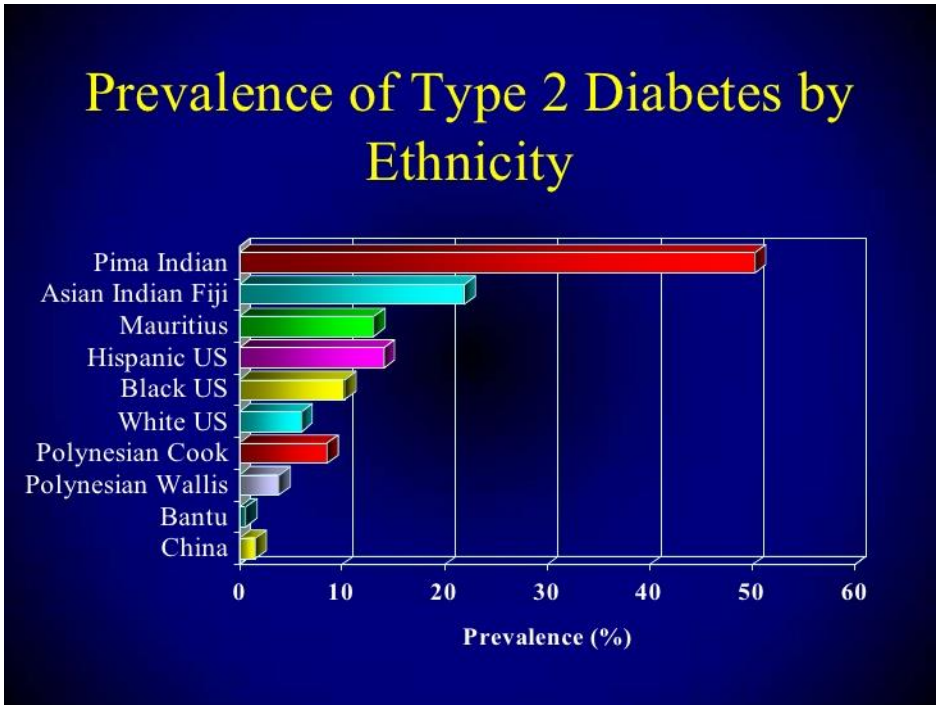
UNC CHARLOTTE

Data Analysis on
PIMA Indian Diabetes Database



Submitted By

Bhavya Joshi (800985776)

UNC CHARLOTTE

Preface

- Objective
- Dataset Description
- Data Exploration
- Handling Missing Values
- Data Modelling
- Observations
- Future Scope
- Definition
- Outliers treatment
- Normalization
- Data Partitioning and Evaluation
- Data Analysis
- Conclusion
- Reference

UNC CHARLOTTE

Objective

- **Applying data analysis techniques, creating visualizations and interpreting the models using histograms, scatter plots etc. to uncover the reason for higher number of diabetic patients amongst the PIMA Indian women.**

- **Predicting whether a patient would be diagnosed with diabetes or not based on parameters available in the dataset**

- **Breaking down the relationships between different parameters of the chosen dataset and find a correlations to better understand the end result.**

UNC CHARLOTTE

Definition

- **Definition:**

  To establish a relationship between parameters such as number of pregnancies, BMI, age, etc., and diabetic diagnosis of the patients based on them.

- **Requirement:**

  A comprehensive understanding of the parameters in relation with the diabetes presented in the database.

UNC CHARLOTTE

Dataset Description

- **PIMA Indians is a group of Native Americans living in an area consisting parts of central and southern Arizona.**
- **In particular, all patients here are females at least 21 years old of Pima Indian heritage.**
- **The utilized dataset was obtained from Kaggle website based on the research conducted by The National Institute of Diabetes and Digestive and Kidney Diseases (NIDDKD)**

UNC CHARLOTTE

Data Snippet

A	B	C	D	E	F	G	H	I
Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1
0	118	84	47	230	45.8	0.551	31	1
7	107	74	0	0	29.6	0.254	31	1
1	103	30	38	83	43.3	0.183	33	0

