

QUALITY OF CARE FOR CHRONIC DISEASE MANAGEMENT

BITS ZG628T: Dissertation

by

Vaibhav Gaikwad

2018HT12597

Dissertation work carried out at

Philips VitalHealth Pvt. Ltd., Mumbai, India



**BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE
PILANI (RAJASTHAN)**

April 2020

QUALITY OF CARE FOR CHRONIC DISEASE MANAGEMENT

BITS ZG628T: Dissertation

by

Vaibhav Gaikwad

2018HT12597

Dissertation work carried out at

Philips VitalHealth Pvt. Ltd., Mumbai, India

Submitted in partial fulfillment of M.Tech. Software Systems degree programme

Under the Supervision of
Aris van Dijk, Philips VitalHealth B.V., Ede, Netherlands




**BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE
PILANI (RAJASTHAN)**

April 2020

CERTIFICATE

This is to certify that the Dissertation entitled Quality of Care for Chronic Disease Management and submitted by Vaibhav Gaikwad having ID-No. 2018HT12597 for the partial fulfilment of the requirements of M.Tech. Software Systems degree of BITS, embodies the bonafide work done by him/her under my supervision.



Signature of the Supervisor

Place: Ede, Netherlands

Aris van Dijk

Date: 20th April 2020

Software Architect, Philips VitalHealth B.V.,

Ede, Netherlands

Birla Institute of Technology & Science, Pilani
Work-Integrated Learning Programmes Division
Second Semester 2019-2020
BITS ZG628T: Dissertation

ABSTRACT

BITS ID No.	: 2018HT12597
NAME OF THE STUDENT	: Vaibhav Gaikwad
EMAIL ADDRESS	: vaibhav.gaikwad@philips.com
STUDENT'S EMPLOYING ORGANIZATION & LOCATION	: Philips VitalHealth Pvt. Ltd., Mumbai, India
SUPERVISOR'S NAME	: Aris van Dijk
SUPERVISOR'S EMPLOYING ORGANIZATION & LOCATION	: Philips VitalHealth B.V. Ede, Netherlands
SUPERVISOR'S EMAIL ADDRESS	: aris.van.dijk@philips.com
DISSERTATION TITLE	: Quality of Care for Chronic Disease Management

Abstract

In the health-care domain Quality of Care (QoC) signifies the effectiveness of the disease management software solution. This paper illustrates the approach to learn about the QoC for the diabetic population. The research utilizes ideas from Data Mining and Machine Learning in identification of data attributes that have significance. The optimization using Random forest over Silhouette method is discussed. It highlights that the risk groups are not well-separated and default clustering does not work. The K-means method adoption in unique way helped to extract patterns from the data for risk-based group identification. The identification of high-risk patients has a

significant value in Diabetes applications because these patients need critical care at all times. QoC depends mainly on the health management of high-risk patients. The result of this study proposes a method for learning about the QoC based clustering of data and identification of the risk-based patient groups. The idea is useful for the improvement of the healthcare protocol. The study opens new avenues for research on methods for the assessment of QoC.



Signature of Student

Date: - 20th April 2020



Signature of Supervisor

Date: - 20th April 2020

Acknowledgements

I am grateful to numerous local and global peers who have contributed towards shaping this project. At the outset, I would like to express my sincere thanks to Mr. Marcel Verhoeve for providing me with the opportunity and facilitating the required permissions needed for this course.

Sincere thanks to Mr. Aris van Dijk for taking the responsibility of being the Mentor and Supervisor. I am grateful for his timely reviews and valuable suggestions, which has helped me a lot for timely completion. I would like to extend my thanks to Mr. Balvvant Singh Bist for being the Additional Examiner.

I am grateful to the Faculty Mentor Dr. Lov Kumar from BITS Pilani, for reviewing the progress and providing valuable feedback.

A special thanks to Ms. Dhvani Shah, a senior from BITS Pilani, for additional guidance.

I am also thankful to Philips VitalHealth Pvt. Ltd. for giving me this opportunity and also making available the resources for this project. I must acknowledge the academic resources that I have acquired from BITS Pilani. Finally, I would like to thank my family for the all the support.

Vaibhav Gaikwad

April, 2020

Table of Contents

1. Introduction.....	1
1.1. Problem statement.....	2
1.2. Proposed solution.....	3
1.3. Objectives.....	3
2. Literature Review.....	5
3. Work on Quality of Care.....	6
3.1. Ideation and design.....	6
3.2. Data collection and analysis.....	9
3.3. Searching for risk groups	14
3.3.1. Elbow analysis	17
3.3.2. Dimensionality reduction.....	18
3.3.3. Silhouette analysis	18
3.3.4. Investigations for risk-based groups formation	20
3.3.5. Attribute selection using Random Forest approach	23
3.3.6. Clustering Confidence Score (CCS)	28
3.4. Design for Quality of Care (QoC).....	29
3.4.1. Approach for Immutable Dataset.....	30
4. Conclusion and Recommendations.....	32
4.1. Conclusion.....	32

4.2. Recommendations	32
5. Future Scope	33
6. Appendices.....	34
6.1. Code: K-means with gap-stat and DBSCAN with kNN	34
6.2. Code: Correlations using Corrplot	35
6.3. Code: K-means with Silhouette.....	36
6.4. Code: Random forest with varImpPlot.....	38
7. References	40
8. Glossary	42

List of Figures

Figure 1: System block diagram	7
Figure 2: Analytics component flow diagram	8
Figure 3: Box plot of raw data	10
Figure 4: Histogram of Insulin.....	11
Figure 5: Correlations in data attributes.....	11
Figure 6: Visualization of correlations in unfiltered data	13
Figure 7: Visualization of correlations in filtered data	14
Figure 8: Gap-stat analysis for K-means	15
Figure 9: kNN analysis for DBSCAN	16
Figure 10: Elbow analysis.....	17
Figure 11: Silhouette analysis for various K values	19
Figure 12: Clustering for risk-based groups formation.....	21
Figure 13: Silhouette for Glucose analysis	22
Figure 14: Random Forest model review.....	24
Figure 15: Attribute selection from Random Forest model	25
Figure 16: Visualization of clusters based on Random Forest	26
Figure 17: K-means Silhouette plot based on Random Forest	27
Figure 18: Visualising trend in clusters size	31

List of Tables

Table 1: Data attributes and descriptions.....	9
Table 2: Observed ranges across clusters	22
Table 3: Random Forest based ranges across clusters	27
Table 4: Risk-based cluster sizes for a year.....	30

1. Introduction

Philips VitalHealth develops a variety of products and custom solutions that helps in managing the care of chronic diseases like Diabetes, Chronic Obstructive Pulmonary Disease (COPD), Asthma, etc. It has developed a platform which helps in rapid application development for any healthcare solution as required, within a short time. The diabetes solution is widely adopted in many parts of the world by different organizations. The application development has focused on the medical workflows and implementation of the care protocol. The data stored in these applications is valuable from a data analysis point-of-view. A thoughtful review of the good-to-have features for the application, has brought to light the topic of data analytics. It is one of the key Unique Selling Point (USP) for the business and its customers. There are various aspects in the domain of analytics; one such point is Quality of Care (QoC), which is the primary focus of this dissertation work.

Healthcare Analytics is the buzz word. The area of work tries to deal with a simple yet challenging topic of healthcare applications, i.e., measuring the QoC. Most of the applications that are built around the protocol to manage chronic diseases fall short of measuring the meaningfulness. Many developed countries have understood the importance of measuring the QoC. A perfect example to support the previous statement is the adoption of the Meaningful Use [1] certification in the United States of America. The certification policy is well documented on by the Centers for Disease Control and Prevention (CDC). The understanding of QoC is vital to the certification policy as it refers to accountable care.

The concept of identification of risk-based groups is an essential aspect of the measurement of QoC. There may be different approaches to identify risk-based groups formations, and one of them is data mining.

Data Mining helps to identify the attributes related to the patient that can be used to learn about the QoC in any chronic disease application. It can also be helpful to compare data attributes between chronic diseases and also to find some co-relations across different chronic diseases. The goal in the current scope is to identify the data attributes for Diabetes Type 2, and then learn about the co-relations among these attributes. It also tries to identify the risk categories in a given population (i.e., high, moderate, and low-risk groups).

Once the groups are observed, the high-risk group can be managed separately, to make sure that the best care is provided to these patients. This is the critical factor for improving QoC.

Machine learning can help us to build a model that can help to predict the meaningful use of an application developed for Diabetes Type 2.

1.1. Problem statement

The Diabetes application lacks data analytics capability. There are existing integrations possibilities to extract the correct set of data into other analytics applications. The current way of performing analytics is not reusable and also not economical. Customers do not have a way to measure the QoC in their installations easily.

This work provides an idea of analysing the QoC for Diabetes solution. The challenges are:

- a. To identify risk groups within the diabetic population.
- b. To deduce data attributes which are promising for QoC.

1.2. Proposed solution

Identification of the risk groups within the Diabetic population is possible with effective implementation of the K-means algorithm (i.e., clustering). In the process, it is essential to identify which data attributes provide better clustering.

The data attributes that can help in learning about the QoC depends on the previous step. The data attributes which can efficiently group the diabetic population into various risk categories are the ones that can help to understand the trend in the patient's health. The identified clusters can be monitored over time to assess the QoC.

1.3. Objectives

The primary goal is to research about the possibilities of using Data Mining and Machine Learning to understand the QoC for a given application based on Diabetes Type 2 protocol. This work is about developing a proposal and testing its feasibility. Software development is taken in the next phase and beyond the scope of this work.

This dissertation work focuses on two topics:

- Data Mining
 - Identification of attribute related to QoC for a Diabetes 2 application.
 - Categorization of patients is done into high, moderate, and low-risk profiles because the high-risk population needs effective and timely care compared to others.
- Machine Learning
 - Design a model on QoC that can then be applied to different population data for the same chronic disease and predict the QoC for newly registered patients.

2. Literature Review

Data Mining in health-care:

1. Elma Kolçe *et al*, 2012, [2] listed down the methods used in health-care domain for data mining. It identifies and evaluates the commonly used data mining methods. The algorithms evaluated are Decision Trees C4.5 and C5, Support Vector Machine (SVM), Artificial neural networks (ANNs) and their Multilayer Perceptron model, Naïve Bayes, Logistic Regression, Genetic Algorithms (GAs) / Evolutionary Programming (EP), Fuzzy Rules. The study shows that it is difficult to recommend a single algorithm as the most suitable for the any specific study on diseases.

