INTRODUCTION:

This report aims on analyzing the Healthy Flow Dataset collected as a part of flow cytometry analysis. Two objectives are applied in the process of evaluating this Dataset:

1. To recognize the groups from the dataset by making use of unsupervised techniques such that the gating number of the population of cells from the Dataset and the groups determined by this objective are somewhat alike.
2. To predict which data will be allocated to which gates that have already been identified reliably by utilizing Supervised methods.

Each objective mentioned before will use a set of methods that have been taught in this course to achieve each goal. A detailed analysis will be performed for each of the methods in this report. Further, graphs and certain outputs of this dataset will be included as a part of this report.

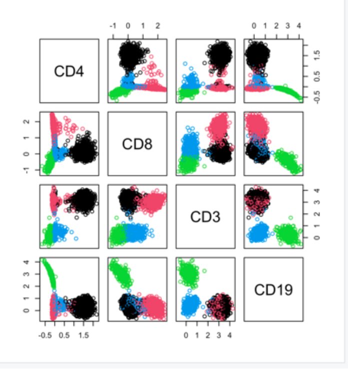
DATASET INTRODUCTION:

The data collected comprises of 19321 populations of lymphocyte cells of a healthy person. Each respective population of cell have unique characteristics compared to the Protein Markers CD3, CD4, CD8, and CD19 which makes the four variables in the dataset. As a part of flow cytometry analysis, population of cells with similar characteristics are assigned to a certain gate number which accounts for the fifth variable in the dataset. The gate numbers are as 1, 2, 3, and 4. The protein markers CD3, CD4, CD8, and CD19 are continuous variables while the gating variable is a categorical variable.

In support of this report, a subset of data has been considered to analyze to understand various matrices and concepts. Therefore, the dataset will include 3864 observations with the same 5 variables. The distribution of each of the variable has been identified using the histograms, boxplots, and density plots. It is observed that CD3, CD8, and CD19 Protein Markers from the Healthy Flow Dataset have outliers and are not normally distributed. [Appendix 1 & 2] On the other hand, CD4 Protein Marker has no outliers but still the data is not normally distributed. Coming to the categorical variable, maximum population of cells have been categorized into gate 1 and the minimum to gate 3. Further, just to check on the correlation which gives the strength and relationship between variables, it is clearly seen that (APPENDIX), CD4 and CD8, CD4 and CD19, CD8 and CD19, CD3 and CD19 are negatively correlated, while a higher correlation has been identified between CD3 and CD4 Protein Markers.

UNSUPERVISED METHODS:

Plotting for the raw data, we can see a plot matrix that gives that there are underlying groups among the first four variables inside the data.

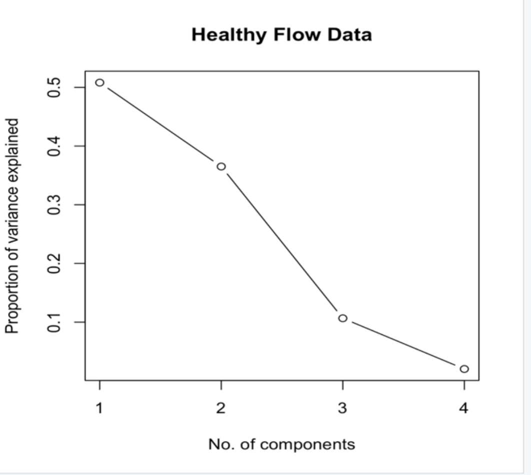


Using just the structure from the dataset with no given labels, Unsupervised learning techniques come into existence for providing the structure or groups in the data. The evaluation can be split into two parts. Firstly, Principal Component Analysis can be used to check different patterns from the data. Second part will focus on subsetting the data using the concepts of Hierarchical Clustering and K-means Clustering.

PRINCIPAL COMPONENT ANALYSIS:

Principal Component Analysis aids in the identifying of the fundamental structure in the data. It, further, finds out the variation and correlation between different variables and provides with a subset of the variables that are not correlated. The focus of this is basically to reduce the dimensions of the dataset and make linear combinations with the native variables which will lead to an ideal set for visualization of the data.

