Statistical Modelling for Prediction of Haemoglobin levels using Non-Invasive Spectrometry

Dhanashri Kanitkar
Hrishita Bapuram
Shraddha Kodavade

**Mentors: Prof. Sarada Samantaray
Prof. Shubham Bhatt**

EzeRx™

EASY FOR PRESCRIPTION

Partha Pratim Das Mahapatra
Founder & CEO

EzeRx is a Med-Tech & BioTech startup founded in 2018 which aims to develop highly progressive medical devices and innovative solutions.

To develop and manufacture innovative **non-invasive medical devices** and solutions that make screening hassle-free and enable regular check-ups for **early detection** of problems so that further treatment can take place, thus bridging the gap between diagnosis and treatment.

Objectives

To build a prediction model to analyze and predict the level of Haemoglobin concentration from available data on light reflected upon the incidence of light at different frequencies using a proprietary device by EzeRx

To understand the use and importance of Principal Component Analysis to optimize the factors and prediction accuracy of thus following models.

To test the accuracy of thus obtained predictions against traditional blood collection computations of Haemoglobin

To explore the effect of factors such as Gender and Age in the improvement of model to fetch results with high accuracy

Project Flow

Raw Data and EDA

Data Cleaning, missing value removal and visualization

Dimension Reduction

Applied PCA and compared results to optimize factors studied

Model Fitting

Applied Linear Regression, SVM and Random Forest algorithms

Model Accuracy and Results

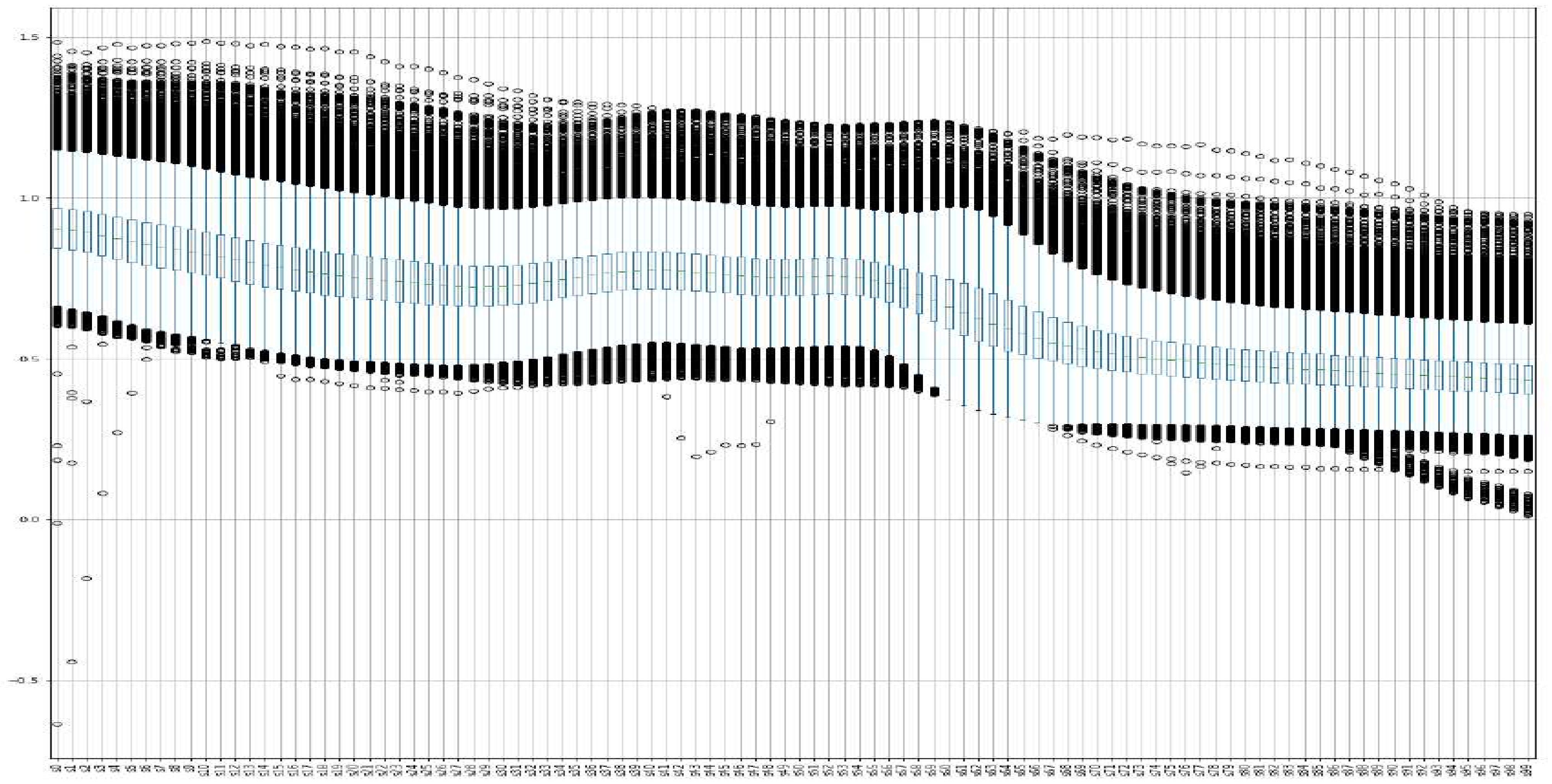
Fit the model with best accuracy, applied cross validation and obtained final results

Data Profile

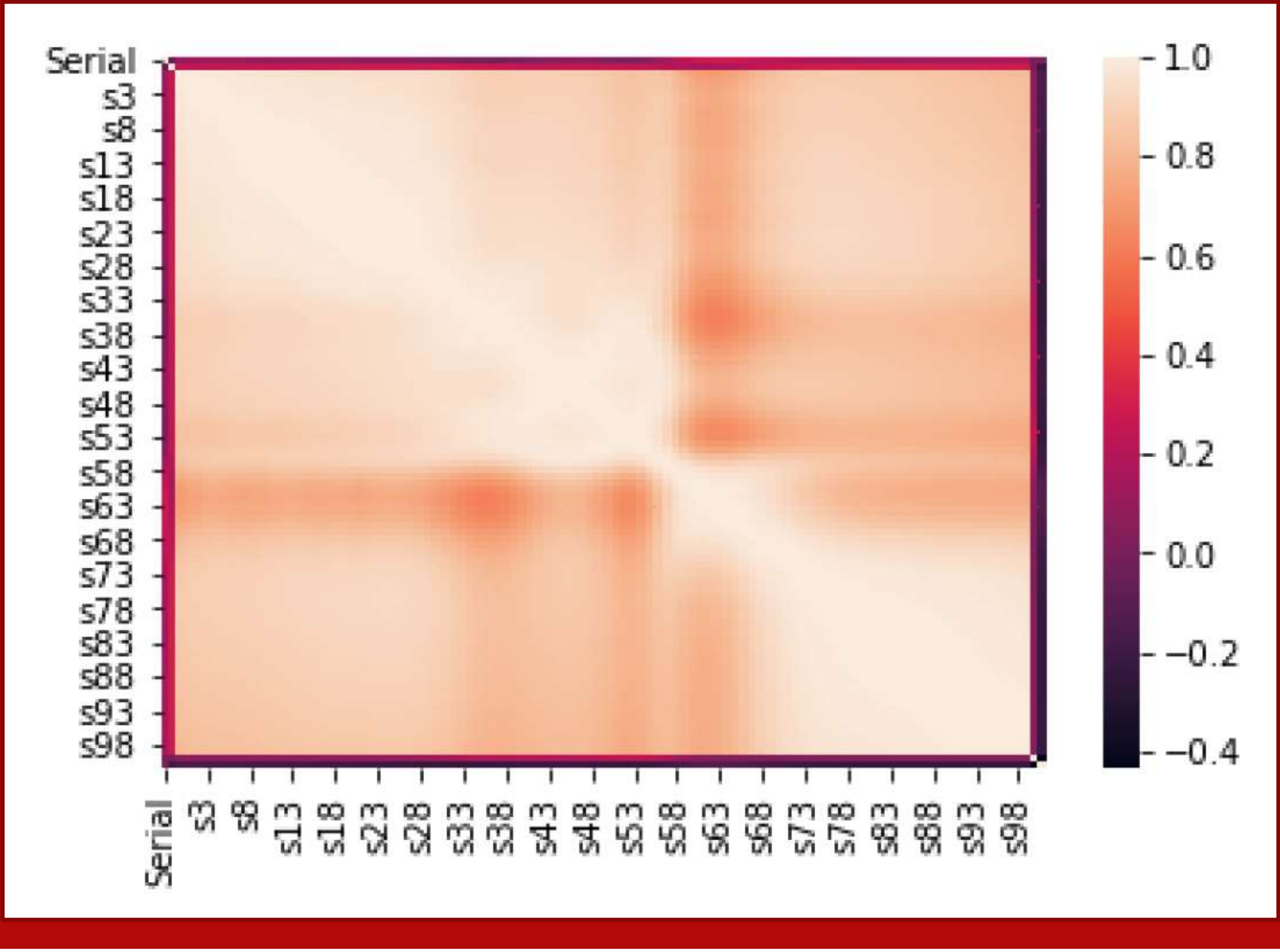
Given Data set of 101765 records consisting of Actual Haemoglobin values computed for 2455 subjects as well as their respective light reflected at 100 different frequencies ranging from 30Hz to 300Hz