UNC CHARLOTTE

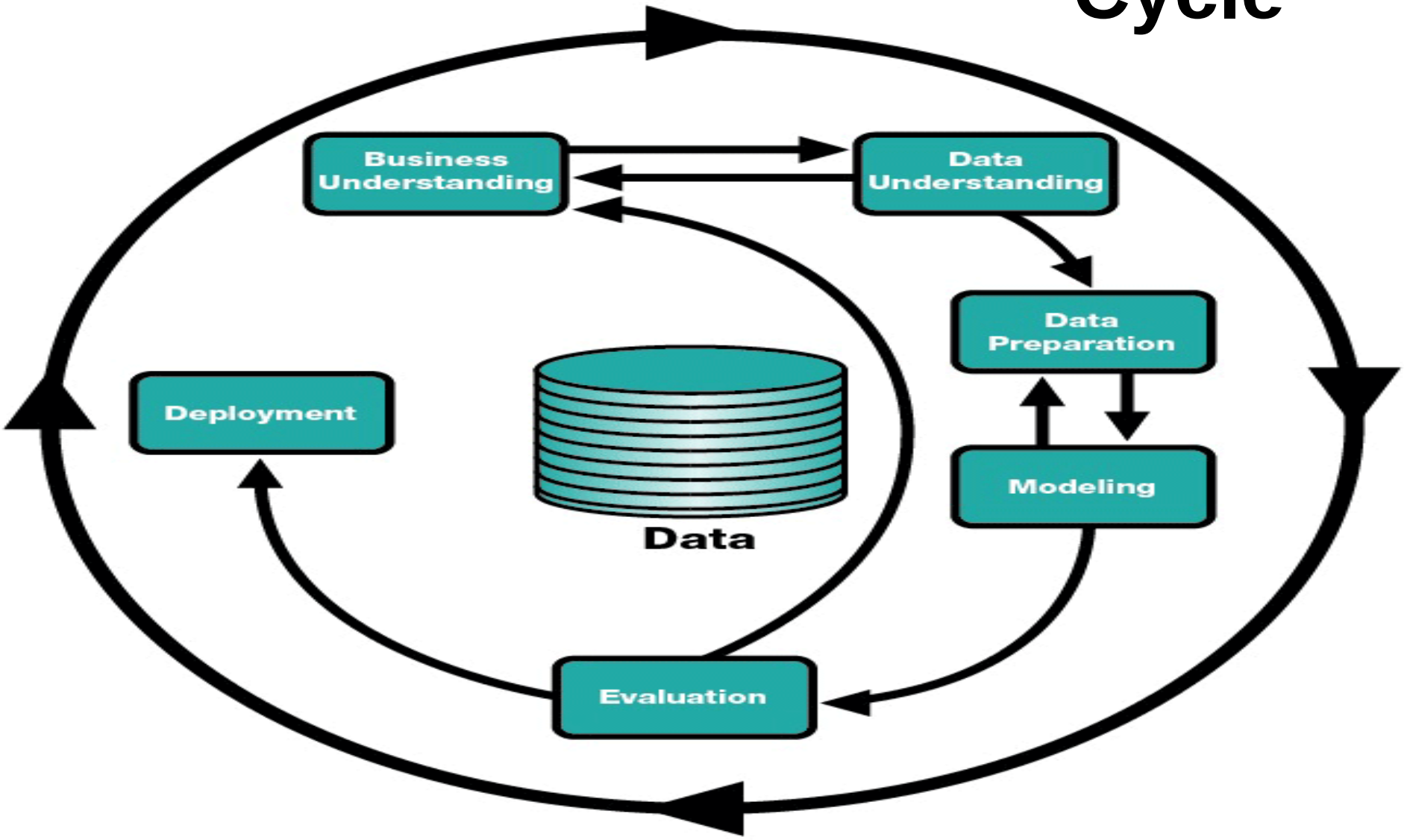
Dataset Description

Attributes Description:

1. **Number of times pregnant**
2. **Plasma glucose concentration a 2-hour in an oral glucose tolerance test**
3. **Diastolic blood pressure (mm Hg)**
4. **Triceps skinfold thickness (mm)**
5. **2-Hour serum insulin (mu U/ml)**
6. **Body mass index (weight in kg/(height in m)^2)**
7. **Diabetes pedigree function**
8. **Age (years)**
9. **Outcome - Class variable (0 or 1)**

UNC CHARLOTTE

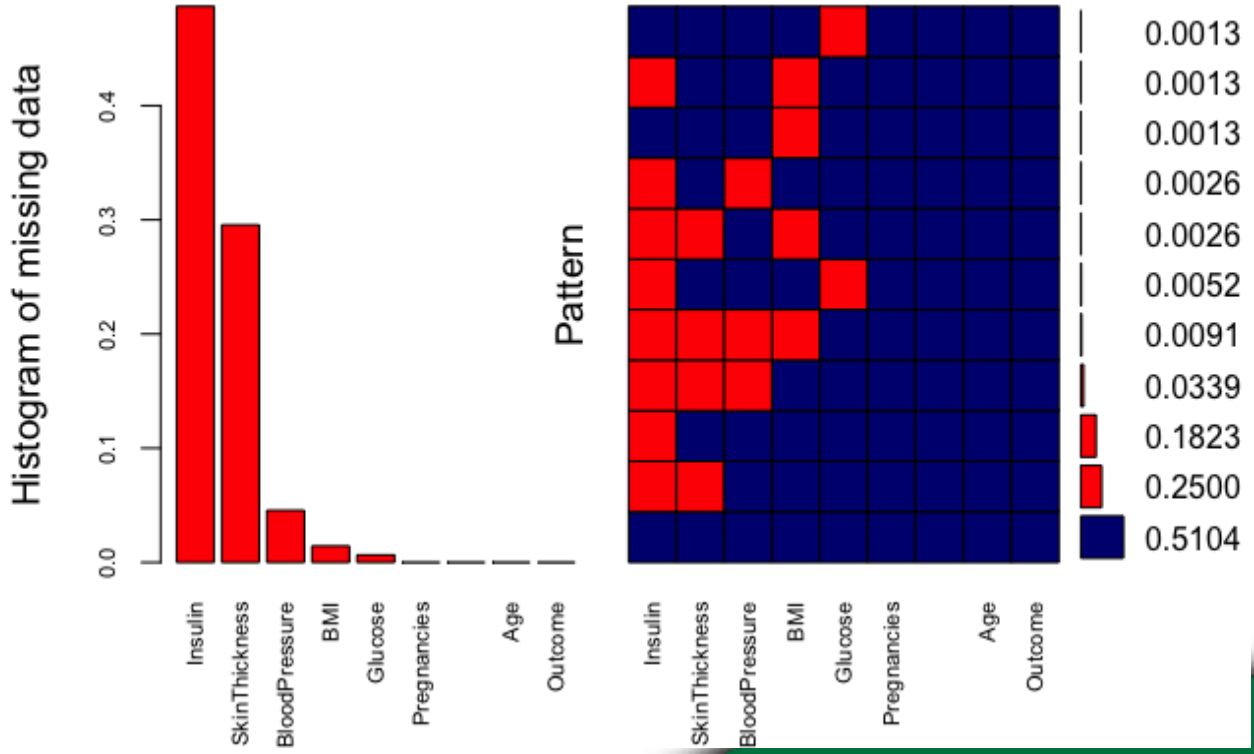
Project Cycle



UNC CHARLOTTE

Data Exploration

- The graph represents missing value proportions for each combination of attributes
- Red depicts missing values, whereas blue depicts valid data



