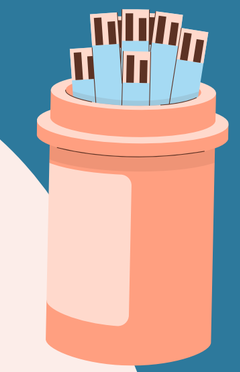
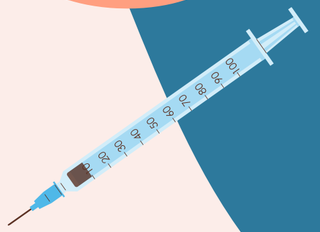
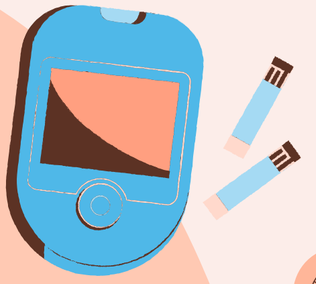




AN INTRODUCTION TO DIABETES

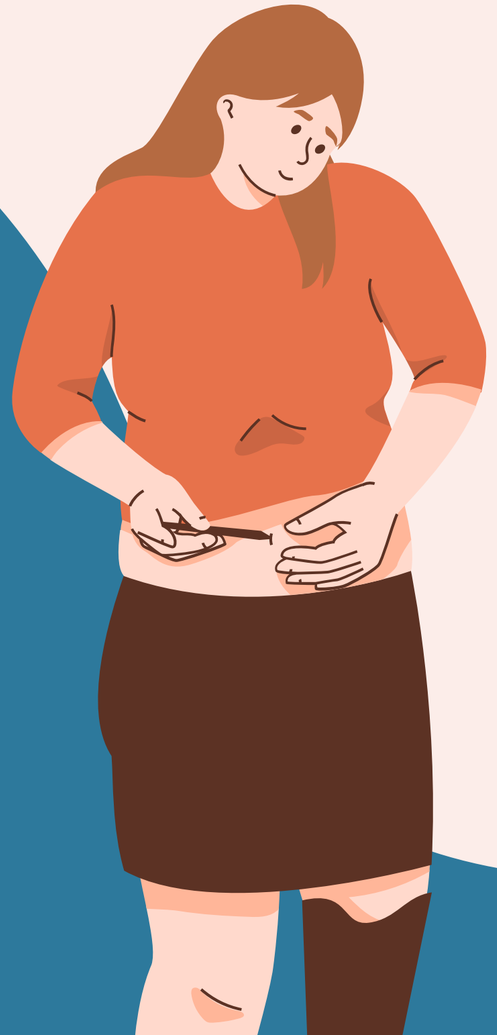
FCEB TEAM3
AUDREY GABRIELLA LUHONGE LEYA
MANIMARAN VASUMATHI
SRIVASTAVA KAVYA



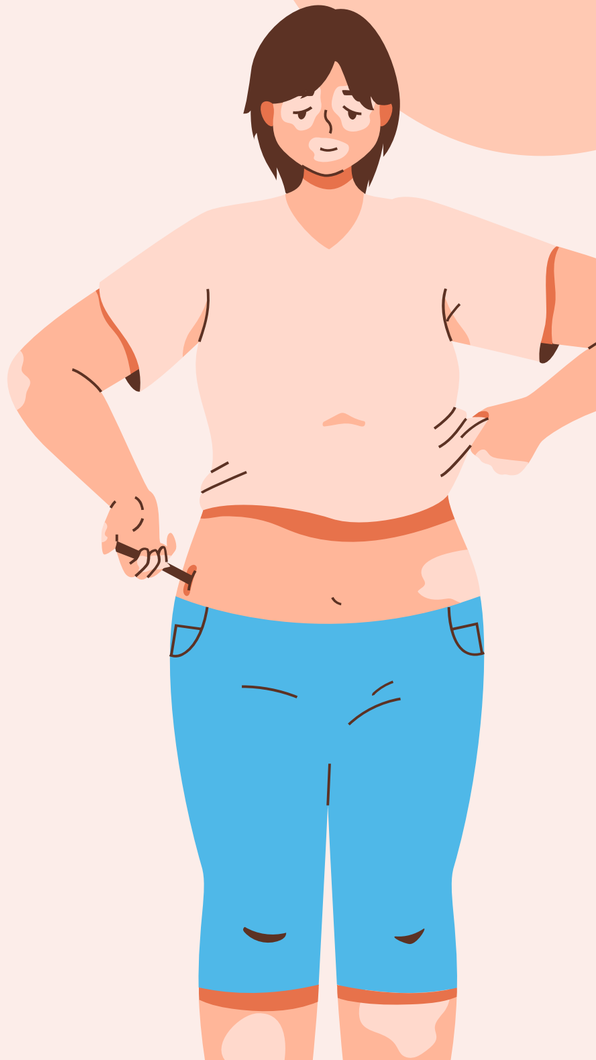
AGENDA :

- Practical Motivation & Problem Definiton
- Data Preparation
- Exploratory Data Analysis
- Classification Model - Decision Trees
- Prediction of Diabetes Type, Risk of DKA and Risk of Hypertension
- Optimisation Model - Random Forest
- Deep Learning

WHAT IS DIABETES?