Though the Healthy Flow Dataset do not have a higher number of dimensions, Principal Component Analysis has still been used in this report to help us to know the trends and correlations among the data. Variables other than the Categorical Gate variable are analyzed in the process of principal component Analysis. All the variables are scaled during this process to minimize the dominance of certain variables in a large dataset like this. The principal components are the eigen vectors of the covariance matrix.

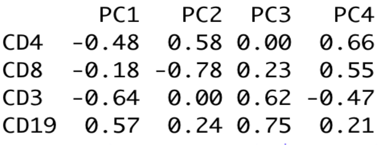


Maximum variation among the Principal Component Analysis has been observed in PC1 accounting to 51% and the other 3 Principal Components sum up to a total variation of 49%. Consequently, standard deviation of 1.42 and 1.20 has been seen from PC1 and PC2 respectively. On further division of the spread among the data, 36% variation has been noticed in PC2, 11% and the least variation with PC4 (2%). The first two components together show 87% of the variation in the data.

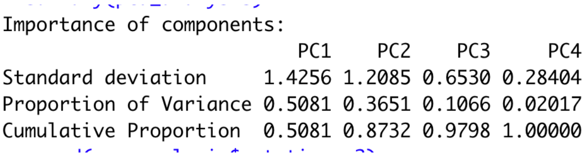
PC1 Interpretation: A high association has been seen with the CD3 Protein Marker. If the CD3 Protein Marker is lower, then CD4 Protein Marker tends to be lower as well, while CD8 also tends to be negatively related but it is not comparatively that low. On Contrast, when the before Protein Markers are higher, the coefficient of CD19 Protein Marker tends to be way higher in comparison. In all, CD3, CD19, and CD4 have a stronger contribution, looking at the loadings of the original variable, in the order when compared to CD8 Protein Marker. To have a higher value of PC1, CD3, CD4 and CD8 must be lower quantities while CD19 must have a higher number.

PC2 Interpretation: The second Principal Component contributes to 36% of the spread among the data with a standard deviation of 1.20. It is perceived that when the value for CD8 Protein Marker in a cell is higher, then CD4 and CD19 Protein Markers tend to have a negative association. Surprisingly, the CD3 Protein has no contribution at all. On the other hand, it is seen that CD8 Protein Marker contributes more to the PC2, then comes CD4. For having a higher value of PC2, CD4 and CD19 Protein Markers must have a higher value while the CD8 must be aimed for low quantity.

The other two principal components are seen as the ones with the noise as they have a lesser variation.

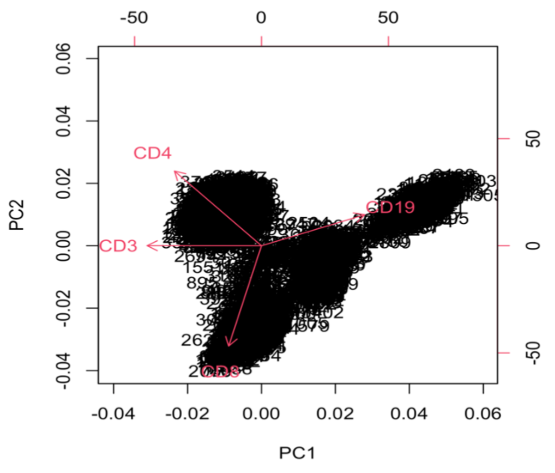


```
       PC1   PC2  PC3   PC4
CD4  -0.48  0.58 0.00  0.66
CD8  -0.18 -0.78 0.23  0.55
CD3  -0.64  0.00 0.62 -0.47
CD19  0.57  0.24 0.75  0.21
```



```
Importance of components:
                          PC1    PC2    PC3     PC4
Standard deviation     1.4256 1.2085 0.6530 0.28404
Proportion of Variance 0.5081 0.3651 0.1066 0.02017
Cumulative Proportion  0.5081 0.8732 0.9798 1.00000
```

A *biplot* has been plotted which gives how the two principal Components are aligned with their observations.t is seen that CD3 contributes most to PC1 while CD8 to PC2. CD4 tends to contribute more or less equally to both PC2 and PC1 but CD19 is more inclined parallelly to PC1. Also, more variability is seen in CD8 and CD19 by the length of the vectors.

