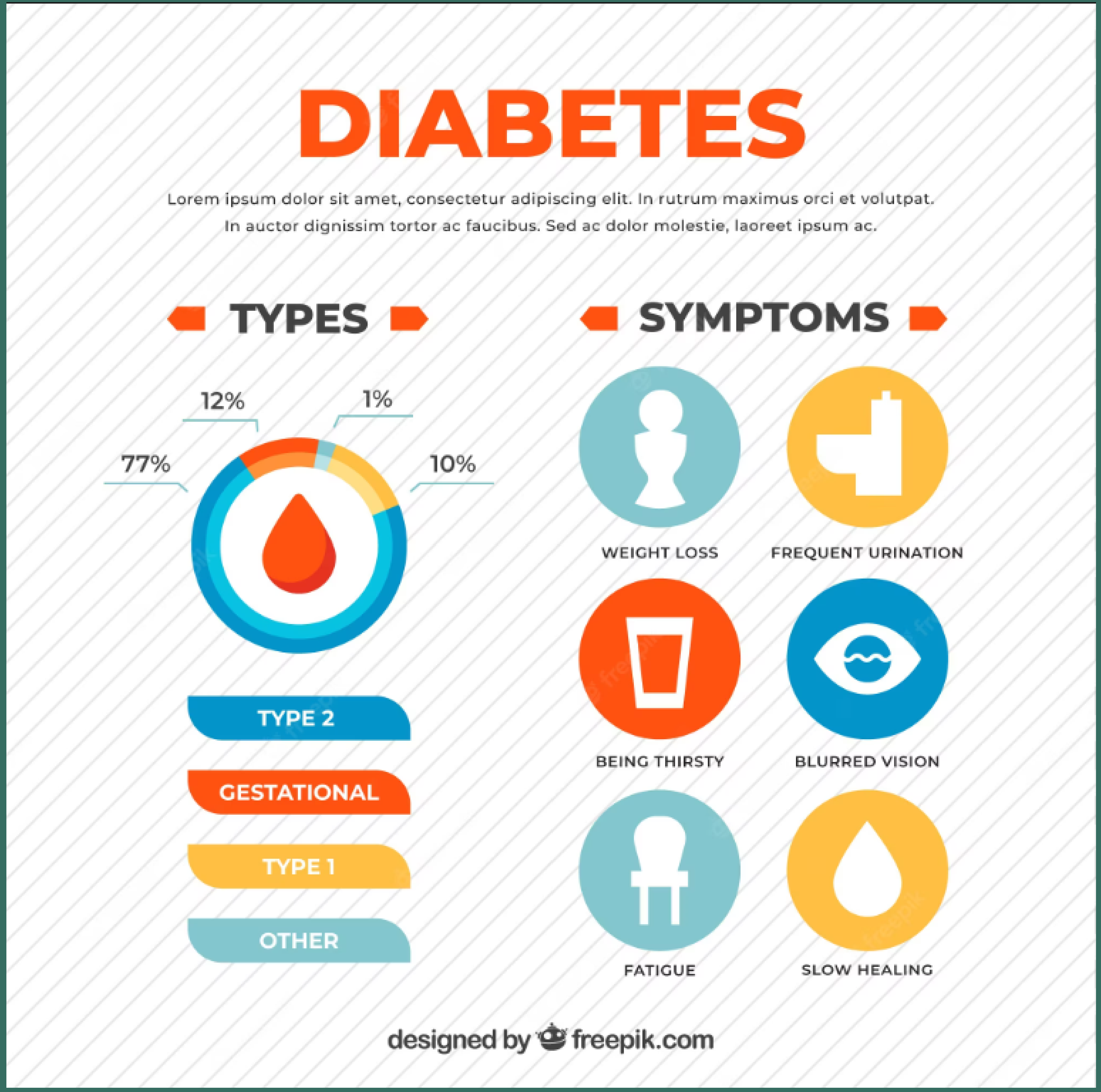
DIABATESE DETECTION

USING MACHINE LEARNING

-Abhay Saini
-Abhay Chauhan
-Riya Sehrawat
-Aadhar Agarwal

INTRODUCTION

- Diabetes is a common chronic disease that can be dangerous.
- Diabetes can be identified when blood glucose is higher than normal level, which is caused by high secretion of insulin or biological effects.
- Diabetes can cause various damage to our body and can disfunction tissues, kidneys, eyes and blood vessels.