UNC CHARLOTTE

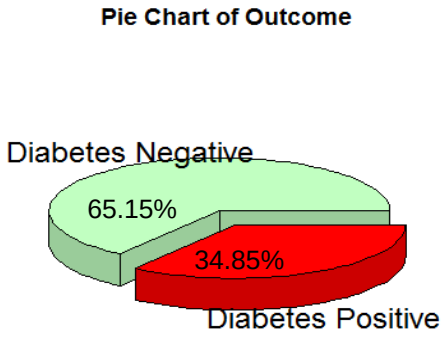
Handling Missing Values

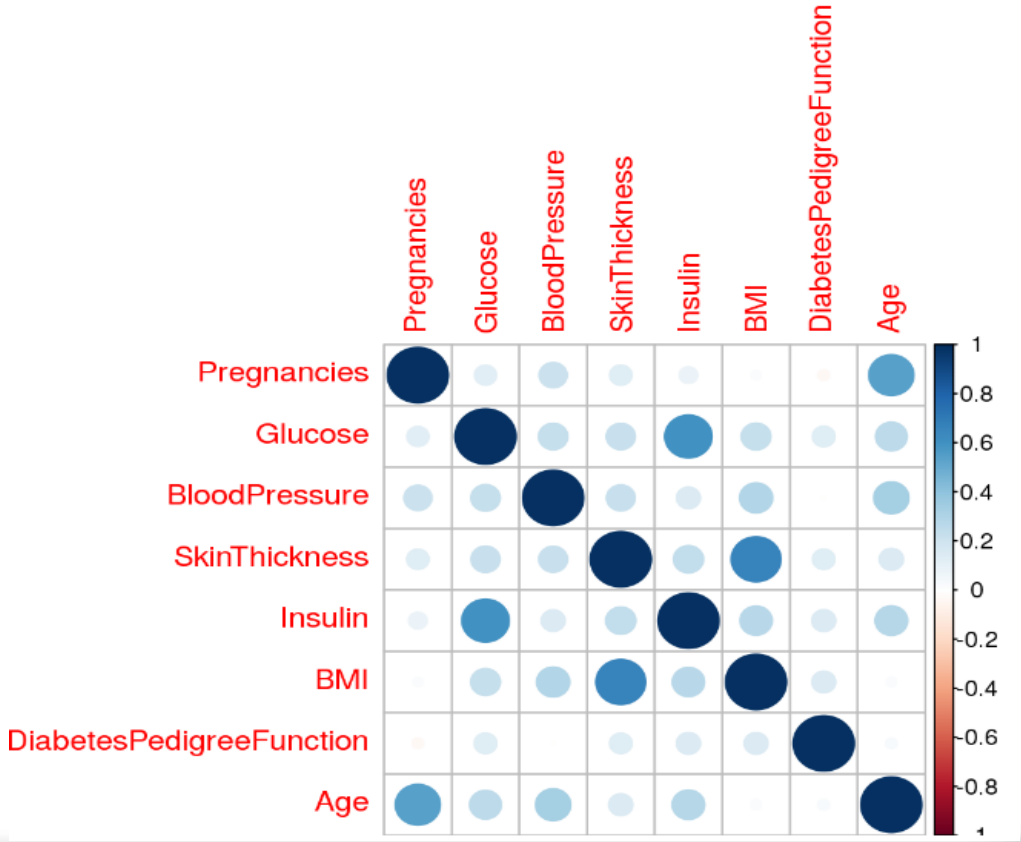
- 392 records were found to have a zero value for different attributes.
- Multivariate Imputation by Chained Equations (MICE) package of R has been used with “norm.predict” method.
- We tried using few MICE methods such as pmm, mean, norm. However, norm.predit gave us a very good imputation of missing data.

UNC CHARLOTTE

Data Analysis

- The Pie Chart represents the ratio of diabetic vs non diabetic PIMA females in the given data





- Left graph represents the correlation across all attributes

- SkinThickness~BMI and Insulin~Glucose were found to have a strong correlation

UNC CHARLOTTE

Observations

		Glucose		
		< 140	>= 140	Total
Diabetic?	False	438 (87.6%)	62 (12.4%)	500
	True	133 (49.6%)	135 (50.37%)	268
	Total	571	197	768

Table for Diabetic patients with Glucose < 140		BMI and Skin Thickness		
		< 26 & < 23	> 26 & > 23	Total
Diabetic?	True	24(10%)	119(90%)	133

		Pregnancies		
		< 5	>=5	Total
Diabetic?	False	356(72%)	144(52%)	500
	True	136(27%)	132(50%)	268
	Total	492	276(35)	768

		BMI		
		<= 25	>25	Total
Diabetic?	False	105(94%)	395(60%)	500
	True	7(6%)	261(99%)	268
	Total	112	656(84)	768