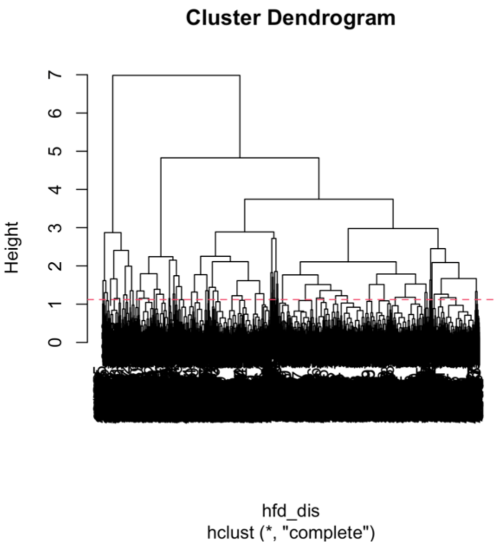


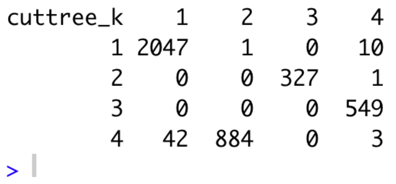
Principal Component Analysis has been seen as an essential tool for dimension reduction from the Dataset. Particularly for this dataset, though trends and association between variables have been noticed clearly, there is no necessity of reducing the variables is seen as most of variation can be observed when considering having three dimensions. Therefore, further analysis of this dataset will be continued without any dimension reduction.

HIERARCHIAL CLUSTERING:

Hierarchical Clustering is a tree-like structure that mainly focusses on identifying different groups in the data based on the distance which is called the dissimilarity among the data points. The bottom of the tree has higher number of groups with less dissimilarity while the top has the opposite with a greater distance. Each branch in the tree like structure which is called as Dendrogram gives the distance between the groups formed. The aim of this is to put all the datapoints that have similar characteristics into one group and maintain dissimilarity between other groups.

All the variables except the Gate variable have been included during the implementation of this method. For merging the data points to see the dissimilarity between them, complete linkage, and maximum dissimilarity, which checks on the farthest distance between two points from two different clusters, has been used here. On note of that, complete linkage gives a good internal similarity between the clusters. Euclidean Distance, which measures a straight-line distance between two points in the space, is used here. Just on a further note, different methods like average linkage, single linkage with other similarities have been checked for creating a *dendrogram* and analyzing the clusters based on this method, single linkage and average linkage with different distance methods is not much of a use as the clusters between heights 1 and 2 are seen as messier. Also, *Rand Index* has been calculated between different method types for checking how each one is like the other. The height for each of the clusters is the dissimilarity between each clustered groups using the Euclidean Distance. Further, to check how many clusters are best choice for the observations from the dataset, a recommended cut off height with formula, $\bar{h}$+3s has been used where $\bar{h}$ is the mean of heights and s is the standard deviation of the heights from the dendrogram. The cut off height form the previous mentioned is 1.1. On contrast, it is seen that, the recommended cut off height doesn't give a proper one for this Healthy flow dataset. The number of clusters the line gives is large which is not a significant way. But a pair wise plot with each of the four Protein markers has been plotted by defining an arbitrary value of number of groups interested as 4. [fig 1] For the same, a table has been created against the gate variable to check how the clusters are grouped using hierarchical Clustering and the gate variable (which we weren't supposed to be using in the unsupervised methods).(table 1) On note of that, it is seen that when the number of groups are taken arbitrarily as 4, only the first cluster satisfies having a greater number of observations correctly. In addition, a pair wise plot has also been formed taking the recommended height which gave a greater number of Hierarchical clusters with fewer observations. [fig 2]





```
cuttree_k    1    2    3    4
        1 2047    1    0   10
        2    0    0  327    1
        3    0    0    0  549
        4   42  884    0    3
>
```

