BAN 620
Data Mining

Team 3

Kalpita Nimje

Koya Sai Teja

Navya Arveti

Kovvuru Saicharanreddy

Sri Vinuta Sai Pinisetty

Problem Statement

1/10
(20-79 years)

37.39mn
diabetic patients

8.3mn
undiagnosed

1 death
every 5 seconds
in 2021

COVID-19

50%
more likely

Source: diabetes.org

Problem Statement

Role of Analytics

Improvement in provisions

Challenge to diagnose

Project Motivation

Improve Diagnosis

<u>Motivation:</u> Current pandemic conditions + role of analytics + skills & knowledge

Research Questions

What? **Prediction**

- Can diabetes be predicted from basic details and vitals?

How? **Methods**

- What methods can be adopted for prediction?

How much? **Accuracy**

- How accurately can we predict?

Why? **Reasons causing Diabetes**

- What vitals and factors in humans make them susceptible to diabetes?

Dataset Description and Details

Dataset	Source	Shape	Datatypes
Diabetes Dataset	kaggle Click to follow	9 attributes 768 records	All variables – Float Outcome is binary (0 or 1)

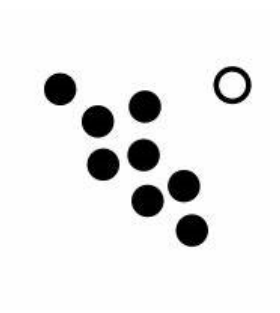
Variable	Description
Pregnancies	Number of times pregnant
Glucose	Plasma glucose concentration at 2 hours in an oral glucose tolerance test
Blood Pressure	Diastolic blood pressure (mm Hg)
SkinThickness	Triceps skin fold thickness (mm)
Insulin	2-Hour serum insulin (mu U/ml)

Variable	Description
BMI	Body mass index (weight in kg/(height in m)^2)
DiabetesPedigreeFunction	Diabetes pedigree function
Age	Age in years
Class (Outcome Variable)	Prediction if the patient has diabetes

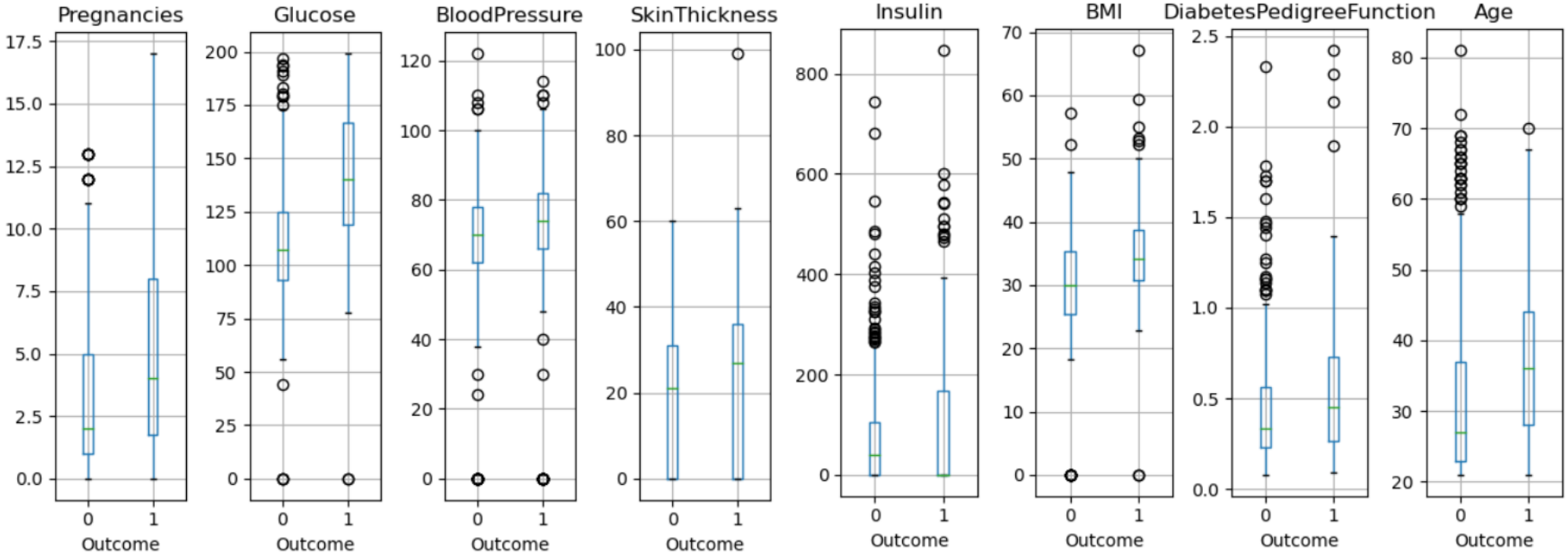
Variables Description

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.00	768.00	768.00	768.00	768.00	768.00	768.00	768.00	768.00
mean	3.85	120.89	69.11	20.54	79.80	31.99	0.47	33.24	0.35
std	3.37	31.97	19.36	15.95	115.24	7.88	0.33	11.76	0.48
min	0.00	0.00	0.00	0.00	0.00	0.00	0.08	21.00	0.00
25%	1.00	99.00	62.00	0.00	0.00	27.30	0.24	24.00	0.00
50%	3.00	117.00	72.00	23.00	30.50	32.00	0.37	29.00	0.00
75%	6.00	140.25	80.00	32.00	127.25	36.60	0.63	41.00	1.00
max	17.00	199.00	122.00	99.00	846.00	67.10	2.42	81.00	1.00