UNC CHARLOTTE

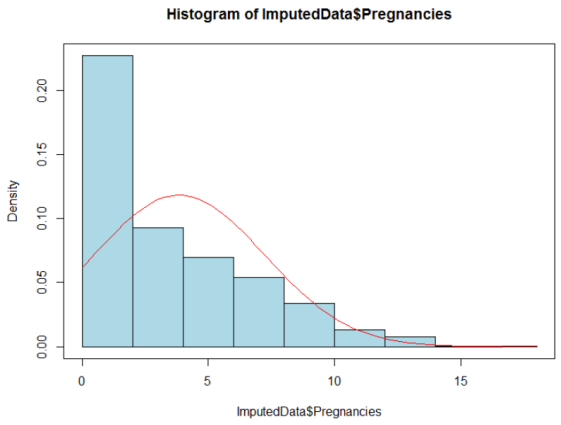
Data Assessment

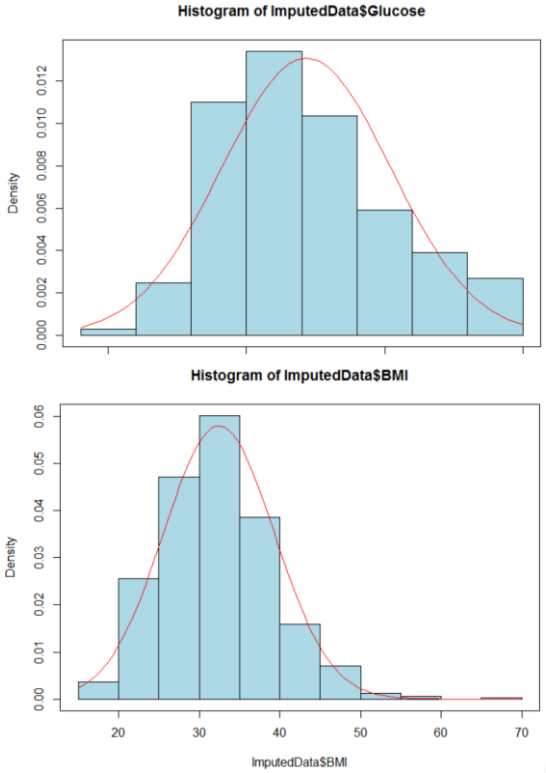
- Although Insulin is a predictor variable towards the outcome, it is not considered for further analysis in this phase since most of its values are imputed
- Although Age is not a strong predictor, we are considering an equal width binning for our analysis
- Glucose is the strongest contributor for our analysis; and we have categorized it as a flag variable

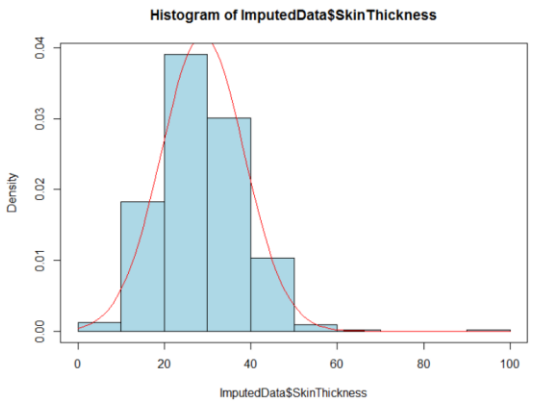
UNC CHARLOTTE

Predictor Variables

- Below graphs represent Histograms for four predictor variables before Transformation, i.e., Pregnancies, Glucose, Skin Thickness and BMI respectively.



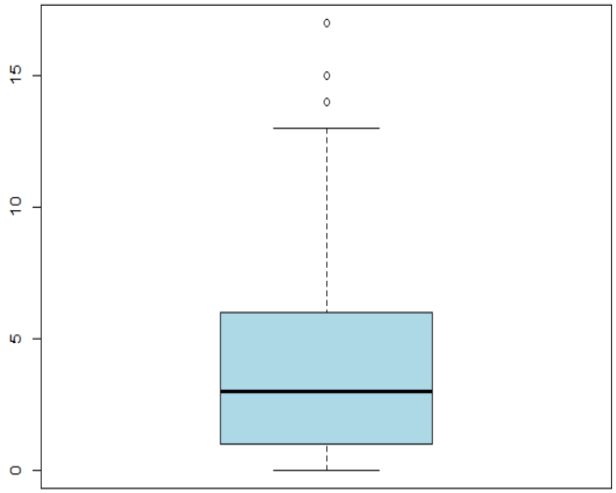


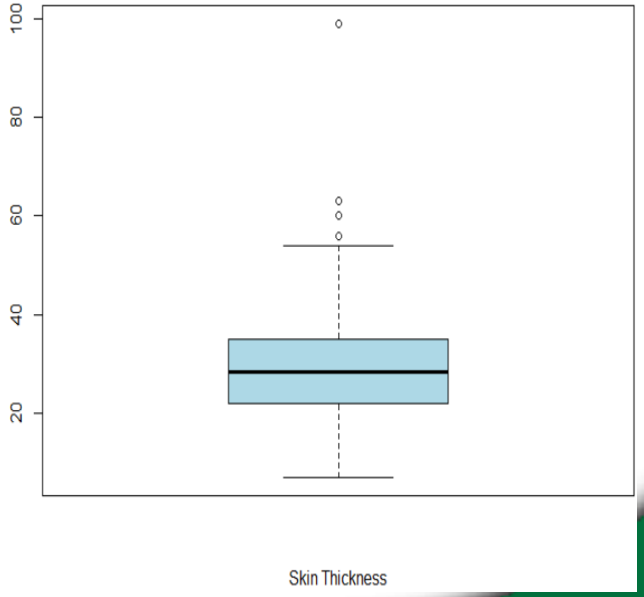


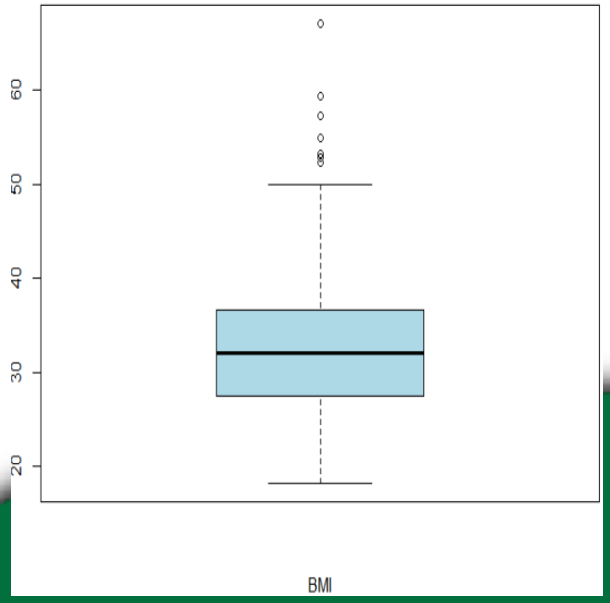
UNC CHARLOTTE

Outlier Detection

- Our Outlier detection and removal focuses only on the four strong predictor variables
- We found that Glucose does not have any Outliers, we checked and removed for the Outliers in the other three variables