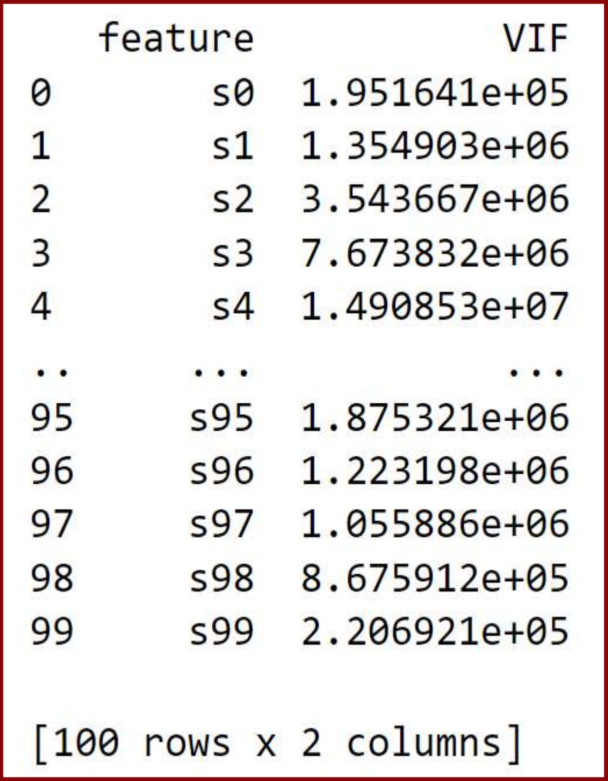
Independent variables	Light reflected at different frequencies	***s1, s2, ... s100***
	Age (discrete)	***Min: 6 years*** ***Max: 95 years***
	Gender (binary)	***Male, Female***
Dependent Variable	Actual HB values **(g/dl)**	***Min: 4*** ***Max: 17.6*** ***Avg: 11.51***

BOX PLOT for Frequency of Light

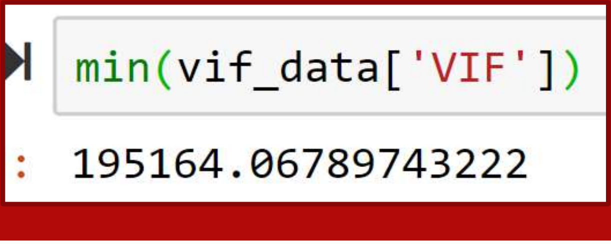


CORRELATION MATRIX for Frequencies of Light (X variables)





```
    feature           VIF
0        s0  1.951641e+05
1        s1  1.354903e+06
2        s2  3.543667e+06
3        s3  7.673832e+06
4        s4  1.490853e+07
..      ...           ...
95      s95  1.875321e+06
96      s96  1.223198e+06
97      s97  1.055886e+06
98      s98  8.675912e+05
99      s99  2.206921e+05

[100 rows x 2 columns]
```



```
min(vif_data['VIF'])
```

```
: 195164.06789743222
```

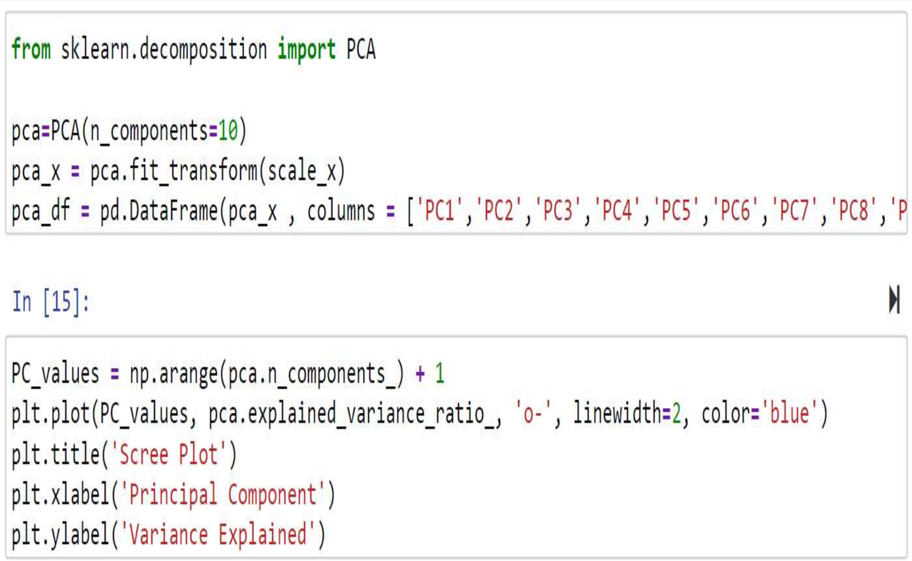
Principal Component Analysis

- Handle “curse of dimensionality” + avoid issues like over-fitting in high dimensional space .
- PCA - method used to reduce number of variables in your data by extracting important one from a large pool.
- Combines highly correlated variables together to form a smaller number of set of variables - “principal components” that account for most variance in the data.

ADVANTAGES:

- Principal components are independent of each other, so removes correlated features.
- PCA helps in overcoming data overfitting issues by decreasing the number of features.