Random Forest for attribute selection:

1. M.J. Hallett *et al*, 2014, [3] studied oral health issues in adults due to quality of life. Random forest method was used to find out the reasons for tooth loss, using set of clinical and genetic data. Amalgamation procedure was applied to limit the number of meaningful prognostic groups. This method was able to remove the bias of the traditional methods used by other researchers. It was able to give high accuracy and efficiency results.
2. Bharatendra Rai, 2017, [4] used Boruta algorithm in random forest model for predictive analysis of 79 various qualitative and quantitative features of people while buying a new house. This method was able to select confirmed specific features thereby minimizing data to 49. The efficiency was studied using coefficient of determination (r-square) and root mean square error (rsme).

There was not enough information present related to data mining and Quality of Care. Above papers helped to perform the feature selection in a better way as compared to the traditional ways of brute-forcing the attributes or correlation based ranking methods. K-means was a good choice to understand effectiveness of unsupervised learning to give knowledge regarding the risk-based groups and then validate it with the real-world cases.

3. Work on Quality of Care

The division of work is listed below:

- a. Ideation and design of the system
- b. Data collection and analysis
- c. Analysis of risk groups
- d. Design for learning on QoC

3.1. Ideation and design

Quality of Care is the measurement of the effectiveness of a healthcare system. A few ideas were brainstormed within a closed group of people to understand the correct approach. The ideas are listed below:

- a. Summarizing the patient's feedback gives information about the QoC. The adoption of such a method does not seem very accurate and feedback is always optional.
- b. Trend analysis of each patient's health can also provide details on the QoC. This method is time-consuming and does not generate insights on the population level.
- c. Comparison of data from previously used applications which have proved better QoC. This is difficult to achieve due to the dependency on finding a trustable existing system.

The discussion sessions on the best tools required for this work was done on the available knowledge and skillsets in the organization. The outcomes of those sessions are listed below:

- a. RStudio [5] was considered for development.
- b. Philips VitalHealth Integrator component acts as a data broker between the main application and the data analytics component.
- c. Data format was Comma Separated Values (CSV) to keep it simple for consumption by other software components.
- d. Data from live applications cannot be used in this work due to data privacy contracts with customers.
- e. Hence for proof-of-concept PIMA dataset [6] was taken from kaggle.com.

Concept diagram of the system

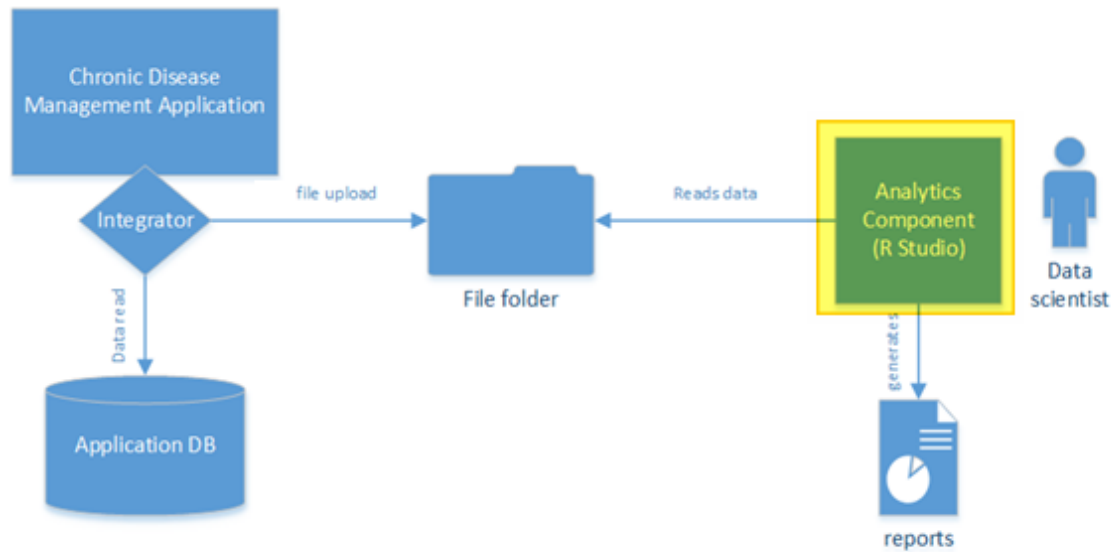


Figure 1: System block diagram

- a. Chronic Disease Management (CDM) applications have an existing dataset for the current analysis.
- b. The integrator component of CDM is responsible for fetching the data from the database and generating CSV files from it, depending on the required data schema.

- c. Data from the application is pulled and stored to CSV file for all patients for the latest records.
- d. The file folder is a shared resource on the network, so that other applications can access and consume it. Data should not contain any Protected Health Information (PHI) [7] attributes in regards to data privacy as governed by the Health Insurance Portability and Accountability Act (HIPAA) regulations.
- e. Analytics application uses the CSV file and performs the required operations to produce reports based on the work that is described in this paper.
- f. If needed, the reports can be pulled back into the application for care providers, but that part is kept out of the scope for now, as the current work needs to undergo a complete development.

The current emphasis is on the modelling of the Analytics Component as highlighted in the above figure (Figure 1).

Following diagram represents the overall work done in the current phase

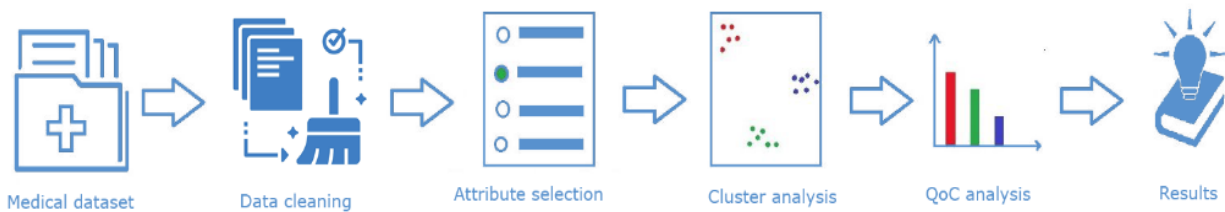


Figure 2: Analytics component flow diagram

Majority of the time was spent on data cleaning, attribute selection and cluster analysis as these are the main processes which were iteratively done to improve on the results and efficiency.

Results are the reports generated from the QoC. The main system development and implementation is out of the scope, but a general idea is discussed.

3.2. Data collection and analysis

Data required for the Diabetes 2 was available on kaggle.com. It was downloaded and used in offline mode as a locally stored CSV file for faster access and reuse.

The dataset contained 8 attributes and 768 records for diabetic and non-diabetic patients.

Data attribute	Description
<i>Pregnancies</i>	Number of times pregnant
<i>Glucose</i>	Plasma glucose concentration (2 hours) oral glucose tolerance
<i>BloodPressure</i>	Diastolic blood pressure (mm Hg)
<i>SkinThickness</i>	Triceps skinfold thickness (mm)
<i>Insulin</i>	2-hour serum insulin (mu U/ml)
<i>DiabetesPedigreeFunction</i>	Score for likelihood of diabetes based on family history
<i>BMI</i>	Body mass index (weight in kg / (height in m) ^2)
<i>Age</i>	Age for a person in years
<i>Outcome</i>	Class variable (0 or 1) whether or not a patient has diabetes

Table 1: Data attributes and descriptions

Data related to *Pregnancies* attribute was not used because the study was not gender-specific.

Analysis of outliers, noise, and missing values was carried on the data.

Box-plot analysis was done for outlier [8] detection, in which it was observed that *Insulin* data showed a higher number of outliers (Figure 3). Further analysis was done using the histogram to find how the data was scattered, and it was observed that the data was skewed. It was clear that most values were 0 (zero). (Figure 4)

The assumption was made that the *Insulin* test was not conducted for many patients, but these records were taken as valid due to proper *Glucose* values. The data was later on filtered to separate out the Diabetic patients so as to understand the correct risk-group formations. This was done by considering records for Outcome = 1 (i.e. Diabetic patient)

