Data Analysis of Blood Transfusion Center using principal component analysis

Mohammed Faizan
INSE 6220: *Advanced Statistical Approaches to Quality*
Montreal, Quebec, Canada
Email: reach2mdfaizan@gmail.com

***Abstract*—With the advent of technology, there have been drastic advancements in the field of healthcare. The data plays an important role in this advancement. Here we are emphasizing data analysis for healthcare organizations. While a doctor's skills and diagnosis are important in medicine, with data science, knowledge of data and prediction plays a crucial role. The donation of blood can save people, especially those who are cancer patients or people who have met with an accident. Therefore, there is a need for a system to track and predict donations which in turn can make sure more people get the blood transfusions they need. Thus, this problem can be seen as a classification task. In particular, we are building a predictive model to determine whether a blood donor gave blood in a certain time based on different characteristics of the input parameters.**

***Keywords*— Healthcare, data analysis, blood donation.**

I. Introduction

In the past two decades, technology has drastically improved storage capabilities. As a result of this improvement, the medical data is now being stored in electronic formats. The medical data could be stored as images, audio, and videos. And more often these stored data are now being used by many healthcare organizations for diagnosis of a patient. The purpose of this paper is to build a predictive model to determine whether a donor gave blood during a certain time.

With the increase of attributes in data, it is a difficult task to locate important features in the data. To overcome that problem, a dimensionality reduction technique can be used for a classification problem. Therefore, in this work, we are using principal component analysis (PCA) to identify important features and do the prediction of the data.

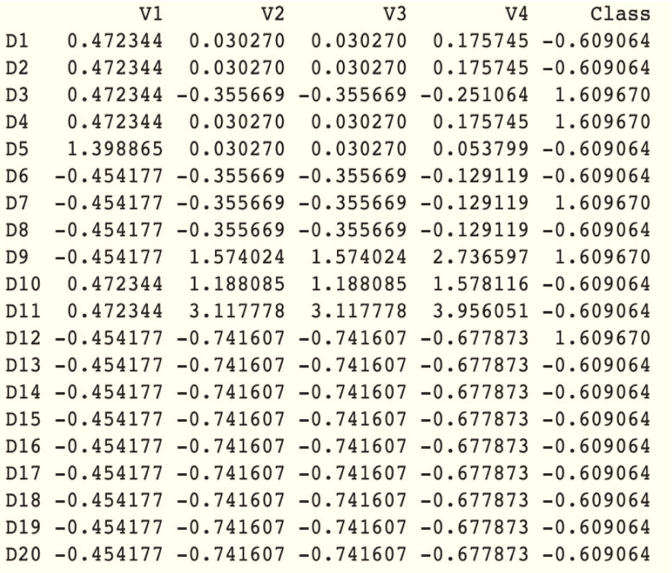
The paper is structured as follows: Section 2 gives detailed information about the data [1]. In section 3, PCA is introduced and analysis is carried out. In section 4, trains and tests the build prediction model and the results of implementations are provided. Finally, section 5 covers the conclusion part of the work.

II. Blood Transfusion Center Data

The data is collected from the OpenML repository [2] and consists of 748 instances with 5 attributes. The area of specialization for the data is the business sector, and the source of the database is from Blood Transfusion Center in Hsin-Chu city in Taiwan [2]. The data set is used in Comma Separated Value (CSV) format and the predictive model is built using Python programming language. The 5 attributes present in the data are:

- V1: Recency – says about months since last blood donation.
- V2: Frequency – it gives information about the total number of donations.
- V3: Monetary – tells the total amount of blood donated in cubic centimeters.
- V4: Time – is a value that describes the number of months since the first donation.
- Class – this is a binary variable attribute representing whether a person has donated blood in March 2007. If the value is 1 then it means donating blood, and 0 denotes otherwise [3].

Figure 1 shows the first 20 rows of this data set, and table I presents the summary of different attributes of the data.



```
           V1        V2        V3        V4     Class
D1   0.472344  0.030270  0.030270  0.175745 -0.609064
D2   0.472344  0.030270  0.030270  0.175745 -0.609064
D3   0.472344 -0.355669 -0.355669 -0.251064  1.609670
D4   0.472344  0.030270  0.030270  0.175745  1.609670
D5   1.398865  0.030270  0.030270  0.053799 -0.609064
D6  -0.454177 -0.355669 -0.355669 -0.129119 -0.609064
D7  -0.454177 -0.355669 -0.355669 -0.129119  1.609670
D8  -0.454177 -0.355669 -0.355669 -0.129119 -0.609064
D9  -0.454177  1.574024  1.574024  2.736597  1.609670
D10  0.472344  1.188085  1.188085  1.578116 -0.609064
D11  0.472344  3.117778  3.117778  3.956051 -0.609064
D12 -0.454177 -0.741607 -0.741607 -0.677873  1.609670
D13 -0.454177 -0.741607 -0.741607 -0.677873 -0.609064
D14 -0.454177 -0.741607 -0.741607 -0.677873 -0.609064
D15 -0.454177 -0.741607 -0.741607 -0.677873 -0.609064
D16 -0.454177 -0.741607 -0.741607 -0.677873 -0.609064
D17 -0.454177 -0.741607 -0.741607 -0.677873 -0.609064
D18 -0.454177 -0.741607 -0.741607 -0.677873 -0.609064
D19 -0.454177 -0.741607 -0.741607 -0.677873 -0.609064
D20 -0.454177 -0.741607 -0.741607 -0.677873 -0.609064
```