Brief Literature Survey:

Quan Zou Method:

Quan Zou worked simultaneously on two datasets. One dataset is the Pima Indians Diabetes Dataset, and another dataset is from a local Hospital in Luzhou, China. The authors employed a two-phase detection method, namely, principal component analysis, minimum redundancy, and maximum relevance. They used three classifiers, that is, decision tree J48, random forest, and neural network.
The authors have achieved an AUC of 0.95

Nishith Kumar's Method:

In this paper, the authors have assumed the medical data to be inherently structured, nonnormal, and nonlinear and therefore made use of three kernel-based Gaussian process classification against naïve Bayes, linear discriminant analysis, and quadratic discriminant analysis. Three kernels are linear, polynomial, and radial basis kernel, and then a comparative analysis of three kernels in the GPC and then the GPC is compared against naïve Bayes, LDA, and QDA.
Accuracy achieved was 81.97%.

Maniruzzaman’s Method:

In this article, the authors adopted four machine learning algorithms, that is, random forest, AdaBoost, naïve Bayes, and decision tree on NHANES (National Health and Nutrition examination survey) dataset.The authors selected the important features from the dataset by making use of the logistic regression model, P value, and odds ratio. The probability of response is calculated from logistic regression [28] by making use of one or more predictions.
Accuracy achieved was 92.54%.

Deepti Sisodia’s Method:

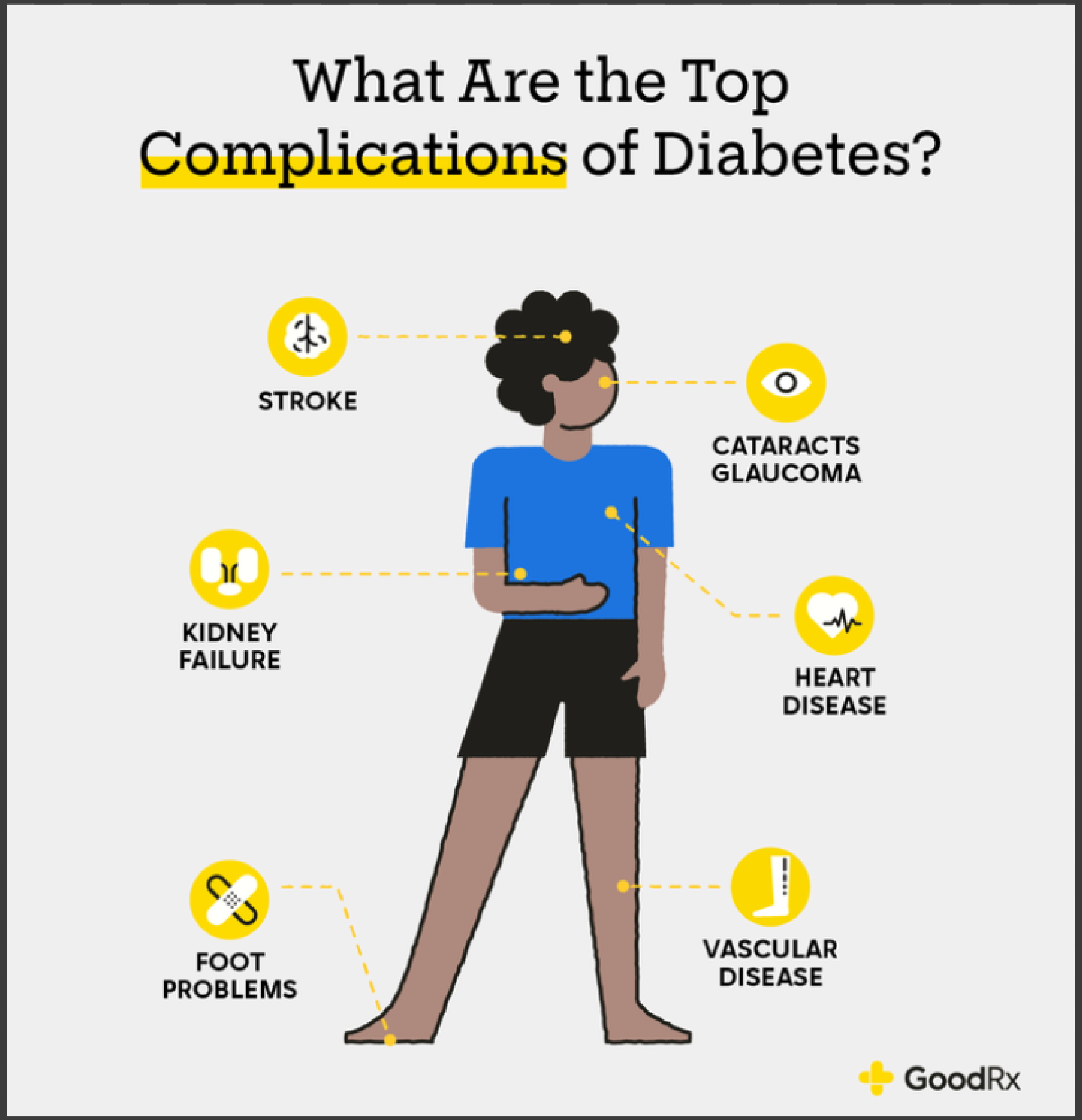
The authors have designed a machine learning model which can maximize the accuracy of the likelihood of diabetes in Indian patients. Three machine learning algorithms, namely, decision tree, that is, support vector machine, and naïve Bayes, are used for the classification of diabetes. The main aim of using Weka by the authors is that, according to the given requirements, the tool can also be personalized.
The authors have a classification accuracy of 77.5%.