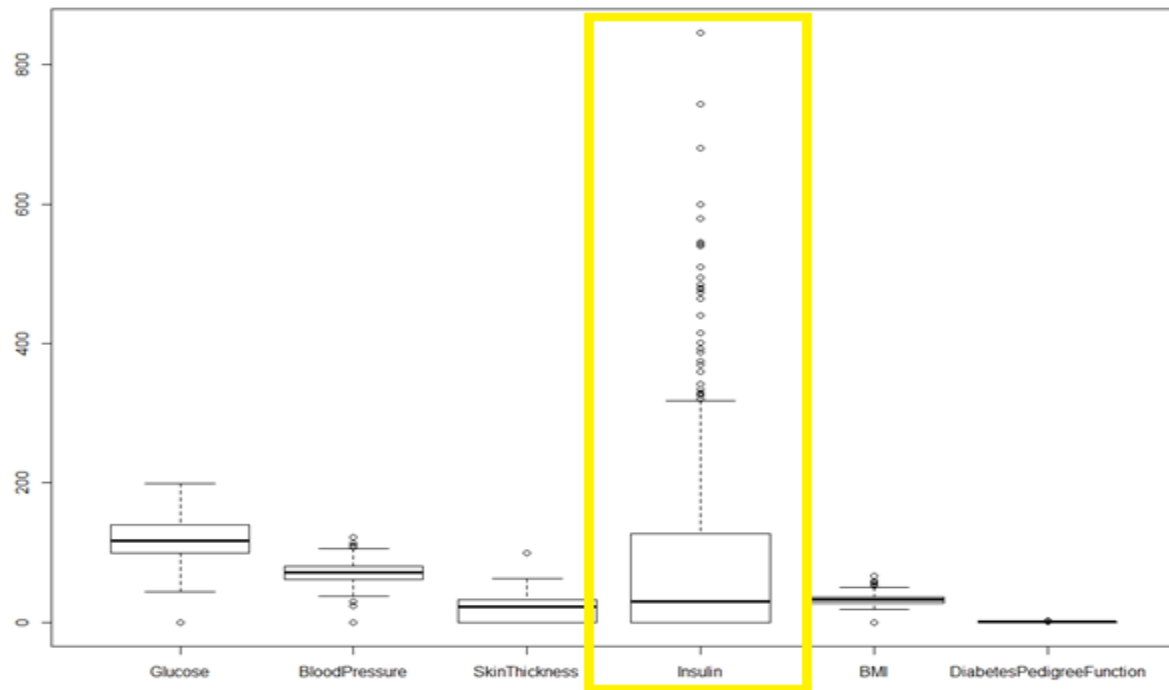


Figure 3: Box plot of raw data

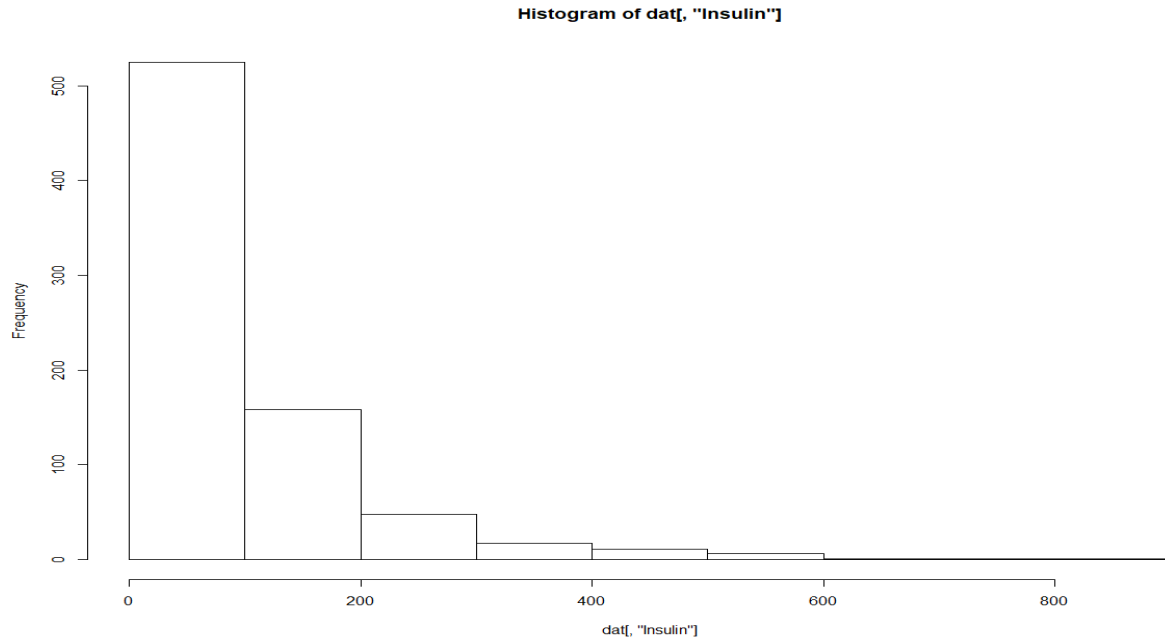


Figure 4: Histogram of Insulin

The cleaning process involved the removal of records with 0 values for *BloodPressure*, *Glucose*, and *BMI* (i.e all three were 0). The dataset was reduced to 249 records after the cleaning process. Data had various ranges, so normalization was considered at the initial stage but was skipped afterwards to avoid side confusion while comparing ranges.

At first, the correlation between the data attributes was analysed. The initial program found the correlations and also stored these to the disk as a CSV file.

	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
Glucose	1	0.152589587	0.057327891	0.33135711	0.221071069	0.1373373	0.26351432
BloodPressure	0.152589587	1	0.207370538	0.088933378	0.281805289	0.041264948	0.239527946
SkinThickness	0.057327891	0.207370538	1	0.43678257	0.392573204	0.183927573	-0.113970262
Insulin	0.33135711	0.088933378	0.43678257	1	0.197859056	0.185070929	-0.042162955
BMI	0.221071069	0.281805289	0.392573204	0.197859056	1	0.140646953	0.03624187
DiabetesPedigreeFunction	0.1373373	0.041264948	0.183927573	0.185070929	0.140646953	1	0.033561312
Age	0.26351432	0.239527946	-0.113970262	-0.042162955	0.03624187	0.033561312	1

Figure 5: Correlations in data attributes

Pearson correlations (strong means > 0.7) method was used to understand the data correlations. It was observed that the data attribute pairs showed some correlations, but not good enough.

- a. *Glucose* and *Insulin* – this was obvious
- b. *SkinThickness* and *Insulin*
- c. *SkinThickness* and *BMI*

Other pairs that were also interesting for the study are

- a. *BloodPressure* and *BMI* (0.28)
- b. *Glucose* and *Age* (0.26)
- c. *BloodPressure* and *Age* (0.239)

Conclusion was made that data attributes do not have good correlations; hence we cannot skip any attribute as this stage.

In general, it was understood that except the *DiabetesPedigreeFunction* attribute, all other attributes should be considered. The approach was kept iteratively to avoid the use of all data attributes which adds more complexity to the model. A visual representation of these correlations was needed. Corrplot [9] was used for visualizations. Corrplot is a R library and stands for correlation plot. The following representation is the same as table shown previously (Figure 5) but easier to understand for a human eye. The size of the circle in Corrplot indicates the effect of correlation while the colour value indicates the direction of the correlation.

- 1. Bigger the circle means higher correlation
- 2. Blue means positive correlation
- 3. Red means negative correlation

Our focus should be on circles which are blue and big.



Figure 6: Visualization of correlations in unfiltered data

Correlations in filtered data are shown below (Figure 7), pairs with strong correlations are listed below

- Glucose* and *Insulin* – this was obvious
- SkinThickness* and *Insulin*
- SkinThickness* and *BMI*

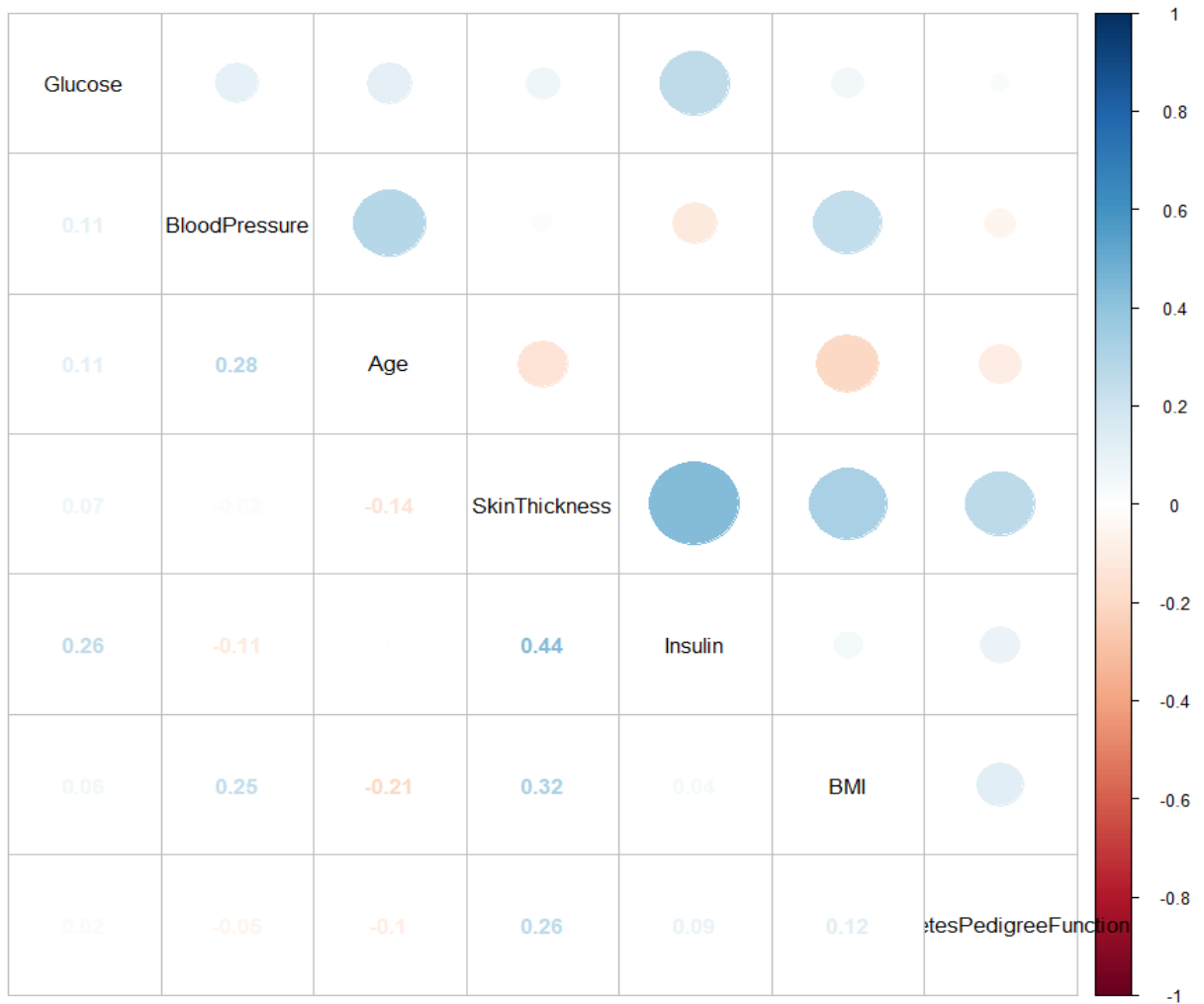


Figure 7: Visualization of correlations in filtered data

3.3. Searching for risk groups

Risk categorization using clustering was a challenging subject. Clustering is an unsupervised learning technique that is used to categorize data, in a way that similar data elements are grouped. Clusters are also called classes. Generated clusters are based on similarities that are intrinsic to the

data attributes and are not known to the person analysing the data. Initially, default K-means and DBSCAN were tested using all data attributes which did not result into any significant results.

