667	0.583333	79.83333333	128.8333333	217	236	1	1	1	1	0	698	773.29167	60	-10	-75.2917	0.0659722	178.0927
14	12	113	1	0	0.25	0.75	116.1666667	163.1666667	197.5	192.6666667	1	1	1	1	0	741	773.29167	60	10	-32.2917	0.0486111	170.6689
15	13	104	-2	0	0.291667	0.708333	110.3333333	164.5	202.1666667	178.8333333	1	1	1	1	0	718	773.29167	60	10	-55.2917	0.0520718	168.0029
16	14	97.5	-3.5	0	1	0	102.5	101.1666667	105.6666667	120	1	1	1	0	0	732	773.29167	60	10	-41.2917	0.0763889	107.6936
17	15	99.5	-1.5	0	0.375	0.625	102.6666667	137.8333333	189	212.5	1	1	1	1	0	736	773.29167	60	10	-37.2917	0.0694444	166.5265
18	16	130	0	0	0.25	0.75	131.5	180.8333333	216.8333333	216.8333333	1	1	1	1	0	760	773.29167	60	10	-13.2917	0.0486111	189.9693
19	17	229.5	1.5	0.166667	0.416667	0.416667	216.1666667	155.1666667	88.5	70.5	1	1	1	1	0	901	773.29167	60	-10	127.7083	0	145.3567
20	18	76	-1	0	0.375	0.625	78	133.3333333	158.1666667	174.6666667	1	1	1	1	0	771	773.29167	60	10	-2.29167	0.0763889	141.3484
21	19	117.5	-1.5	0	0.291667	0.708333	122.1666667	177.8333333	200.1666667	157.3333333	1	1	1	0	0	733	773.29167	60	10	-40.2917	0.0520833	167.2223
22	20	129.5	-1.5	0	0.291667	0.708333	133	180	193.3333333	152.1666667	1	1	1	0	0	720	773.29167	60	10	-53.2917	0.0520949	166.6852
23	21	81	-1	0	1	0	83	96.66666667	90.5	93	1	1	0	0	0	741	773.29167	60	10	-32.2917	0.03125	90.95718
24	22	121	1	0	0.291667	0.708333	122.6666667	154.8333333	195.3333333	174	1	1	1	1	0	725	773.29167	60	10	-48.2917	0.0590278	164.3445
25	23	70.5	1.5	0.041667	0.708333	0.25	82.33333333	133.6666667	141.3333333	107.1666667	1	1	1	1	0	842	773.29167	60	-10	68.70833	0.0590278	119.3337
26	24	120.5	-1.5	0	0.291667	0.708333	123	166.3333333	201.6666667	195.6666667	1	1	1	1	0	727	773.29167	60	10	-46.2917	0.0590278	174.8576
27	25	82	-1	0	1	0	86.33333333	120.3333333	125.1666667	103.3333333	1	1	1	0	0	730	773.29167	60	10	-43.2917	0.0416667	110.0596
28	26	87.5	1.5	0.041667	0.958333	0	83.66666667	91.83333333	113	85.16666667	1	1	1	0	0	755	773.29167	60	10	-18.2917	0.0520833	94.51279
29	27	147	0	0	0.666667	0.333333	141.1666667	135.5	135.8333333	117.8333333	1	1	0	0	0	763	773.29167	60	10	-10.2917	0.0034722	133.149
30	28	115	-2	0	0.375	0.625	115.1666667	145	170	176.1666667	1	1	1	0	0	722	773.29167	60	10	-51.2917	0.0798611	153.772
31	29	157.5	3.5	0	0.541667	0.458333	149.6666667	97.5	121	153	1	1	1	0	0	744	773.29167	60	10	-29.2917	0.0798611	132.5784
32	30	109	-6	0	1	0	122.3333333	107.6666667	94.83333333	80.66666667	1	1	1	0	0	714	773.29167	60	10	-59.2917	0.0138889	102.7888