Saumendra Mohapatra’s Method:

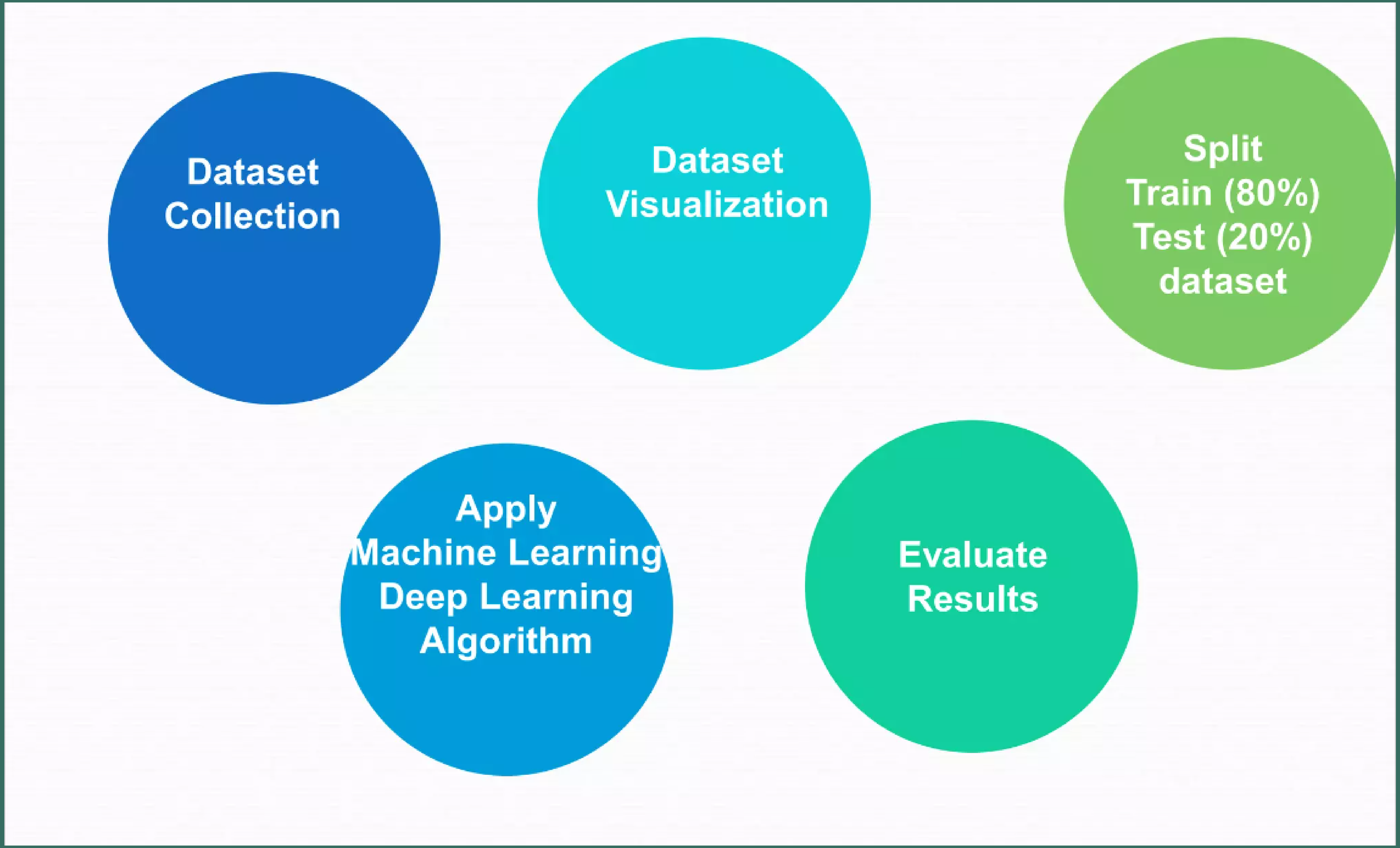
The authors have applied five classification methods, that is, random forest, k-nearest neighbour, support vector machine, naïve Bayes, and decision tree on a dataset which is downloaded from UCI repository and contains 15 attributes and 2500 values, although they took only 768 values for testing purpose.If the difference between the correlated values of the two attributes is large, the attribute is considered less significant. By using this correlation, the best attribute from the dataset is selected and arranged in a significant order.
The authors have achieved the highest ***specificity of 98.2% and 98%*** through decision trees and random forest, respectively.

