Quality of Care

for Chronic Disease Management

VAIBHAV GAIKWAD // 2018HT12597

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE
PILANI (RAJASTHAN)

April 2020

Nobody ever figures out what life is all about, and it doesn't matter. Explore the world. Nearly everything is really interesting if you go into it deeply enough.

- RICHARD P. FEYNMAN

Introduction

- In the healthcare domain, the applications are built around the idea of providing care.
- Philips VitalHealth has various chronic disease management solutions build for diseases (e.g. Diabetes, COPD, Asthma, etc.).
- Philips culture supports innovation to add more value to the products for staying ahead of the competition.
- One of the widely talked Unique Selling Point (USP) is about Quality of Care (QoC) in the solutions.
- The dissertation work focuses on the research of building a model for the assessment of QoC.

Motivations

Analysis of Quality of Care (QoC) is about gaining information about the effectiveness of care management for the patients.

- QoC is stated good when the health of the population is stable or improving.
- QoC is also a measure of the effectiveness of a treatment protocol.
- QoC in a business case means proving the ROI (return on investments) to customers.
- QoC measurement is a must for accountable care as per CDC's Meaningful Use policy.

Observations & Objectives

Observations

- Current solutions have little or no way to assess the QoC.
- Knowledge related to QoC is observed in the people but is not yet implemented.
- Ad-hoc methods are adopted for QoC report generation on a need basis.
- Data needed for assessing QoC is present in the existing system.