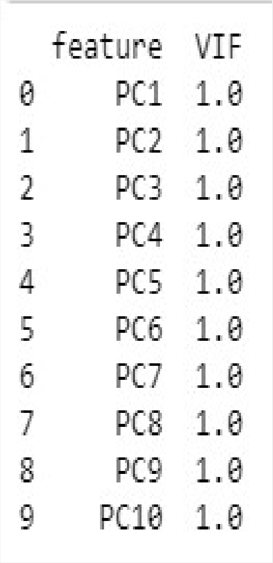
Principal Component Analysis



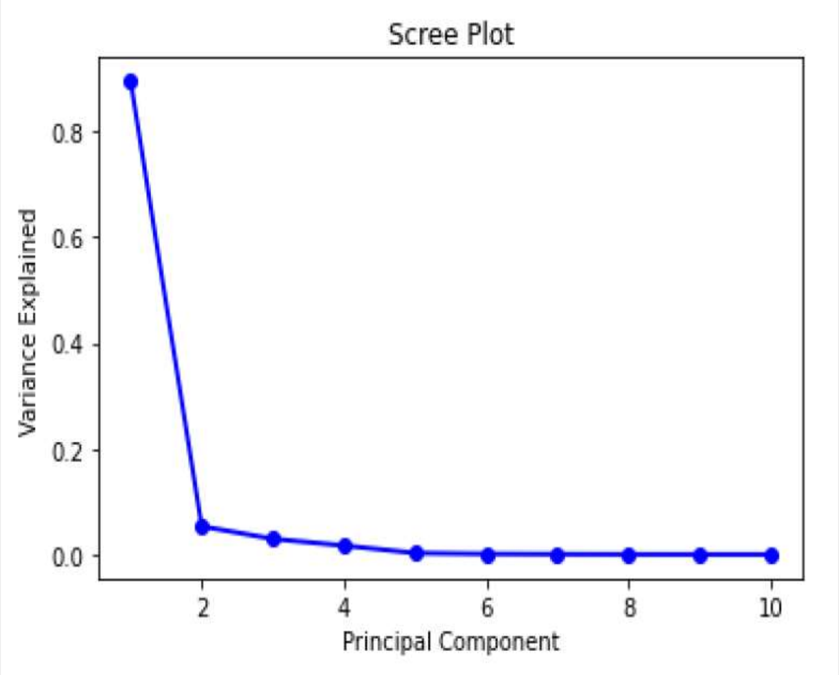
```
from sklearn.decomposition import PCA


pca=PCA(n_components=10)
pca_x = pca.fit_transform(scale_x)
pca_df = pd.DataFrame(pca_x , columns = ['PC1','PC2','PC3','PC4','PC5','PC6','PC7','PC8','P
```

```
In [15]:
```

```
PC_values = np.arange(pca.n_components_) + 1
plt.plot(PC_values, pca.explained_variance_ratio_, 'o-', linewidth=2, color='blue')
plt.title('Scree Plot')
plt.xlabel('Principal Component')
plt.ylabel('Variance Explained')
```

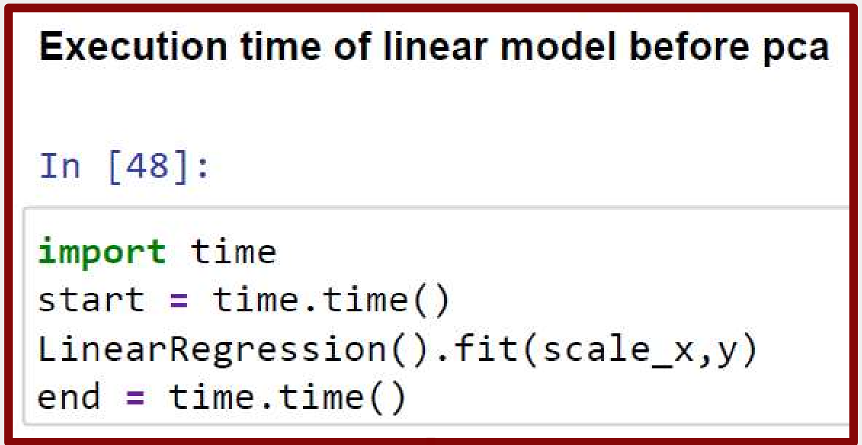


```
  feature  VIF
0     PC1  1.0
1     PC2  1.0
2     PC3  1.0
3     PC4  1.0
4     PC5  1.0
5     PC6  1.0
6     PC7  1.0
7     PC8  1.0
8     PC9  1.0
9    PC10  1.0
```



TOTAL VARIANCE EXPLAINED: 99.97%

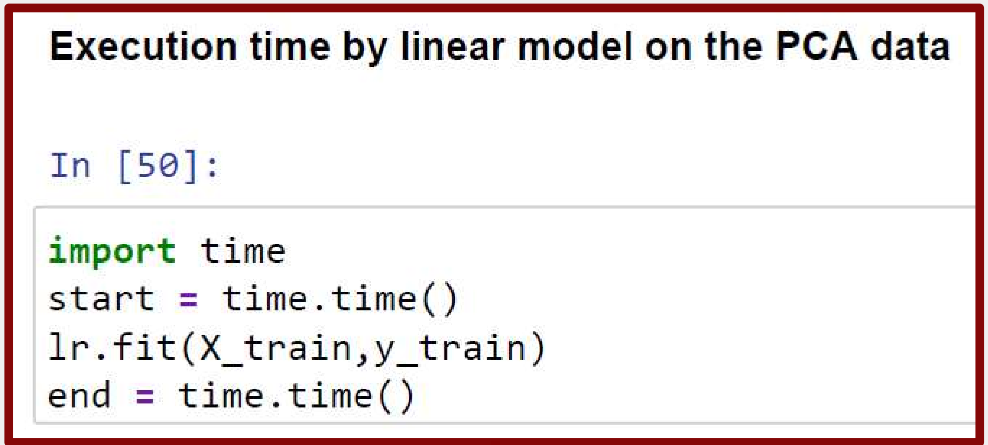
Time Saved by PCA



Execution time of linear model before pca

```
In [48]:
```

```
import time
start = time.time()
LinearRegression().fit(scale_x,y)
end = time.time()
```



Execution time by linear model on the PCA data

```
In [50]:
```

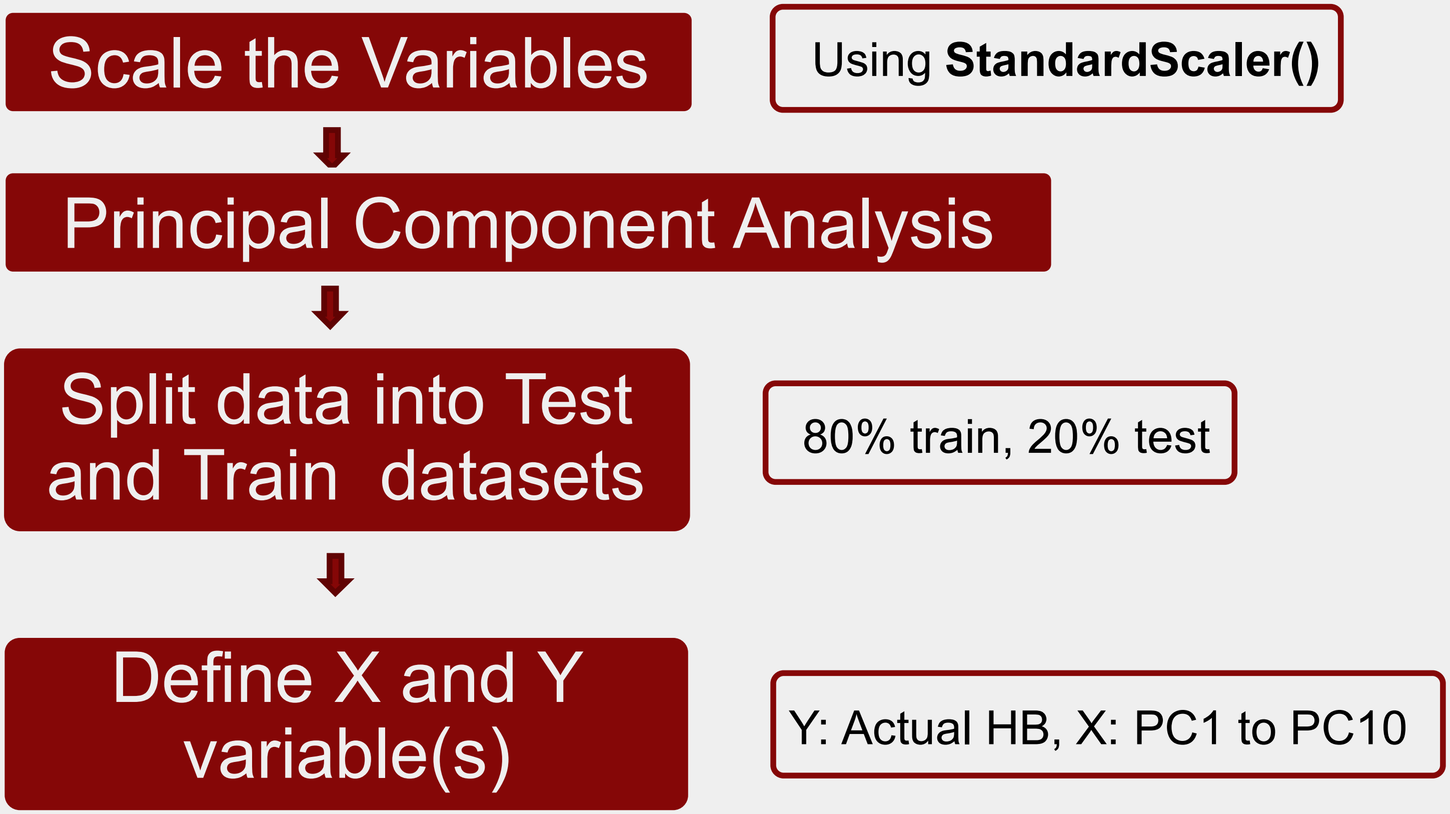
```
import time
start = time.time()
lr.fit(X_train,y_train)
end = time.time()
```