Problem formulation:

- Diabetes prediction is important for proper treatment to avoid further complications of the disease.
- Numerous ML (machine learning) algorithms have been utilized, according to a recent study, to identify and forecast diseases.
- However, according to the research history, none of them have been able to attain good accuracy, i.e., more than 80%.
- Diabetics patient medical record and different types of algorithms are added in dataset for experimental analysis.Performance and accuracy of the applied algorithms is discussed and compared.



Methadology:


Dataset Collection
Dataset Visualization
Split Train (80%) Test (20%) dataset
Apply Machine Learning Deep Learning Algorithm
Evaluate Results

Requirements

Hardware Requirements:

Processor : Any Processor above 500 MHz.

Ram : 4 GB

Hard Disk : 4 GB

Input device : Standard Keyboard and Mouse.

Output device : VGA and High Resolution Monitor.

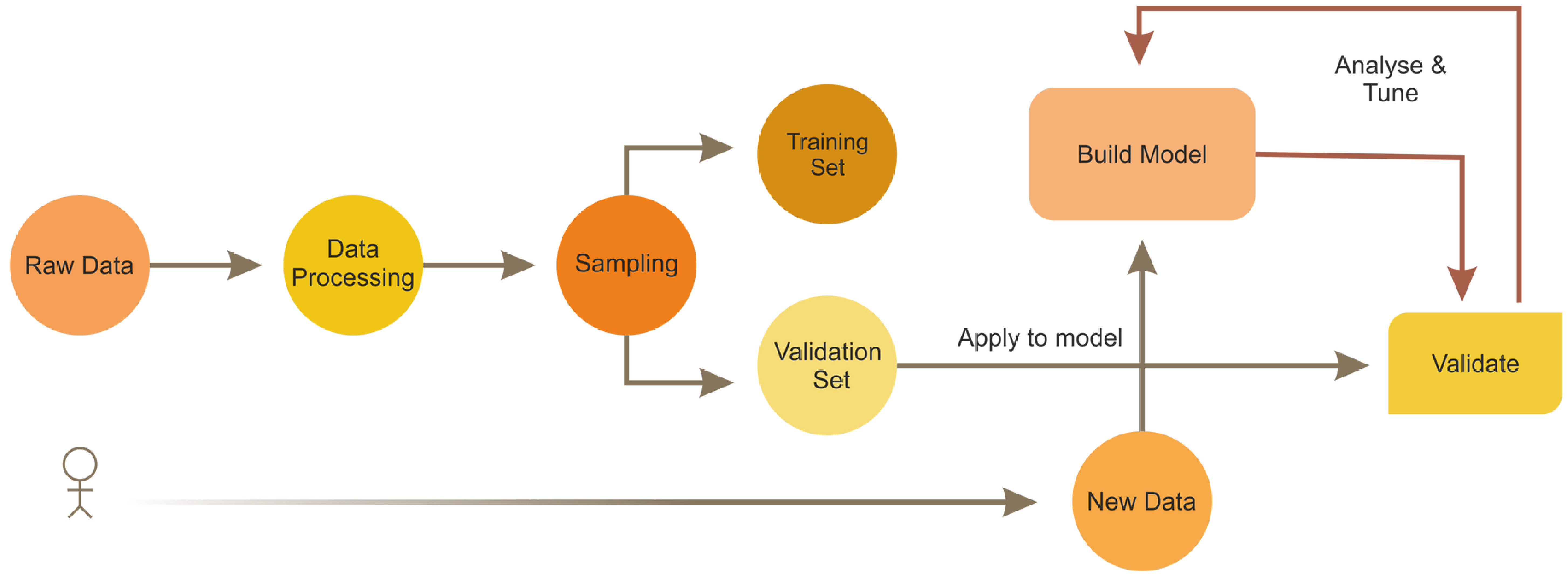
Software Requirements:

Operating System : Windows 7 or higher

Programming : Python 3.6 and related libraries

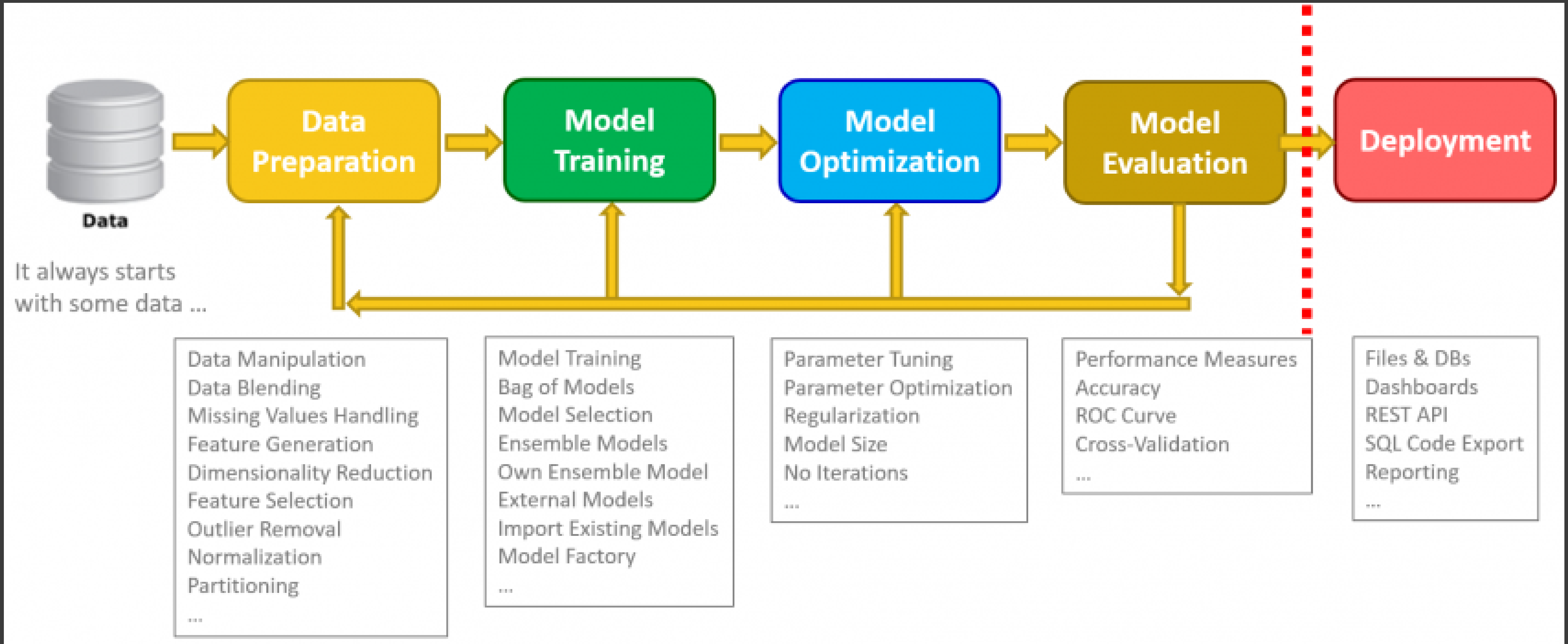
MACHINE LEARNING

Block Diagram :



- We can define the machine learning workflow in 5 stages.