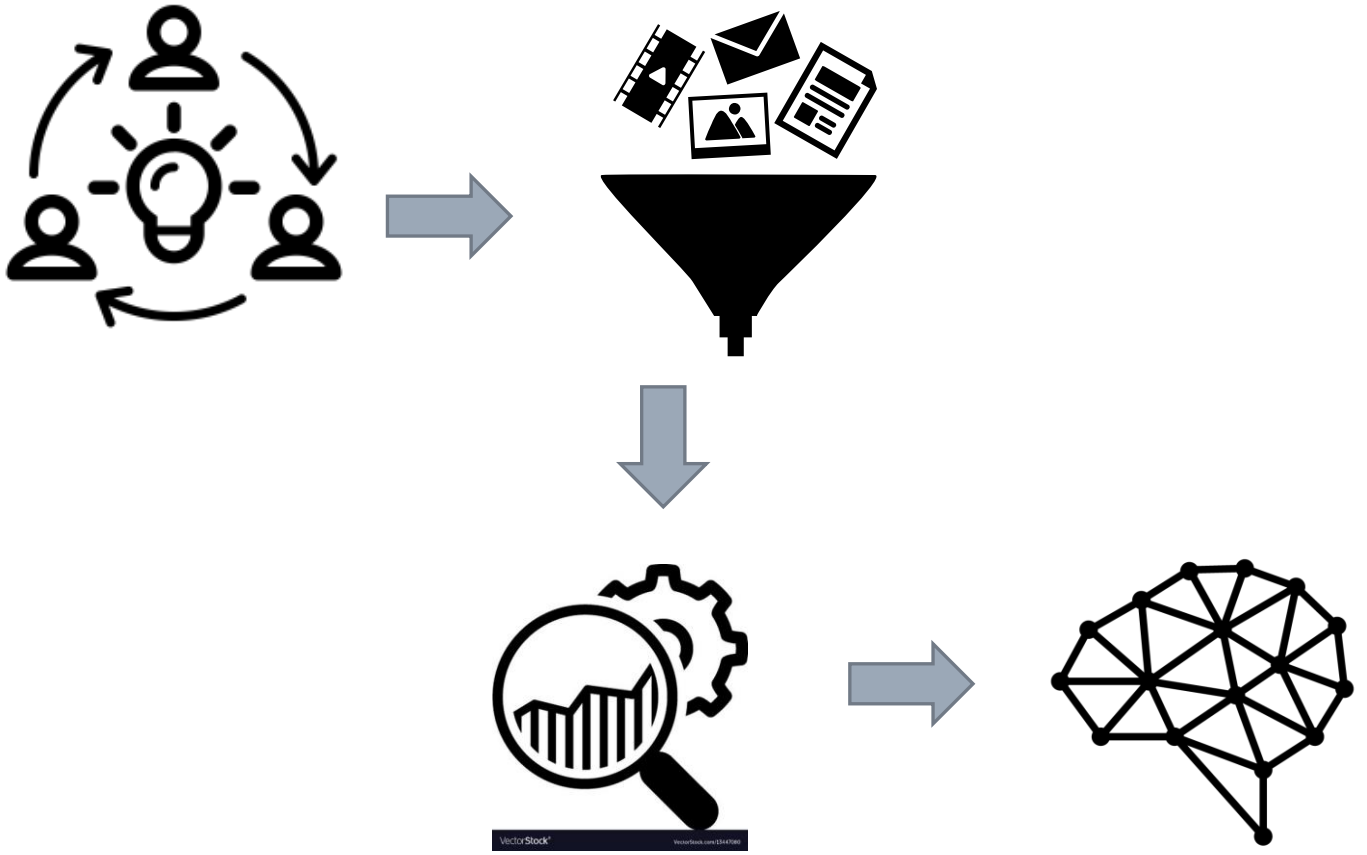
Objectives

- Define a way to find QoC attributes for an application
- Propose a model to assess the QoC using those attributes

Work Summary

Division of work

- Ideation and design
- Data collection and analysis
- Searching for risk groups
- Design for Quality of Care



Ideation

Brainstorming sessions helped to understand the ways to find QoC.

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

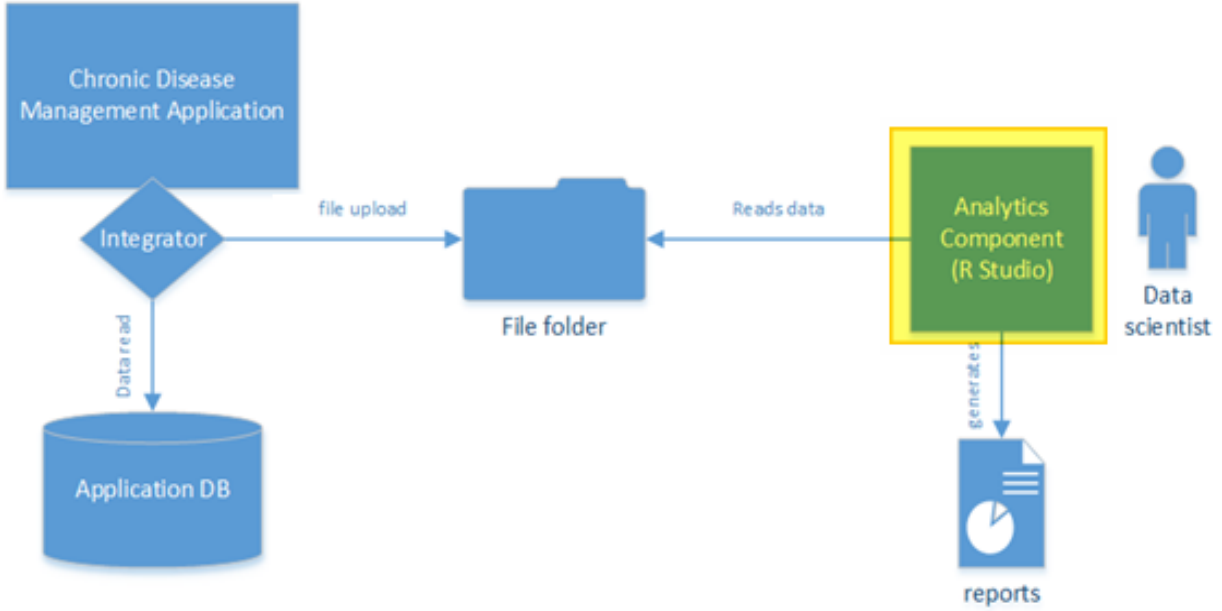
Research requirement was gathered, and a proposal was made with the objectives.