Table 1

fig 1

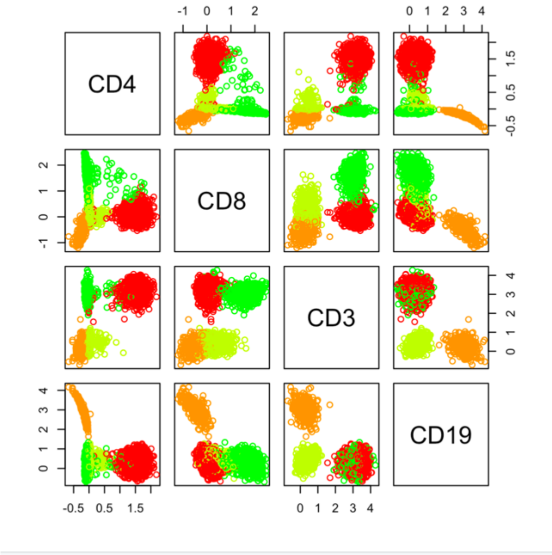
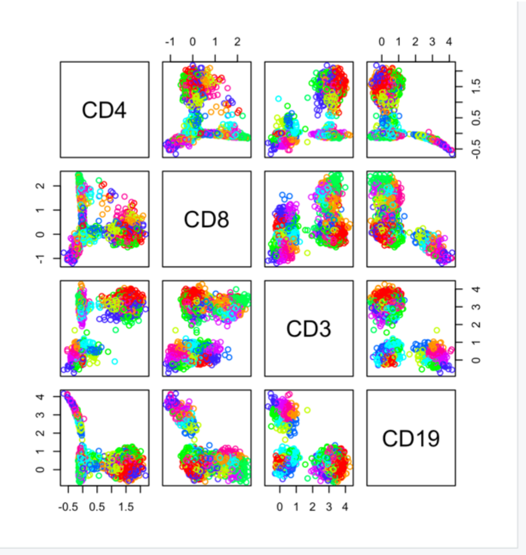


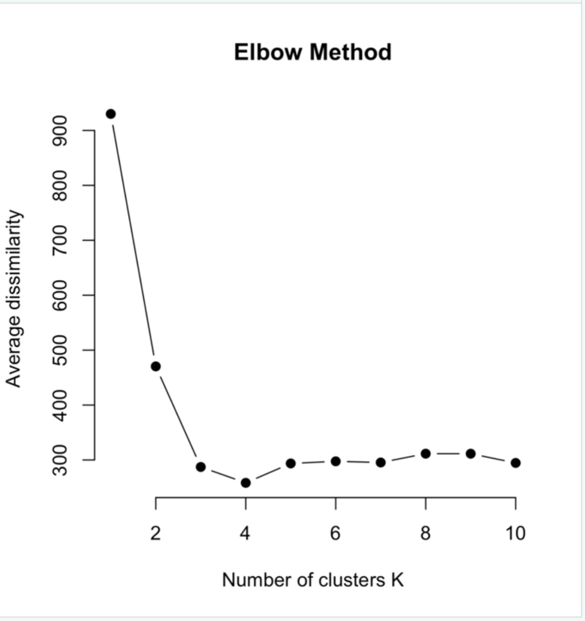
fig 2



K-MEANS CLUSTERING:

K-means Clustering is the third type of method that is used to identify groups in the dataset without labels provided. The main principle of this method is used to divide the data into k distinct groups and make sure that the observations in each group are similar while there is a dissimilarity between observations from different other groups. This is based on identifying the centers from the clusters. Observations from the dataset will be grouped to their closest centroid using Euclidean Distance. This method is an iterative process until it finds the best solution. For the implementation of k-means or k-medoids clustering we must know the feasible number of clusters the dataset can have. Just to mention, Gower distance has also been used to check the clustering with k-medoids.