Diabetes is a **chronic metabolic disorder** characterized by **elevated levels of glucose** (sugar) in the bloodstream. It occurs when the body either cannot produce enough insulin (a hormone that regulates blood sugar levels) or cannot use insulin effectively. Insulin is crucial for allowing glucose to enter cells and be used for energy.



kaggle

AKSHAY DATTATRAY KHARE · UPDATED 2 YEARS AGO

1108 | New Notebook | Download (9 kB)

Diabetes Dataset

Diabetes Patients Data

loodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
72	35	0	33.6	0.627	50	1
66	29	0	26.6	0.351	31	0
64	0	0	23.3	0.672	32	1
66	23	94	28.1	0.167	21	0
40	35	168	43.1	2.288	33	1
74	0	0	25.6	0.201	30	0
50	32	88	31	0.248	26	1
0	0	0	35.3	0.134	29	0
70	45	543	30.5	0.158	53	1
96	0	0	0	0.232	54	1
92	0	0	37.6	0.191	30	0
74	0	0	38	0.537	34	1
80	0	0	27.1	1.441	57	0
60	23	846	30.1	0.398	59	1
72	19	175	25.8	0.587	51	1
0	0	0	30	0.484	32	

Data Card | Code (441) | Discussion (11) | Suggestions (0)

About Dataset

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.2
From the data set in the (.csv) File We can find several variables, some of them are independent (several medical predictor variables) and only one target dependent variable (Outcome).

Usability
10.00

License
CC0: Public Domain

Expected update frequency
Annually

Tags

Tabular

Data Visualization

PROBLEM DEFINITION

- Determination of type of Diabetes
- Prediction DKA Risk
- Prediction of Hypertension Risk

DATA PREPARATION & CLEANING

Removing irrelevant columns

```python
# List of columns to remove
columns_to_remove = ['SkinThickness', 'Age']

# Drop the specified columns
cleaned_df = thedata.drop(columns=columns_to_remove)
cleaned_df.head()
```

	Pregnancies	Glucose	BloodPressure	Insulin	BMI	DiabetesPedigreeFunction	Outcome	Type
0	6	148	72	0	33.6	0.627	1	0
1	1	85	66	0	26.6	0.351	0	0
2	8	183	64	0	23.3	0.672	1	0
3	1	89	66	94	28.1	0.167	0	0
4	0	137	40	168	43.1	2.288	1	0

Missing values

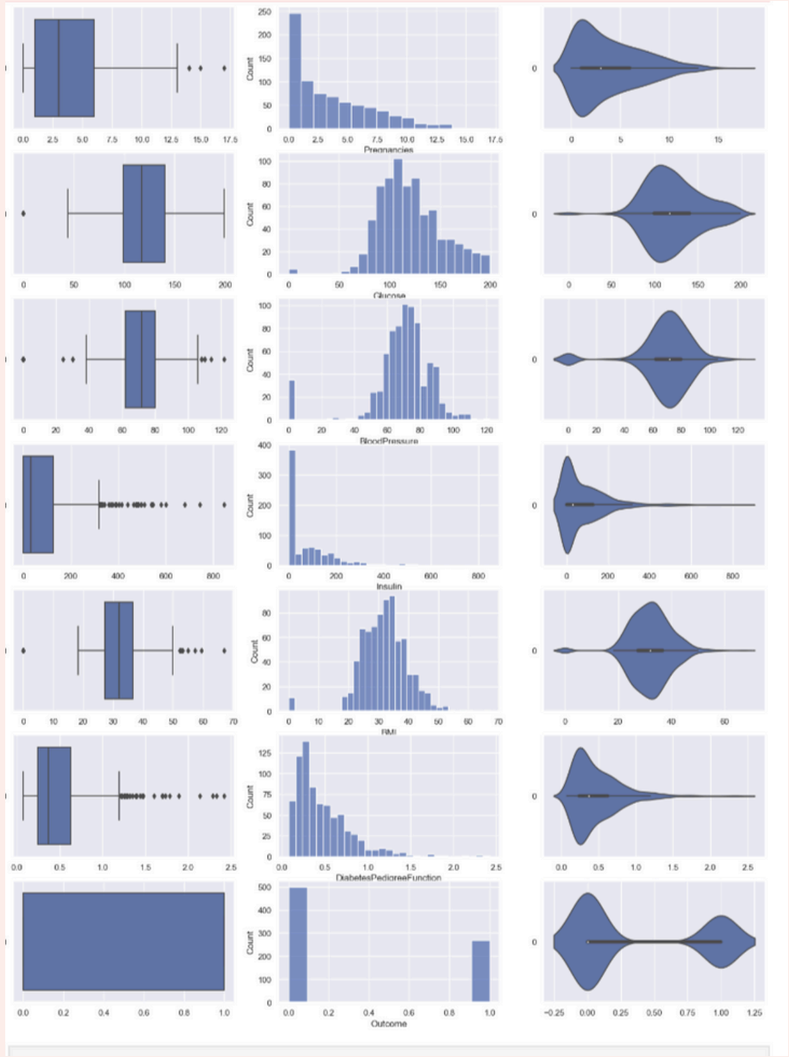
```
Pregnancies                 0
Glucose                     0
BloodPressure               0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
Outcome                     0
dtype: int64
```