Preprocessing Data

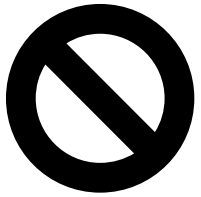


Outliers



Null Values

There are **0 null values** in the dataset

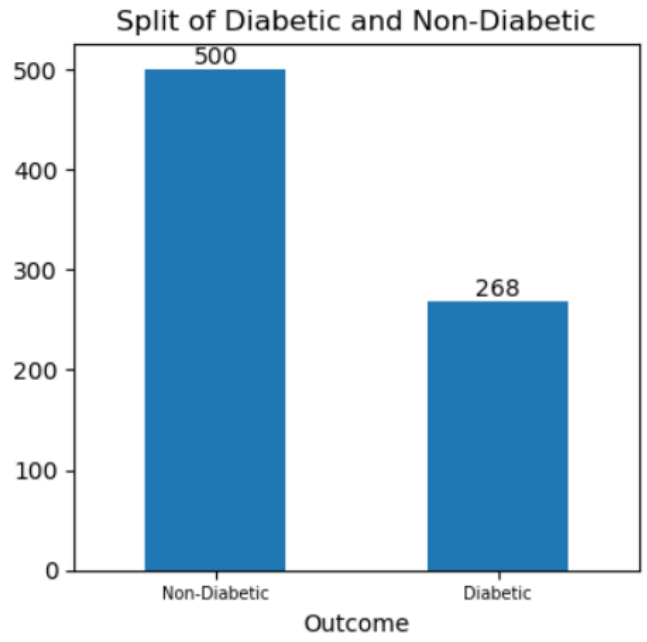


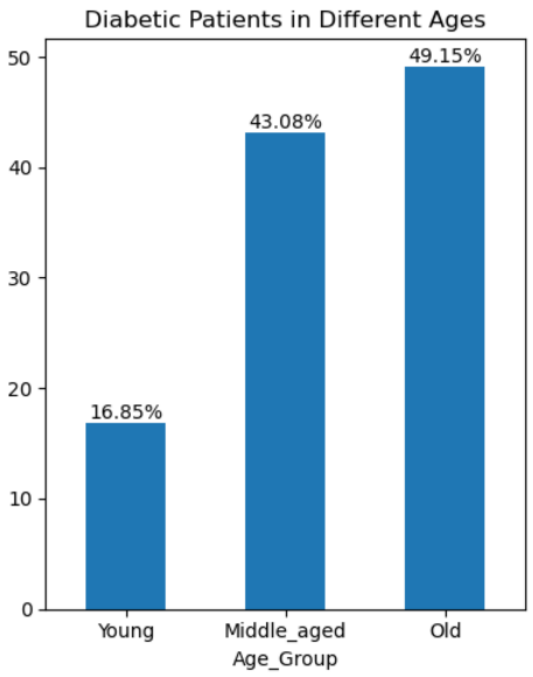
Inappropriate Values

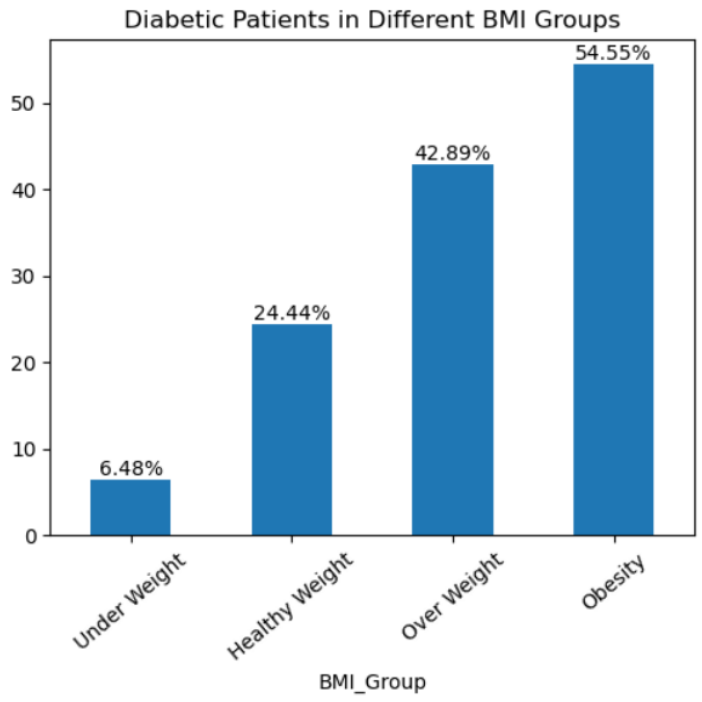
Pregnancies	0
Glucose	5
BloodPressure	35
SkinThickness	227
Insulin	374
BMI	11
DiabetesPedigreeFunction	0
Age	0
Outcome	0