Gathering data ► Data pre-processing ► Researching the model ► Training and testing the model ► Evaluation



It always starts with some data ...

Data Manipulation
Data Blending
Missing Values Handling
Feature Generation
Dimensionality Reduction
Feature Selection
Outlier Removal
Normalization
Partitioning
...

Model Training
Bag of Models
Model Selection
Ensemble Models
Own Ensemble Model
External Models
Import Existing Models
Model Factory
...

Parameter Tuning
Parameter Optimization
Regularization
Model Size
No Iterations
...

Performance Measures
Accuracy
ROC Curve
Cross-Validation
...

Files & DBs
Dashboards
REST API
SQL Code Export
Reporting
...

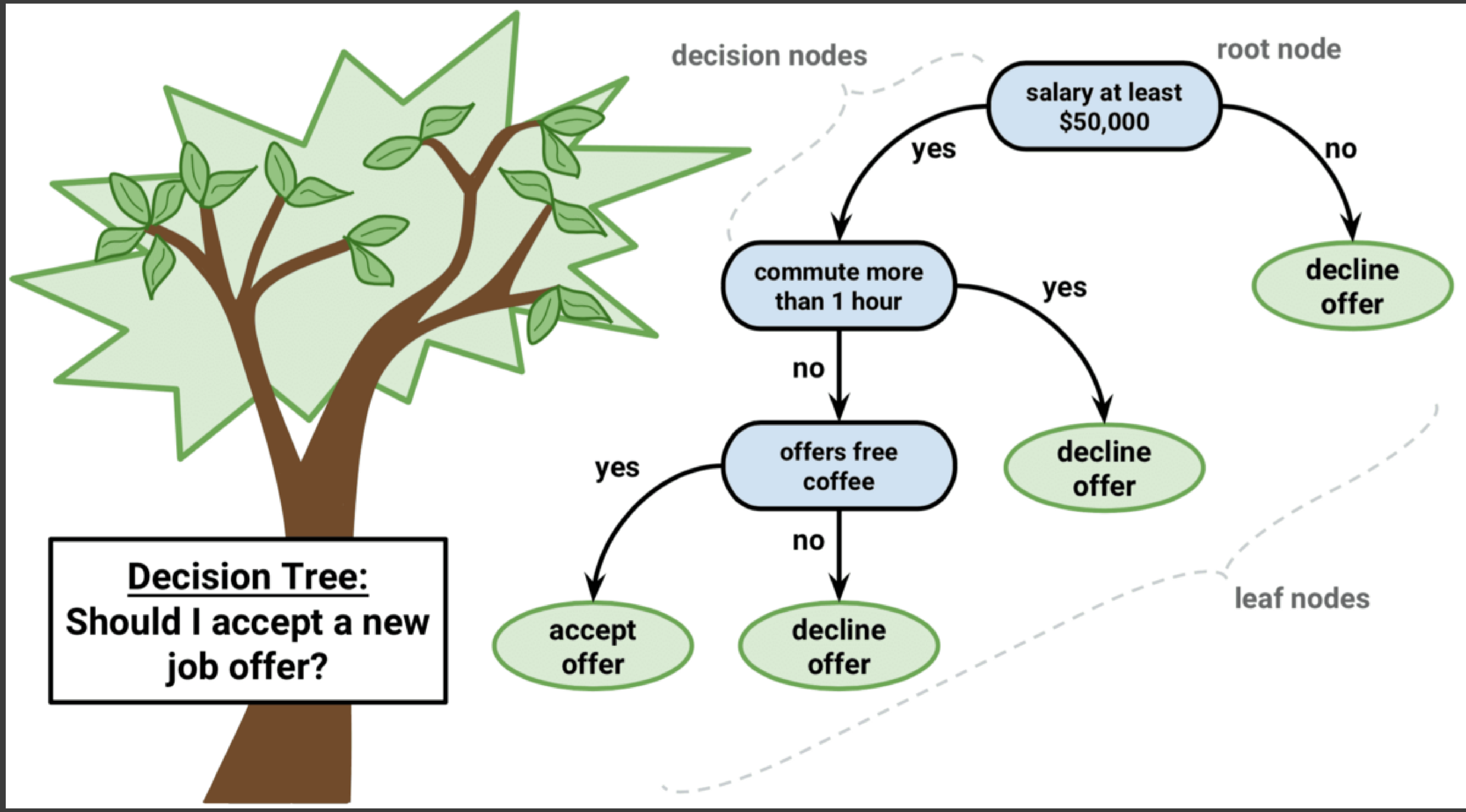
Training and Testing the model