EXPLORATORY ANALYSIS

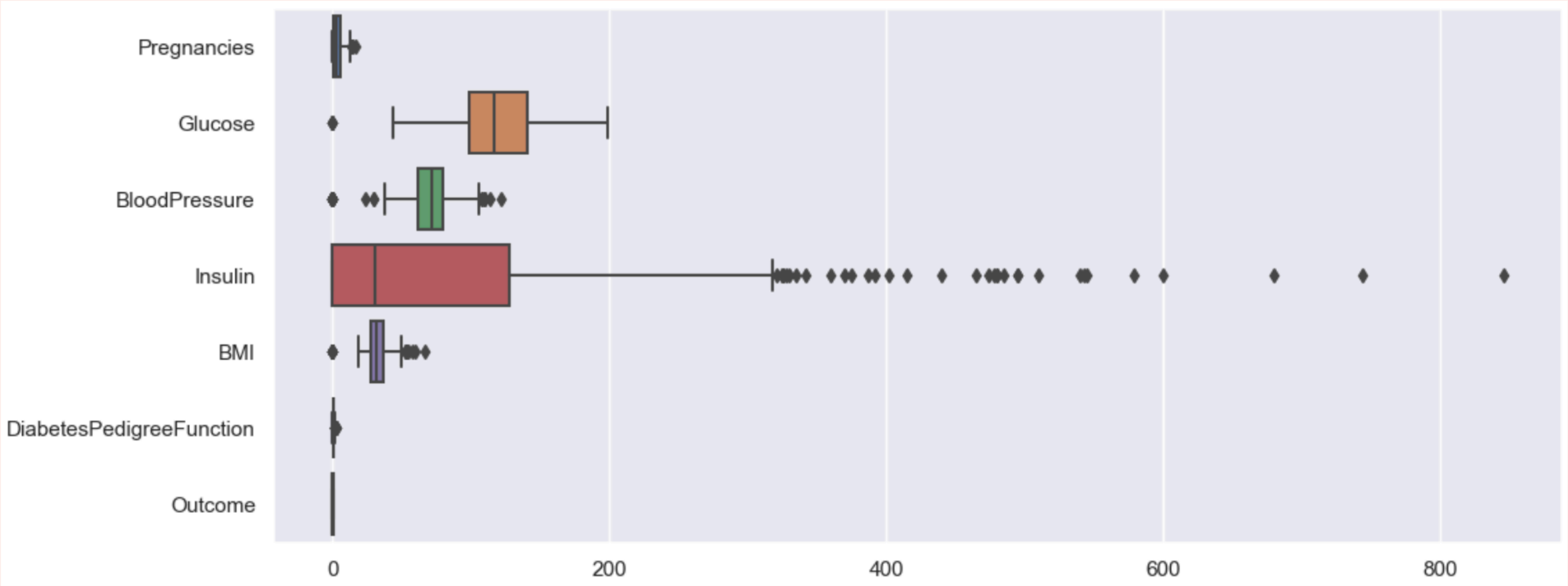
Summary statistics

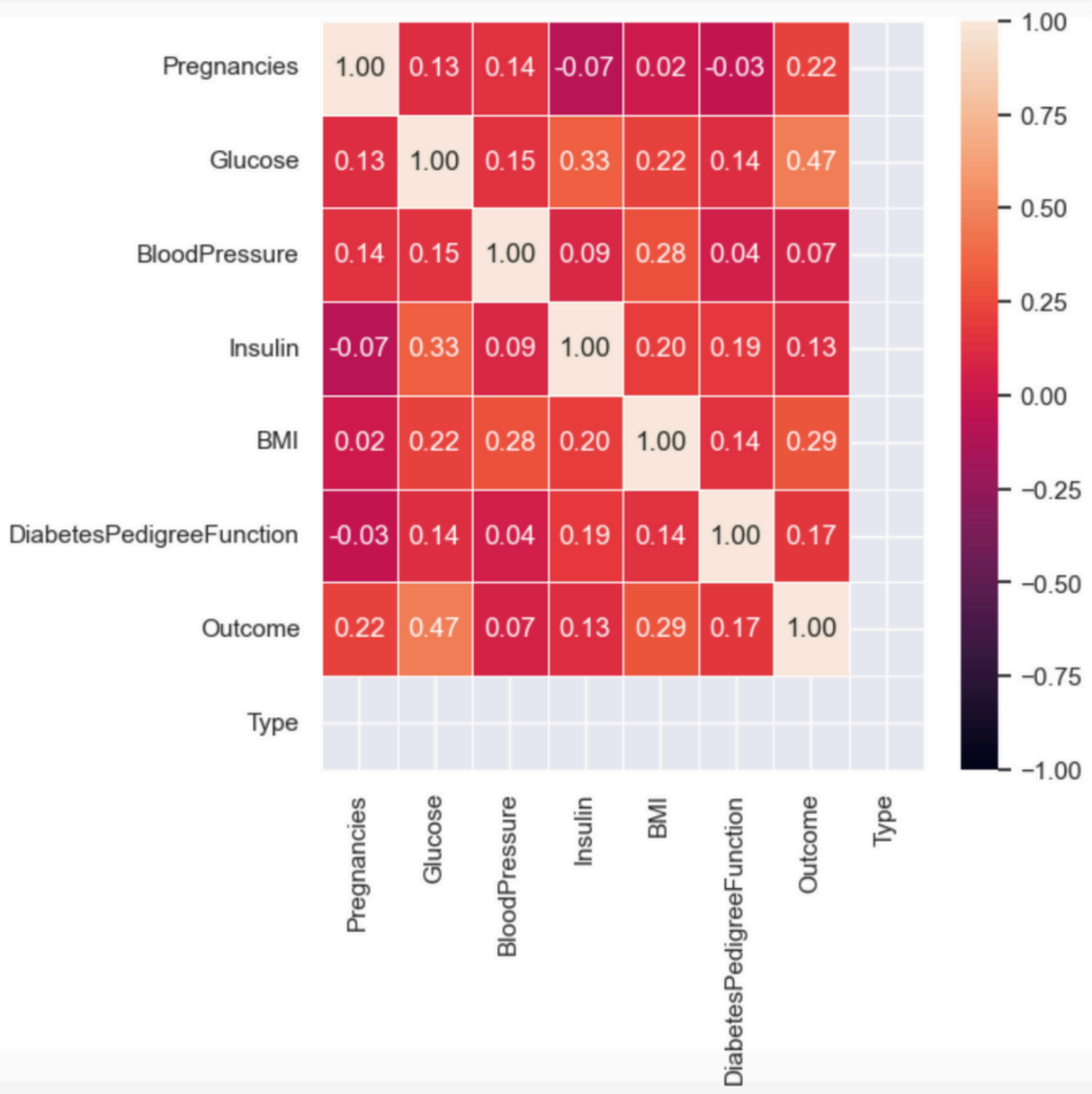
	Pregnancies	Glucose	BloodPressure	Insulin	BMI	DiabetesPedigreeFunction	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	79.799479	31.992578	0.471876	0.348958
std	3.369578	31.972618	19.355807	115.244002	7.884160	0.331329	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	27.300000	0.243750	0.000000