3.044158

0.106074

Reduction in time by nearly 96.5%

Pre Model Fitting



Scale the Variables

Using **StandardScaler()**

Principal Component Analysis

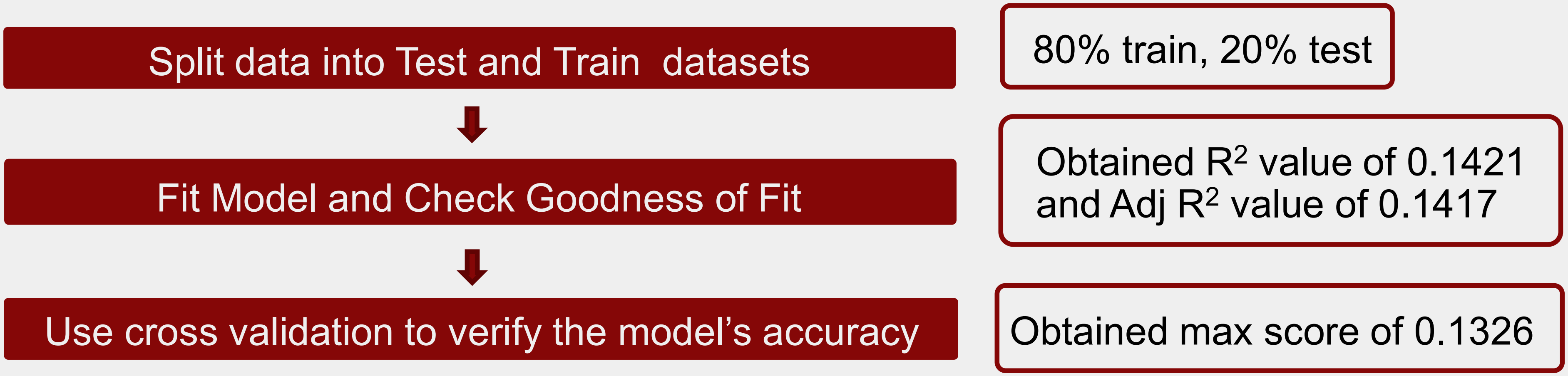
Split data into Test and Train datasets

80% train, 20% test

Define X and Y variable(s)

Y: Actual HB, X: PC1 to PC10

Linear Regression



Split data into Test and Train datasets

Fit Model and Check Goodness of Fit

Use cross validation to verify the model's accuracy

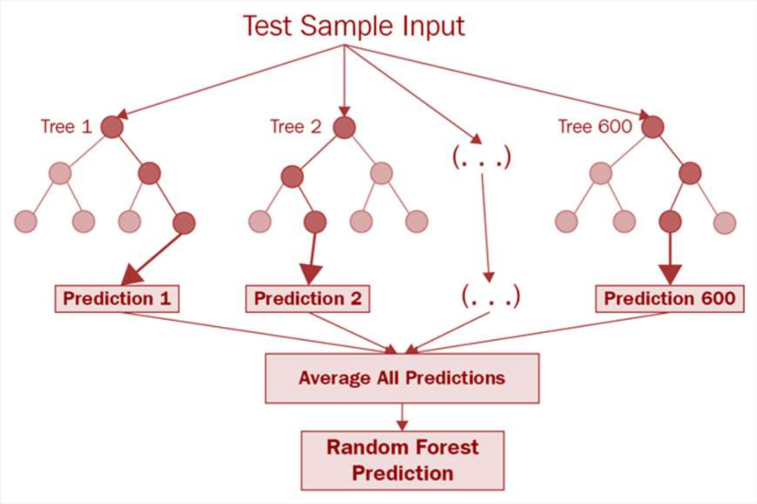
80% train, 20% test

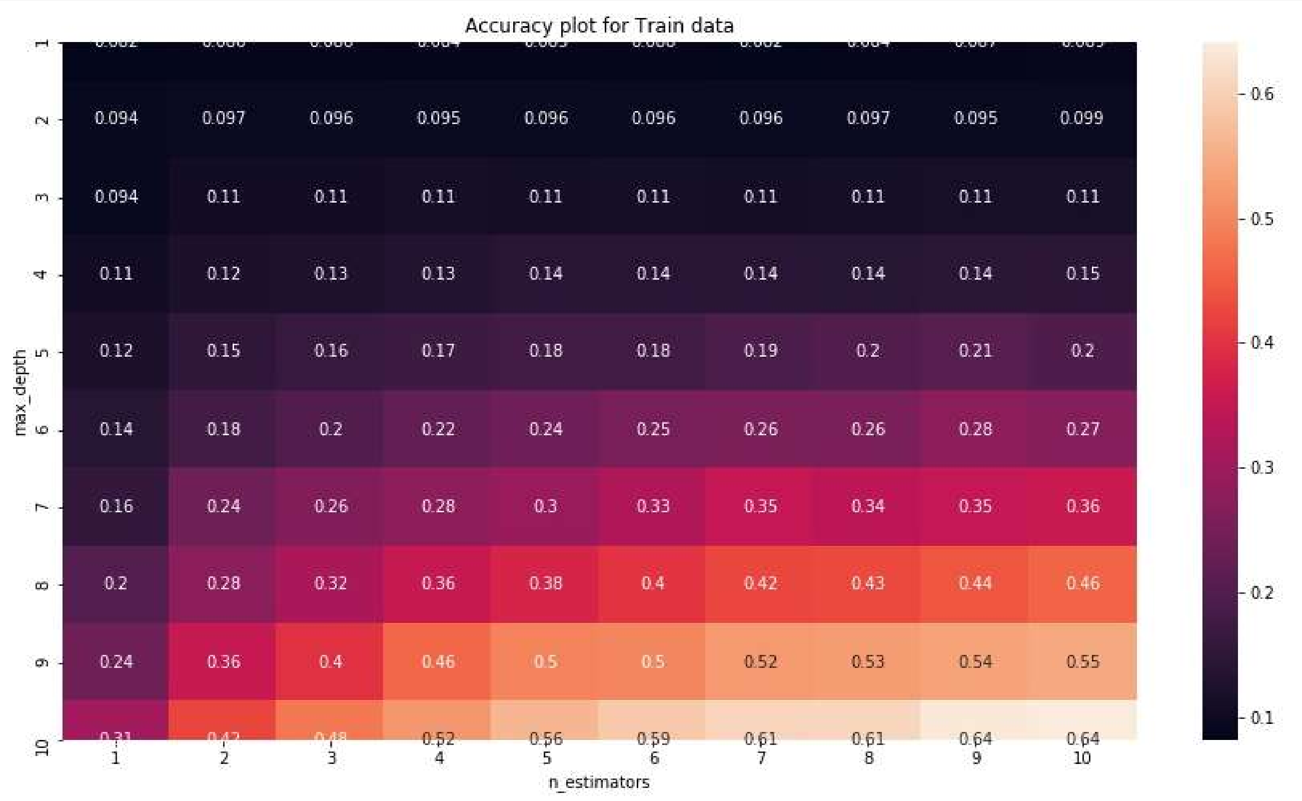
Obtained R^2 value of 0.1421 and Adj R^2 value of 0.1417