UNC CHARLOTTE

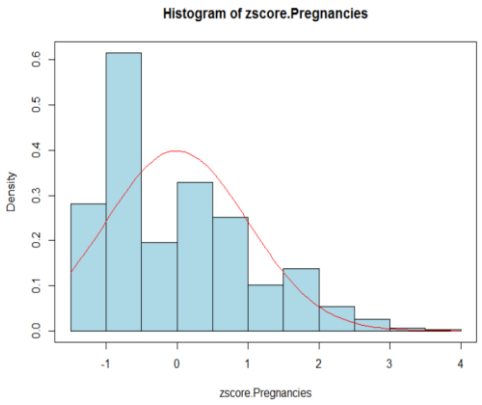
Outlier Removal

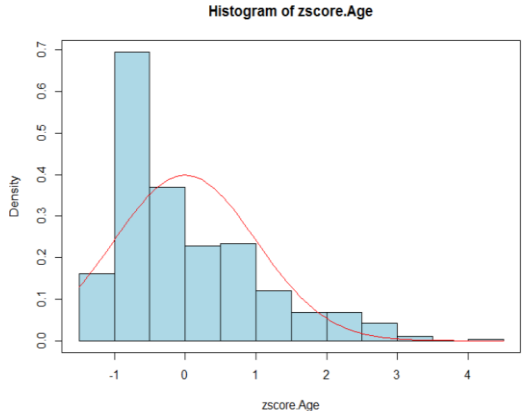
```
(BMI <- which(ImputedData$BMI %in% boxplot.stats(ImputedData$BMI)$out))
(Glucose <- which(ImputedData$Glucose %in% boxplot.stats(ImputedData$Glucose)$out))
(SkinThickness <- which(ImputedData$SkinThickness %in% boxplot.stats(ImputedData$SkinThickness)$out))
(Pregnancies <- which(ImputedData$Pregnancies %in% boxplot.stats(ImputedData$Pregnancies)$out))

all_outliers <- Reduce(union, list(BMI,Glucose,SkinThickness,Pregnancies))
all_outliers

removeOutliers <- function(data, indicesOfAllOutliers) {
  for (x in 1:nrow(indicesOfAllOutliers)){
    i = indicesOfAllOutliers[x,]
    data <- data[-c(i), ]
  }
  return(data)
}

transformedData<-removeOutliers(ImputedData, as.data.frame(all_outliers))
```

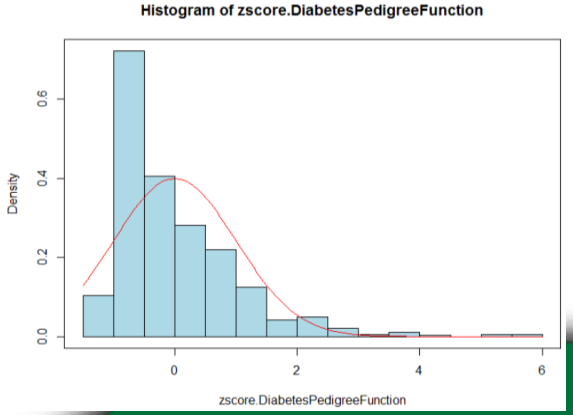
UNC CHARLOTTE

Normalization

- The following three variables were not in normalized form
  - Pregnancies, Age, Diabetes Pedigree Function
- Hence we employed Z-Score Transformation on these variables
- Below graphs represent Z-Score Transformed histograms for the above variables







UNC CHARLOTTE

Data Partitioning and Evaluation

- The Dataset was split into Train and Test with the Sample. Split function was provided in the caTools library. Split Ratio of 70% was employed.
- Below t-test was performed on the partitions to validate the partitioning;
  - T-test for Glucose : p-value = 0.57
  - T-test for SkinThickness: p-value = 0.41
  - T-test for BMI: p-value = 0.89
  - T-test for Pregnancies: p-value = 0.16
- Based on the results above we conclude the partition is valid (p-value>0.05).