50%	3.000000	117.000000	72.000000	30.500000	32.000000	0.372500	0.000000
75%	6.000000	140.250000	80.000000	127.250000	36.600000	0.626250	1.000000
max	17.000000	199.000000	122.000000	846.000000	67.100000	2.420000	1.000000

EXPLORATORY ANALYSIS



ANALYTIC VISUALIZATION





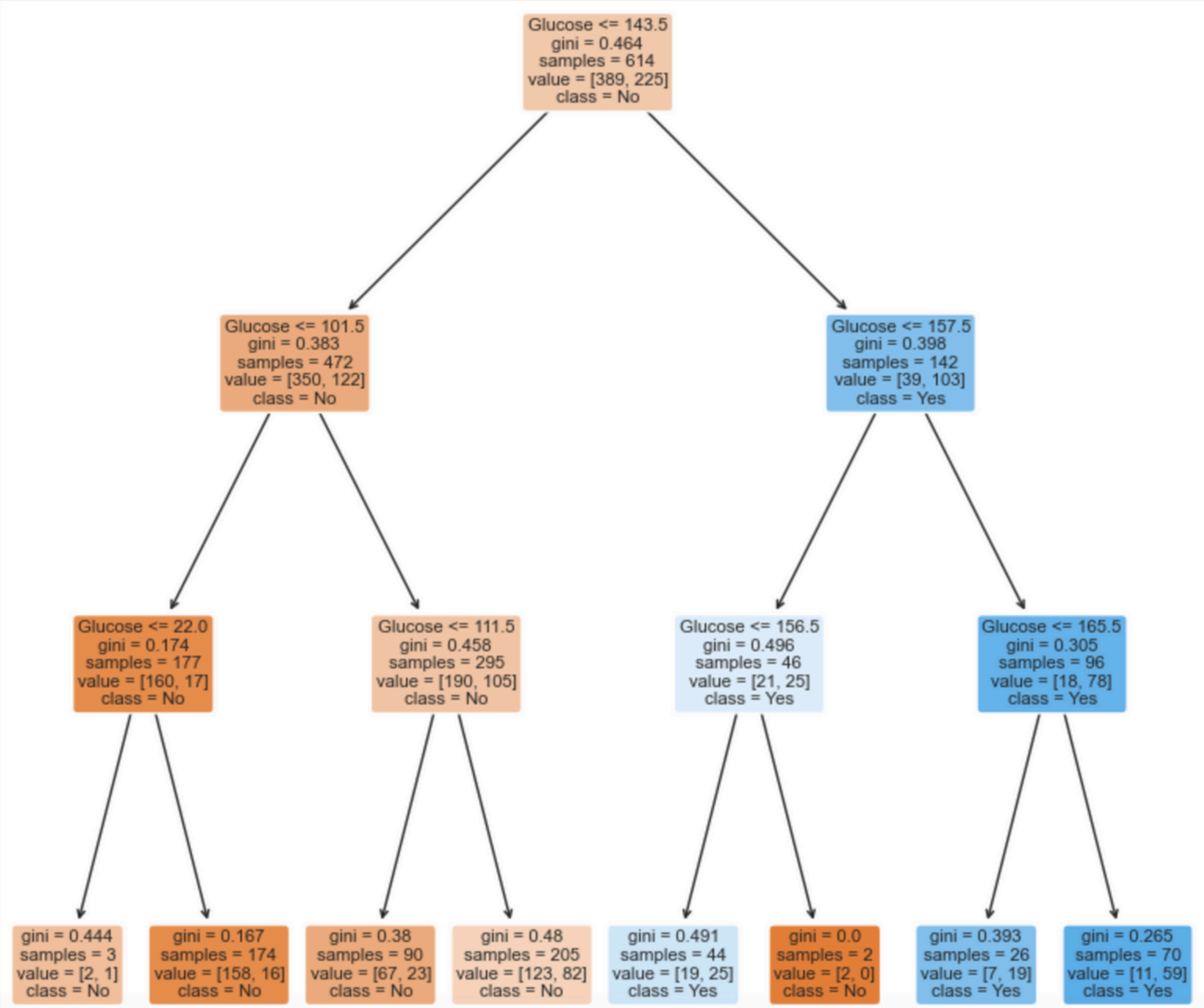
Diabetes treatment: Using insulin to manage blood sugar