Default K-means used gap-stat analysis to find the optimal number of clusters

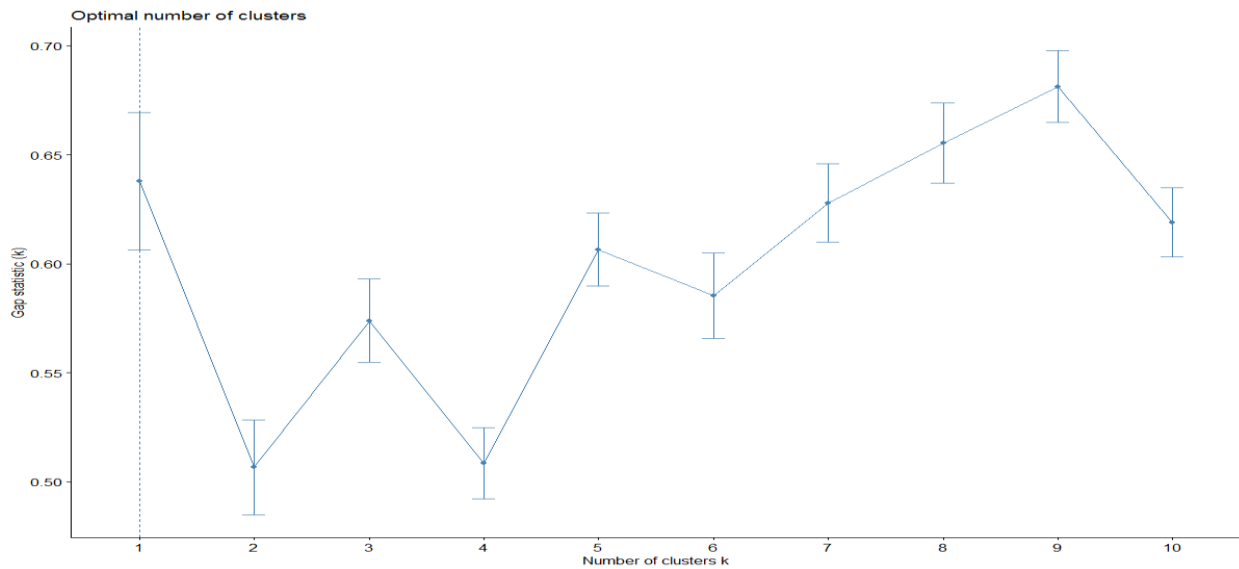


Figure 8: Gap-stat analysis for K-means

Default DBSCAN used kNN to understand the value for “eps” which was close to 60 for $K = 3$ but did not generate any risk-based groups

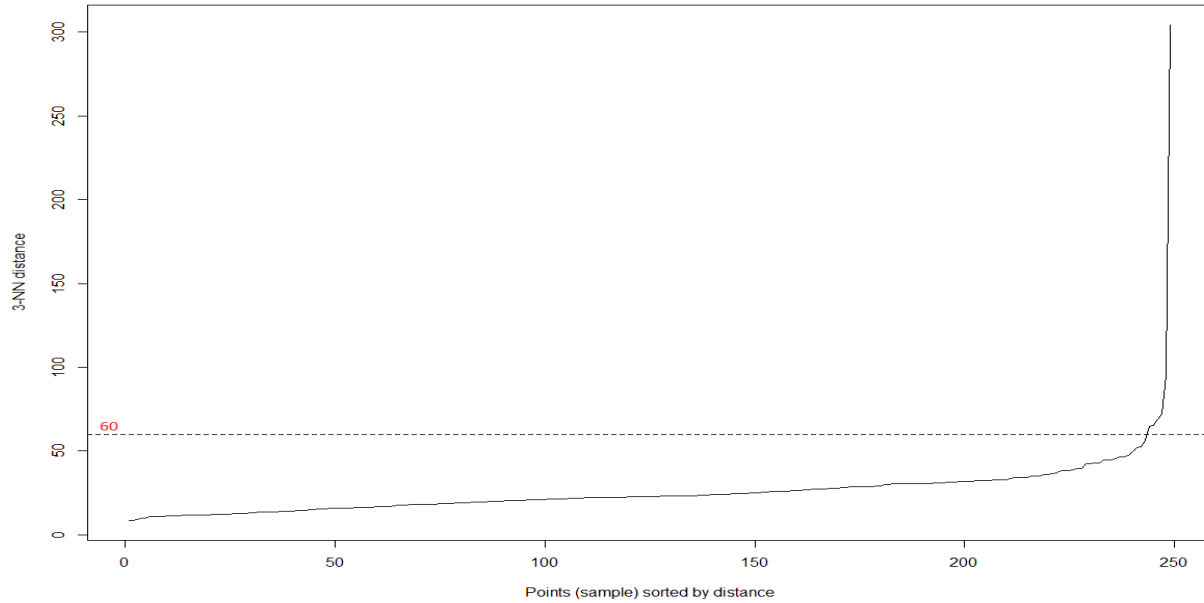


Figure 9: kNN analysis for DBSCAN

In this study, K-means was used, and the approach was slightly tweaked to custom fit the philosophy of this research. In K-means, the alphabet K denotes the number of clusters that are expected to be formed after the process is executed on the dataset. For example, $K = 5$ means 5 clusters are created from a given dataset, and each cluster is a group of similar data points. The efficiency of K-means algorithm is dependent on the inter-cluster and intra-cluster distances. The distance of a point within the cluster and its centre is called as intra-cluster distance. The inter-cluster distance is the average distance between the point in cluster from the centres of the other clusters. The goal is to minimise the intra-cluster distances (high-dense clusters) and maximise the inter-cluster distances (well-separated clusters), and find the optimal value of K.

This analysis intends to search for 3 groups, and then validate if those groups represent any patterns related to the formation of risk-based group (i.e., low, moderate, and high). The algorithm

K-means was used and capped at $K = 3$. This way, we intentionally try to see if we can find the risk-based groups (low, medium, and high) by limiting the clusters. The way K-means is adopted for this analysis is different from the normal process. In normal process, K-means is executed for the best value of K , that is derived from the Elbow method. In the current approach $K = 3$ was fixed.

3.3.1. Elbow analysis

Elbow analysis was performed, and it was observed that K values between 3 and 6 would give better results for clustering. The result of Elbow analysis was represented as a graph of K in relation to “Within Groups Sum of Squares” (WGSS). Minimising the WGSS is an optimization goal in this method. It is observed from (Figure 8) that at $K = 13$ the WGSS is at the lowest. This conveys that K-means for current dataset will give better results at $K = 13$. This is a classic example of overfitting, so it was neglected.

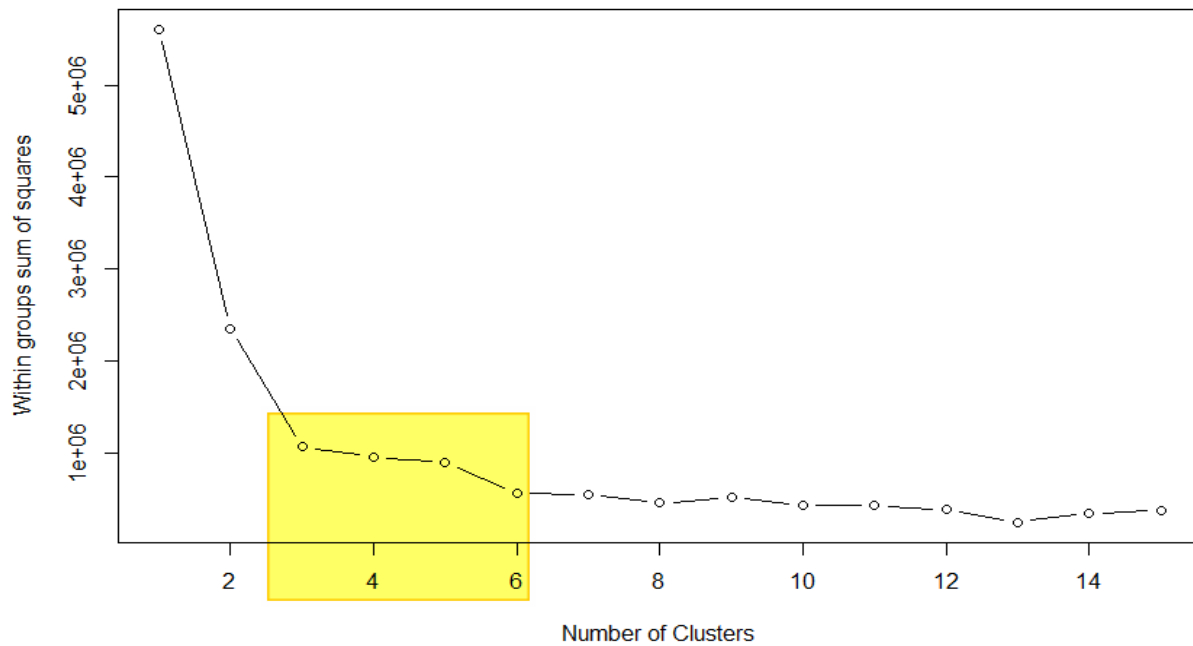


Figure 10: Elbow analysis

The next step is reduction on the dimensionality with $K = 3$.

3.3.2. Dimensionality reduction

Currently, there are seven data attributes (i.e. seven dimensions) in the dataset; it's essential to find the minimum number of data attributes required to get the optimal clustering results. This process to reduce the number of data attributes based on their significance in the clustering is referred as dimensionality reduction.

The best possible clusters at $K = 3$ and the minimum set of data attributes defines the best fit solution for the current clustering. The start should be with minimum required data attributes, and then iteratively add new data attribute, till the efficiency in the clustering is improved. In case, the addition of more attributes does not show improvements or reduce the existing efficiency, then the additional attributes can be omitted. The Silhouette [10] analysis helps to find the efficiency of clustering.

3.3.3. Silhouette analysis

The analysis helps to study the separation distance between the resulting clusters. The silhouette plot displays a measure of closeness for each point in one cluster relative to points in the neighbouring clusters. It is called the Silhouette width (SW) or Silhouette co-efficient and it ranges between $[-1 \leq SW \leq 1]$

- a. $SW = 0$ suggests the point is on the decision boundary, i.e. cannot be assigned to one single cluster
- b. $SW > 0$ and SW nearing the value of 1 suggests point is far away from the neighbouring clusters.
- c. $SW < 0$ hints that the point is assigned to wrong clusters.

The Average Silhouette Width (ASW) of a cluster is the average of the Silhouette Widths for the points belonging to that cluster. It is expected to have ASW above 0.5 to confirm the well separation of clusters.

Using the principle of Silhouette analysis, tests were done iteratively by varying the value of K from 2 to 6, and it was observed that K = 3 gave the optimum results. During these trials, various data attribute combinations were applied to see the best fit for clustering. The variation in K values from 2 to 6 was done to confirm that the value of K = 3 is the optimal one.

The following figure (Figure 9) shows a table for incremental values of K from 2 to 6 (i.e., the ideal range as per the Elbow analysis) and the respective average Silhouette Width (SW).

The column “Silhouette” indicates the Average Silhouette Width at each testing stage.