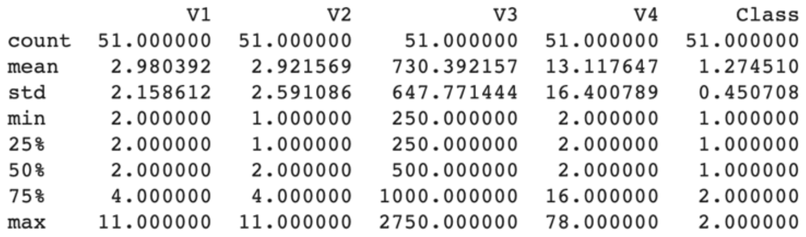
Fig. 1. Top 20 rows of blood transfusion center data set after normalization.

Figure 2 shows the box plot diagram for all the attributes of the data. The box plot gives a standardized distribution of the data. It indicates that there are 12 outliers in total, with Time attribute having the highest number of outliers (4) and Recency having the lowest number of outliers (2). Also, notice that the Class attribute does not have any outliers as it is a binary variable as mentioned above.

To have a better understanding of the relationship between the attributes, the scatter plot matrix is used as shown in figure 3. The scatter plot matrix seems to suggest a linear relationship between the two variables, V2 and V3.

Classification can be of two separate methods – it could be a binary classification or multiclass classification. After thoroughly examining the data with the help of figures and information, the problem observed would be a binary classification problem, as we are required to identify whether a person has donated blood on a particular date, which enables classifiers to predictions using the Boolean output of either 1 or 0.

TABLE I. - Summary of Blood Transfusion Center data



```
              V1         V2           V3         V4      Class
count  51.000000  51.000000    51.000000  51.000000  51.000000
mean    2.980392   2.921569   730.392157  13.117647   1.274510
std     2.158612   2.591086   647.771444  16.400789   0.450708
min     2.000000   1.000000   250.000000   2.000000   1.000000
25%     2.000000   1.000000   250.000000   2.000000   1.000000
50%     2.000000   2.000000   500.000000   2.000000   1.000000
75%     4.000000   4.000000  1000.000000  16.000000   2.000000
max    11.000000  11.000000  2750.000000  78.000000   2.000000
```

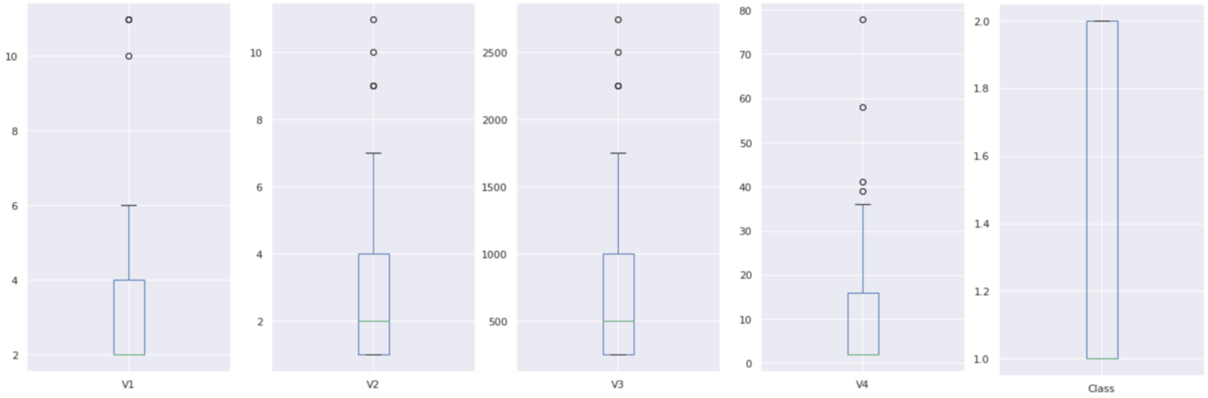


Fig. 2. Box Plot diagram.