- Research a way to find QoC attributes for an application.
- Propose a model to assess the QoC using those attributes.

Ideation

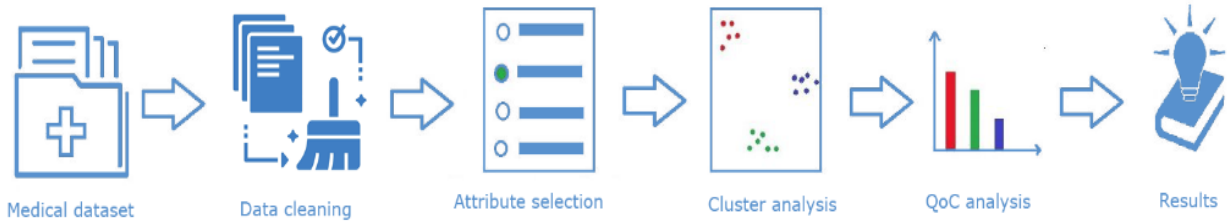
System block diagram

- The focus is on the design of the Analytics Component.
- Rest of the integrations will be considered during the actual implementation.



Ideation

Analytics Component



- Chronic Disease Management (CDM) application holds the medical data, and can be exported.
- Data cleaning removes insignificant data from the exported dataset.
- Attribute selection helps to find the valuable data attributes helpful for clustering.
- Cluster analysis helps to understand the risk-based group formations.
- QoC analysis will help in understanding if the care management is effective or not.
- Results will be the reports that are generated for the care organisation.

Data Analysis

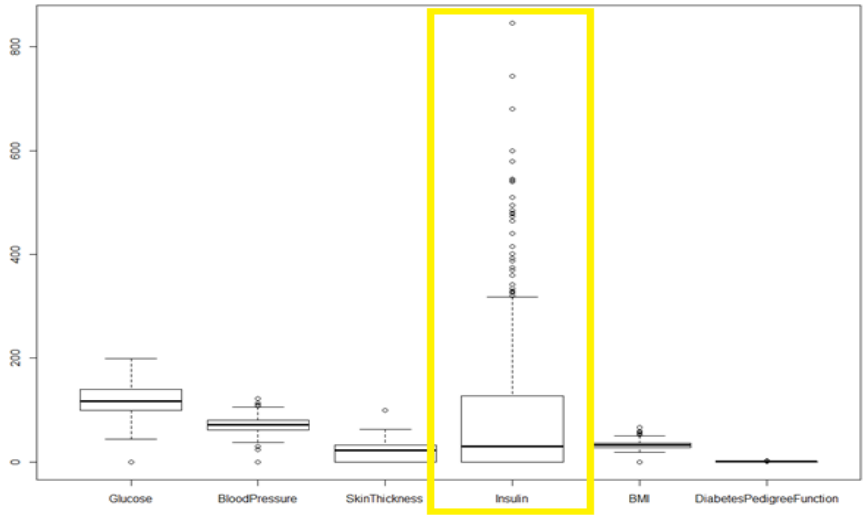
Collection

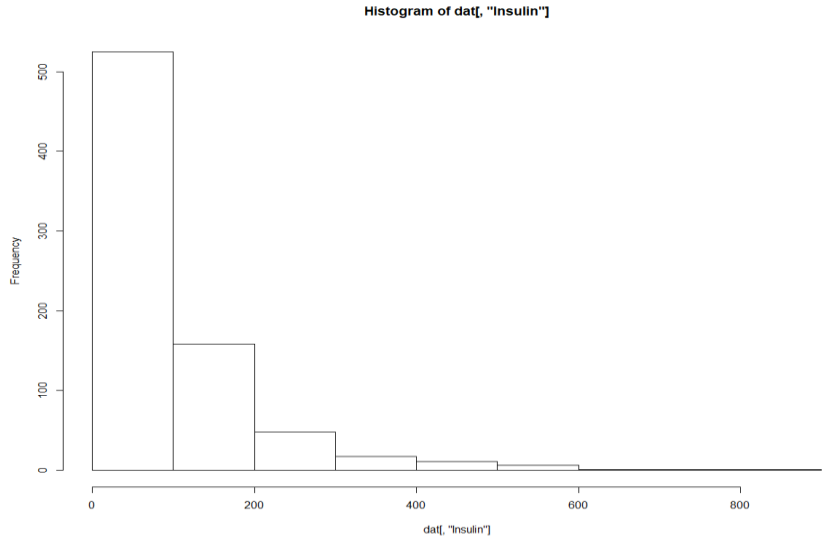
- Privacy contracts made it difficult to use applications with live data.
- Dataset from Kaggle.com was used referred to as PIMA dataset, which had 786 records for diabetic and non-diabetic female patients.
- Following eight data attributes were present as listed in the table.

