**Data Modeling**

As the aim of the project is to estimate the blood glucose level from NIR-spectrum data, so finding the best model that suits the data set with highest accuracy is the major part of the project. This part is discussed step by step below:

1. First, Partial least square regression model (PLSR) was used splitting the samples into 60% for training and 40% for validation. And optimum number of components for least MSE (mean square error) was calculated. But the accuracy and coefficient of determination was not good.
2. As, PLSR was not able to predict accurately, Support Vector Machine (SVM), a supervised machine learning modeling, was applied on the NIR- data set by classifying them in ten classes each signifying defined glucose levels. The dataset was split by 60% for training and 40% for testing. In the dataset, total no of samples taken is 100 which includes ten samples for each class and total number of features was 156 which are the wavelengths from 1200 nm – 2400 nm. But the result was not up to the mark.
3. So, to overcoming high dimensionality of the model, Principal component analysis (PCA) was performed on the dataset prior to using SVM modeling. As PCA is affected by scale, so scaling the features in the data was done before applying PCA.
4. After performing scaling and PCA modeling, linear kernel SVM classification was applied on the dataset by splitting 60% for training and 40% for testing. And the score, accuracy, precision was observed.

All the data processing, modeling and results were performed in PyCharm (version 2018.2) integrated development environment using python 3.7 programming language.

**Results:**

* **PLSR method:**

To get the best accuracy from this method applying in the dataset, the optimum number of PLS components had to be measured. So, it was found that, the number of PLS component is one for the least value of the MSE.

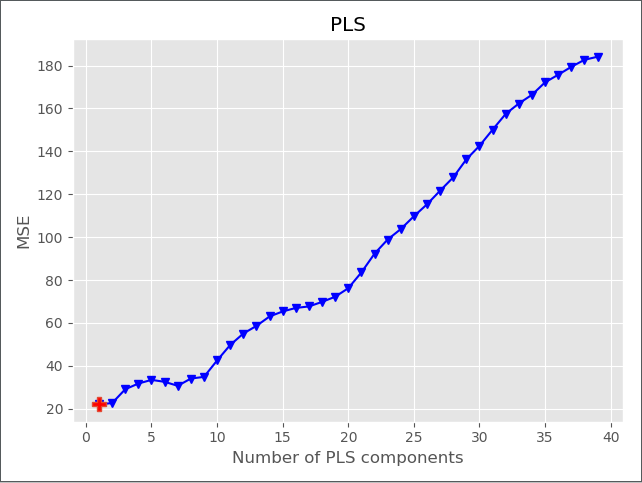
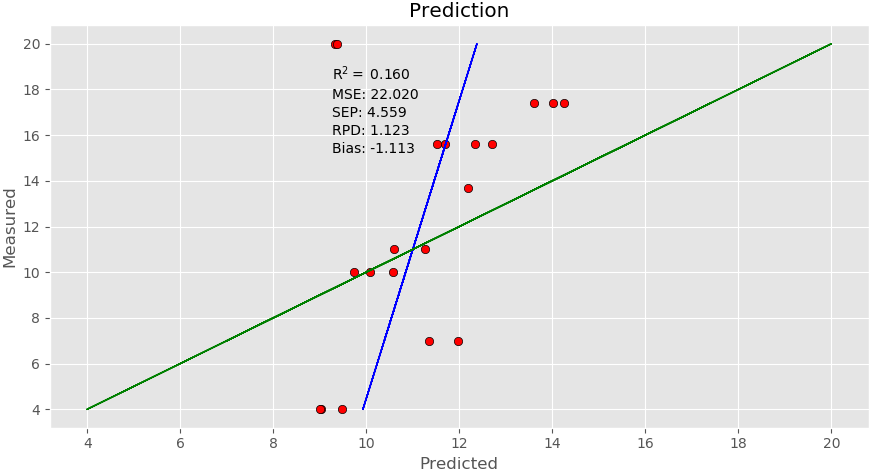


Figure: MSE vs Number of PLS components

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| R2 | MSE | SEP: | RPD | Bias: |
| 0.160 | 22.020 | 4.559 | 1.123 | -1.113 |

 Figure: Measured vs Predicted by PLS regression method

Comments:

The coefficient of determination (R2) of this method is **0.16** with MSE 22.02 which is not so good for prediction.

* **SVM (kernel-linear):**

Linear kernel supervised support vector machine method was used to model the dataset. The 156 different wavelengths had been considered as features and different predefined glucose levels was labeled as classes for building the model.