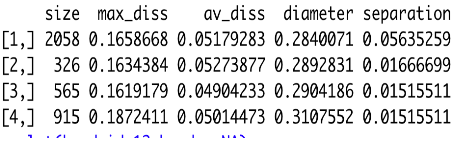
Coming to this Healthy Flow Dataset, to continue to form the clusters using k-medoids and k-means, the elbow method has been used to determine the number of clusters are better for this data. The k values are taken for evaluation here are from 1 to 10. In the elbow method, we aim for a kink in the line graph which tells us that any greater number of clusters formed will not make much of a difference considering the dissimilarity. It is identified that *k = 4* number of clusters is a better choice. Subsequently, after implementation of clustering with 4 number of clusters, we can analyze the following from using the Gower distance:



2058 observations are observed in the first cluster. The average dissimilarity between the observations in the cluster is 51%. The first cluster takes the highest number of observations. Coming to cluster 2, there are 326 observations with an average dissimilarity of 52%. Cluster 3 includes 565 observations of cells and has an average dissimilarity of 49%. Lastly, the third cluster contains 915 observations accounting for 50% average dissimilarity. The *silhouette width* for the first three clusters is above 0.71 indicating that those clusters somewhat fit better. [fig 5] On further analysis with the gating variables, it is observed that the gating number 2 has been fit properly with all the protein makers other than CD4 containing few outliers in each. [fig 3]

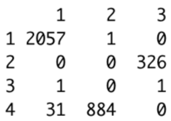
Forming a table with the clusters formed and the gating variable, the gate 1 has been split into the first and fourth cluster with 2057 observations in the first cluster. Most of the observations of gating number 2 is identified in the cluster 4. [fig 4] Thirdly, all the observations except one of gating number 3 are fitte been preciselv fit into the cluster 3. [fig 6]

fig 3



```
     size  max_diss     av_diss  diameter separation
[1,] 2058 0.1658668 0.05179283 0.2840071 0.05635259
[2,]  326 0.1634384 0.05273877 0.2892831 0.01666699
[3,]  565 0.1619179 0.04904233 0.2904186 0.01515511
[4,]  915 0.1872411 0.05014473 0.3107552 0.01515511
```

fig 4



```
     1    2    3
1 2057    1    0
2    0    0  326
3    1    0    1
4   31  884    0
```

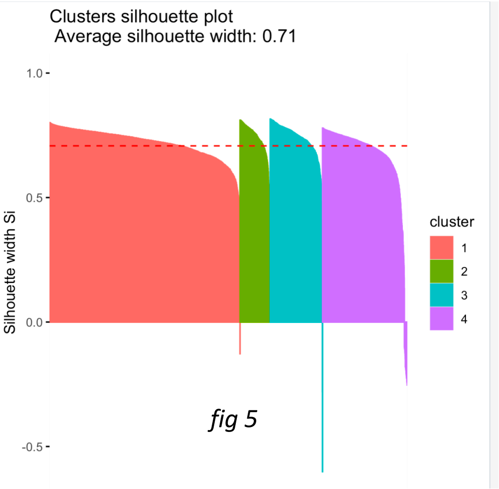


fig 5

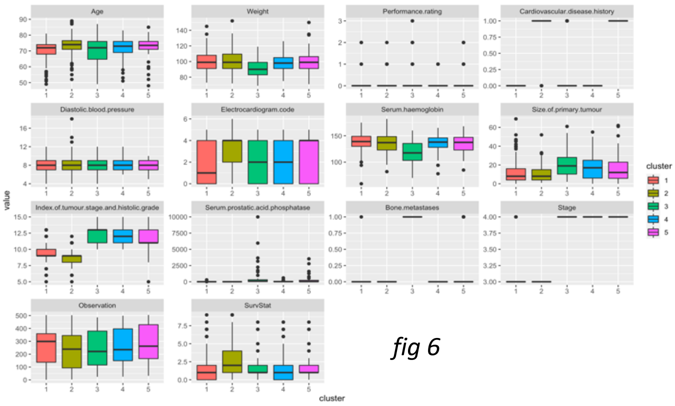


fig 6

Further, k-means has been used to form different clusters using the same k value that tends to be feasible from the elbow method. When a table is formed against the gating variable from the dataset, most of the gating variables seem to be fitted into each of the clusters properly other than the gate number 1 which is split between cluster 1 and 4 with majority in the cluster 1. A plot with the four clusters is formed against the Principal Components 1 and 2. [fig 7] The average silhouette width has been checked to assess the relative compactness of the cluster solution which came out to be 0.9. Clusters 1, 2, and 4 have higher silhouette width than the average silhouette width indicating they are fit better while the cluster 3 has a silhouette width less. [fig 9] On a further analysis, a centroid of the four centers and using the K-means clustering has been formed where the observations of the protein markers are kind of tightly packed around the centroids. [fig 8]