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

Data Analysis
Cleaning

- As study was not gender specific “Pregnancies” data was neglected.
- Removal of records with 0 value for Glucose, Blood Sugar, BMI.
- Number of were reduced to 249 after the data cleaning process.





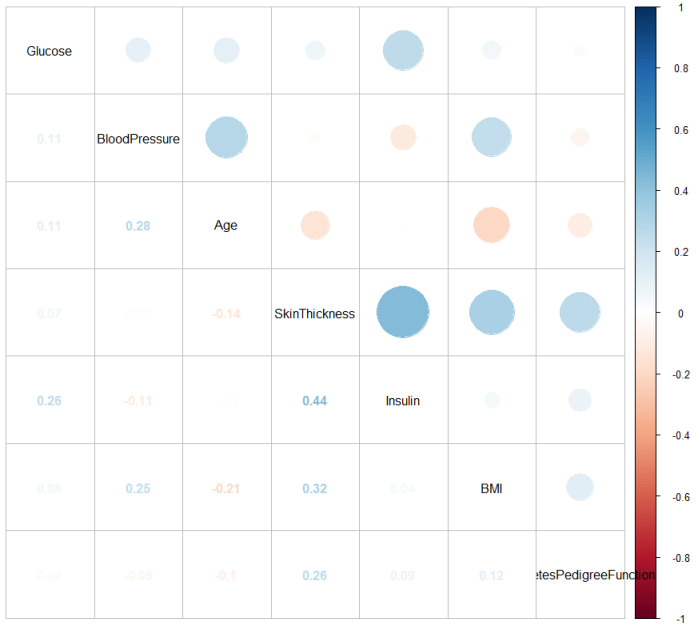
Data Analysis

Correlations among attributes

Correlations in filtered data with strong correlations are listed below

- *Glucose* and *Insulin* (obvious)
- *SkinThickness* and *Insulin*
- *SkinThickness* and *BMI*

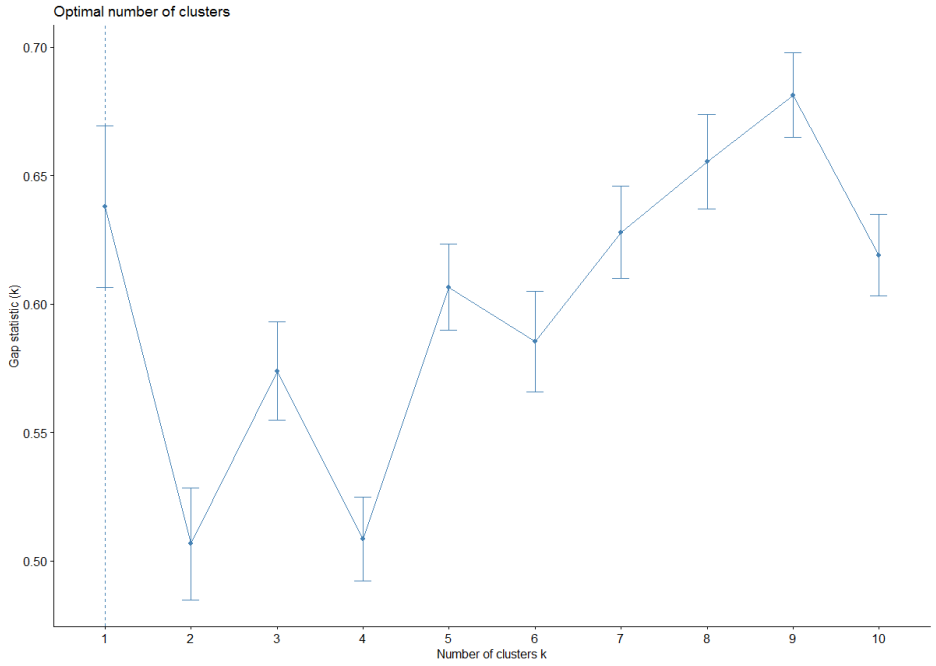
	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
Glucose	1	0.152589587	0.057327891	0.33135711	0.221071069	0.1373373	0.26351432
BloodPressure	0.152589587	1	0.207370538	0.088933378	0.281805289	0.041264948	0.239527946
SkinThickness	0.057327891	0.207370538	1	0.43678257	0.392573204	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