Obtained max score of 0.1326

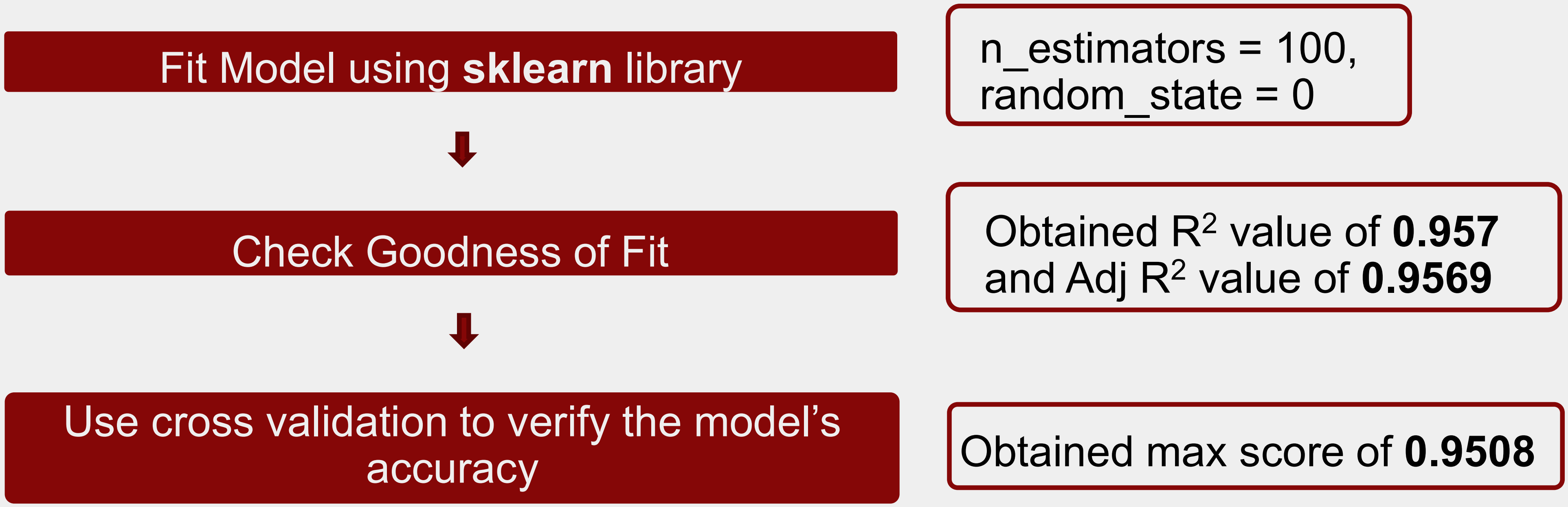
HB = 11.51 + (0.0227*PC1) + (0.1836*PC2) + (-0.1424*PC3) + (-0.2829*PC4) + (-0.1626*PC5) + (0.1172*PC6) + (0.1001*PC7) + (0.7588*PC8) + (0.1467*PC9) + (-0.9694*PC10)

Random Forest





Random Forest



Fit Model using **sklearn** library

n_estimators = 100, random_state = 0

Check Goodness of Fit

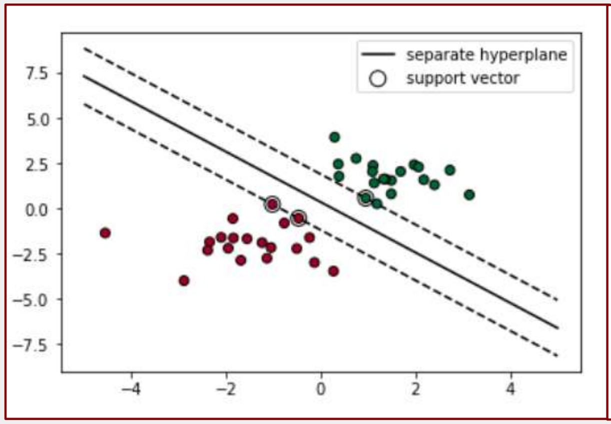
Obtained R^2 value of **0.957** and Adj R^2 value of **0.9569**

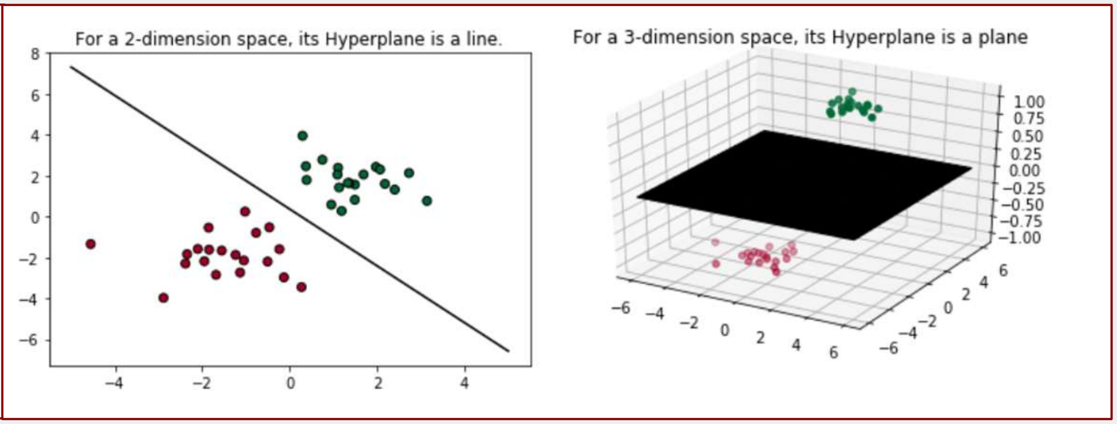
Use cross validation to verify the model's accuracy

Obtained max score of **0.9508**

Support Vector Machines

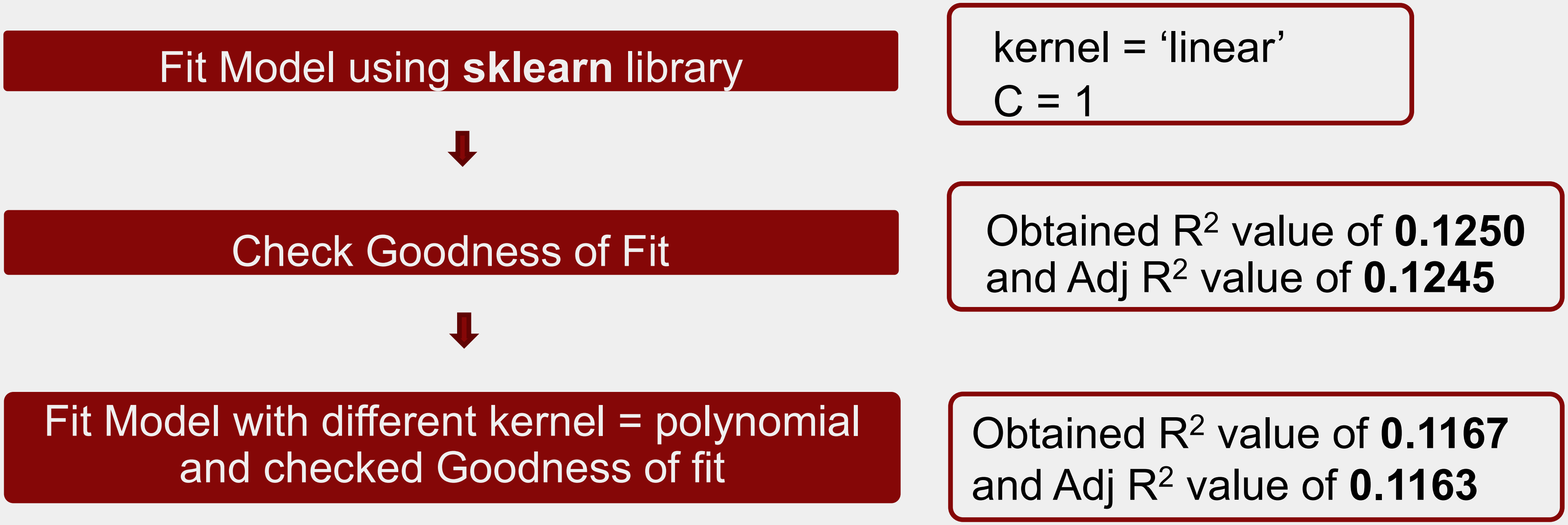
- Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems.
- The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes.