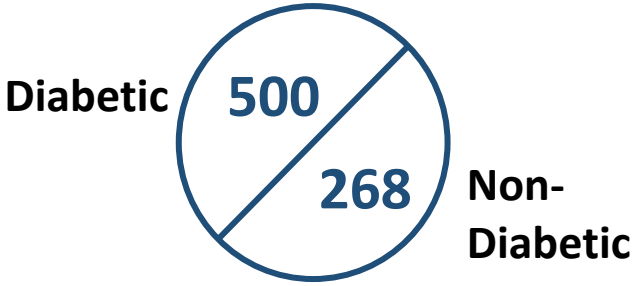
Replaced with **mean**

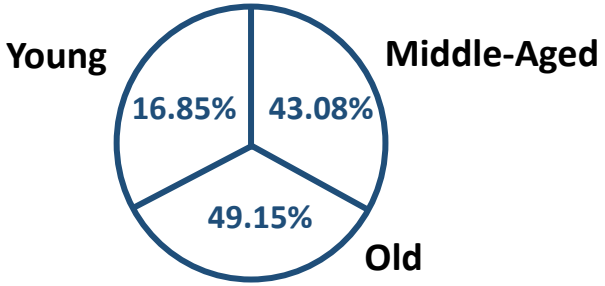
Initial Analysis

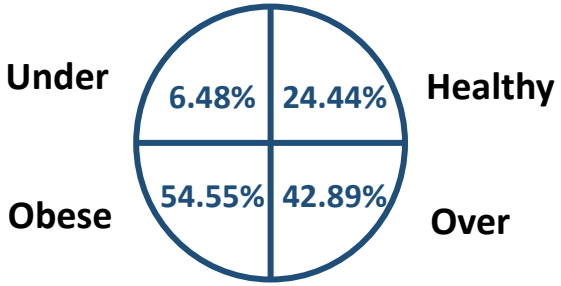




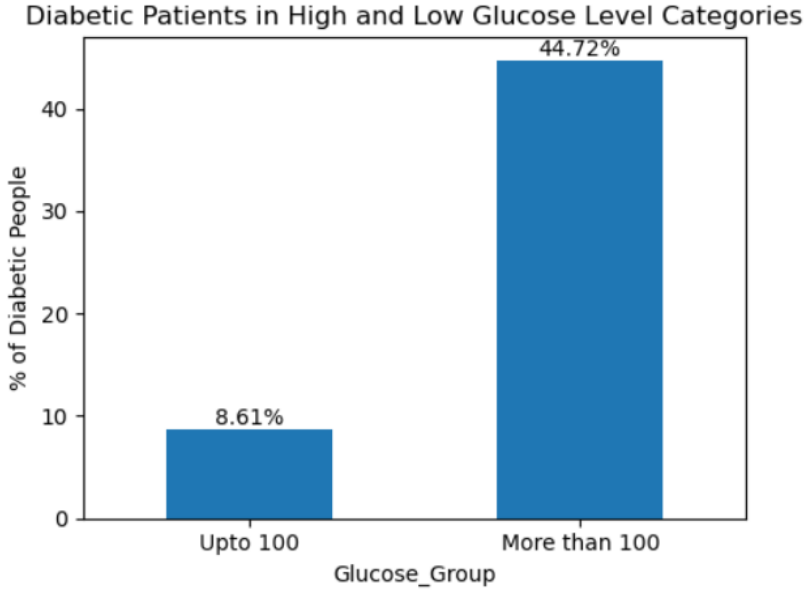


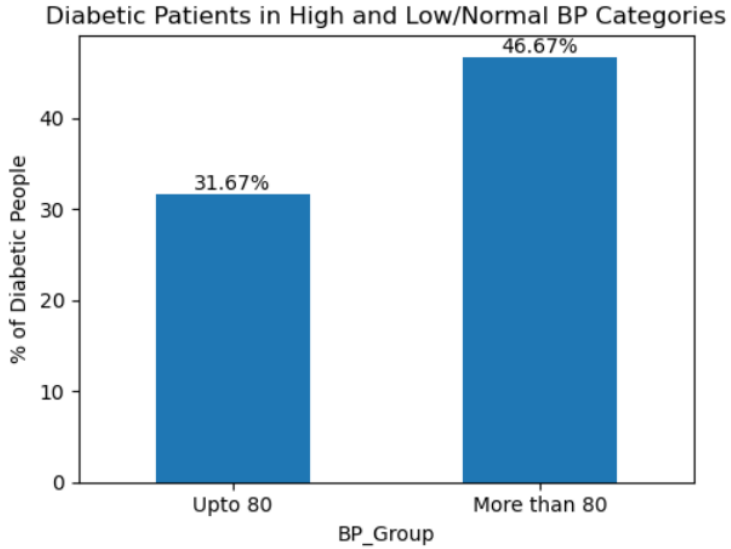


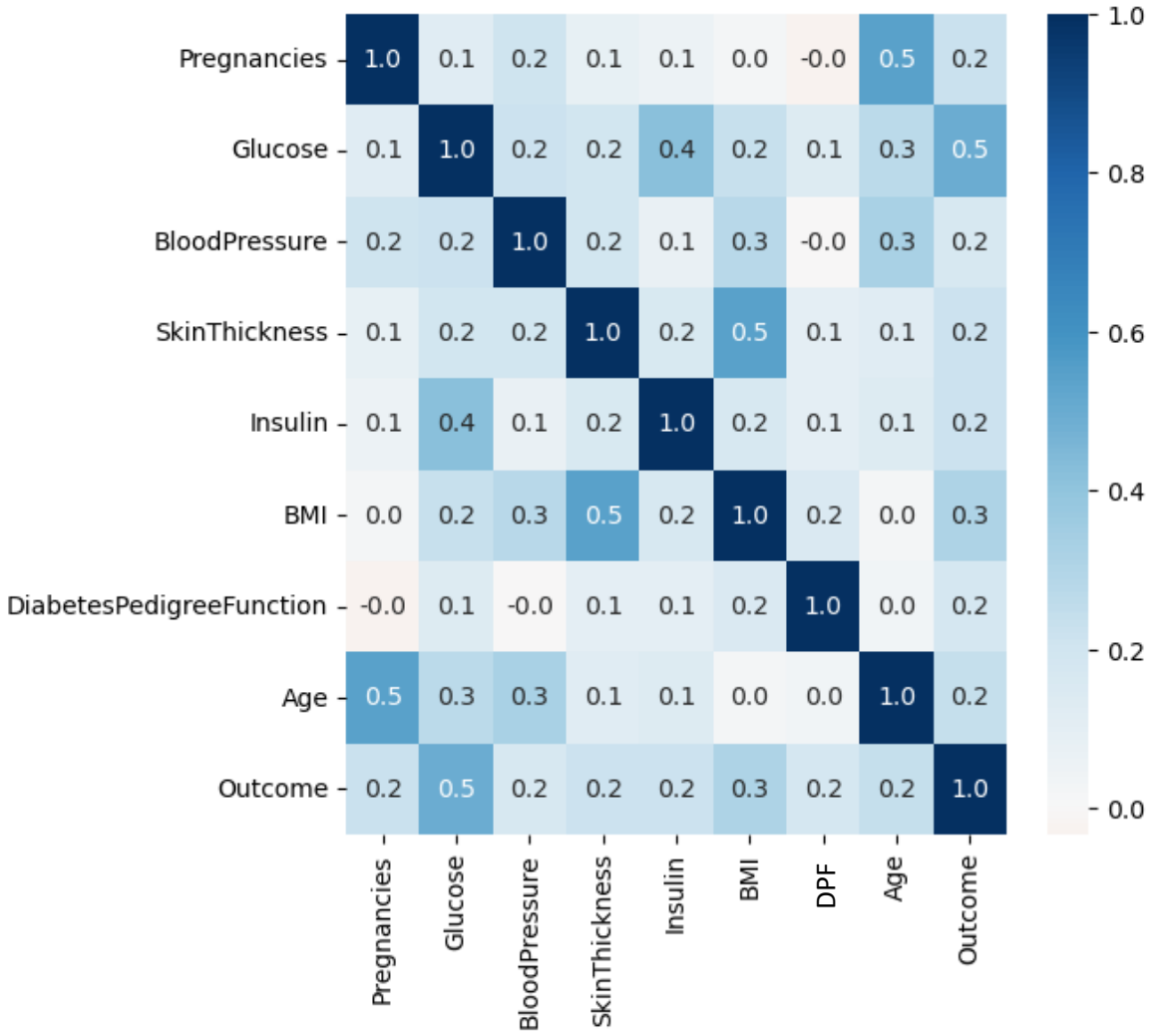




Initial Analysis





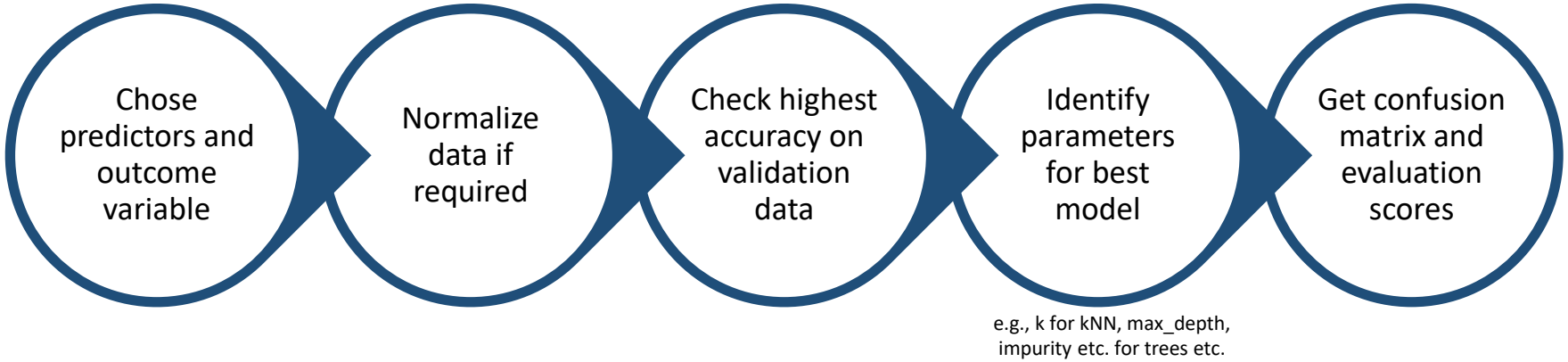


Insights:

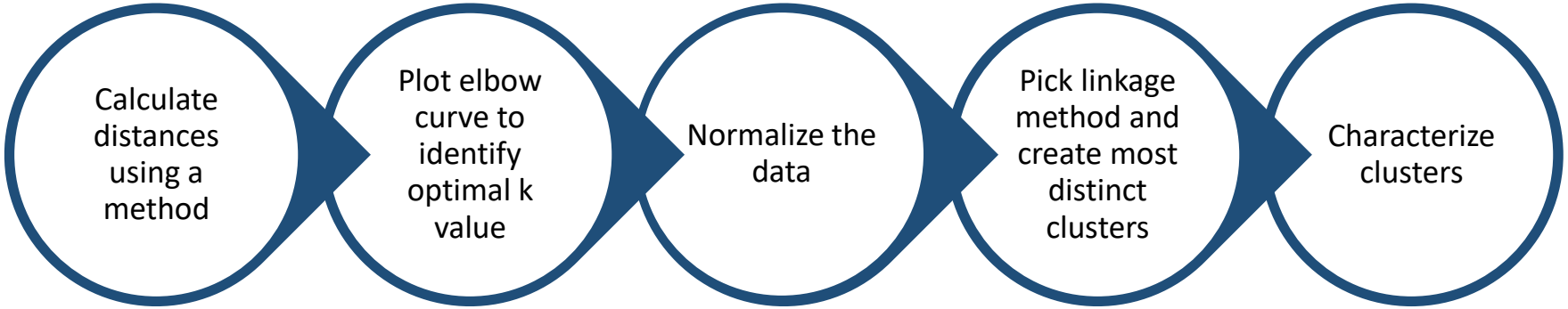
1. **Bar Plot:** Diabetes is prevalent in patients with high glucose and BP levels
2. **Map:** Age <-> no. of pregnancies; Skin Thickness <-> BMI; Glucose <-> Insulin

Analysis and Results – Model Building

kNN Classifier, Decision Trees, Random Forest, Boosted Trees, Neural Networks, Logistic Regression



Hierarchical Clustering



kNN and Decision Trees

1.1

kNN Classifier

Best accuracy in validation data is obtained at **k = 13**

Actual	Prediction	
	0	1
0	95	12
1	16	31

Accuracy Score: 81.82%
Precision Score: 72.09%
Recall Score: 65.96%
f1 Score: 68.89%

Prediction: For any new patient, we would identify the 13 nearest neighbors and predict if the patient is based on the majority among these 13 neighbors

2.1

Decision Tree with maximum depth = 2

Actual	Prediction	
	0	1
0	88	11
1	21	34

Accuracy Score: 79.22%
Precision Score: 75.56%
Recall Score: 61.82%
f1 Score: 68.0%

2.2

Tree with minimum impurity= 0.01

Actual	Prediction	
	0	1
0	87	12
1	16	39

Accuracy Score: 81.82%
Precision Score: 76.47%
Recall Score: 70.91%
f1 Score: 73.58%