								Filtered data	Raw data
Glucose	Insulin	BMI	Age	SkinThick	BloodPressure	D P Func	K	Silhouette	Silhouette
	x	x	x				2	0.61	0.66
	x	x	x			x	2	0.6	0.63
	x	x	x		x		2	0.59	0.64
	x	x	x	x			2	0.56	0.6
x	x	x	x				2	0.61	0.66
	x	x	x				3	0.7	0.65
	x	x	x			x	3	0.68	0.6
	x	x	x		x		3	0.67	0.62
	x	x	x	x			3	0.61	0.56
x	x	x	x				3	0.7	0.65
	x	x	x				4	0.7	0.63
	x	x	x				5	0.66	0.62
	x	x	x				6	0.62	0.44

Figure 11: Silhouette analysis for various K values

It is seen from the results (Figure 9) that the combination of *Insulin*, *BMI*, and *Age* gives good results for clustering with $ASW = 0.7$. The other close choices were neglected as extra adding attributes showed no significant change in the ASW.

In the following step, based on the findings from the Silhouette analysis, clustering was performed using the combination of [*Insulin*, *BMI*, *Age*], and further investigations were carried out to understand the risk-based groups formation.

3.3.4. Investigations for risk-based groups formation

Risk-based groups in clusters are validated based on range formations, using combination [*Insulin*, *BMI*, *Age*]. In a diabetic population, the risk-based groups are primarily done using *Glucose* value range. A blood sugar level less than 140 mg/dL is normal. A reading of more than 200 mg/dL after two hours indicates diabetes. A reading between 140 and 199 mg/dL indicates prediabetes. The dataset had the range [78 - 199] for *Glucose* values, that indicates some patients with normal glucose records were wrongly classified as diabetic. If the values [78 - 199] are properly spread across the three clusters and ranges are observed, then it can be concluded that risk-based group formation is possible.

Observations from clustering: [*Insulin*, *BMI*, *Age*]

The clustering results on the filtered data and using data attributes [*Insulin*, *BMI*, *Age*] are shown in the plot (Figure 10).

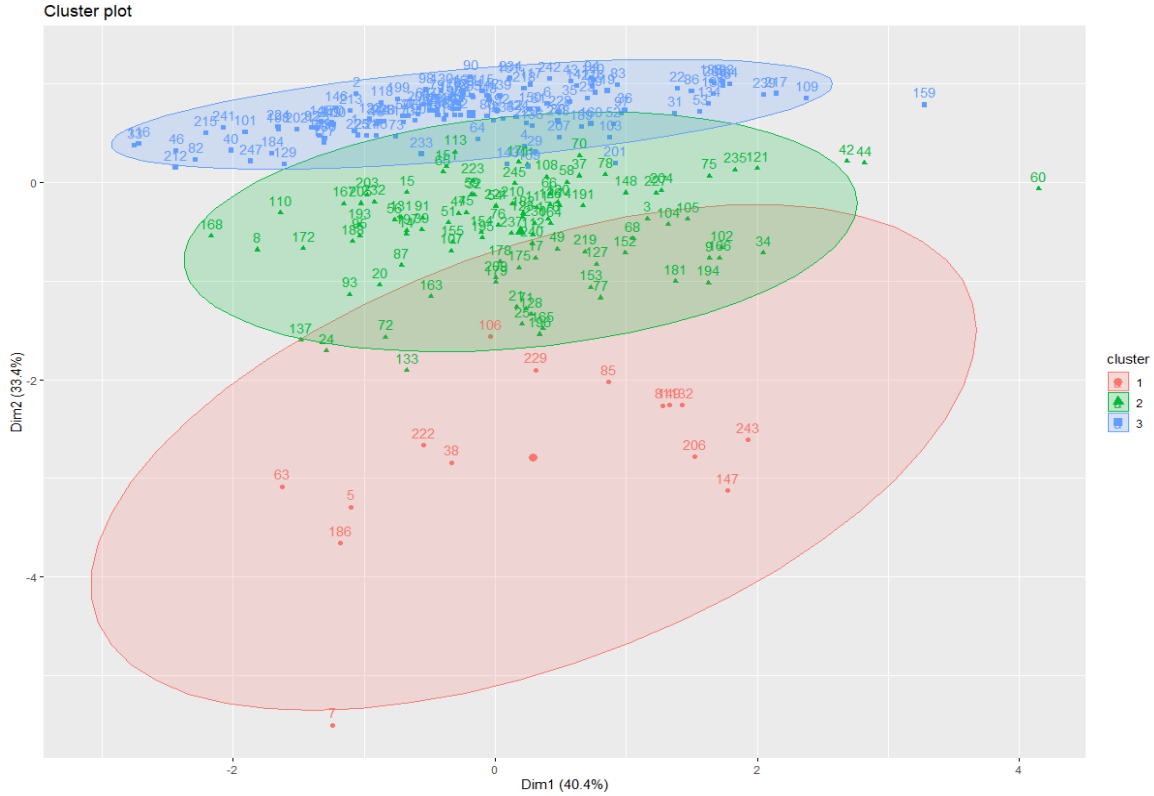


Figure 12: Clustering for risk-based groups formation

The plot in (Figure 10) contains 3-dimensional data and so it is difficult to observe clear separation of clusters using a 2-dimensional plotting method. The ASW (0.7) assures that the three clusters are well separated.

Silhouette analysis indicating 0.7 as the average silhouette width is shown in the following plot (Figure 11).

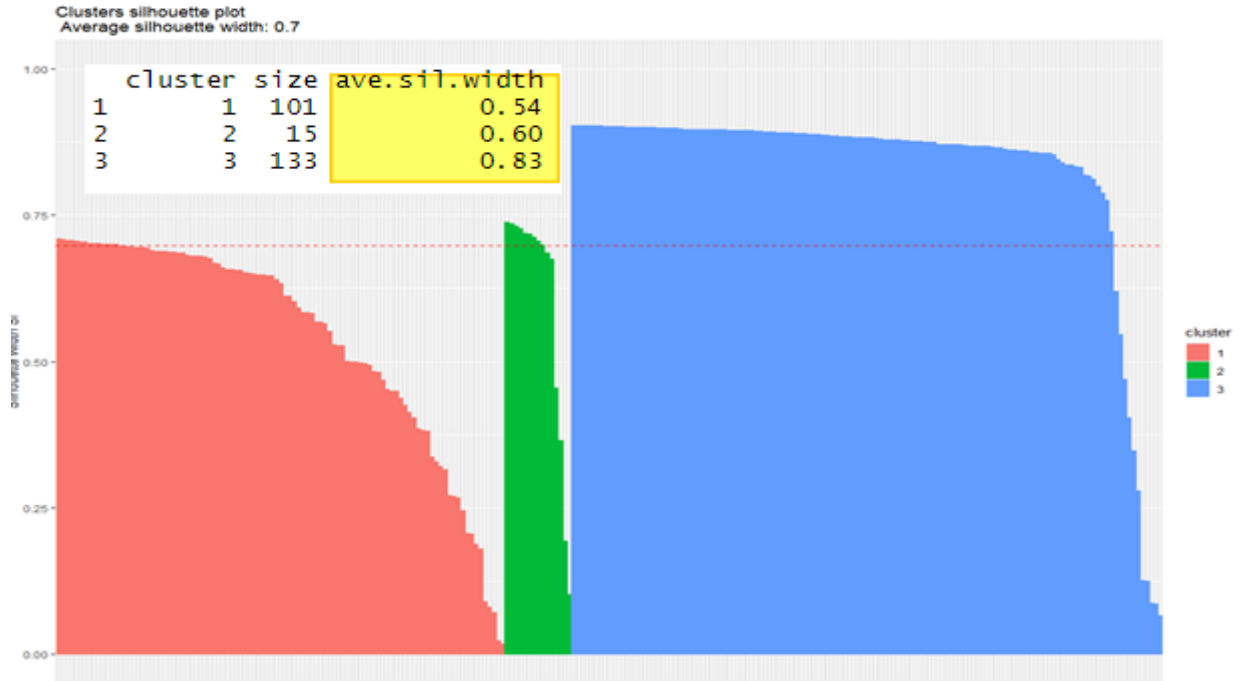


Figure 13: Silhouette for Glucose analysis

Next step is to check clusters for meaningful data related to risk-based groups. (Table 2)

Cluster No.	<i>Glucose</i>	<i>BloodPressure</i>	<i>Insulin</i>	<i>BMI</i>	<i>Age</i>
Cluster 1	88 - 198	30 - 110	96 - 328	23.4 - 67.1	21 - 58
Cluster 2	124 - 197	50 - 90	360 - 846	28 - 46.2	21 - 60
Cluster 3	78 - 199	50 - 114	0 - 91	22.9 - 59.4	21 - 70

Table 2: Observed ranges across clusters

Range formations for *Glucose*, *BloodPressure*, *BMI* or any other attributes are not observed because these were irregularly spread over the whole population. It was observed that some

patients with normal *Glucose* value were also marked diabetic in the source data, so dataset may have false values.

The *Insulin* based range formation is visible. Cluster 3 has a low range [0-9], Cluster 1 has the moderate range [96 - 328] and Cluster 2 has the high range [360 - 846]. The range formation is an interesting observation and these patterns are very useful in this work. This supports the assumption that range formation is possible when optimal clustering is performed and clusters are well-separated.

3.3.5. Attribute selection using Random Forest approach

The results from the previous approach are not promising in relation to the goals. The *Insulin* based risk-groups is a good start but it was important to see formation of *Glucose* based risk-groups from the clusters. The next approach uses Random Forest algorithm to understand the feature selection and revisit the clustering processing [3] [4].

Method:

- a. Execute Random Forest to generate a model for prediction using a dataset partition (training:70% & testing:30%)
- b. Test the prediction accuracy till we get the accuracy around acceptable range. This case it was more than 80% (Observed was 80.34%).
- c. Find the top three or four attributes from that model using varImpPlot [11].
- d. Use the combinations in K-means and observe the results for clustering

Random Forest accuracy was observed as 80.34% with generated trees count 500 and 2 attributes selected at time. The plot analysis showed that optimal value for trees count was between 20 – 50, as shown in the following graph (Figure 12).