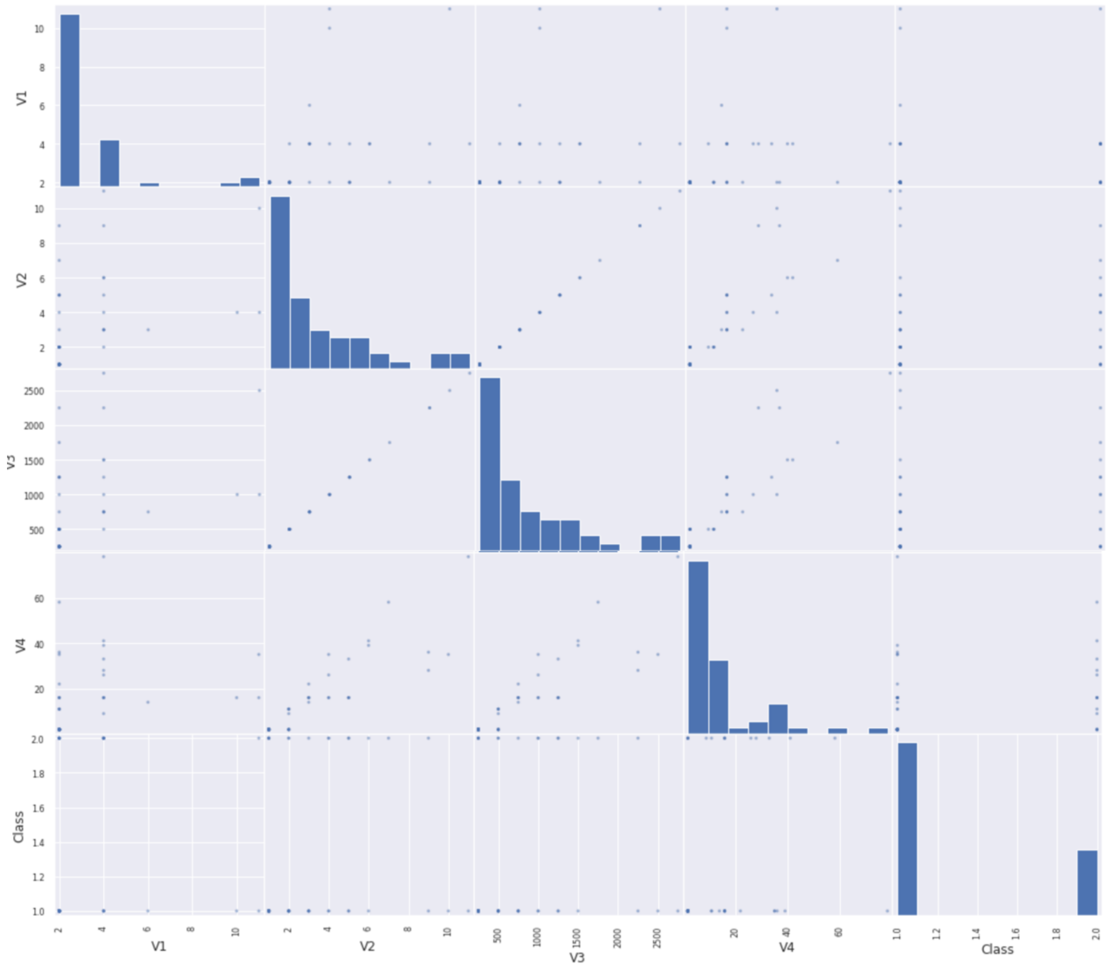


Fig. 3. Scatter Plot Matrix.

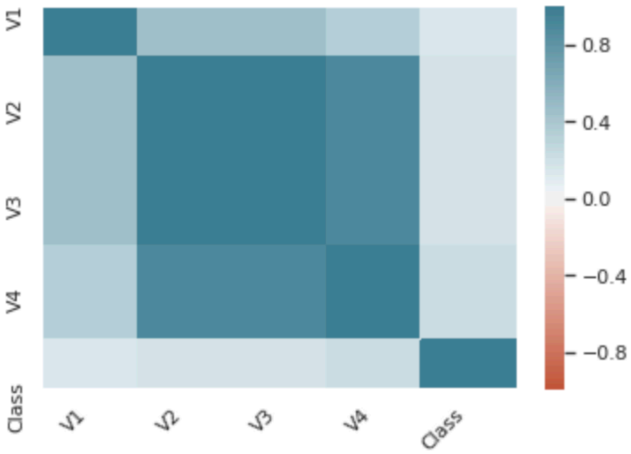


Fig. 4. Correlation Matrix

To check the relationship among columns of the data, a correlation matrix as shown in figure 4 can be used. By this, we can interpret that there is no negative correlation (red shade) between the columns. Also, there is a strong positive correlation between V2 and V3 (blue shade), which means if one of them increases, the other will also increase. In addition, we note that there exists a weak positive correlation between the variables, V2 and V4. This is likely due to the presence of anomalies in the data.

III. Principal Component Analysis

An analysis method is needed to identify important features in data. Principal Component Analysis is a technique that transforms multivariate data into uncorrelated variables called principal components [4].

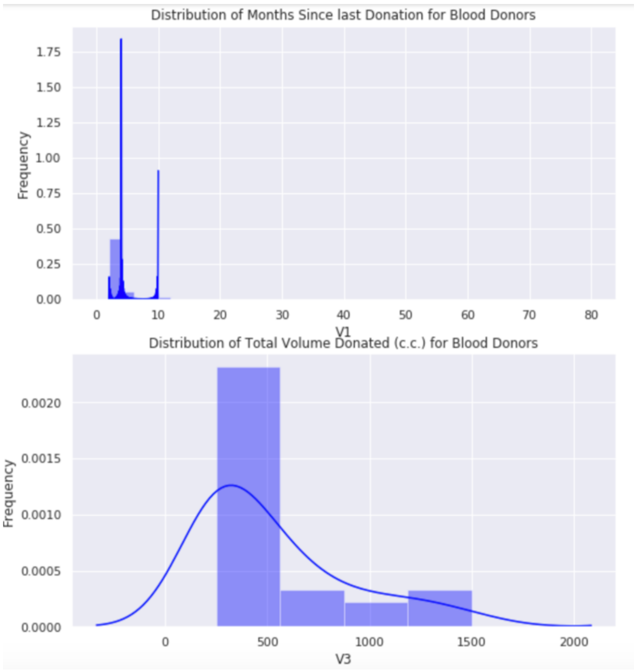


Fig. 5. Distribution Plot for V1 and V2

Figure 5 portrays the distribution of months since the last donation and total volume of blood donated in cubic centimeters. It can be interpreted from the above figure that people have frequently donated blood in the past.