Learning how insulin affects your blood sugar can help you better manage your condition.

By Mayo Clinic Staff

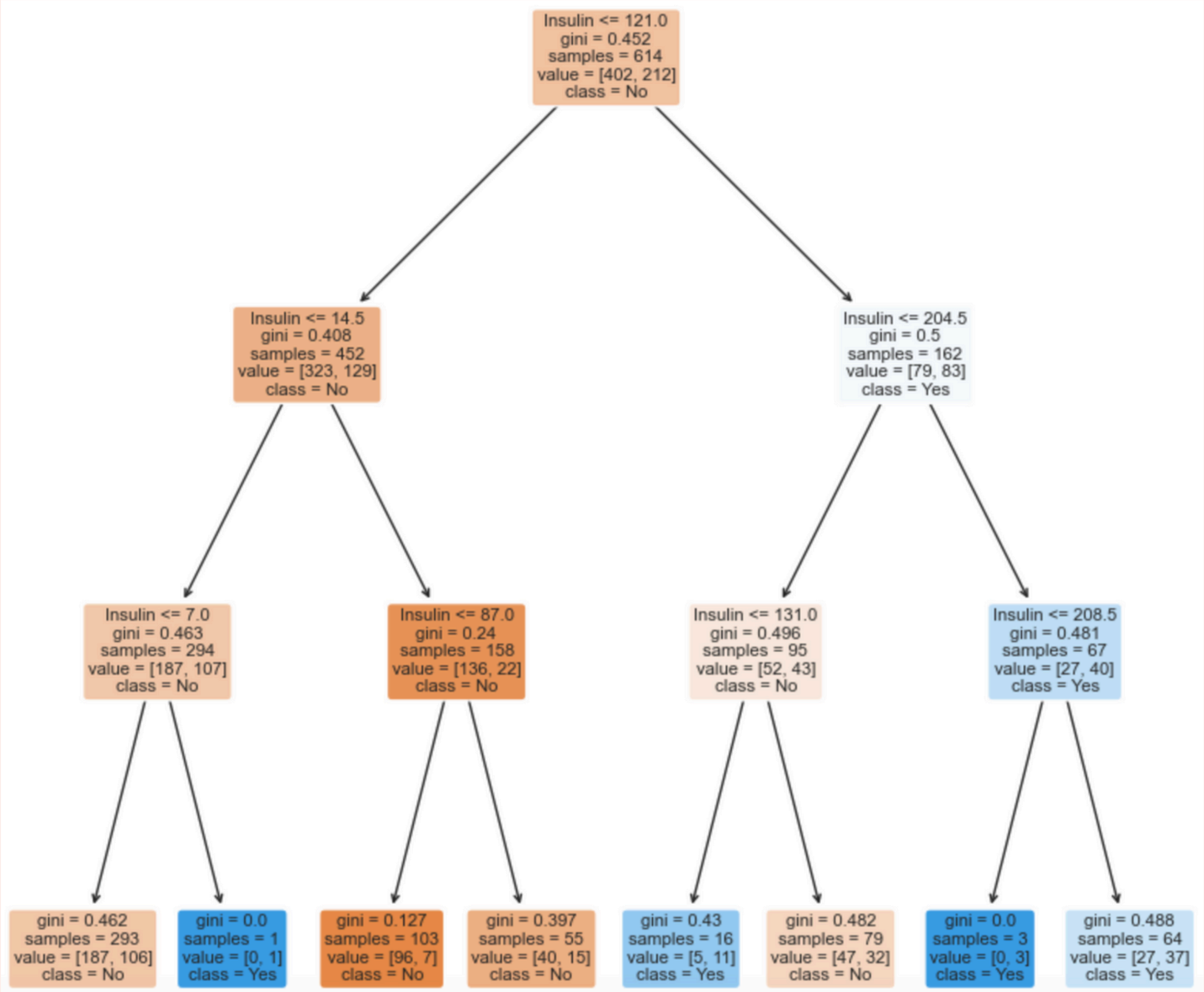
Insulin therapy often is an important part of diabetes treatment. It helps keep blood sugar under control and prevents diabetes complications. It works like the hormone insulin that the body usually makes.