UNC CHARLOTTE

Data Modeling

Unsupervised Learning Method

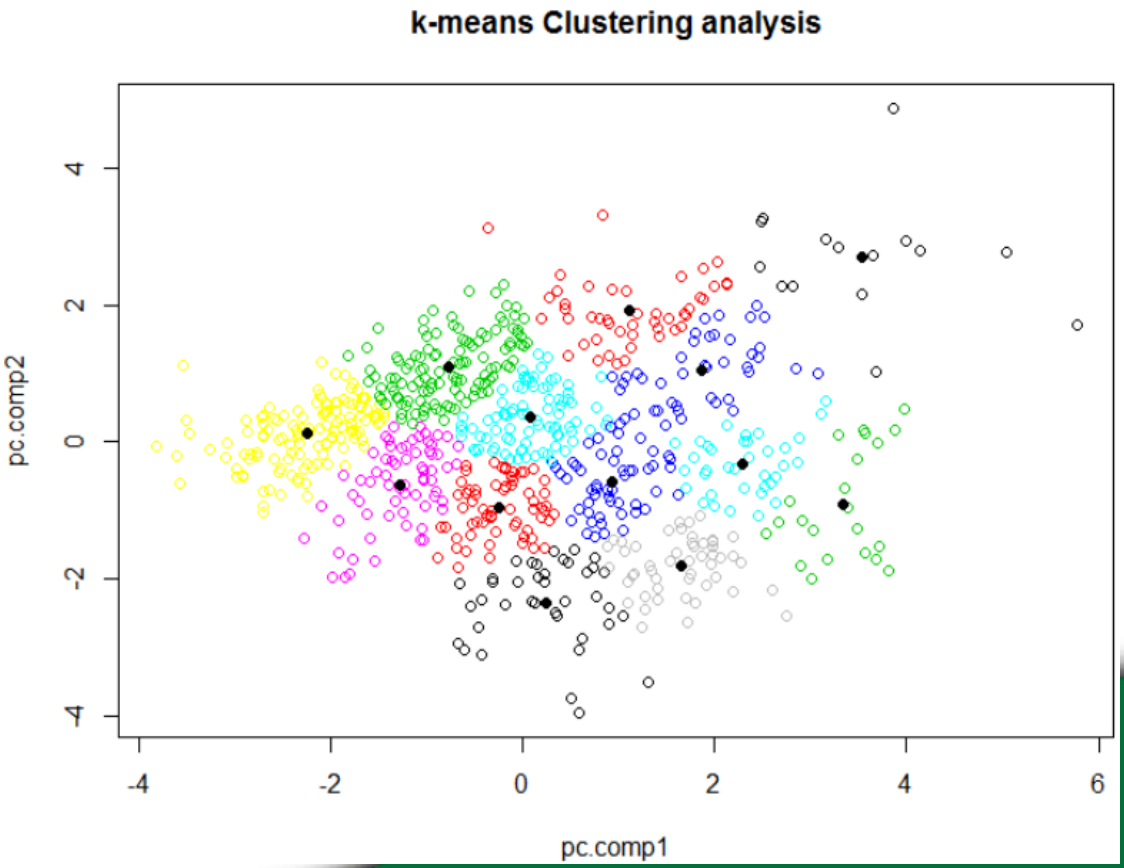
- K-means Clustering

➢ Cluster data based on their similarity.

➢ OUTCOME column removed for this technique; and the algorithm just tries to find patterns in the data.

➢ For the convenience of visualization, we have taken the first two principal components as the new feature variables and conduct k-means only on these two dimensional data.

➢ Figure shows the resulting scatter plot with different clusters in different colors. The solid black circles are the centers of the clusters.

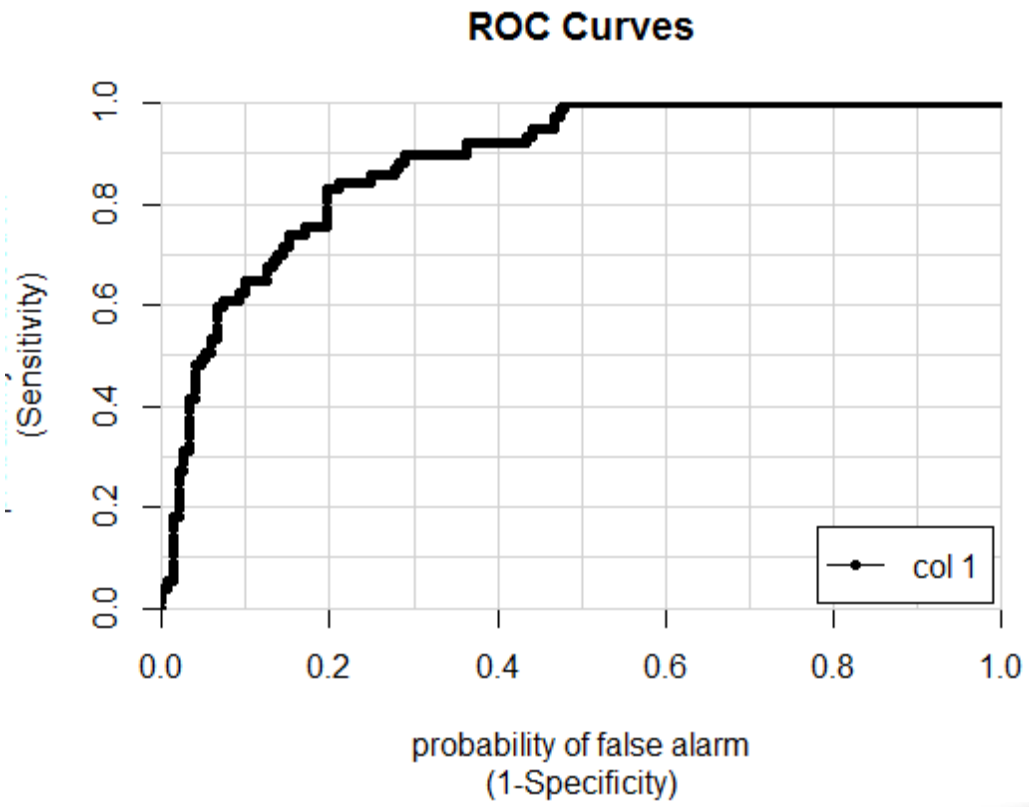


UNC CHARLOTTE

Data Modeling (Contd..)