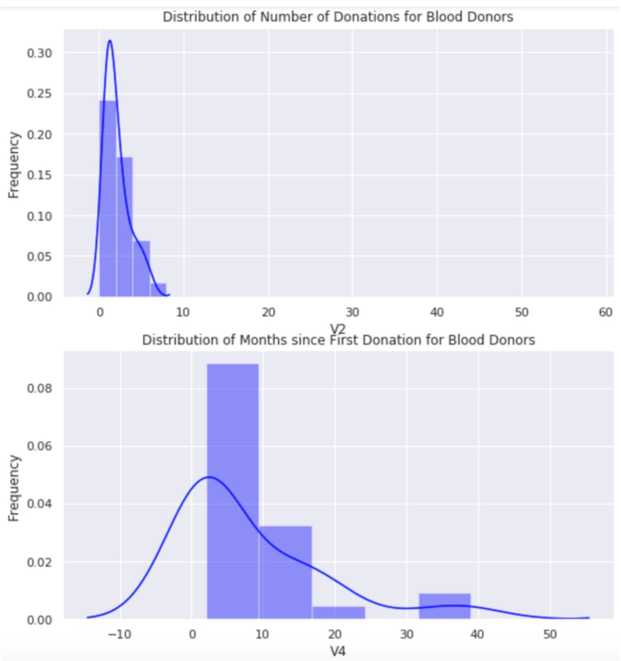


Fig. 6. Distribution Plot for V3 and V4

Figure 6 shows that V4 which is the months since the first donation feature is particularly not informative of whether a blood donation occurred. In addition, there are several outliers in the data which in turn could reduce the prediction accuracy.

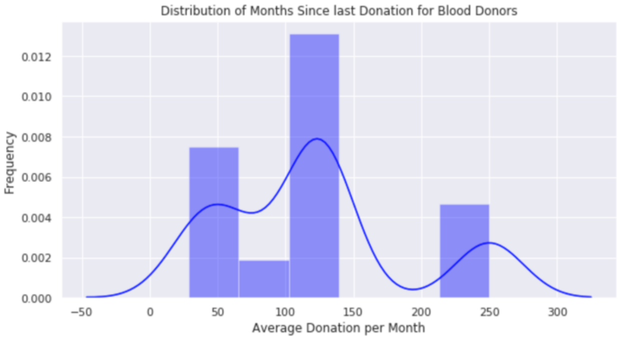


Fig. 7. Distribution Average number of donations per month

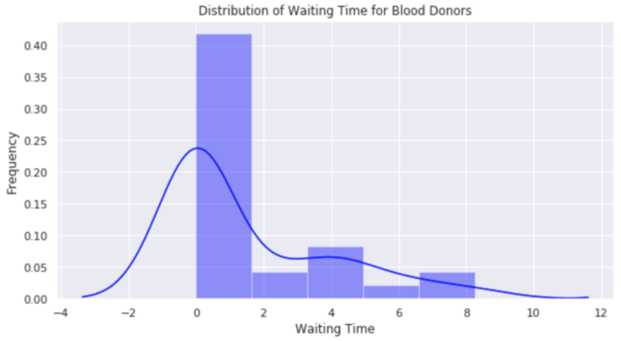
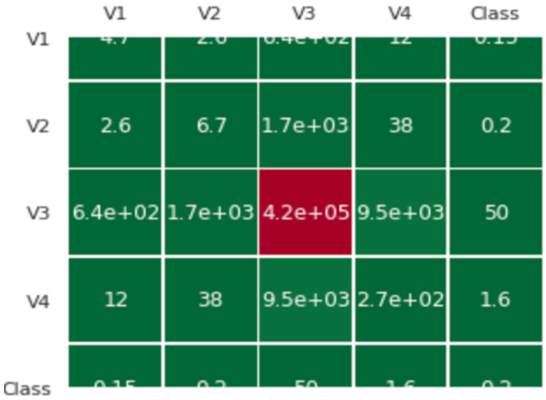


Fig. 8. Distribution of waiting time.

From the plot in figure 7, we can conclude that on average, there is a donation of around 100ml of blood per month. We used a new variable called waiting time, to figure out the frequency of donation of blood. As shown in figure 8, there are a plethora of one-time blood donors, and people do not donate blood after their first time. Now that we have an understanding of the data, we proceed by first normalizing the dataset and apply principal component analysis.