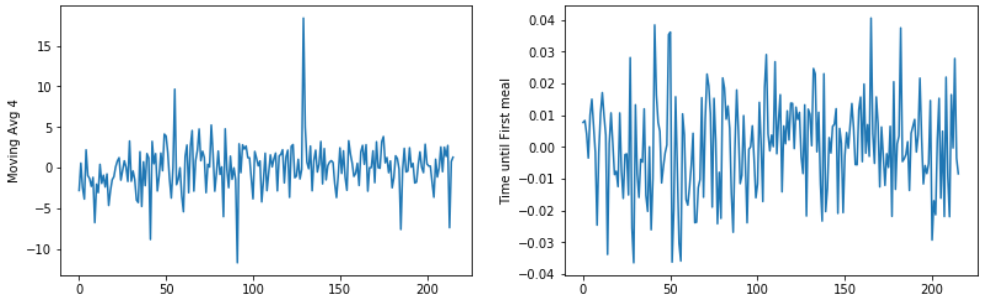
file.csv

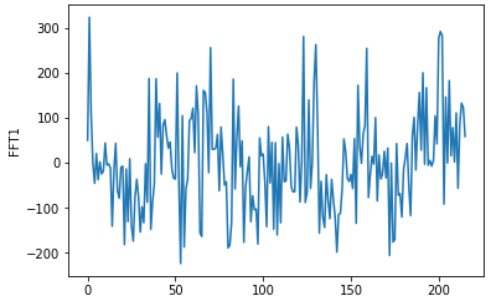
1.5 Provide this feature matrix to PCA and derive the new feature matrix. Chose the top 5 features and plot them for each time series. (5 points)

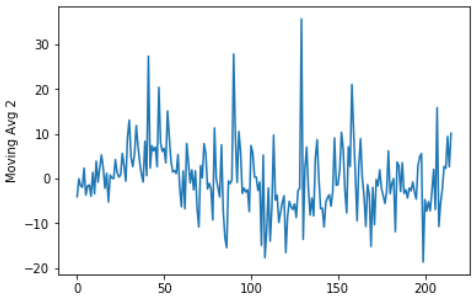
Principal Component Analysis is a decomposition technique, which reduces the dimensions. It will break a matrix in eigen-value matrix with corresponding eigen-vector matrix. The output matrix also consists of the features given as input, but the feature space is changed and the output is sorted as per eigen-values, where each eigen-value is sorted in the order. The order is determined by the variance ratio, which is the ratio of variance and mean. More the ratio, more dispersion and better the feature extracted. Thus, the importance of the feature is more if the variance ratio is more.

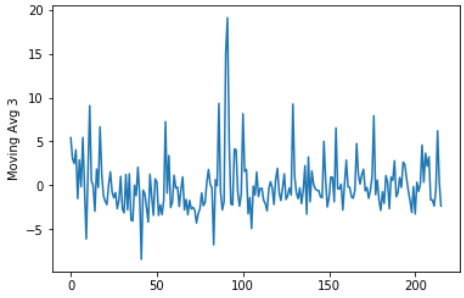
	0	1	variance ratio
0	PC0	Moving Avg 4	46.57374
1	PC1	Time of First meal	36.25960
2	PC2	FFT1	13.21609
3	PC3	Moving Avg 2	2.67021
4	PC4	Moving Avg 3	0.64203
5	PC5	Regular Meal	0.33672
6	PC6	Moving Avg 1	0.22660
7	PC7	RMS	0.04032
8	PC8	FFT2	0.03311
9	PC9	Var 3	0.00078
10	PC10	Var 4	0.00036

Top 5 features are: Moving average (Window 4), Time for first Meal, Fast Fourier Transform (Coefficient 1), Moving average (Window 2), Moving average(Window 3)









Top 5 distinct features extracted are: Moving average, Time of First Meal, Fast Fourier Transform, Root Mean Square, Comparison of range and Standard Deviation.

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

Moving Average(window 4): The average of blood glucose in later part of series shows much deviation, due to which the variance is high which results in high variance ratio. Compared to the starting part which remains similar for all other series.

Time of First Meal: The meal time for the patient varies much with the mean. As the covariance is having high positive value, it is because of the fact that the variance is very high compared to mean, and the data can be analyzed.

Fast Fourier Transform: The fast fourier transform converts the signal time series data into the frequency domain. As the fast fourier takes very less time to convert the data from one domain to another, it is one of the good feature extraction techniques and due to the fact that the covariance is high compared to others, it gives the best results.

Moving Average(window 2): The average of blood glucose in later part of series shows much deviation, Compared to the starting part which remains similar for all other series.

Moving Average(window 3): The average of blood glucose in later part of series shows much deviation, Compared to the starting part which remains similar for all other series.

The subsequent different features in the decreasing order of importance are:

Root Mean Square: The root mean square values are high but on a scale compared to other features, the variation is less, hence there is less dispersion and the feature does not hold as much importance as other.

Comparison of Range and Standard Deviation: As the output is a boolean value, the variance value will be low, which will result in low variance ratio and thus, PCA output gives it low importance.