Decision Trees

2.3

Tree with user defined parameters

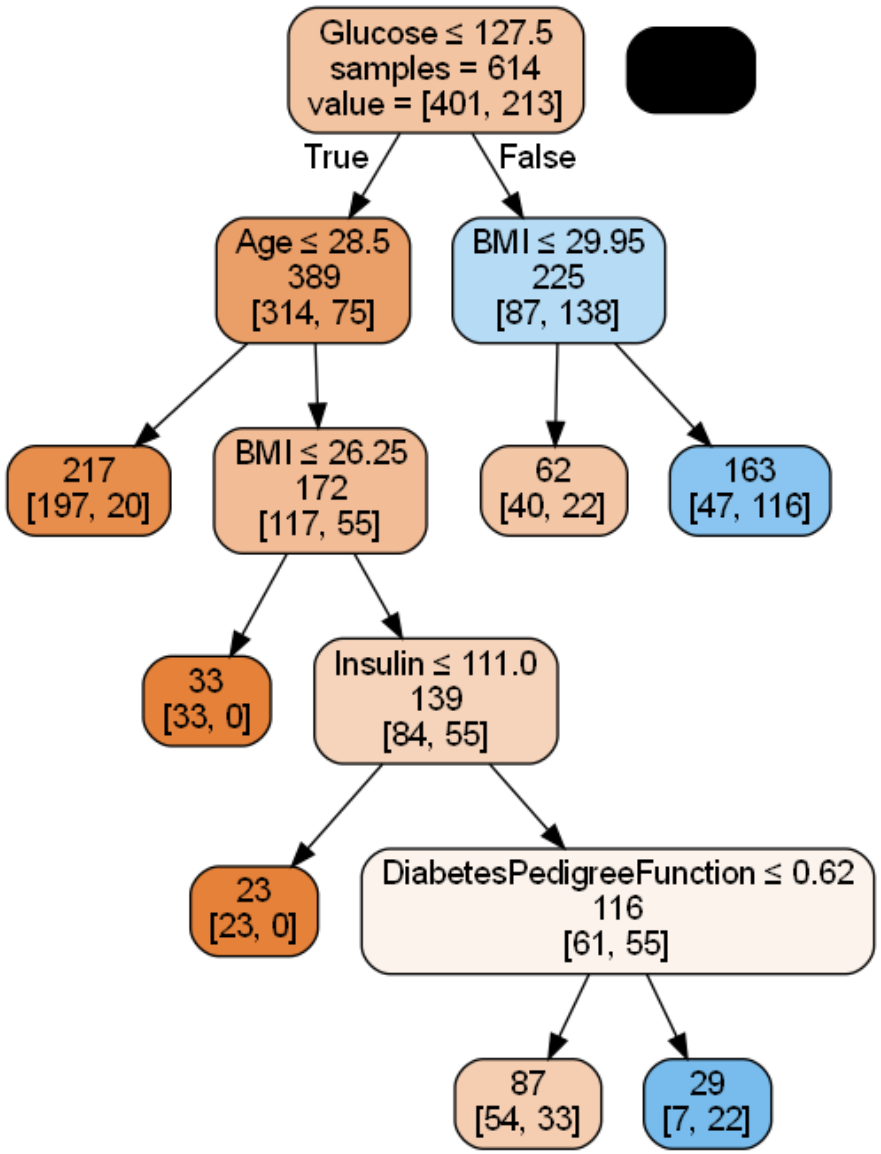
Max depth = 8
Minimum impurity decrease = 0.01
Minimum number of samples = 30

Actual	Prediction	
	0	1
0	87	12
1	16	39

Accuracy Score: 81.82%
Precision Score: 76.47%
Recall Score: 70.91%
f1 Score: 73.58%

Observation: The tree obtained by this method is the same as obtained in 2.2. This is also, by far, the best prediction model

Tree with user defined parameters



Decision Trees

2.4

Best Tree

Best tree is obtained at:
Max depth = 11
Minimum impurity decrease = 0
Minimum number of samples = 76

Actual	Prediction	
	0	1
0	83	16
1	15	40

Accuracy Score: 79.87%
Precision Score: 71.43%
Recall Score: 72.73%
f1 Score: 72.07%

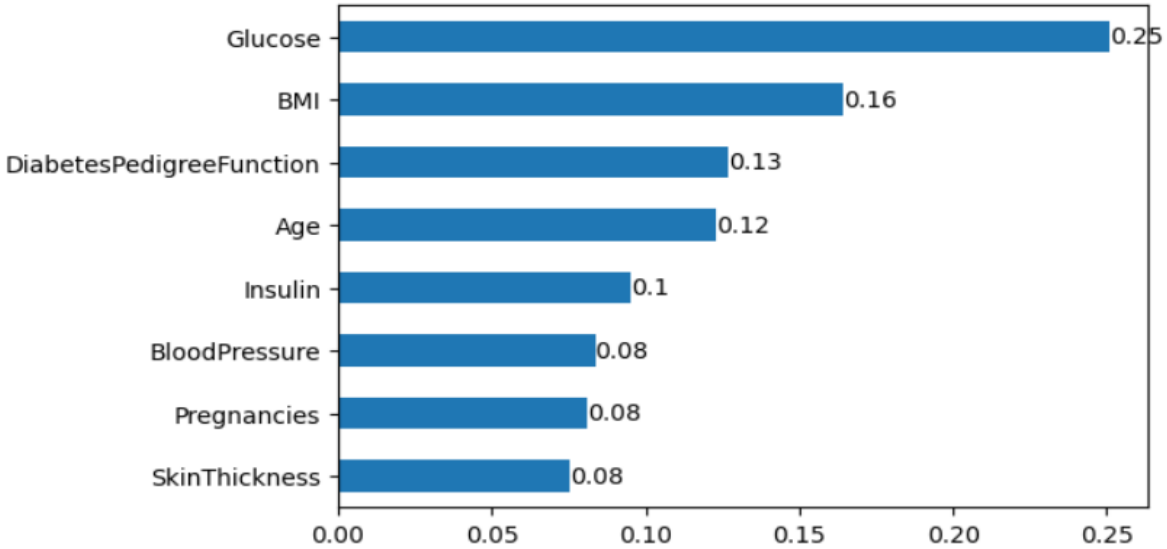
3

Random Forest

Actual	Prediction	
	0	1
0	89	10
1	20	35

Accuracy Score: 80.52%
Precision Score: 77.78%
Recall Score: 63.64%
f1 Score: 70.0%