	V1	V2	V3	V4	Class
V1	4.7	2.6	6.4e+02	12	0.15
V2	2.6	6.7	1.7e+03	38	0.2
V3	6.4e+02	1.7e+03	4.2e+05	9.5e+03	50
V4	12	38	9.5e+03	2.7e+02	1.6
Class	0.15	0.2	50	1.6	0.2

Fig. 9. Covariance Matrix with centered data.

After normalizing the data, we have drawn a covariance matrix with centered data as shown in figure 9, we go ahead with the scaling of data to achieve the best quality of results for the prediction model.

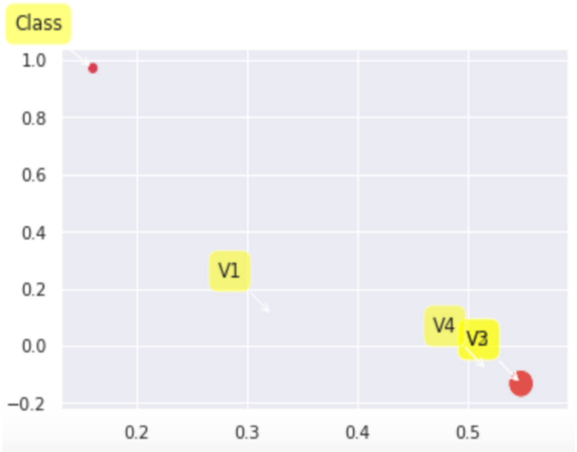


Fig. 10. Eigen Value and Vectors representation.

The next step in PCA analysis is the calculation of eigenvalues and eigenvectors of the covariance matrix generated. The eigenvector is the best fit line, while the eigenvalues are the values that tell us how the data is spread out on this line as shown in figure 10. Thus, we have calculated eigenvalues which tell us more about variance in particular dimensions.

$$\lambda = \begin{bmatrix} 3.11612 \\ 0.9501 \\ 0.7576 \\ 0.1310 \\ 0.0005 \end{bmatrix}$$

After calculating the eigenvalues and eigenvectors, we can calculate the principal component analysis as follows:

$$Z = XA, \qquad (1)$$

Finally, the most important aspect of using the principal component analysis is to identify which attribute has the highest contribution. To implement this step, we plot principal component coefficients against each other as shown in figure 11. The following points can be inferred from observing the diagrams-

- V2 and V3 have similar coefficients for A_1.
- Class is the feature that contributes to A_2 more than others.
- Some features have negative coefficients like V2, V3, and V4, whereas V1 and Class have positive coefficients.

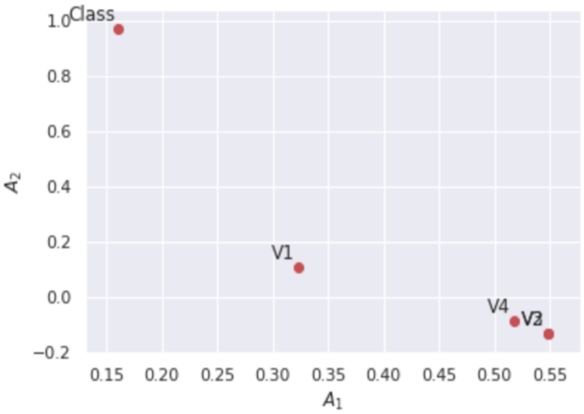


Fig. 11. Scatter plot of A1 and A2 coefficients.

To figure out how many components should be considered; we calculate the explained variance for each component as follows:

$$L_j = \frac{\lambda_j}{\sum_{j=1}^{J} \lambda_j} \times 100\%, \; j = 1, \dots, J \qquad (2)$$

Using the above equation, we calculated the explained variance ratio for the components in blood transfusion center data set, and the values are as follows:

$$L_1 = 63.2247\% \quad L_2 = 19.0026\% \quad L_3 = 15.1525\%$$
$$L_4 = 0.2620\% \quad L_5 = 0.0009\%$$

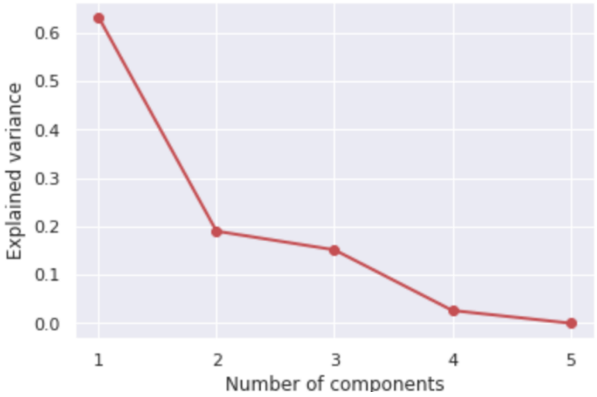


Fig. 12. Explained Variance Plot or Scree plot.

Based on the values represented in the Scree plot figure 12, we can say that the first 2 components contribute to more than 82% of the variance in blood transfusion center data. The minimum dimension used in the representation of the plot is d=1.

From the Pareto plot shown in figure 13, we infer the cumulative explained variance which in turn helps in identifying important features of the dataset. By looking at the diagram, it can be easily observed that the first two components account for most of the variance in data.