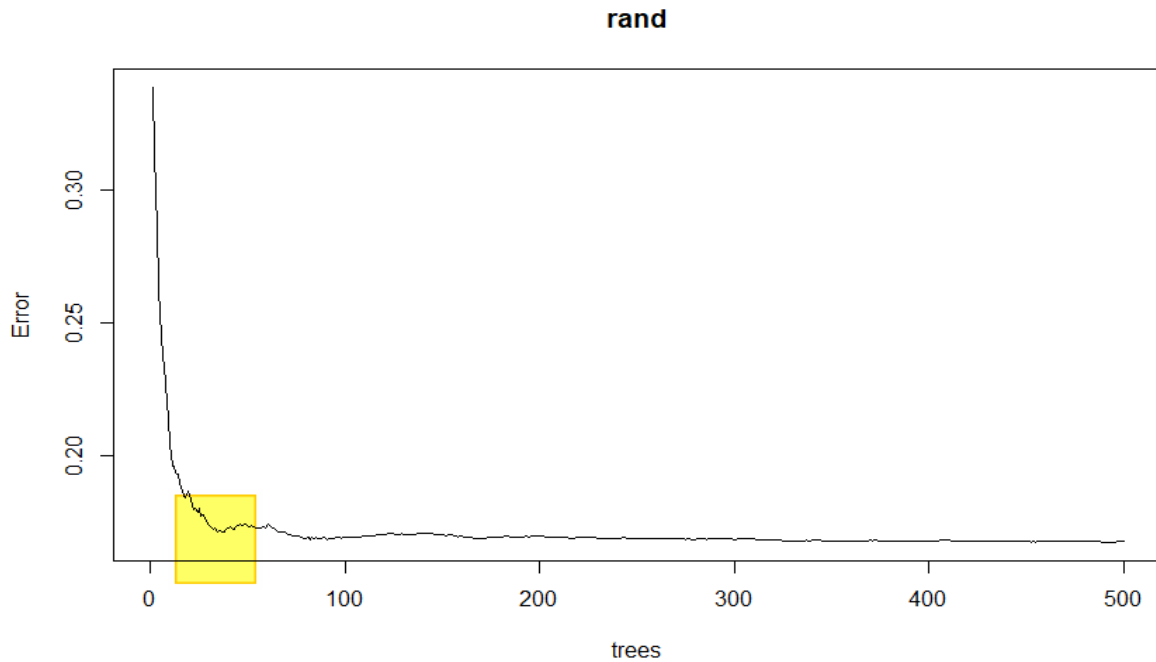


Figure 14: Random Forest model review

Attribute selection from the Random Forest model was done using “varImpPlot” method from R. The function is important to understand which attributes helped in getting the accuracy of the model to that desired level. In this case the plot will indicate the precedence of the data attributes based on the purity factor. It is recommended to start with the higher purity value data attributes. It is helpful to get a visual understanding on the attributes that helped to get the accuracy. The following plot (Figure 13) shows the data attributes that generated the accuracy of 80.34% for the Random Forest model.

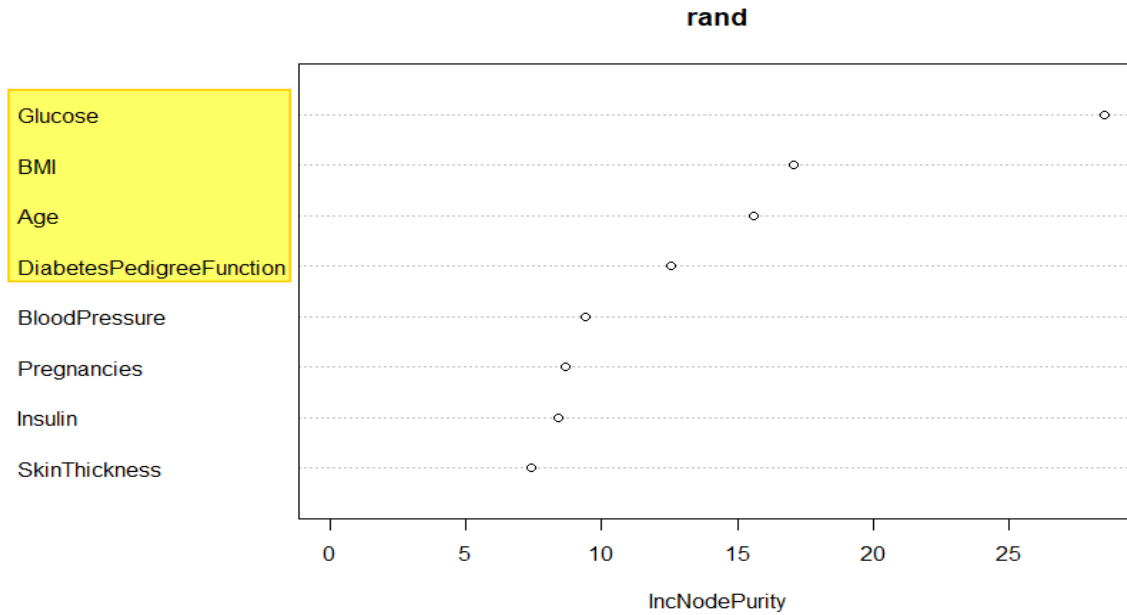


Figure 15: Attribute selection from Random Forest model

The suggestions from the attribute selection step gave the hint to use the following combinations for K-means:

- a. *Glucose, BMI, Age*
- b. *Glucose, BMI, Age, DiabetesPedigreeFunction*

Observations from clustering: [*Glucose, BMI, Age*]

K-means clusters were generated as before, but data attribute combination [*Glucose, BMI, Age*] was different. Visual representation looked intertwined and more cluttered. It is good to find better R libraries for visualising 3-dimensional data for future scope.

The following plot (Figure 14) shows the cluster formations. Cluster colours have no significance other than portraying visual separation in the plot.

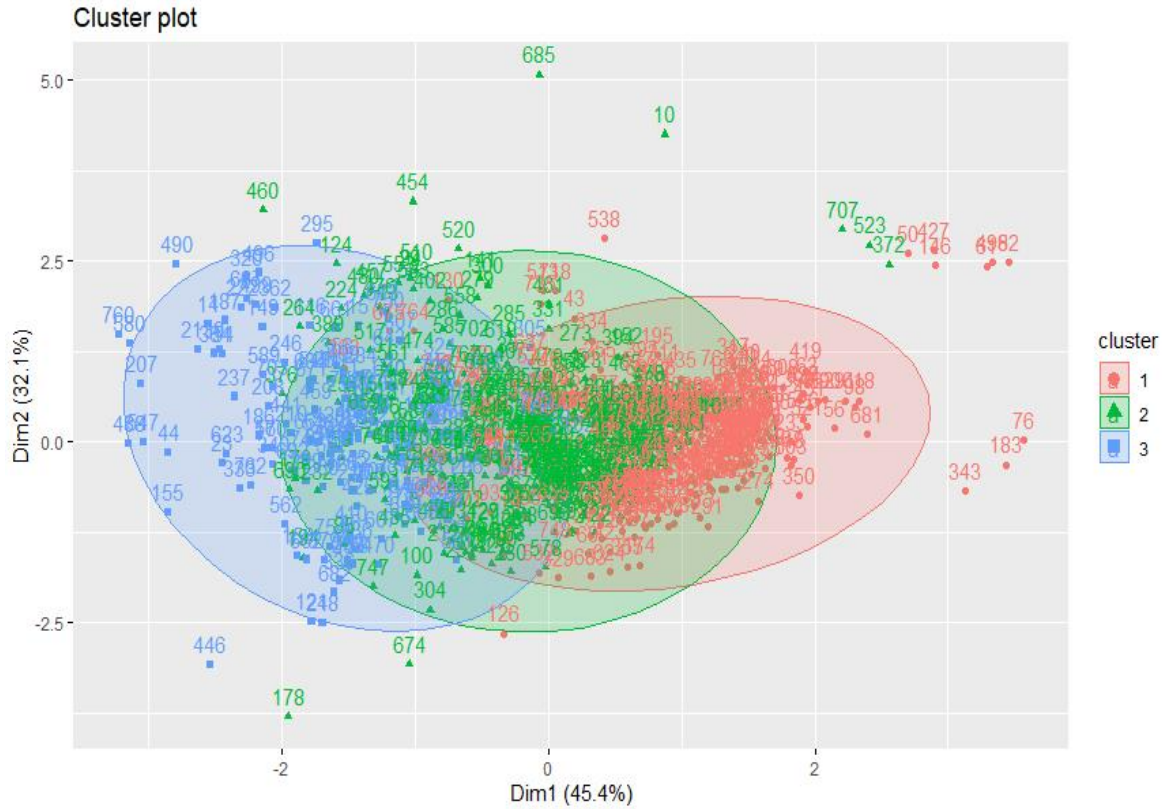


Figure 16: Visualization of clusters based on Random Forest

Next step was to understand the ASW value and, to compare the change with the previous results.

The ASW was 0.37 which was lower compared to the previous value (i.e. $ASW = 0.7$) and also below the good scale (i.e. $ASW > 0.5$), it meant the inter-cluster distances were reduced.

The ASW was significantly reduced but was still above zero and that was a positive note on which the further risk-based analysis was carried.

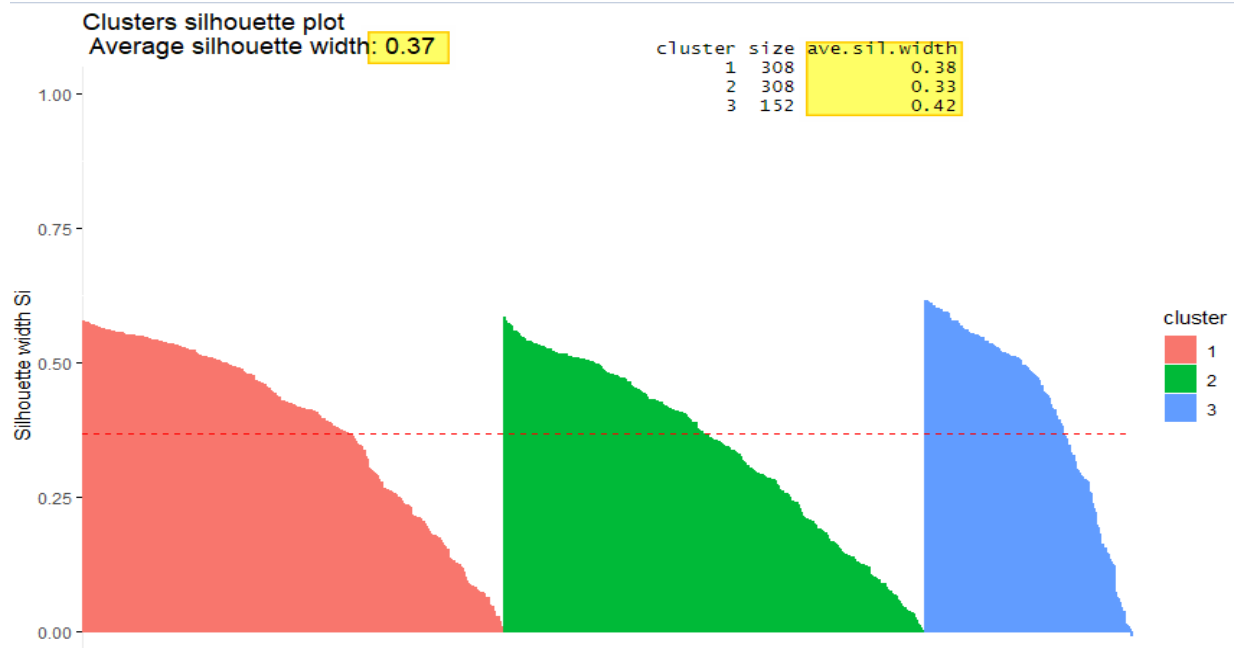


Figure 17: K-means Silhouette plot based on Random Forest

The risk-based group formation showed excellent results that were close to the goals of this research. The following table (Table 3) shows the *Glucose* range formation and the previously observed Insulin range is also retained.