Search Risk Groups

Standard K-means (gap-stat)

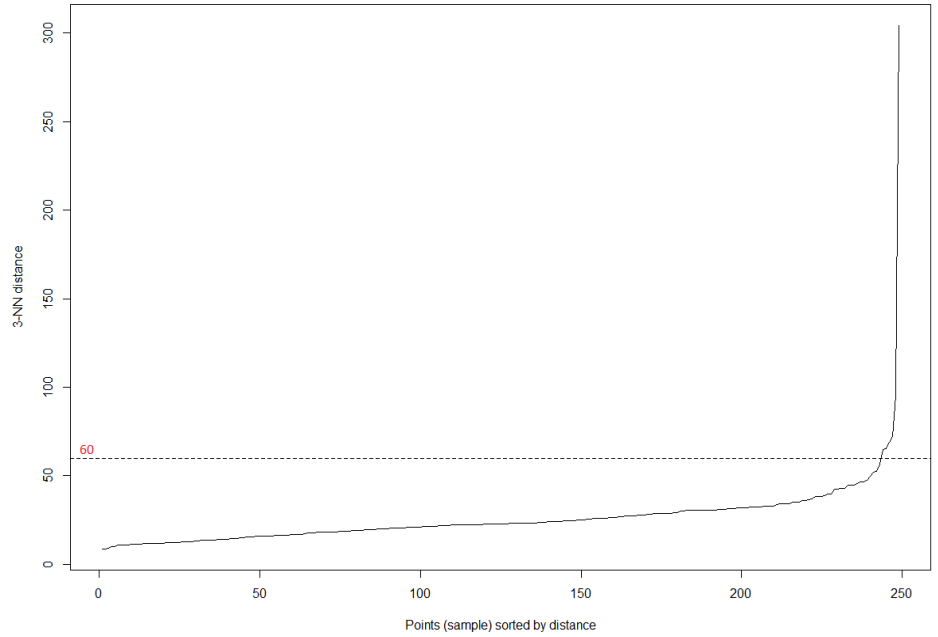
- K-means execute with gap-stat indicated the data is not separable, as optimal K value was 1.
- K-means with all eight data attributes for K = 3 was not useful.
- Conclusion: Data was not easily separable into groups based on risk using default K-means.



Search Risk Groups

Standard DBSCAN (kNN)

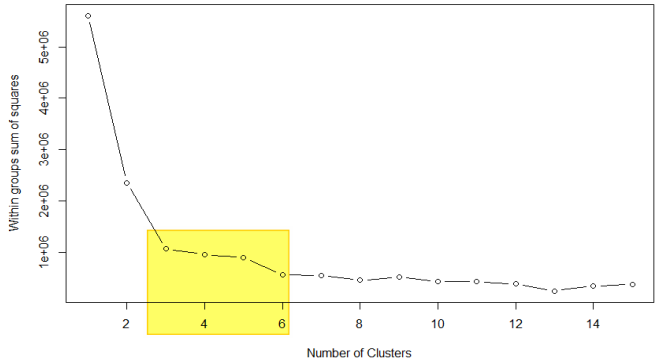
- DBSCAN was tested with kNN for K = 3 and min eps distance was observed to be between 60 to 80
- DBSCAN with eps = 60 and MinPts = 5, generated 2 clusters but did not generate any useful results in regards to the risk groups.
- Conclusion: DBSCAN is not a correct choice to proceed, as data is not easily separable.