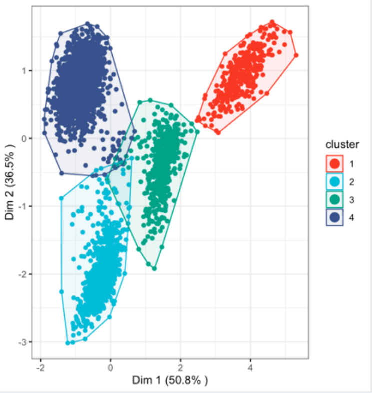


fig 7

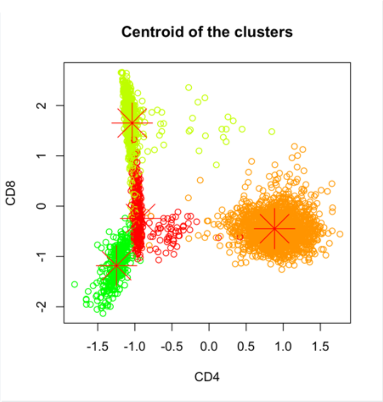


fig 8

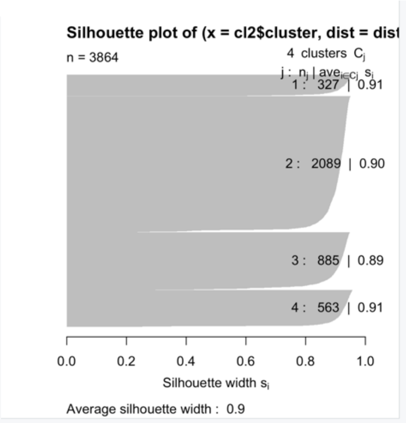


fig 9

CONCLUSION:

Starting with the PCA analysis, when compared to the gating variables, it is seen that there is some good separation between the groups identified in accordance with the Principal Components 1 and 2 while the other observations are all over. [Appendix 3] Coming to the hierarchical clustering when the number of groups to be formed are taken arbitrarily as 4, 4 groups can be clearly seen with the CD3 and CD8 Protein Markers while taking the recommended height to form a plot gives out just a random scatter plot between each of the protein markers. K-medoids and K-means have given a similar kind of output with gate variables 3 and 4 fit almost correctly into each separate cluster when compared to the gating variable by using the number of groups as 4. A further implementation gave a plot how each variable has been put into different clusters using the gating variables.

SUPERVISED METHODS:

Supervised learning methods aids in accurately predicting the data that have already been labelled into some groups. This part of the report deals with the identification of how precisely the gates have been grouped from the data. This is when the supervised learning methods contrast with unsupervised learning methods. For this, the data must have a training subset of data where the prediction refines based on that subset by using the target variable. The data will be classified based on the training subset and the target variable, here it will be the gate variable.

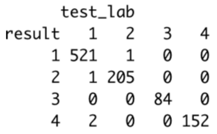
To predict the data that is predicted to the identified gates accurately, this report will delve into the concepts of K-nearest neighbors, Quadratic Discriminant Analysis.

K-NEAREST NEIGHBORS:

This method basically considers the Euclidean distance between the observations and on taking the k closest neighbors, the unlabeled data will be then classified based on its closest neighbors' class.