For a 2-dimension space, its Hyperplane is a line.

For a 3-dimension space, its Hyperplane is a plane

Support Vector Machines



Fit Model using **sklearn** library

kernel = ‘linear’
C = 1

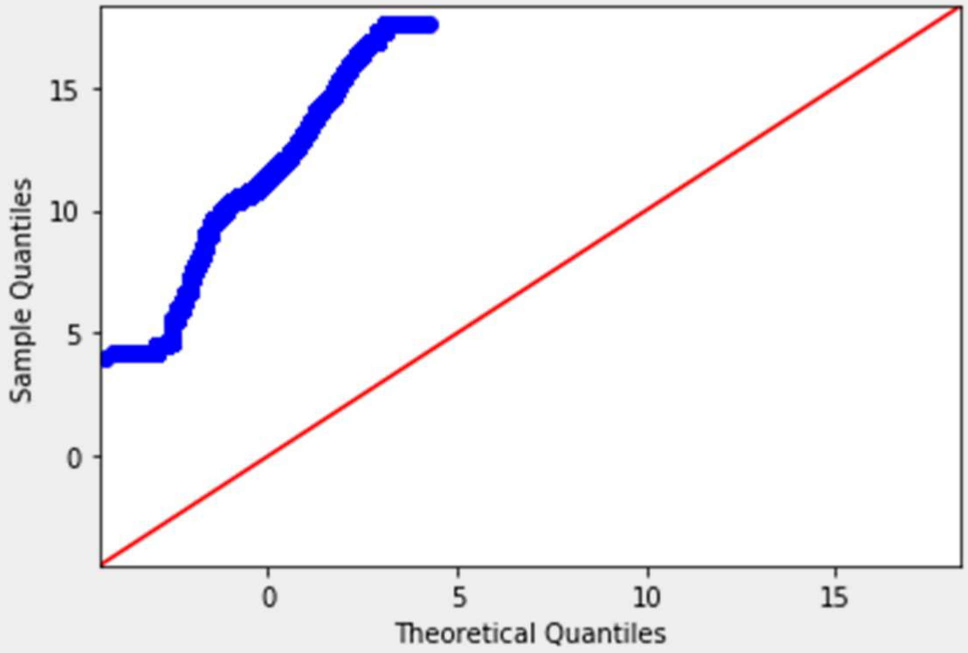
Check Goodness of Fit

Obtained R^2 value of **0.1250** and Adj R^2 value of **0.1245**

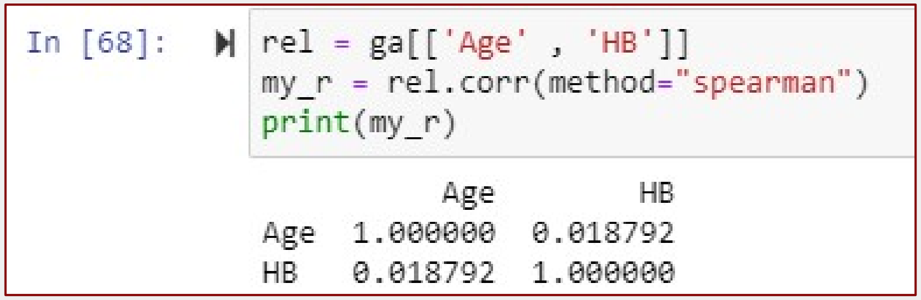
Fit Model with different kernel = polynomial and checked Goodness of fit

Obtained R^2 value of **0.1167** and Adj R^2 value of **0.1163**

Age

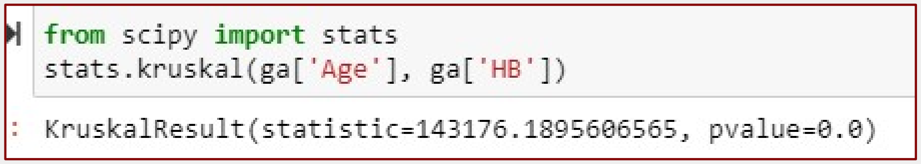


Correlation between Age and HB



```
In [68]:   rel = ga[['Age' , 'HB']]
           my_r = rel.corr(method="spearman")
           print(my_r)

                Age        HB
           Age  1.000000  0.018792
           HB   0.018792  1.000000
```

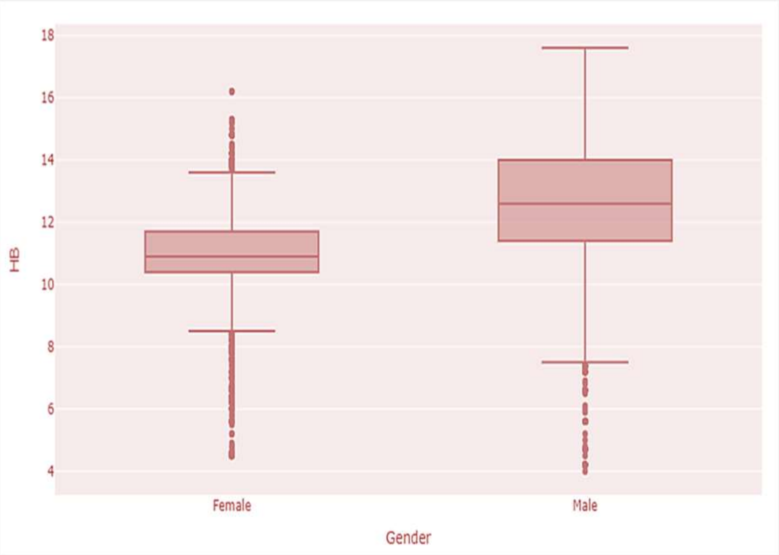
Kruskal-Wallis Test of Significance for Age

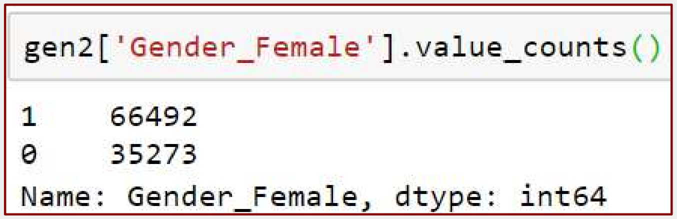


```
from scipy import stats
stats.kruskal(ga['Age'], ga['HB'])

: KruskalResult(statistic=143176.1895606565, pvalue=0.0)
```