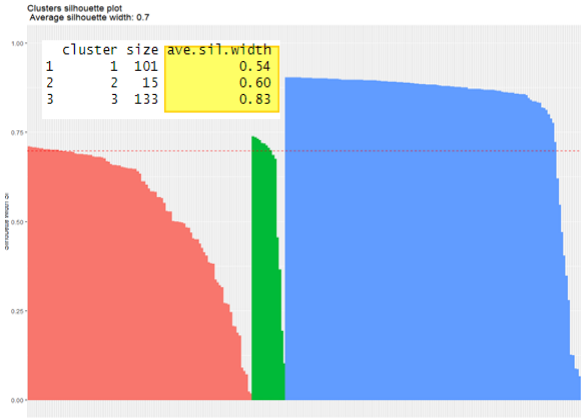
Search Risk Groups

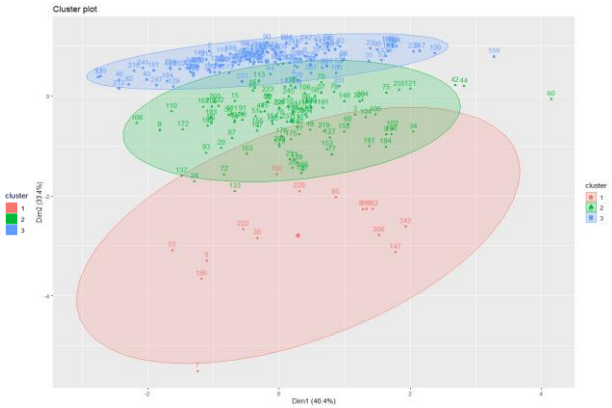
K-means clustering & Silhouette analysis

- Elbow analysis confirmed K = 3 lies in optimal section.
- Silhouette analysis iterations were carried for K [2-6] to find the attributes which resulted in good clustering in K-means.
- Attributes [Insulin, BMI, Age] were found promising at value 0.7 for Average Silhouette Width (ASW).



								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





Search Risk Groups

Range formation

The Insulin based range formation was visible.

- Cluster 3 has a low range [0-9].
- Cluster 1 has the moderate range [96 - 328].
- Cluster 2 has the high range [360 - 846].

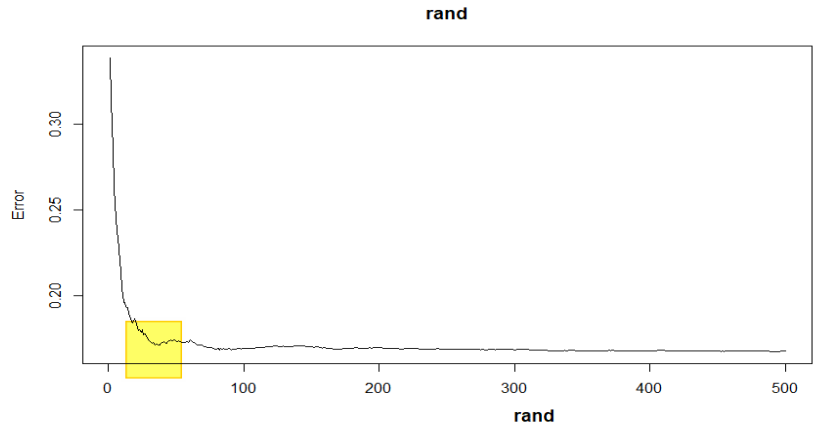
The result was not promising so attribute selection with Random Forest method was performed and clustering was redone.