Cluster No.	<i>Glucose</i>	<i>BloodPressure</i>	<i>Insulin</i>	<i>BMI</i>	<i>Age</i>
Cluster 1	78 - 125	30 - 100	0 - 258	22.9 - 55	21 – 62
Cluster 2	160 - 199	50 - 110	0 - 846	23.3 - 59.4	21 – 66
Cluster 3	123 - 159	40 - 114	0 - 600	23.8 - 67.1	21 - 70

Table 3: Random Forest based ranges across clusters

The *Glucose* based range formation is visible. Cluster 1 has a low range [78-125], Cluster 3 has the moderate range [123 - 159] and Cluster 2 has the high range [160 - 199].

The next data attribute combination [*Glucose, BMI, Age, DiabetesPedigreeFunction*] was not required to be tested as group formation was successful.

This concludes the topic of extraction of data attributes for QoC. A noteworthy learning from this part suggests that the clustering results are more efficient based on Attribute Selection as compared with the method based on higher correlations and Average Silhouette Width. Next it is important to make a design to measure the QoC based on the knowledge from clustering.

3.3.6. Clustering Confidence Score (CCS)

The design on QoC is established by finding the correct high-risk patients. The next step is about the derivation of QoC based on these key data attributes identified till this stage.

The study defines a new variable as Clustering Confidence Score (CCS). It has a range [1-10] which indicated the level of confidence that the range formations are correct. Here CCS = 1 means “no confidence” and 10 means “complete confidence”. The assignment of CCS to assert the accuracy for range formations is authorised only for a medical professional. This can be automated if the application can configure the expected ranges.

Each data attribute which is significant in the context of the disease is taken into account and its ranges are configured. Note: $CCS \leq 7$ should not be acceptable.

$CCS = [Number\ of\ correctly\ identified\ data\ points\ in\ all\ significant\ data\ attributes\ in\ the\ respective\ ranges] / [Total\ number\ of\ data\ points]] \times 10$

3.4. Design for Quality of Care (QoC)

There are two scenarios for which the measure for QoC can be determined. One for an existing system that has lots of existing data, and others for a newly implemented system in which data is getting generated. Both scenarios have a way to measure the QoC.

a. **Scenario 1:** Existing system with a high volume of past data (immutable)

In this scenario, risk-based clusters can be generated at specific time intervals, and the trend should be observed on high-risk population clusters. The expected results can be used to learn about the general QoC. A general decline in the size of the high-risk population cluster indicated a good QoC.

b. **Scenario 2:** New system with less volume but growing data (mutable)

In a scenario of a newly installed system, the data is less to do the complete analysis. Clustering can be performed, and then the high-risk population can be monitored. If computing resources are available at ease, then a real-time trend analysis on the high-risk population cluster can provide a better understanding of QoC. Furthermore, development can be done to learn patterns about the effectiveness of specific treatment protocols.

Current work focuses on approach for “Scenario 1” where the dataset does not change and analysis is done on existing data from the past. The similarity in both mentioned cases is that these scenarios can be analysed using trend analysis on the size of the high-risk population cluster.

3.4.1. Approach for Immutable Dataset

Assumptions

1. Data is filtered to choose the patients who are valid throughout the year. This is done by finding the intersection set of patients in first month and last month.
2. The clustering for risk-based groups formation is accurate and correct glucose-based range formations are observed.
3. Dataset does not change during the process.
4. Clustering Confidence Score is more than acceptable value. This configuration is done by the medical staff.
5. One year of past data is considered and QoC is observed on it.
6. Data points are the cluster count values for each month. There will be 3 values each month.

Given below is a hypothetical example: Data for 2019 with three clusters count for whole year (Total: 200 patients, C1: high risk group, C2: moderate risk group and C3: low risk group)

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
C1	85	86	85	82	75	78	74	70	65	64	60	56
C2	39	39	40	43	47	44	46	49	53	54	56	60
C3	76	75	75	75	78	78	80	81	82	82	84	84

Table 4: Risk-based cluster sizes for a year

The data trend can be better visualised to understand the QoC. Here are the plots for the following data with high risk patients indicated in red line, moderate risk with orange colour and low risk with green colour.

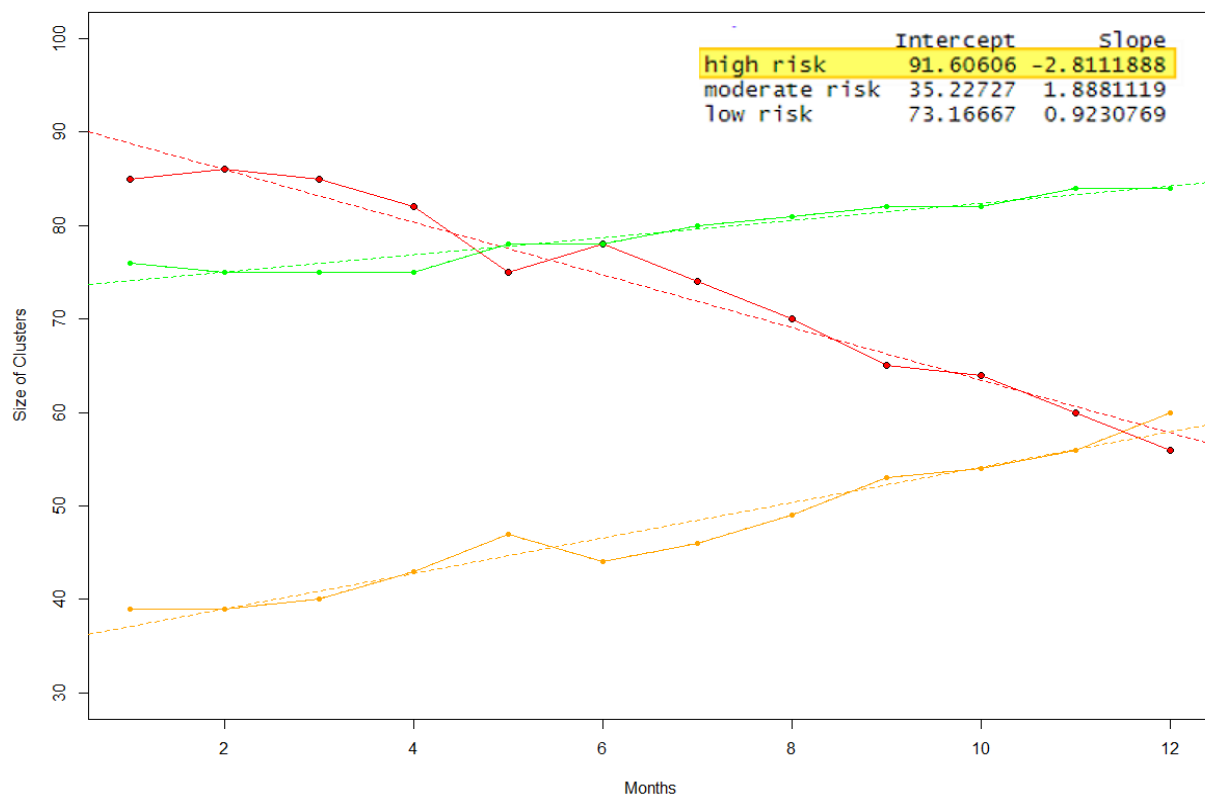


Figure 18: Visualising trend in clusters size

The focus is on the high-risk cluster size and so that is indicated with red colour for emphasis. The decreasing trend in the red line indicates good QoC. The increase in the moderate risk group size does not necessarily indicate poor QoC if the low risk group is having a stable trend. Simple linear regression can be applied to understand the trend lines. The negative slope value for the high risk trend line is good indicator of better QoC.

4. Conclusion and Recommendations

4.1. Conclusion

- Real life data on diseases like Diabetes is not easily separable into clusters using default method of K-means or DBSCAN method. Assumption related to good separated clusters should generate risk-based groups was found to be incorrect.
- Unsupervised learning (K-means clustering) helps in analysis of risk-based groups when we find the right set of data attributes.
- Random Forest provides better results for attribute selection compared to Average Silhouette Width analysis.
- Cluster Confidence Score plays an important role to validate the learning from the model from a medical professional.
- Risk-based groups formation was observed after taking the attributes provided by Random Forest approach when prediction accuracy was acquired around 80%.
- QoC for existing and new systems do have similar models based on trend of size of clusters.

4.2. Recommendations

- Invest more time in collection of valid and extensive dataset so as to get better results in the clustering and risk-based groups formation.
- K-medoids can be evaluated instead of K-means to observe the difference in groups formation and the efficiency of clustering. It can even help with larger datasets.

5. Future Scope

- Constraint-based clustering and initializing the centroids could be a promising way for further research. When the data is well scattered, then use the 3 mean values of risk-based groups can be used to initialize the cluster centres.
- QoC analysis for an individual patient can be studied by finding the variation from the baseline (i.e., normal value line or range). The deviation away from the normal range based on configured threshold, can be taken as poor QoC.
- Significance of QoC attributes over time for the same application. To find out if the initially chosen QoC attributes can be applied in future.
- Philips VitalHealth application (especially Co-ordinate which captures information on Diabetes population) was studied briefly and the data attributes from this study were also found in the application. This will help to build a model for it.