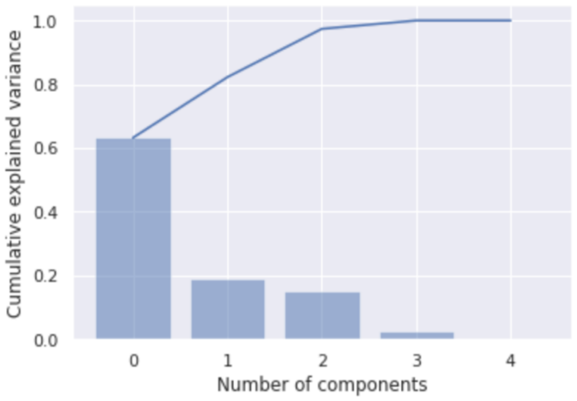


Fig. 13. Pareto Plot.

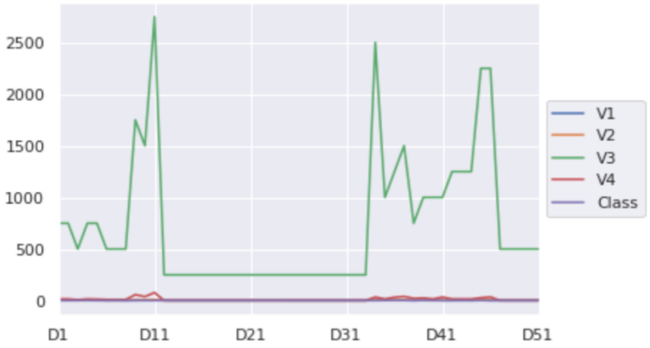


Fig. 14. Profile plot.

To estimate the marginal mean and examine the relative behavior of all variables in the data set, we use the profile plot, which is shown in figure 14. The interpretation that can be done by looking at the profile plot is that the mean and standard deviation for V3 is very high compared to the other variables.

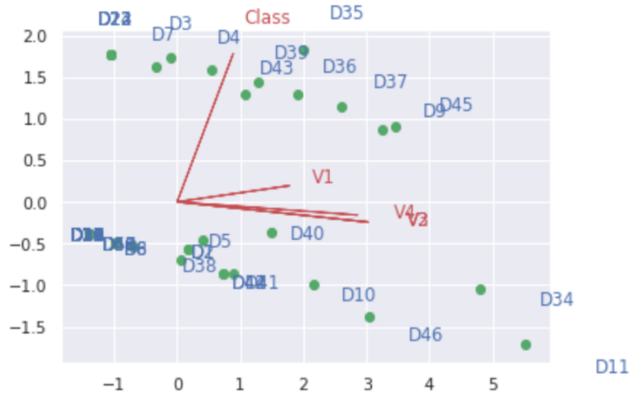


Fig. 15. 2D biplot.

To comprehend the information about PCA scores for observations, and to see how each attribute contributes to principal components, we use Biplots. A 2-Dimensional bi-plot is shown in figure 14. In the figure, the axis of the plot is the principal component. The following information can be inferred from the above plot:

- V2, V3, and V4 are negatively skewed for PC1, whereas V1 and Class are positively skewed.
- For PC2, all attributes are positively skewed, but some observations are less than zero.

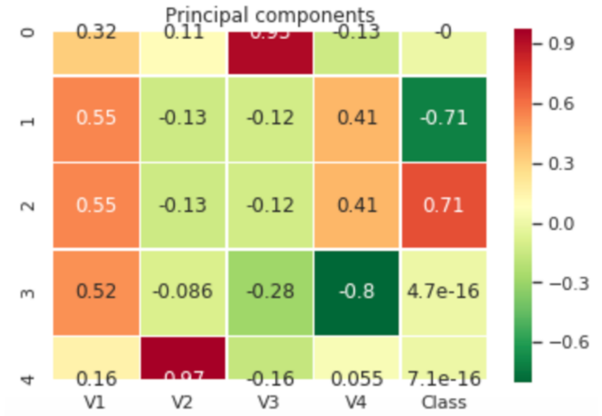


Fig. 16. Principal Components.

We have successfully applied principal component analysis technique for the blood transfusion data set, and have adopted the first two-component, which corresponds to 82% variability in the data set. In the next section of the paper, we are going to use the above information of the analysis and build a model to predict whether a donor has donated blood at a particular time.

IV. Results And Discussion

The following accuracy scores for the classifier (note that these output values are directly taken from the python console).

With the help of PCA, we were able to find V2, and V3 are important features. But we have taken the 'Class' attribute (binary variable) as the target attribute so that we can predict the donation of blood.