Feature Importance



Trees, Neural Networks

4

Boosted Trees

Actual	Prediction	
	0	1
0	86	13
1	19	36

Accuracy Score: 79.22%
Precision Score: 73.47%
Recall Score: 65.45%
f1 Score: 69.23%

5

Neural Networks

Actual	Prediction	
	0	1
0	430	70
1	127	141

Accuracy Score: 74.35%
Precision Score: 66.82%
Recall Score: 52.61%
f1 Score: 58.87%

Observations:

- In neural network, the model cannot be explained. Also, the accuracy we obtain is not the highest.
- The model is run on the complete dataset and not on training data. The prediction is done on the new data directly.

Logistic Regression

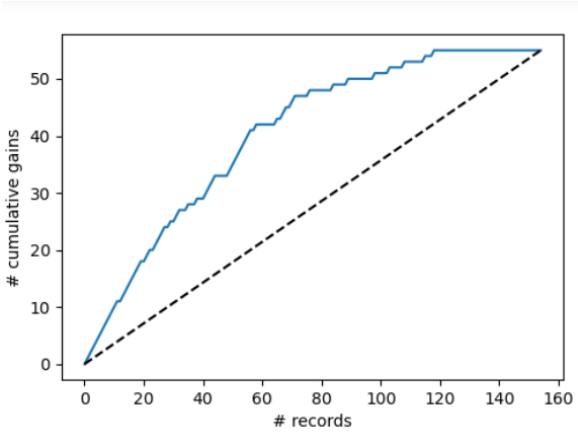
6.1

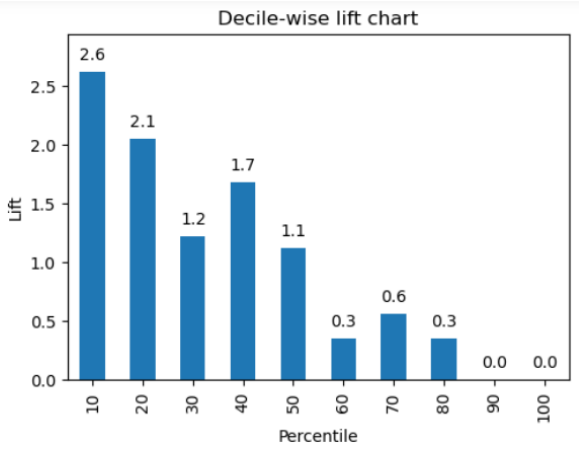
	coef	std err	z	P>|z|	[0.025	0.975]
Pregnancies	0.1338	0.029	4.561	0.000	0.076	0.191
Glucose	0.0225	0.003	7.030	0.000	0.016	0.029
BloodPressure	-0.0595	0.007	-8.493	0.000	-0.073	-0.046
SkinThickness	-0.0075	0.011	-0.676	0.499	-0.029	0.014
Insulin	-0.0003	0.001	-0.320	0.749	-0.002	0.002
BMI	0.0195	0.015	1.336	0.181	-0.009	0.048
DiabetesPedigreeFunction	0.3733	0.249	1.497	0.134	-0.115	0.862
Age	-0.0023	0.009	-0.264	0.792	-0.020	0.015

Observation: The coefficients pregnancies, glucose and BP are statistically significant

Actual	Prediction	
	0	1
0	88	11
1	24	31

Accuracy Score: 77.27%
Precision Score: 73.81%
Recall Score: 56.36%
f1 Score: 63.92%





6.2

Logistic Regression-different cutoff

Cut off chosen is 0.4

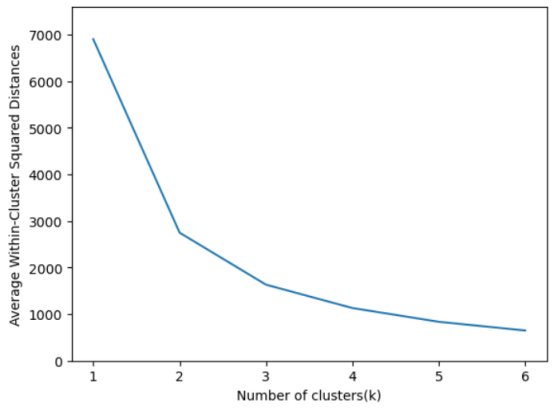
Actual	Prediction	
	0	1
0	84	15
1	15	40

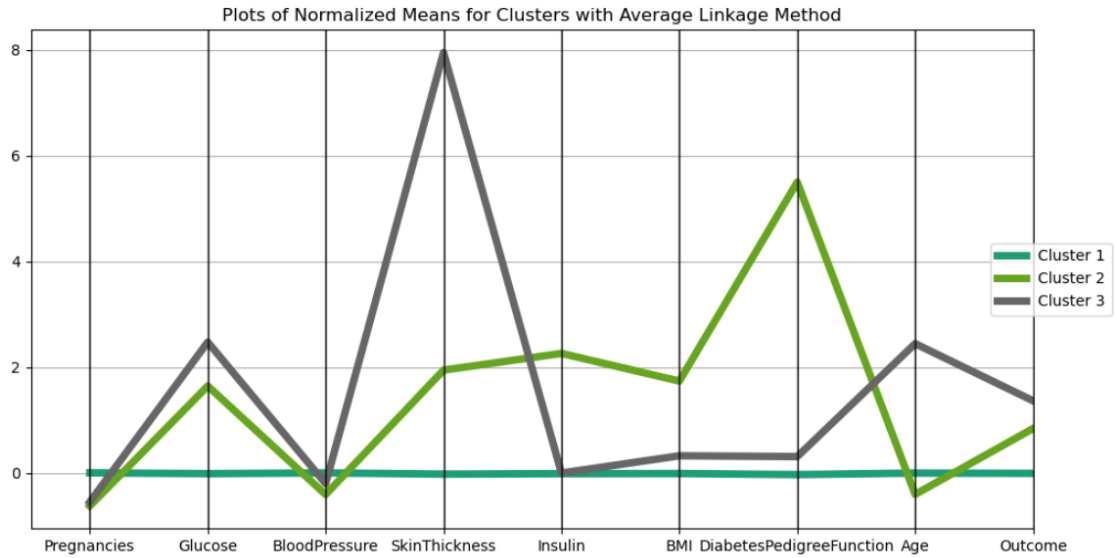
Accuracy Score: 80.52%
Precision Score: 72.73%
Recall Score: 72.73%
f1 Score: 72.73%