6. Appendices

6.1. Code: K-means with gap-stat and DBSCAN with kNN

```
library("factoextra")

source("read_diab_file.r")
source("get_filtered_data.r")
# columns
# 1 Pregnancies,
# 2 Glucose,
# 3 BloodPressure,
# 4 SkinThickness,
# 5 Insulin,
# 6 BMI,
# 7 DiabetesPedigreeFunction,
# 8 Age,
# 9 Outcome

raw = 0
no_of_clusters = 3

if(raw == 1)
{
  dat = read_diab_file()
}

if(raw == 0)
{
  dat = get_filtered_data()
}

dat <- dat[,1:8]
# Finding optimal no. of clusters using k-means
fviz_nbclust(dat, kmeans, method = "gap_stat")

# result we see optimal number is 1, in our case that will not make any significant results

km.res <- kmeans(dat, no_of_clusters, nstart = 25)
fviz_cluster(km.res, dat, geom = "point",
             ellipse= FALSE, show.clust.cent = FALSE,
             palette = "jco", ggtheme = theme_classic())

km.df1<-data.frame(dat,cluster=km.res$cluster)
```

```

for(xx in 1:no_of_clusters)
{
  dfc <- filter(km.df1, cluster == xx)

  cat("c",xx,":- Glucose range[" , range(dfc$Glucose), "] BP range[" ,range(dfc$BloodPressure)
, "] Insulin range[" ,range(dfc$Insulin) , "] BMI range[" ,range(dfc$BMI), "] Age
range[" ,range(dfc$Age), "] ST range [" ,range(dfc$SkinThickness), "] DPF range
[" ,range(dfc$DiabetesPedigreeFunction),"]\r\n")
}
# Now DBSCAN

set.seed(123)
res <- dbscan::kNNdistplot(dat, k = 3)
abline(h = 60, lty = 2)

db <- fpc::dbscan(dat, eps = 60, MinPts = 6)
print(db)
fviz_cluster(db, data = dat, stand = FALSE,
  ellipse = FALSE, show.clust.cent = FALSE,
  geom = "point", palette = "jco", ggtheme = theme_classic())

summary(db$cluster)

km.df2<-data.frame(dat,cluster=db$cluster)
for(xx in 1:2)
{
  dfc <- filter(km.df2, cluster == xx)

  cat("c",xx,":- Glucose range[" , range(dfc$Glucose), "] BP range[" ,range(dfc$BloodPressure)
, "] Insulin range[" ,range(dfc$Insulin) , "] BMI range[" ,range(dfc$BMI), "] Age
range[" ,range(dfc$Age), "] ST range [" ,range(dfc$SkinThickness), "] DPF range
[" ,range(dfc$DiabetesPedigreeFunction),"]\r\n")
}
# results are not interesting in terms of risk based range formation

```

6.2. Code: Correlations using Corrplot

```

#find best correlations
# columns
# 1 Pregnancies,
# 2 Glucose,
# 3 BloodPressure,
# 4 SkinThickness,
# 5 Insulin,

```

```

# 6 BMI,
# 7 DiabetesPedigreeFunction,
# 8 Age,
# 9 Outcome

library(corrplot)
library(tidyverse)
library(BBmisc)

source("read_diab_file.r")
source("get_filtered_data.r")

raw = 0

if(raw == 1)
{
  dat = read_diab_file()
}
if(raw == 0)
{
  dat = get_filtered_data()
}
#
# dim(dat)
#
# hist(dat[, "Insulin"])

correlations = cor(dat[, 2:8])
write.csv(correlations, "./correlations.csv")

corrplot.mixed(correlations, order="hclust", tl.col="black")

```

6.3. Code: K-means with Silhouette

```

# k-means - https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/kmeans
# columns
# 1 Pregnancies,
# 2 Glucose,
# 3 BloodPressure,
# 4 SkinThickness,
# 5 Insulin,
# 6 BMI,
# 7 DiabetesPedigreeFunction,

```

```

# 8 Age,
# 9 Outcome

library(tidyverse) # data manipulation
library(cluster) # clustering algorithms
library(factoextra) # clustering algorithms & visualization
source("read_diab_file.r")
source("get_filtered_data.r")

raw = 0
no_of_clusters = 3

if(raw == 1)
{
  dat = read_diab_file()
}

if(raw == 0)
{
  dat = get_filtered_data()
}

filter1 <- c(5,8,6) # Insulin, Age, BMI > had good correlation with Glucose - 0.71
filter2 <- c(5,8,6,3) # Insulin, Age, BMI, BP > had good correlation with Glucose - 0.66
filter3 <- c(5,8,6,4) # Insulin, Age, BMI, Skin Thickness > had good correlation with Glucose - 0.68
filter4 <- c(5,8,6,2) # Insulin, Age, BMI, Glucose

filter5 <- c(5,8,6,7) # Insulin, Age, BMI, DPF

filter6 <- c(3,5,6) # BP, Insulin, BMI

filter7 <- c(3,8,4,5) # BP, Insulin, BMI

filter8 <- c(2,8,5) # BP, Insulin, BMI

filter9 <- c(2,6,8) # Glucose, BMI, Age - (prime filter - after random forest)

filter10 <- c(2,6,8,7) # Glucose, BMI, Age, DPF

filtered_data = dat[,filter9]

run_kmeans <- function(xtimes)
{
  kmeans <- lapply(seq_len(xtimes), function(i){
    results <- kmeans(filtered_data, no_of_clusters)
  })
  return(kmeans)
}

```

```

}

kmeans_all_results = run_kmeans(10)

perf <- sapply(kmeans_all_results, function(d) as.numeric(d["tot.withinss"]))
index <- which.min(perf)

kmeans_results = kmeans_all_results[[index]]

fviz_cluster(kmeans_results, data = filtered_data, iter.max = 10, nstart = 1, algorithm = "Hartigan-
Wong", ellipse.type = "norm") #+
#scale_colour_manual(values = c(cluster_colors[1], cluster_colors[2], cluster_colors[3])) +
#scale_fill_manual(values = c(cluster_colors[1], cluster_colors[2], cluster_colors[3]))

c_len = length(kmeans_results$cluster)

df1<-data.frame(dat,cluster=kmeans_results$cluster)

for(xx in 1:no_of_clusters)
{
  dfc <- filter(df1, cluster == xx)

  cat("c",xx,":- Glucose range[" , range(dfc$Glucose), "] BP range[" ,range(dfc$BloodPressure) ,"]
  Insulin range[" ,range(dfc$Insulin) ,"] BMI range[" ,range(dfc$BMI),"] Age
  range[" ,range(dfc$Age),"] ST range [" ,range(dfc$SkinThickness),"] DPF range
  [" ,range(dfc$DiabetesPedigreeFunction),"]\r\n")
}

sil <- silhouette(kmeans_results$cluster, dist(filtered_data))

fviz_silhouette(sil)

```

6.4. Code: Random forest with varImpPlot

```

# random forest

library(randomForest)

library(caret)

library(tidyverse) # data manipulation

source("read_diab_file.r")

```



```

source("get_filtered_data.r")

raw = 1

if(raw == 1) {
  dat <- read_diab_file();
}

if(raw == 0) {
  dat <- get_filtered_data();
}

ind <- sample(2, nrow(dat), replace = TRUE, prob = c(0.7, 0.3))

train_data <- dat[ind == 1,]
test_data <- dat[ind == 2,]

nrow(train_data)

nrow(test_data)

set.seed(234)

rand <- randomForest(Outcome~., data=train_data) #, method = 'class' , parms = list(split =
"information")

pred1 <- predict(rand, test_data[-9],type="class" )

pred_new <- sapply(pred1, function(d) round(d, digits = 0))

confMatrix <- table(test_data$Outcome,pred_new)

accuracy <- sum(diag(confMatrix))/sum(confMatrix)

cat("accuracy:",accuracy,"%")

plot(rand)

varImpPlot(rand)

```

7. References

- [1] “Meaningful Use,” [Online]. Available:
<https://www.cdc.gov/ehrmeaningfuluse/index.html>.
- [2] P. F. 2. Elma Kolçe (Çela) 1, “A Literature Review of Data Mining Techniques Used in Healthcare Databases,” pp. 1-8, January 2012.
- [3] M. J. Hallett, J. J. Fan, X. G. Su, R. A. Levine and M. E. Nunn, “Random forest and variable importance rankings for correlated survival data, with applications to tooth loss,” *Statistical Modelling*, pp. 523-547, 2014.
- [4] B. Rai, “Feature Selection and Predictive Modeling of Housing Data Using Random Forest,” *International Journal of Business and Economics Engineering*, vol. 11, no. 4, pp. 919-923, 2017.
- [5] “R Studio,” R Studio, [Online]. Available: <https://rstudio.com/>.
- [6] “PIMA Indians Diabetes Database,” National Institute of Diabetes and Digestive and Kidney Diseases, [Online]. Available: <https://www.kaggle.com/uciml/pima-indians-diabetes-database>.
- [7] “Summary of HIPAA Privacy Rule,” U.S. Department of Health & Human Services, [Online]. Available: <https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html>.

- [8] “Outlier,” [Online]. Available: <https://en.wikipedia.org/wiki/Outlier>.
- [9] Y. Su, “Correlation Plot,” [Online]. Available:
<https://www.rdocumentation.org/packages/arm/versions/1.10-1/topics/corrplot>.
- [10] “Silhouette (clustering),” [Online]. Available:
[https://en.wikipedia.org/wiki/Silhouette_\(clustering\)](https://en.wikipedia.org/wiki/Silhouette_(clustering)).
- [11] C. Chapman and E. M. Feit, “R for Marketing Research and Analytics,” in *R for Marketing Research and Analytics*, Springer Nature, 2015, pp. 331,332,333.

8. Glossary

ASW: Average Silhouette Width.....	19
CCS: Clustering Confidence Score.....	28
CDC : Centers for Disease Control and Prevention.....	1
CDM : Chronic Disease Managment.....	7
COPD: Chronic Obstructive Pulmonary Disease	1
CSV : Comma Seperated Values	7
DBSCAN: Density-based spatial clustering of applications with noise	15
HIPAA: Health Insurance Portability and Accountability Act.....	8
kNN: k Nearest Neighbours.....	15
PHI: Protected Health Information	8
QoC : Quality of Care	1
SW : Silhouette Width	19
USP : Unique Selling Point.....	1