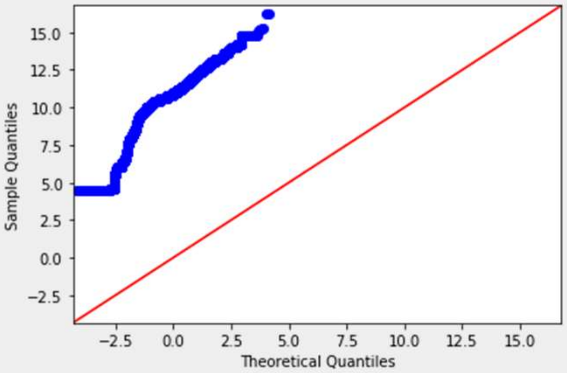
Gender

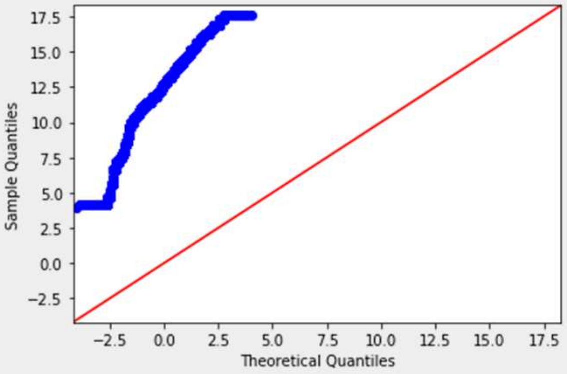
BOX-PLOT OF GENDER VS. HB



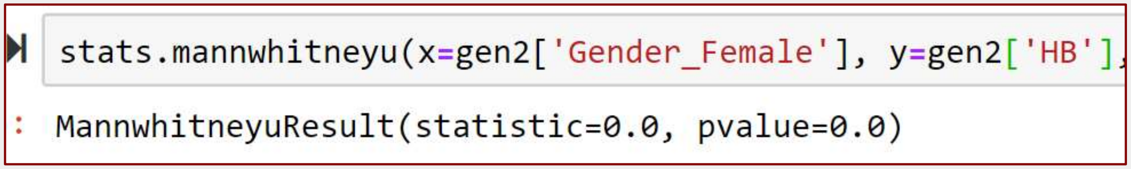


```
gen2['Gender_Female'].value_counts()

1    66492
0    35273
Name: Gender_Female, dtype: int64
```





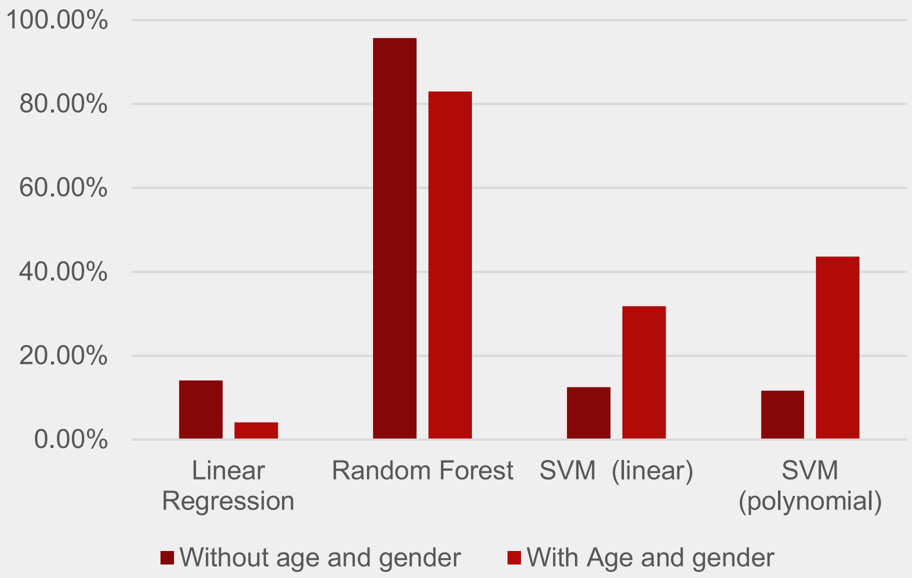
Mann Whitney U Test of Significance for Gender

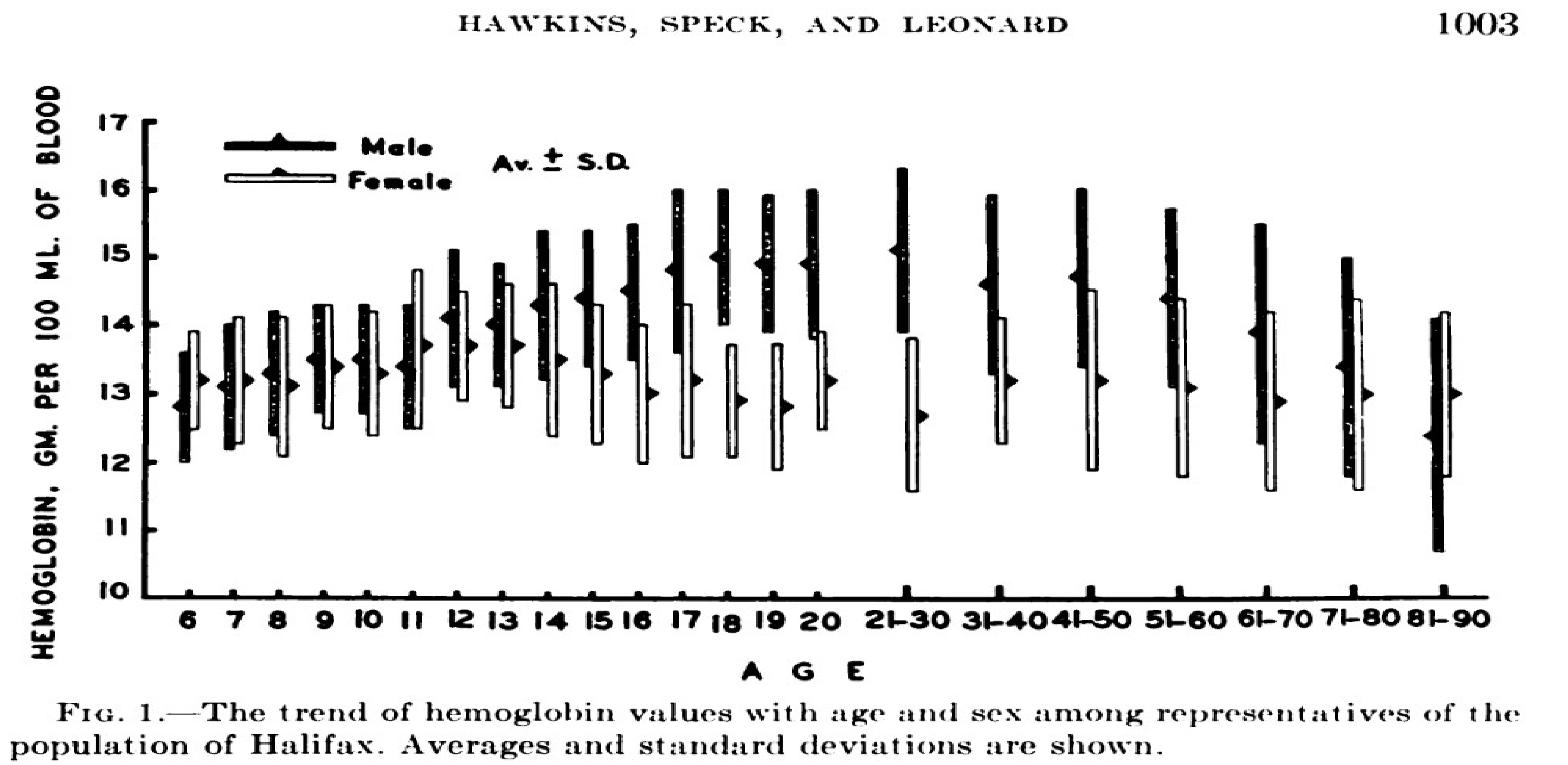


```
stats.mannwhitneyu(x=gen2['Gender_Female'], y=gen2['HB'],

: MannwhitneyuResult(statistic=0.0, pvalue=0.0)
```

Comparison of Model Results

Model	Without age and gender	With Age and gender
Linear Regression	14.17%	4.169%
Random Forest	95.69%	82.93%
SVM (linear)	12.50%	3.18%
SVM (polynomial)	11.67%	4.36%`