Supervised Learning Method

- Logistic Regression



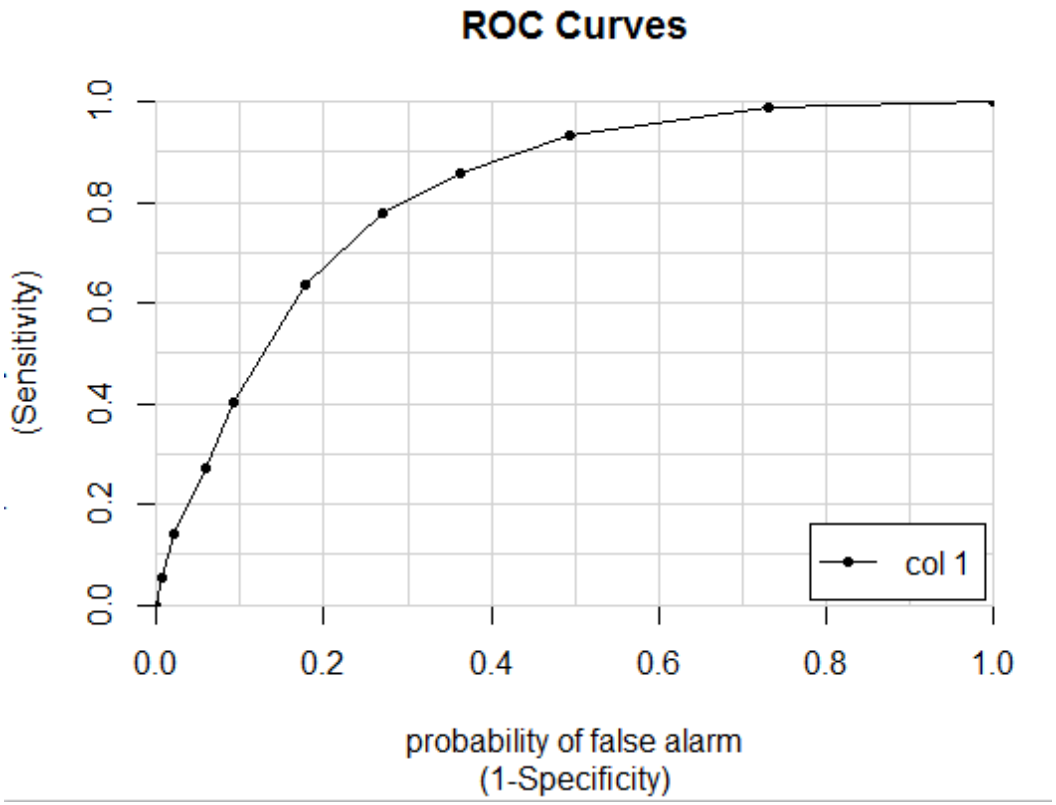
Result from Logistic Regression
AUC = 0.835292 (not bad)

Accuracy : 0.7719
Sensitivity : 0.5417
Specificity : 0.8782
Pos Pred Value : 0.6724
Neg Pred Value : 0.8059
Balanced Accuracy : 0.7099

UNC CHARLOTTE

Data Modeling (Contd..)

- k-Nearest Neighbor



Result from k-Nearest Neighbor

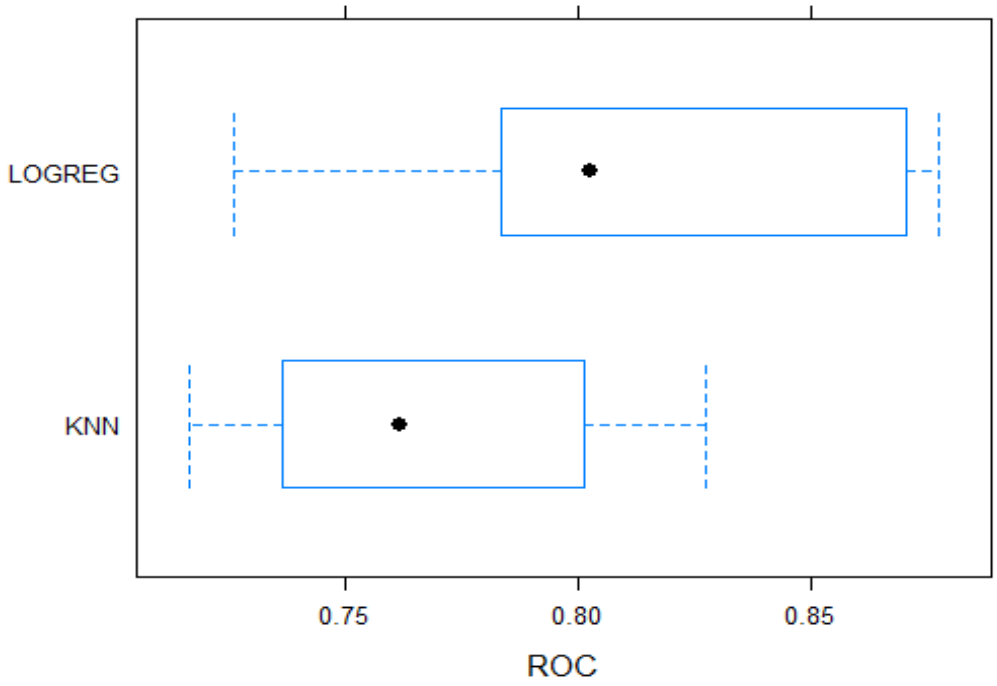
Accuracy : 0.7193
Sensitivity : 0.6111
Specificity : 0.7692
Pos Pred Value : 0.5500
Neg Pred Value : 0.8108
Balanced Accuracy : 0.6902

UNC CHARLOTTE

Data Modeling (Contd..)