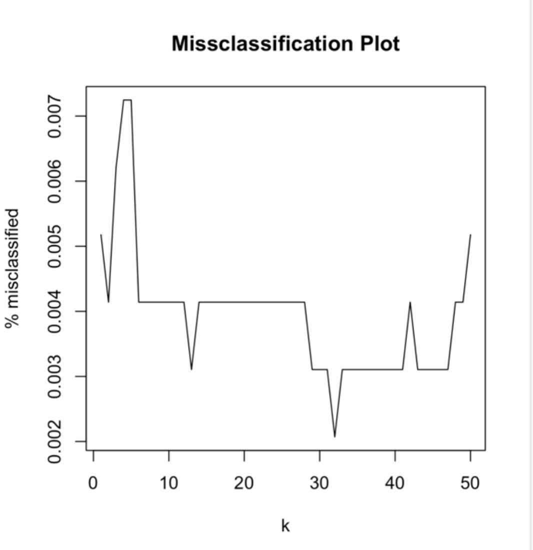
To start with, choosing k that will be giving the good predictive performance for this method of classification is the major task. For the same, splitting the data into Training set, where the labels of observations from this subset will be used for classifying the unlabeled observations, Test Set, within this subset it checks on misclassification of the data using the training subset and target label, and Validation set, checks on how well the k-nearest neighbors predict on new data points. Consequently, this method checks how the data has been classified incorrectly with a set of k closest neighbors and checks for a low misclassification rate.

On note of this report, the data is split into 50% as the *training* subset of data, 25% of the data as the *test* subset and 25% as the *validation* subset of the data after all the protein markers are scaled for the analysis. The training data subset will be compared with the true classification (the gate variable) for the training dataset which will then be used to classify the test and the validation subsets.



```
      test_lab
result   1   2   3   4
     1 521   1   0   0
     2   1 205   0   0
     3   0   0  84   0
     4   2   0   0 152
```

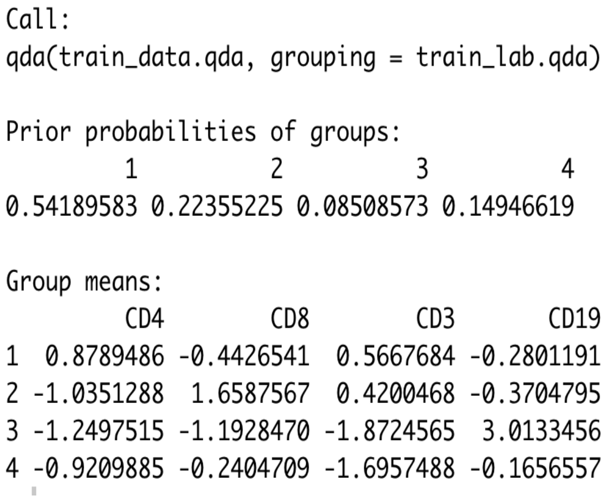
Firstly, the training and the testing subsets of the data are passed and checked with the K-nearest neighbors taking the closest neighbors arbitrarily as 4. For this classification of the healthy flow data, 962 observations, around 0.4%, from the test subset of the data have been *classified correctly* according to the gate variable.



The main part of the k-nearest neighbors is to check the misclassification rate among the groups. Values of k from 1 to 50 have been taken to check the optimal value for the closest neighbors. The optimal value is chosen when the misclassification rate is lower than the other. It is observed that when k=32 the *misclassification* rate is at the minimum (0.2%). From the previous statement, 99.8% of the test data is classified correctly using the gate variable. Finally, the validation subset of the data is passed on with 32 closest neighbors and it gives those 963 observations have been classified correctly with one observation from group 1 being in the gate number 2 and 2 observations of the gate variable 1 is seen in group 2 leaving the misfit percentage to be approximately 0.3%. Surprisingly, when the misclassification rate has been checked for k values in the range of 1 to 10, it leaves out the minimal misclassification rate, approx. 0.35%, against the 4 nearest neighbors for this dataset.

QUADRATIC DISCRIMINANT ANALYSIS:

This is the second supervised learning method that will be used in this report to check the accuracy of the prediction of the healthy flow dataset into gate variables. This technique uses the labelled data to classify the unlabeled observations. The main contrast seen with the first type of classification method, k-nearest neighbors, is that the QDA takes the distribution of the variables into account which leads to the numerical values about how uncertain the data is. In this method, the main assumption taken into consideration is the observation from a particular group follows a Multivariate Normal Distribution with mean mu which presents the scatter of the observations in the plot in its location and covariance sigma is the reason for the ellipsoidal shape. The new observations are then classified based on the posterior probability (aim for the largest). Further, prior probabilities, probability of an observation belonging to a class, are assumed to be either known or estimated. For this dataset, we take that the covariances for each of the groups are different which leaves with the QDA rather than the LDA method.



```
Call:
qda(train_data.qda, grouping = train_lab.qda)

Prior probabilities of groups:
         1          2          3          4 
0.54189583 0.22355225 0.08508573 0.14946619 

Group means:
         CD4        CD8        CD3       CD19
1  0.8789486 -0.4426541  0.5667684 -0.2801191
2 -1.0351288  1.6587567  0.4200468 -0.3704795
3 -1.2497515 -1.1928470 -1.8724565  3.0133456
4 -0.9209885 -0.2404709 -1.6957488 -0.1656557
```

For performing this method analysis on the Healthy Flow Dataset, the data is split into 80% subset of training data and 20% subset of testing data. The analysis shows that the gate number 1 has the highest prior proportion in the training subset of the data and the least being the gates 3. Further, estimated means of each of the protein markers against each gate has been observed. Further, the prediction of the QDA model on the testing subset is performed giving the classification of the data.

CONCLUSION:

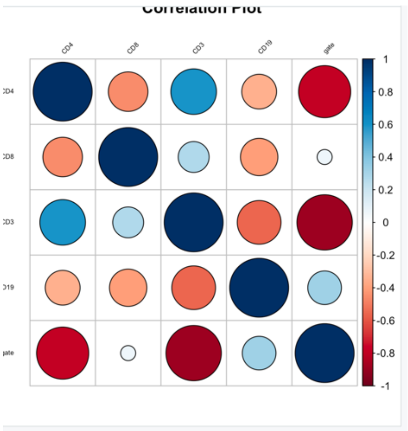
To conclude, an analysis of the Healthy flow dataset has been performed with the supervised learning methods using the split of the data. It came to notice that the K Nearest Neighbors have accurately predicted 99.7% of Healthy Flow Dataset correctly into groups taking the closest nearest neighbors as 32.

REFERENCES:

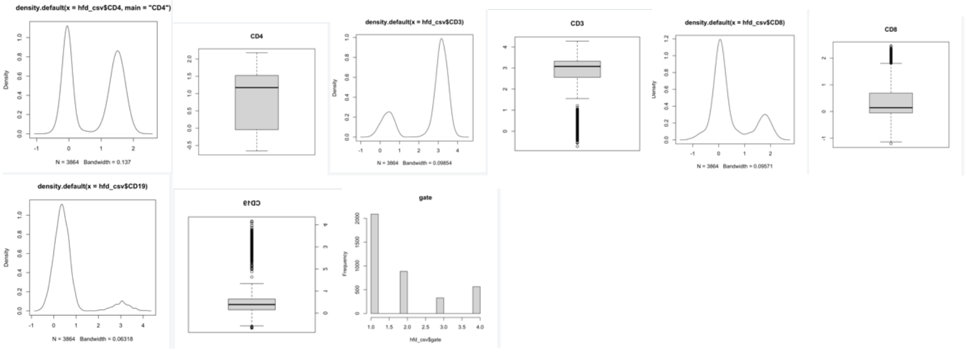
1. https://www.geo.fu-berlin.de/en/v/soga/Geodata-analysis/Principal-Component-Analysis/principal-components-basics/Interpretation-and-visualization/index.html
2. https://www.datanovia.com/en/blog/k-means-clustering-visualization-in-r-step-by-step-guide/

APPENDIX:

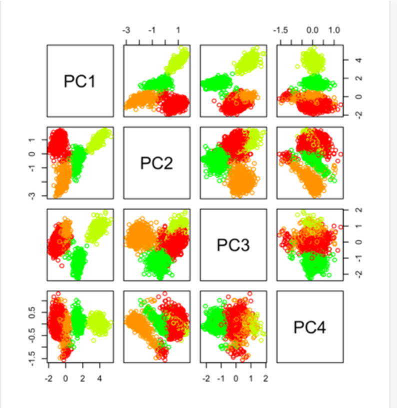
1. Correlation Plot



2. Distribution Plots:



3. Principal Component Analysis:


PC1
PC2
PC3
PC4