Glucose X Outcome



- Classification Accuracy, training set : 0.741042345276873
- Classification Accuracy, test set : 0.8181818181818182
- Between 19-24% of False Negative

Insulin X Outcome

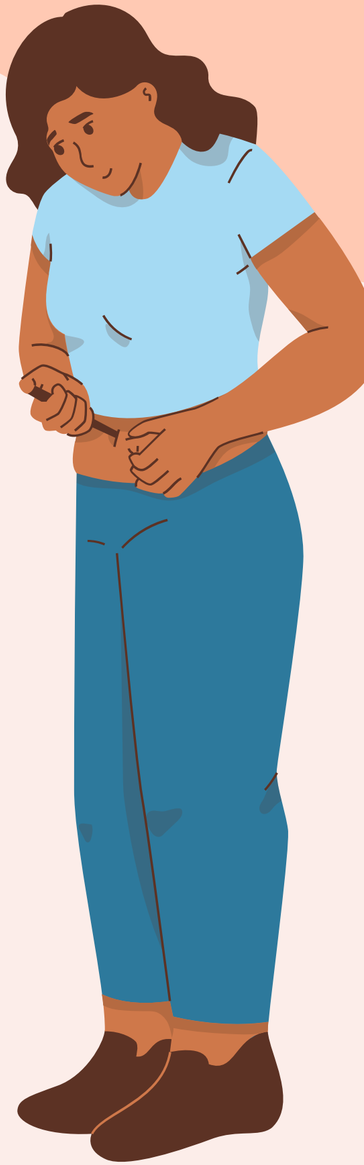


- Classification Accuracy, training set : 0.6514657980456026
- Classification Accuracy, test set : 0.7792207792207793
- Between 14-18% of False Negative

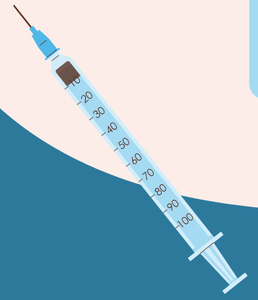
TYPE 1 OR TYPE 2?

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome	Type	DKAprone	Hypertension
1	6	148	72	35	0	33.6	0.627	50	1	2	0	0
2	1	85	66	29	0	26.6	0.351	31	0	0	0	0
3	8	183	64	0	0	23.3	0.672	32	1	2	0	0
4	1	89	66	23	94	28.1	0.167	21	0	0	0	0
5	0	137	40	35	168	43.1	2.288	33	1	2	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...
764	10	101	76	48	180	32.9	0.171	63	0	0	0	0
765	2	122	70	27	0	36.8	0.34	27	0	0	0	0
766	5	121	72	23	112	26.2	0.245	30	0	0	0	0
767	1	126	60	0	0	30.1	0.349	47	1	1	0	0
768	1	93	70	31	0	30.4	0.315	23	0	0	0	0

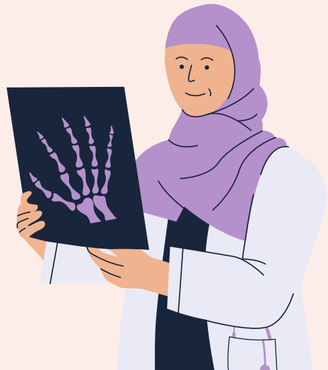
768 rows × 12 columns



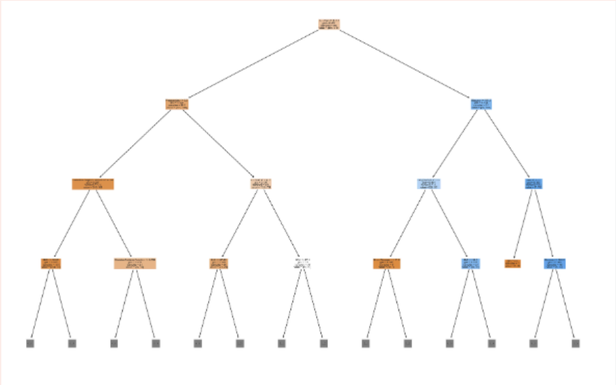
SHOULD YOU BE WORRIED?