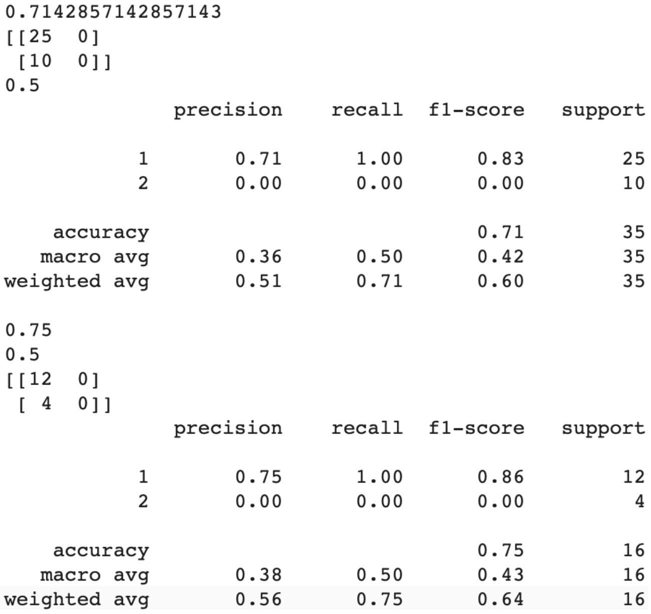
The model is built in the following steps:

1. Preparation of the data set.
2. Descriptive statistics and statistical analysis of the data set.
3. Create a training and testing set (including normalization, validation, etc.)
4. Scale training the test data.
5. Building a predictive model by introducing the algorithms.
6. Applying a suitable algorithm to yield the best possible prediction.
7. Testing the performance of the final prediction model.
8. Comparing the model results with other classifiers.

The initial classifier algorithm that we have used is the k-nearest neighbor (KNN) Classifier. The other classifiers used for the prediction and comparison are:

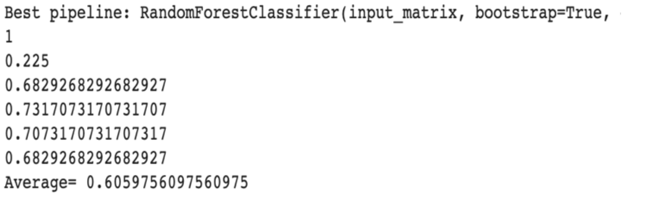
- Random Forest Classifier.
- Support Vector Machine (SVC).
- Perceptron
- Decision Tree Classifier.
- Bernoulli Naïve Bayes Classifier.

We have observed from the results that the k-nearest neighbor classifier yields the best performance results for the prediction of the donation of blood.

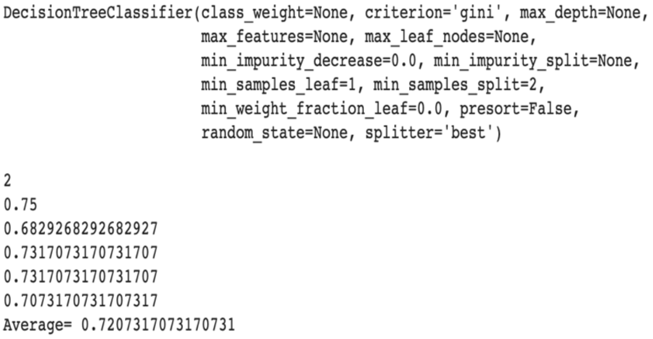


```
0.7142857142857143
[[25  0]
 [10  0]]
0.5
              precision    recall  f1-score   support

           1       0.71      1.00      0.83        25
           2       0.00      0.00      0.00        10

    accuracy                           0.71        35
   macro avg       0.36      0.50      0.42        35
weighted avg       0.51      0.71      0.60        35

0.75
0.5
[[12  0]
 [ 4  0]]
              precision    recall  f1-score   support

           1       0.75      1.00      0.86        12
           2       0.00      0.00      0.00         4

    accuracy                           0.75        16
   macro avg       0.38      0.50      0.43        16
weighted avg       0.56      0.75      0.64        16
```

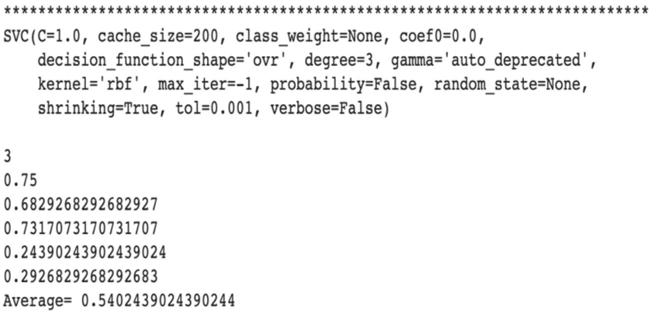
Fig. 17. KNN Classifier model results.



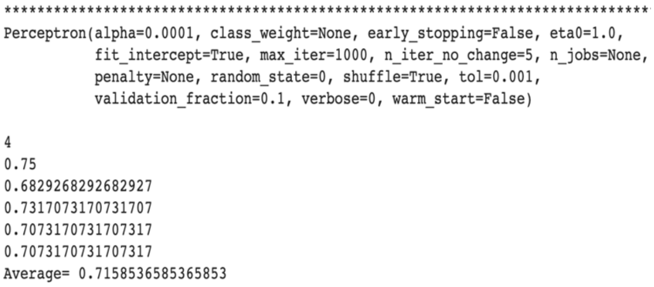
```
Best pipeline: RandomForestClassifier(input_matrix, bootstrap=True,
1
0.225
0.6829268292682927
0.7317073170731707
0.7073170731707317
0.6829268292682927
Average= 0.6059756097560975
```