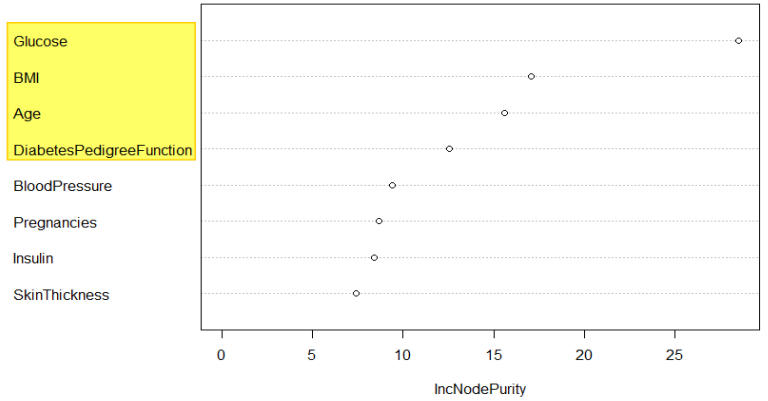
Cluster No.	Glucose	BloodPressure	Insulin	BMI	Age
Cluster 1	88 - 198	30 - 110	96 - 328	23.4 - 67.1	21 - 58
Cluster 2	124 - 197	50 - 90	360 - 846	28 - 46.2	21 - 60
Cluster 3	78 - 199	50 - 114	0 - 91	22.9 - 59.4	21 - 70

Search Risk Groups

Random Forest and Attribute Selection

- Random Forest indicated the optimal attribute combination was [Glucose, BMI, Age]. If range formation is not observed then DPF attribute can be added.
- Positive results were seen using the [Glucose, BMI, Age] after clustering.
- ASW = 0.37 was observed and was comparatively low but still in the valid range (positive).





Cluster No.	Glucose	BloodPressure	Insulin	BMI	Age
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

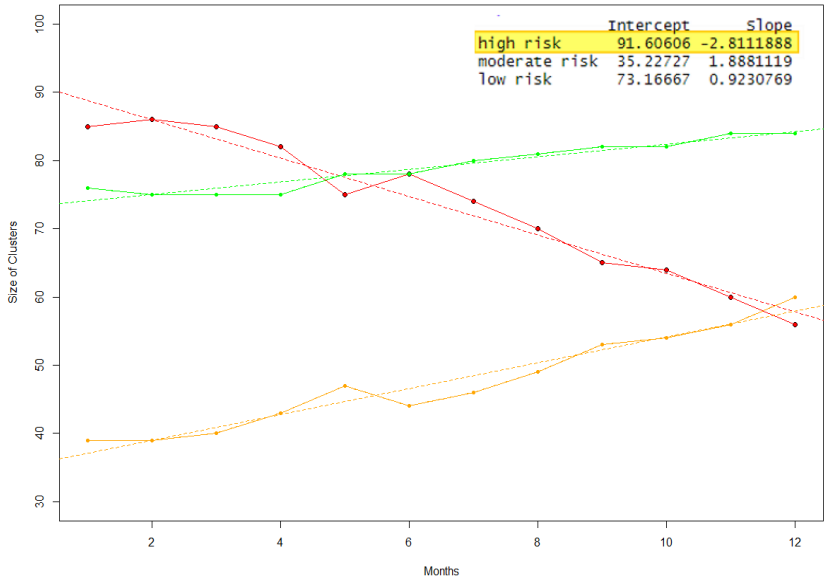
QoC design

- Clustering Confidence Score (CCS) – was defined as new terms which gives the percentage of valid data points in range to the actual world. Values lies in range [0 - 10]. Score above 7 is reliable.

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

- Design was based on trend analysis of cluster sizes that are generated from the previous steps.
- The focus is on the high-risk cluster size,
  decreasing trend of high-risk cluster indicates good QoC.
- Following example data is plotted (C1: high risk cluster)
- Simple linear regression (colour: red) line is added + slope details
- Negative slope: Good QoC (for high risk)

	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



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.

References

- 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.
- 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.
- 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.

THANKS