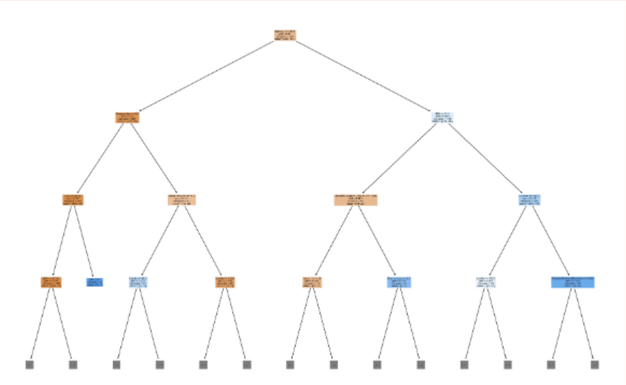


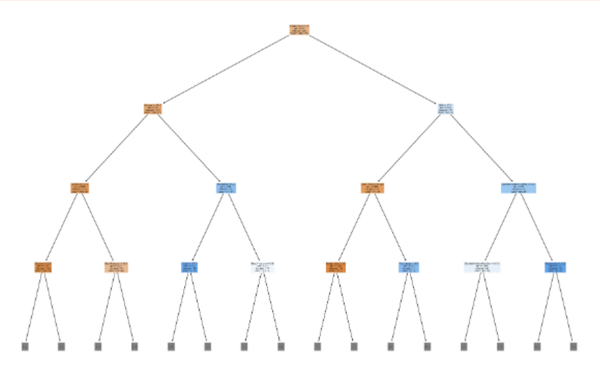
	Type	DKAprone	Hypertension
count	768.000000	768.000000	768.000000
mean	0.601562	0.042969	0.083333
std	0.863635	0.202919	0.328297
min	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000
50%	0.000000	0.000000	0.000000
75%	2.000000	0.000000	0.000000
max	2.000000	1.000000	2.000000