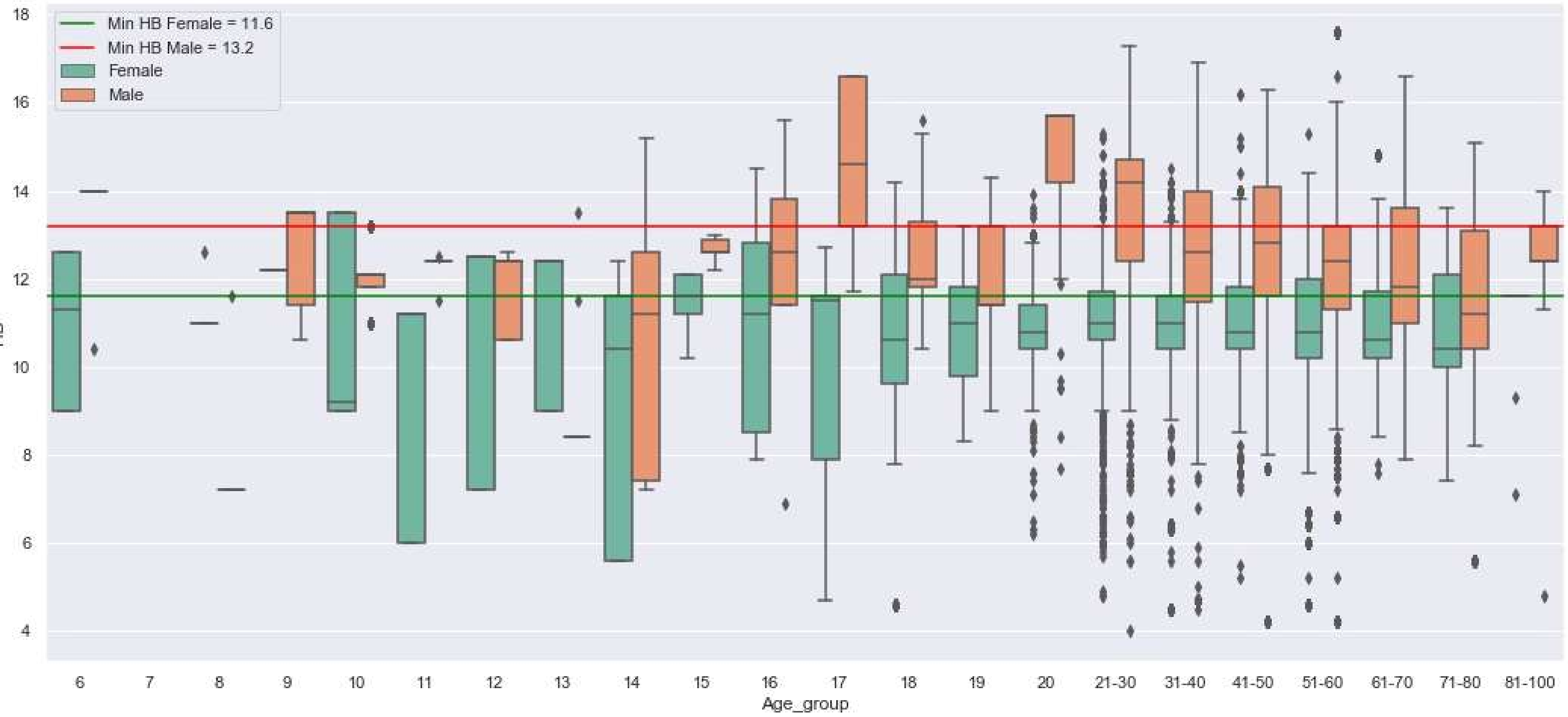


HAWKINS, SPECK, AND LEONARD 1003

FIG. 1.—The trend of hemoglobin values with age and sex among representatives of the population of Halifax. Averages and standard deviations are shown.

W. W. HAWKINS, ElaLY5 SPECK AND VERNA G. LEONARD [1952-53] "Variation of the Hemoglobin Level with Age and Sex"

BOX-PLOT for Age and Gender Vs. Actual HB



Conclusions

1. The Random Forest Algorithm was **successful** in predicting the Haemoglobin Blood Concentration levels with a prediction accuracy of **almost 96%** without the inclusion of Age and Gender as predictors.

2. The use of Principal Component Analysis has been an important step in making the predictor variables of light reflected (at different frequencies) independent principal components therefore **removing the multicollinearity** present in Xs. Moreover, it greatly improved model efficiency by reducing the **computational time** in execution.

3. Although Age and Gender are factors of **statistical significance** to HB, they negatively impact model accuracy due to the niche sample group from which the data was collected.

Link to code: https://github.com/hbapuram/IDS_EzeRx

FUTURE WORK

- The necessary adjustments based on factors [device dependent variables] need to be investigated before generalising the results. The incandescent light's properties also present another variable for future exploration.
- Limited Haemoglobin Range has been considered which contributed to a skewed data set, hence a broader coverage can be expected to normalise the data leading to uniform results. Patients with blood disorders should also be inclusive.
- To enforce proper finger placement and reduce movement, a finger cuff can be designed to centre the finger in the device.
- Apart from Hemoglobin (HB) we are trying to model SpO2, BL, CR, RBS, Sugar - FS and PP, HB A1c, Sickle Cell, Urea, Uric Acid, CHOL, TG, HDL, LDL, NA+, K+, CAL, PHOS etc. Looks like sky is the limit...
- A research paper is being worked upon keeping in mind the quality of data sets in order to publish the results.

Acknowledgement

- We want to thank the team of EzeRx for their help with the coordination and data collection during the clinical study and extending the same to us . This helped us gain insights on new developments in medical avenues.
- In particular, We would like to acknowledge and express our gratitude to our Mentor, Prof Sarada Samantaray sir, for not only helping us structure, format and understand the concepts employed in this project, but to also provide us such an enriching experience and a wonderful opportunity to get a sense of the real time data.
- We would also like to thank our Introduction to Data Science Professor, Shubham Bhatt for guiding us through the project.
- This work is performed under the approval from the NMIMS deemed to be University, Navi Mumbai , School of Science, who provided us with necessary tools and resources.

References

1] Edward Jay Wang , William Li , Doug Hawkins , Terry Gernsheimer , Colette Norby-Slycord , Shwetak N. Patel [2016] " HemaApp: Noninvasive Blood Screening of Haemoglobin using Smartphone Cameras." UBICOMP '16, SEPTEMBER 12–16, 2016, HEIDELBERG, GERMANY.

2] W. W. HAWKINS, ElaLY5 SPECK AND VERNA G. LEONARD [1952-53] "Variation of the Hemoglobin Level with Age and Sex".

3] Bushra Alsunaidi 1 , Murad Althobaiti 1 , Mahbubunnabi Tamal 1 , Waleed Albaker 2 and Ibraheem Al-Naib [2021] "A Review of Non-Invasive Optical Systems for Continuous Blood Glucose Monitoring" Sensors 2021, 21, 6820

4] Caje Pinto, Jivan Parab, Gourish Naik [2020] "Non-invasive hemoglobin measurement using embedded platform" Sensing and Bio-Sensing Research Volume 29, August 2020, 100370

5] Giovanni Dimauro * , Danilo Caivano , Pierangelo Di Pilato, Alessandro Dipalma and Mauro Giuseppe Camporeale [2020] "A Systematic Mapping Study on Research in Anemia Assessment with Non-Invasive Devices" Appl. Sci. 2020, 10, 4804

6] Selim Suner,James Rayner ,Ibrahim U. Ozturan ,Geoffrey Hogan, Caroline P. Meehan ,Alison B. Chambers,Janette Baird, Gregory D. Jay [2021], "Prediction of anaemia and estimation of hemoglobin concentration using a smartphone camera"

7] Joon-myoung Kwon, MD ,Younghoon Cho, MD ,Ki-Hyun Jeon, MD ,Soohyun Cho, MD Kyung-Hee Kim, MD,Seung Don Baek, MD, Soomin Jeung, MD, Jinsik Park, MD, Byung-Hee Oh, MD [2020] "A deep learning algorithm to detect anaemia with ECGs: a retrospective, multicentre study" VOLUME 2, ISSUE 7, E358-E367

Our Team

Dhanashri Kanitkar

Hrishita Bapuram

Shraddha Kodavade

THANK YOU!