Clustering

7

Hierarchical Clustering



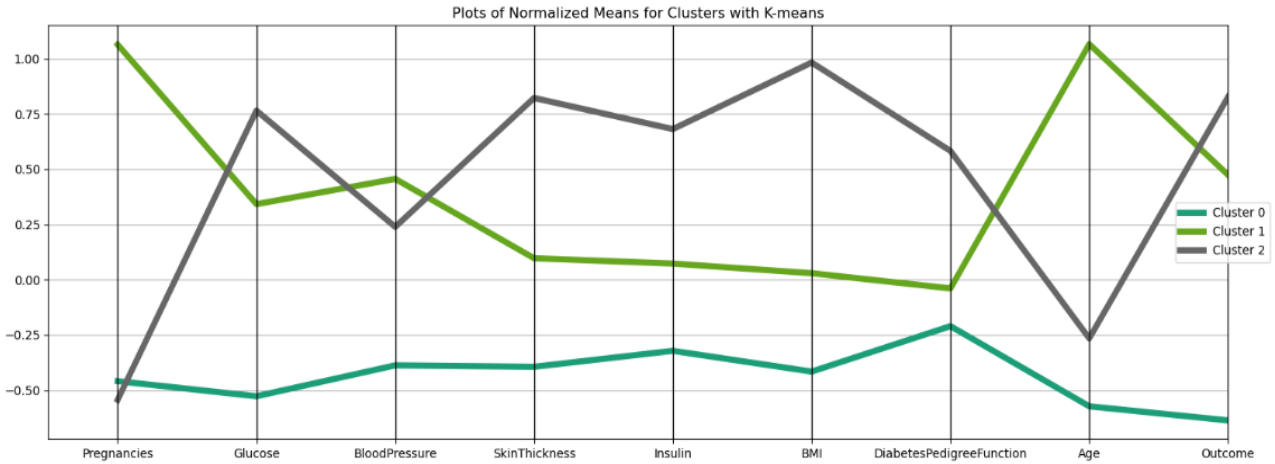


Analysis: Clusters don't help in good characterization

8

Non-hierarchical Clustering

K-Means Clustering



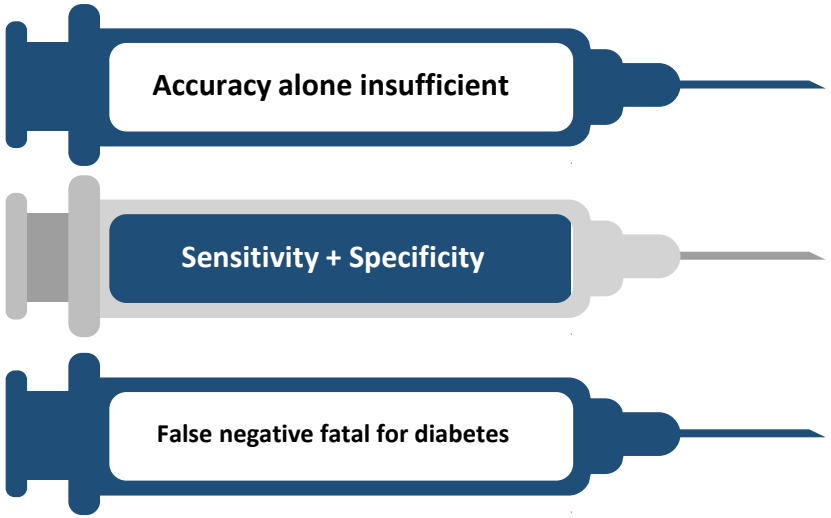
Analysis: Good characterization of clusters

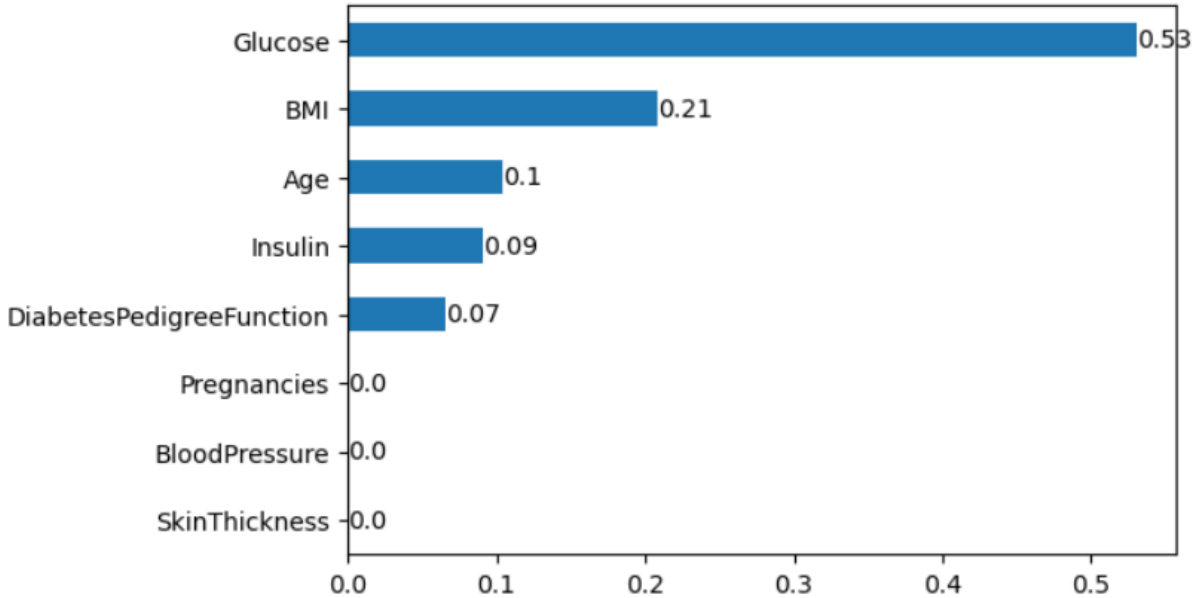
Cluster 0: Low across all values – not diabetic prone
Cluster 1: Aged, high number of pregnancies (female dominated), high BP – diabetic
Cluster 2: Young, high skin thickness, high insulin, high dpf, obese segment – highly diabetic