RANDOM FOREST OPTIMISATION MODEL

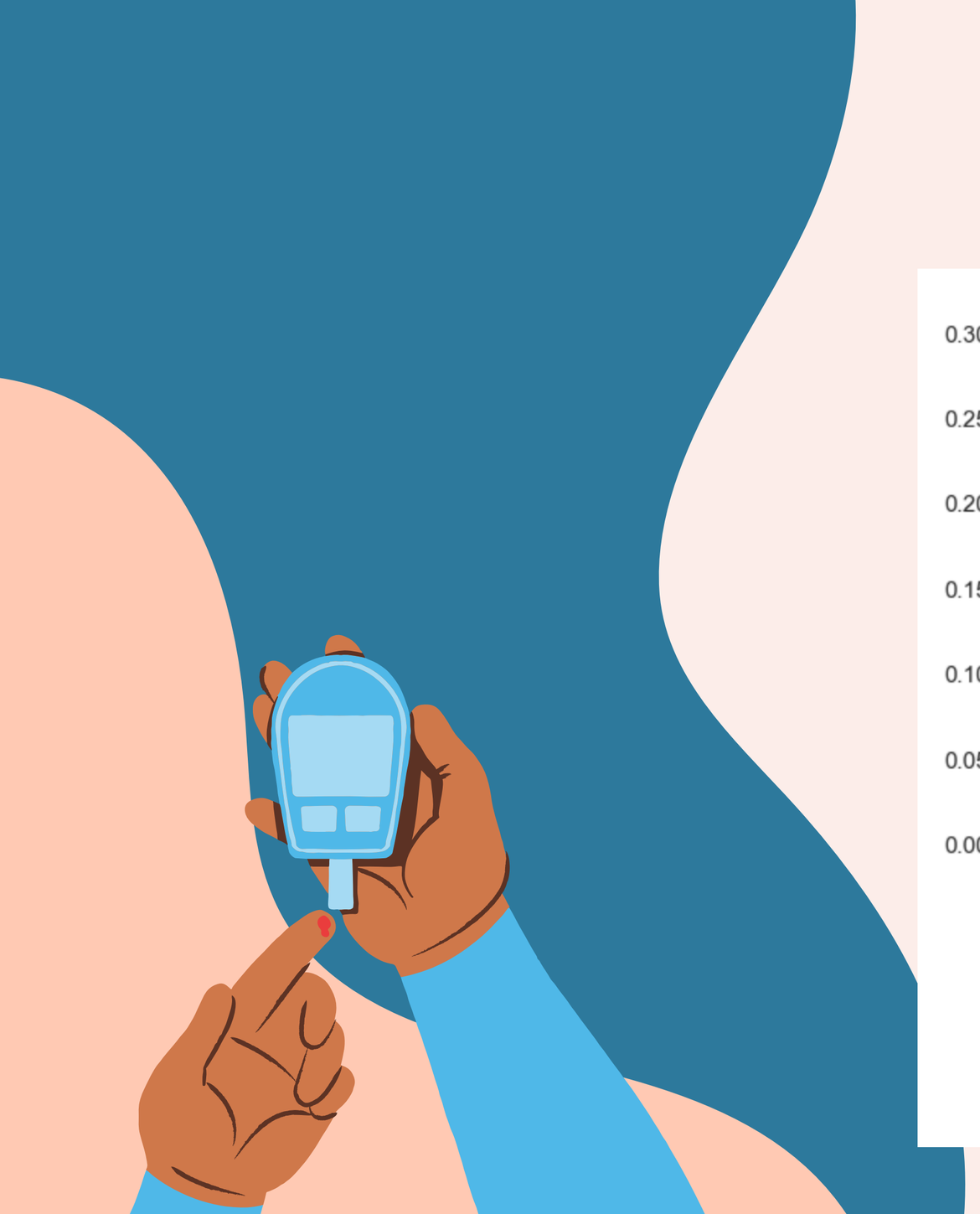


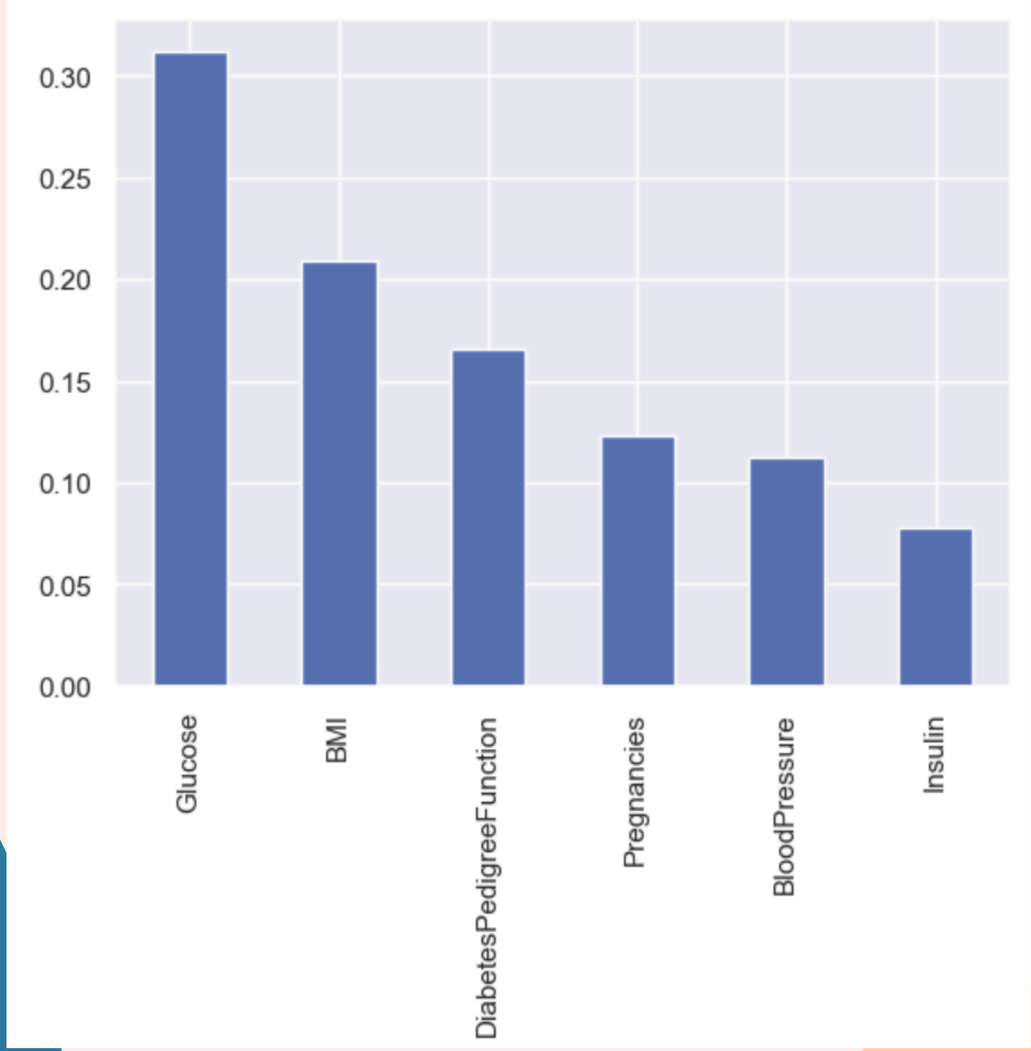




```
Best estimator: RandomForestClassifier(max_depth=13, n_estimators=454)
Best hyperparameters: {'max_depth': 13, 'n_estimators': 454}
```

GLUCOSE IS THE MOST IMPORTANT CONTRIBUTING FACTOR





```
Model: "sequential_4"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_9 (Dense)             (None, 256)               2304

 dense_10 (Dense)            (None, 156)               40092

 dense_11 (Dense)            (None, 1)                 157

=================================================================
Total params: 42,553
Trainable params: 42,553
Non-trainable params: 0
_________________________________________________________________
```

Evaluating

[31]:
```
y_pred_test= model.predict(xtest)
y_pred_test= [0 if val<0.5 else 1 for val in y_pred]
```

```
4/4 [==============================] - 0s 6ms/step
```

[32]:
```
# checking the accuracy score for the test dataset
print("Accuracy score (test): " , accuracy_score(ytest,y_pred_test))
```

```
Accuracy score (test):  0.7241379310344828
```

[33]:
```
# evaluating the model for train set
y_pred_train= model.predict(xtrain)
y_pred_train= [0 if val<0.5 else 1 for val in y_pred_train]
```

```
21/21 [==============================] - 0s 2ms/step
```

[34]:
```
# checking the accuracy score for the train dataset
print("Accuracy score (test): " , accuracy_score(ytrain,y_pred_train))
```

```
Accuracy score (test):  0.8773006134969326
```

DEEP LEARNING

CONCLUSION

- We began by investigating the dataset with EDA
- Classification techniques helped determine the contribution of each factor
- Random forest helped us recognise that Glucose is the key factor
- Created a deep learning model
- Accuracy scores for DL model were higher than what we obtained in classification
- With more refinement outcomes can be improved drastically