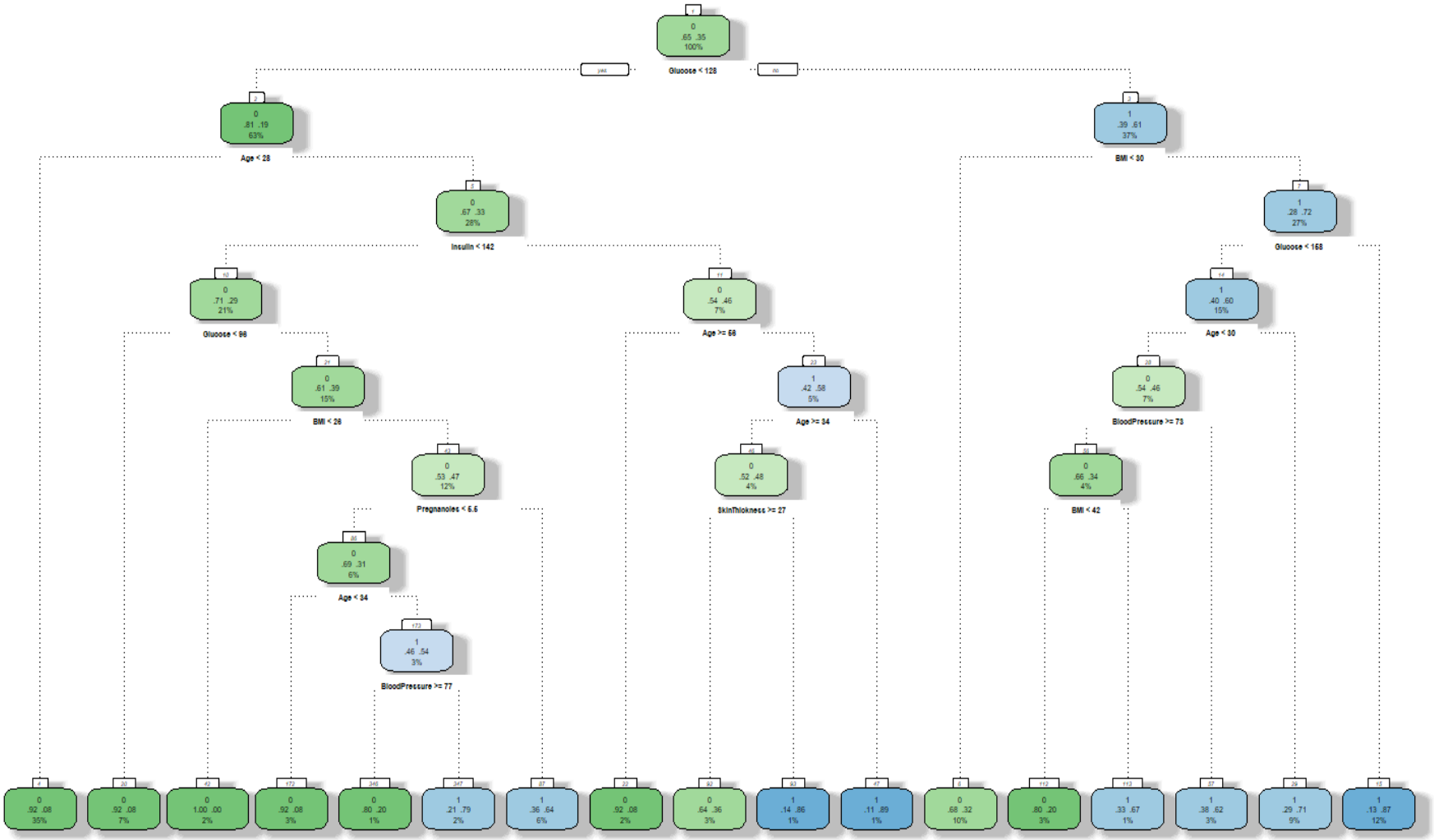
Models Comparison

```
Models: LOGREG, KNN
Number of resamples: 10

ROC
         Min. 1st Qu. Median   Mean 3rd Qu.   Max. NA's
LOGREG 0.7221  0.7973 0.8529 0.8325  0.8784 0.9056    0
KNN    0.7051  0.7410 0.7894 0.7884  0.8213 0.8943    0

Sens
         Min. 1st Qu. Median   Mean 3rd Qu.   Max. NA's
LOGREG 0.7143  0.8088 0.8693 0.8634  0.9273 0.9706    0
KNN    0.7353  0.7721 0.7971 0.7965  0.8176 0.8571    0

Spec
         Min. 1st Qu. Median   Mean 3rd Qu.   Max. NA's
LOGREG 0.4211  0.5572 0.6500 0.6124  0.6875 0.7368    0
KNN    0.2632  0.5250 0.6408 0.6153  0.6961 0.9000    0
```



UNC CHARLOTTE

Decision Tr



Rattle 2017-Nov-25 19:10:36 Sharath

UNC CHARLOTTE

Data Modeling (Contd..)

Association*(Using "aPriori" algorithm)*

- {Age=[21.0,32.5), Outcome=1} => {Pregnancies=[0.00, 4.19)}
- {Glucose=[150,199], BMI=[29.2,38.0)} => {Outcome=1}
- {Glucose=[44,112), Outcome=1} => {Insulin=[-20.8,158.7)}
- {Glucose=[150,199], DPF=[0.476,1.030)} => {Outcome=1}
- {Pregnancies=[4.19, 9.02), DPF=[1.030,2.420]} => {Outcome=1}
- {Insulin=[365.2,846.0], Age=[48.5,81.0]} => {Outcome=1}

UNC CHARLOTTE

Conclusion

- We have developed four models. It gives us the results of prediction and accuracy – shows which are the most important factors to have diabetes.
- Higher the number of pregnancies count, higher the probability of getting the diabetes.
- As per the observations, as a person reaches a near diabetic phase, it is found that BMI and ST increase in tandem. This leads to insulin resistance

UNC CHARLOTTE

Future Scope

- If we get and correlate the cardiac history of the patients we can build a better model with better accuracy
- With a better (more complete) Insulin records, we can build a much better model
- Try to build different machine learning models and possibly have a stacked one

UNC CHARLOTTE

References

1. https://en.wikipedia.org/wiki/Pima_people
2. http://www.srpmic-nsn.gov/history_culture/
3. http://care.diabetesjournals.org/content/29/8/1866
4. https://www.kaggle.com/uciml/pima-indians-diabetes-database
5. https://rpubs.com/ikodesh/53189
6. https://www.youtube.com/watch?v=pN4HqWRybwk
7. http://www.diabetes.co.uk/diabetes_care/blood-sugar-level-ranges.html
8. http://machinelearningmastery.com/
9. https://www.r-bloggers.com/self-organising-maps-for-customer-segmentation-using-r/
10. https://onlinecourses.science.psu.edu/stat857/node/125