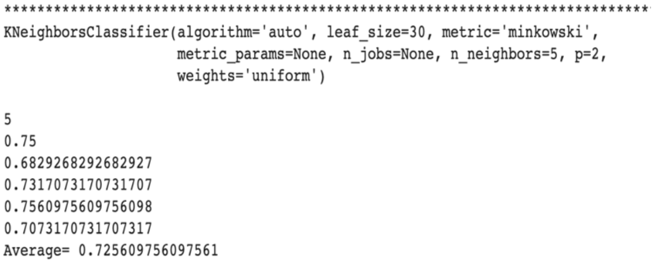
```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
                       max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, presort=False,
                       random_state=None, splitter='best')

2
0.75
0.6829268292682927
0.7317073170731707
0.7317073170731707
0.7073170731707317
Average= 0.7207317073170731
```



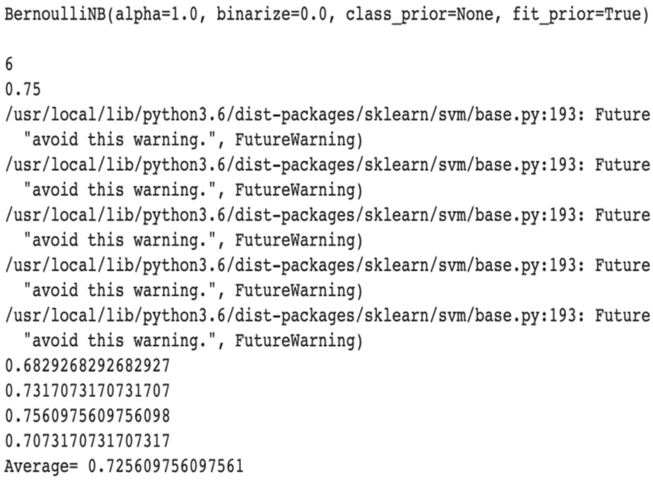
```
***************************************************************************
SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,
    decision_function_shape='ovr', degree=3, gamma='auto_deprecated',
    kernel='rbf', max_iter=-1, probability=False, random_state=None,
    shrinking=True, tol=0.001, verbose=False)

3
0.75
0.6829268292682927
0.7317073170731707
0.24390243902439024
0.2926829268292683
Average= 0.5402439024390244
```



```
***************************************************************************
Perceptron(alpha=0.0001, class_weight=None, early_stopping=False, eta0=1.0,
           fit_intercept=True, max_iter=1000, n_iter_no_change=5, n_jobs=None,
           penalty=None, random_state=0, shuffle=True, tol=0.001,
           validation_fraction=0.1, verbose=0, warm_start=False)

4
0.75
0.6829268292682927
0.7317073170731707
0.7073170731707317
0.7073170731707317
Average= 0.7158536585365853
```



```
***************************************************************************
KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
                     metric_params=None, n_jobs=None, n_neighbors=5, p=2,
                     weights='uniform')

5
0.75
0.6829268292682927
0.7317073170731707
0.7560975609756098
0.7073170731707317
Average= 0.725609756097561
```



```
BernoulliNB(alpha=1.0, binarize=0.0, class_prior=None, fit_prior=True)

6
0.75
/usr/local/lib/python3.6/dist-packages/sklearn/svm/base.py:193: Future
  "avoid this warning.", FutureWarning)
/usr/local/lib/python3.6/dist-packages/sklearn/svm/base.py:193: Future
  "avoid this warning.", FutureWarning)
/usr/local/lib/python3.6/dist-packages/sklearn/svm/base.py:193: Future
  "avoid this warning.", FutureWarning)
/usr/local/lib/python3.6/dist-packages/sklearn/svm/base.py:193: Future
  "avoid this warning.", FutureWarning)
/usr/local/lib/python3.6/dist-packages/sklearn/svm/base.py:193: Future
  "avoid this warning.", FutureWarning)
0.6829268292682927
0.7317073170731707
0.7560975609756098
0.7073170731707317
Average= 0.725609756097561
```

V. Conclusion

Using the data analysis methods, we identified the relationship between different attributes of blood transfusion center data set, used principal component analysis to know the important features in the data set, and lastly, built a predictive model for the donation of the blood. The k-nearest neighbor classifier gives the best results compared to other classifier techniques. Also, the Area Under the ROC Curve (AUC) would be the best error metric for this binary classification problem, because there exists a class imbalance in the data set. We can have better performance for the model if we resolve the class imbalance issue.

References

[1] Xia, J., D. I. Broadhurst, M. Wilson, D. S. Wishart, et al. Metabolomics (2013) 9: 280. https://doi.org/10.1007/s11306-012-0482-9.

[2] "Blood-transfusion-service-center" https://www.openml.org/d/1464.

[3] D. Bahel, P. Ghosh, A. Sarkar, M. A. Lanham, "Predicting Blood donations using Machine learning techniques", unpublished.

[4] I. T. Jolliffe, "Principal Component Analysis", 2nd ed., New York, Springer-Verlag New York, 2002, pp. 1-27, Available: https://www.springer.com/gp/book/9780387954424.