% of diabetic patients in clusters	
Cluster 0	4.52%
Cluster 1	57.50%
Cluster 2	74.34%

Comparison of Different Models

Model	Accuracy	Recall	Precision	F1 Score
kNN Classifier	81.82%	65.96%	72.09%	68.89%
Decision Tree with max depth = 2	79.22%	61.82%	75.56%	68.0%
Tree with minimum impurity = 0.01	**81.82%**	**70.91%**	**76.47%**	**73.58%**
Best Tree	79.87%	72.73%	71.43%	72.07%
Random Forest	80.52%	63.64%	77.78%	70.0%
Boosted Trees	79.22%	65.45%	73.47%	69.23%
Logistic Regression	77.27%	56.36%	73.81%	63.92%
Logistic Regression with cut off = 0.04	80.52%	72.73%	72.73%	72.73%
Neural Networks	74.35%	52.61%	66.82%	58.87%

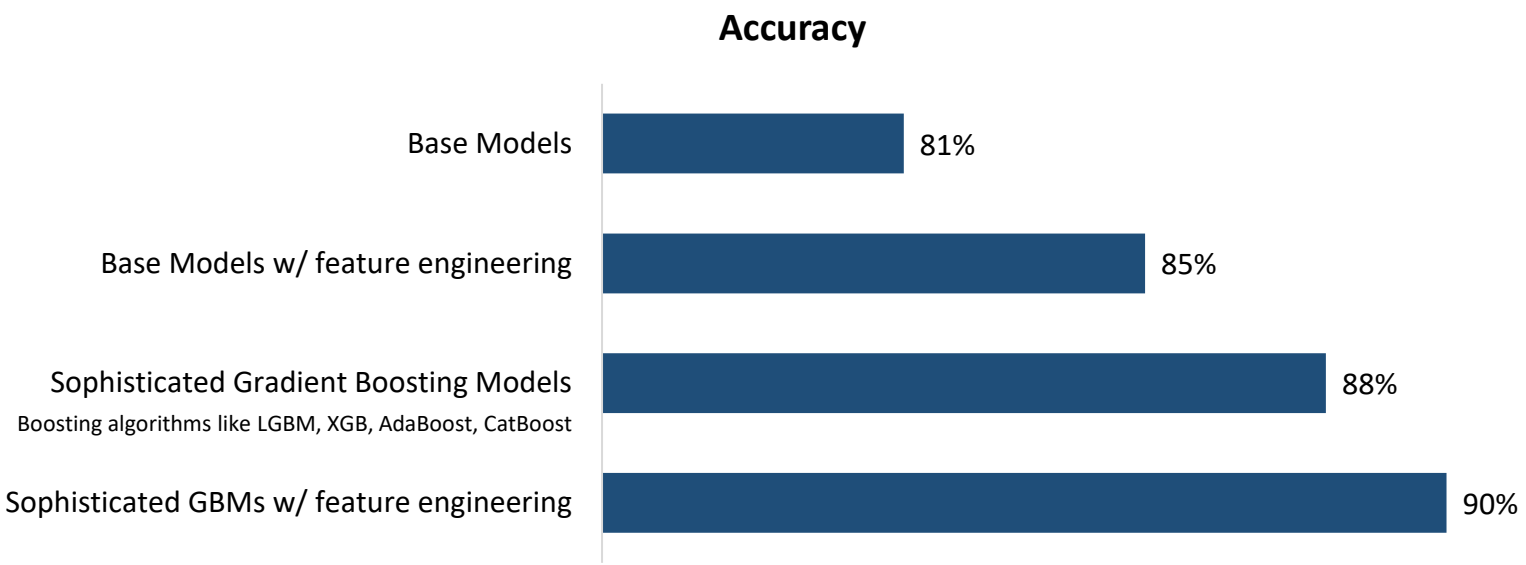




Feature Importance for our chosen model

Comparison with Kaggle Models

Kaggle Models



<u>Final Verdict</u>

Our model - Decision Tree with minimum impurity of 0.01 performs better than the base models available on Kaggle

Our Learnings from Project

- Visualize
- Data Preprocessing
- Bigger Training Data
- Seed Value
- Accuracy alone may not suffice
- Goal Clarity
- Multiple other algorithms
- Choice Depends

Conclusion and Recommendations

Conclusions

- Understand most important factors
- Important early diagnosis
- Glucose
- Awareness and accessibility to track glucose levels
- BMI
- Obesity
- Increase in age

Recommendations

- Promote healthier ways - glucose levels low (<100)
- Increasing glucose level – inform of healthier lifestyle
- Age increased, increased caution
- People with hereditary diabetes – care from early stgae

Diabetes is a disease which can be **controlled** by **keeping a check on vitals** such as **glucose level, BMI** and if proper care is taken from an **early stage in life** and by **maintaining healthy lifestyle** including **food habits** and **physical exercise**