Confusion matrix:

[5 0 0 0 0 0 0 0 0 0]

[0 3 0 0 0 0 0 0 0 0]

[0 0 4 0 0 0 0 0 0 0]

[0 0 0 2 0 0 0 0 0 0]

[0 0 0 0 3 0 0 0 0 0]

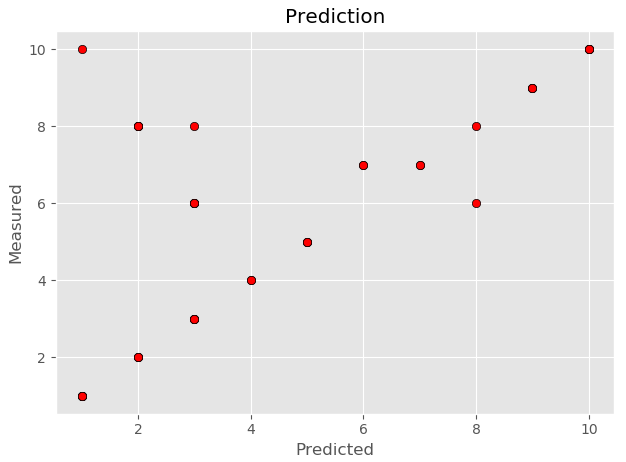
[0 0 4 0 0 0 0 1 0 0]

[0 0 0 0 0 2 2 0 0 0]

[0 4 1 0 0 0 0 1 0 0]

[0 0 0 0 0 0 0 0 3 0]

[1 0 0 0 0 0 0 0 0 4]

 Figure : Measured vs predicted in SVM method

Classification report:

Accuracy: **0.675**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | precision | Recall | F1-score | support |
| 1 | 0.83 | 1.00 | 0.91 | 5 |
| 2 | 0.43 | 1.00 | 0.60 | 3 |
| 3 | 0.44 | 1.00 | 0.62 | 4 |
| 4 | 1.00 | 1.00 | 1.00 | 2 |
| 5 | 1.00 | 1.00 | 1.00 | 3 |
| 6 | 0.00 | 0.00 | 0.00 | 5 |
| 7 | 1.00 | 0.50 | 0.67 | 4 |
| 8 | 0.50 | 0.17 | 0.25 | 6 |
| 9 | 1.00 | 1.00 | 1.00 | 3 |
| 10 | 1.00 | 0.80 | 0.89 | 5 |
| weighted avg | 0.68 | 0.68 | 0.64 | 40 |

Comments:

So, the model predicted with 67.5% accuracy. Class 2, 3, 6 were predicted badly in this model.

* Improving SVM by using PCA:

SVM method showed better estimation than PLS regression method in comparison. The accuracy of the SVM method was improved by introducing principal component analysis (PCA). The preprocessing of the dataset before using SVM modeling is discussed:

Standardizing the data:

Feature scaling through standardization is an important preprocessing step for many machine learning algorithms. Standardization involves rescaling the features such that they have the properties of a standard normal distribution with a mean of zero and a standard deviation of one. PCA is affected by scale so it was needed to scale the features in the data before applying PCA.

Choosing the of Principal Components:

While choosing the number of principal components, 95% variance was used that means, it chose the minimum number of principal components such that 95% of the variance was retained. In this case, 95% of the variance amounts to **35** principal components. Now, SVM was applied with 35 variables and 100 samples and the result was observed.

Confusion matrix:

[5 0 0 0 0 0 0 0 0 0]

[0 3 0 0 0 0 0 0 0 0]

[0 0 4 0 0 0 0 0 0 0]

[0 0 0 2 0 0 0 0 0 0]

[0 0 0 0 3 0 0 0 0 0]

[0 0 3 0 0 2 0 0 0 0]

[0 0 0 0 0 2 2 0 0 0]

[0 2 1 0 0 0 1 2 0 0]

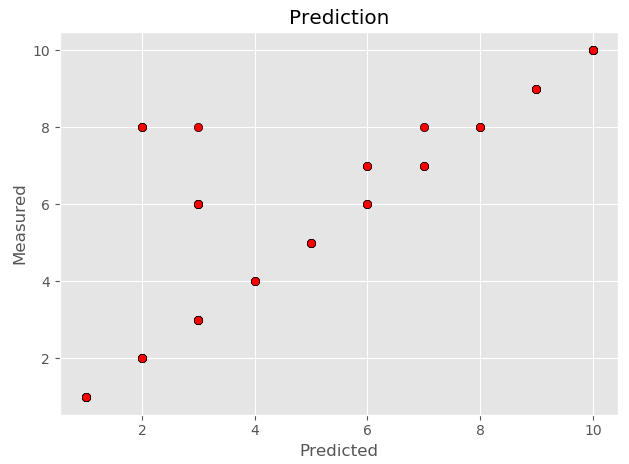
[0 0 0 0 0 0 0 0 3 0]

[0 0 0 0 0 0 0 0 0 5]

Classification report:

Accuracy: **0.775**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | precision | recall | f1-score | support |
| 1 | 1.00 | 1.00 | 1.00 | 5 |
| 2 | 0.60 | 1.00 | 0.75 | 3 |
| 3 | 0.50 | 1.00 | 0.67 | 4 |
| 4 | 1.00 | 1.00 | 1.00 | 2 |
| 5 | 1.00 | 1.00 | 1.00 | 3 |
| 6 | 0.50 | 0.40 | 0.44 | 5 |
| 7 | 0.67 | 0.50 | 0.57 | 4 |
| 8 | 1.00 | 0.33 | 0.50 | 6 |
| 9 | 1.00 | 1.00 | 1.00 | 3 |
| 10 | 1.00 | 1.00 | 1.00 | 5 |
| weighted avg | 0.82 | 0.78 | 0.76 | 40 |

 Figure: Measured vs Predicted in PCA+SVM model

Comments:

So, the model predicted with 77.5% accuracy. Class 2, 3, 6 were predicted better than previous. The overall precession was also improved to 82 %