- Training is the most important part, where we train our model using the data available and make the machine learn and understand the data.
- When the model has learned from the data, we provide the model with another dataset to evaluate how good our model is performing, if it is performing well, we then test the model using test data, where we get to know the final performance of our model, which can be measure using various metrics, such as Accuracy, recall, precision, and through classification report.
- This whole process of building and deploying a model is done using 3 different datasets which are split using train_test_split(), which are ‘Training data’, ‘Validation data’, and ‘Testing data’.

Algorithms Used

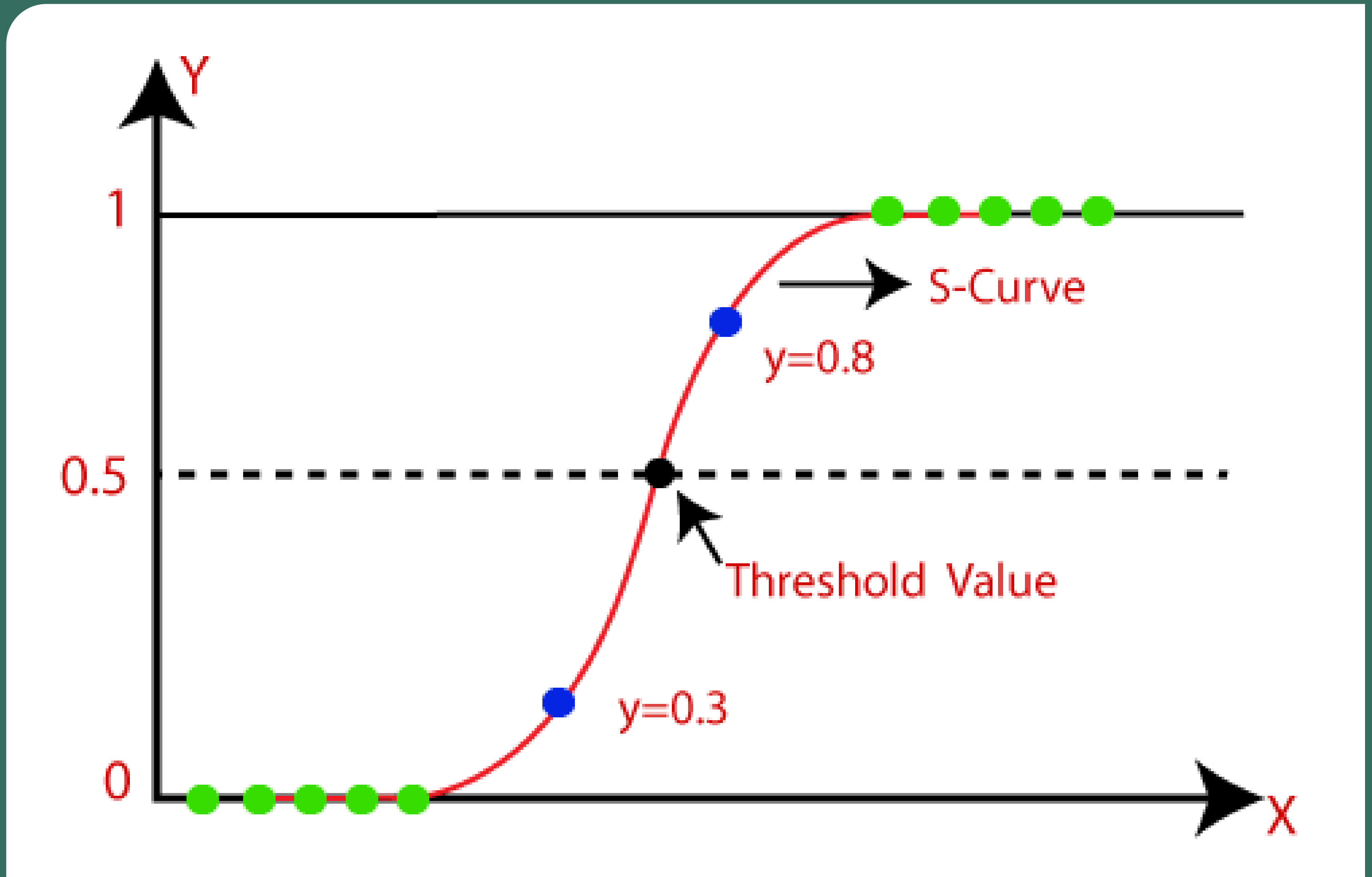
Decision Tree:

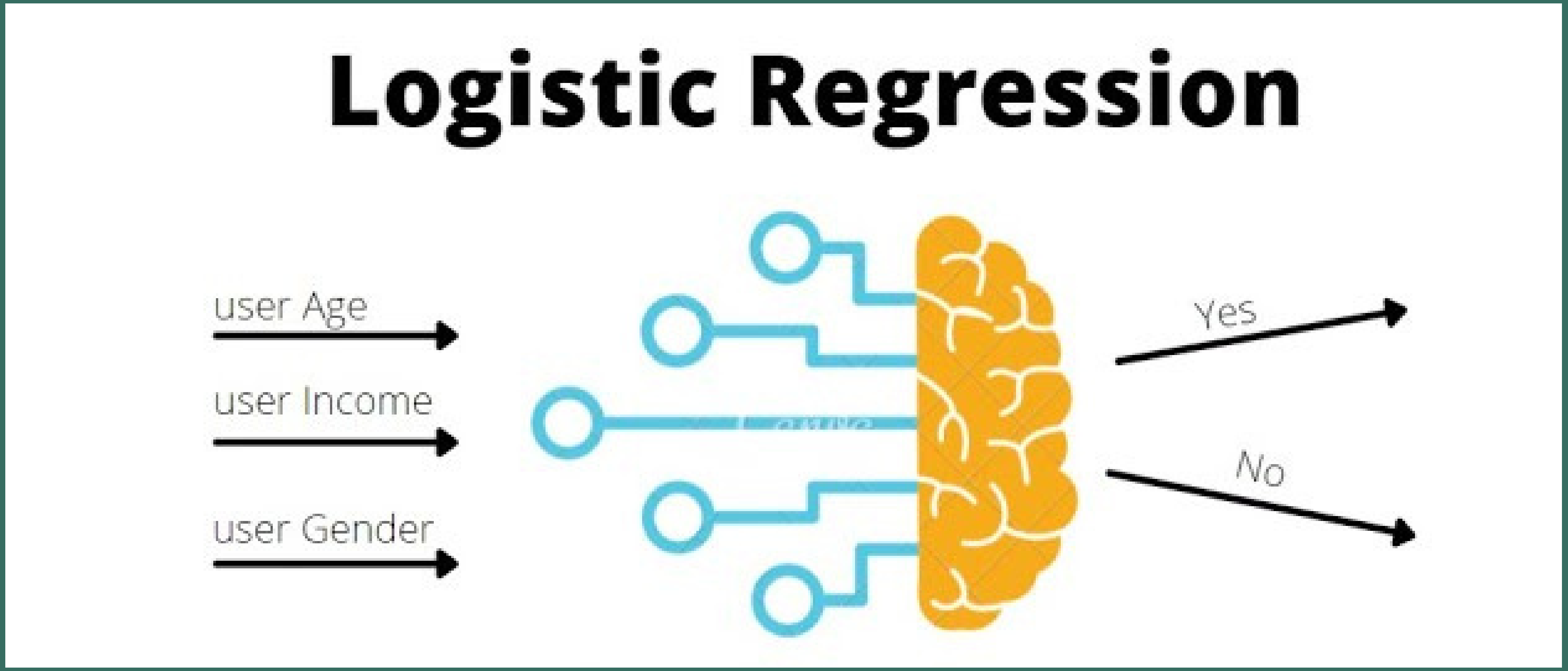
- Decision tree, as the name suggests, creates a branch of nodes.
- Where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and the last nodes are termed as the leaf nodes.
- Leaf node means there cannot be any nodes attached to them, and each leaf node (terminal node) holds a class label.



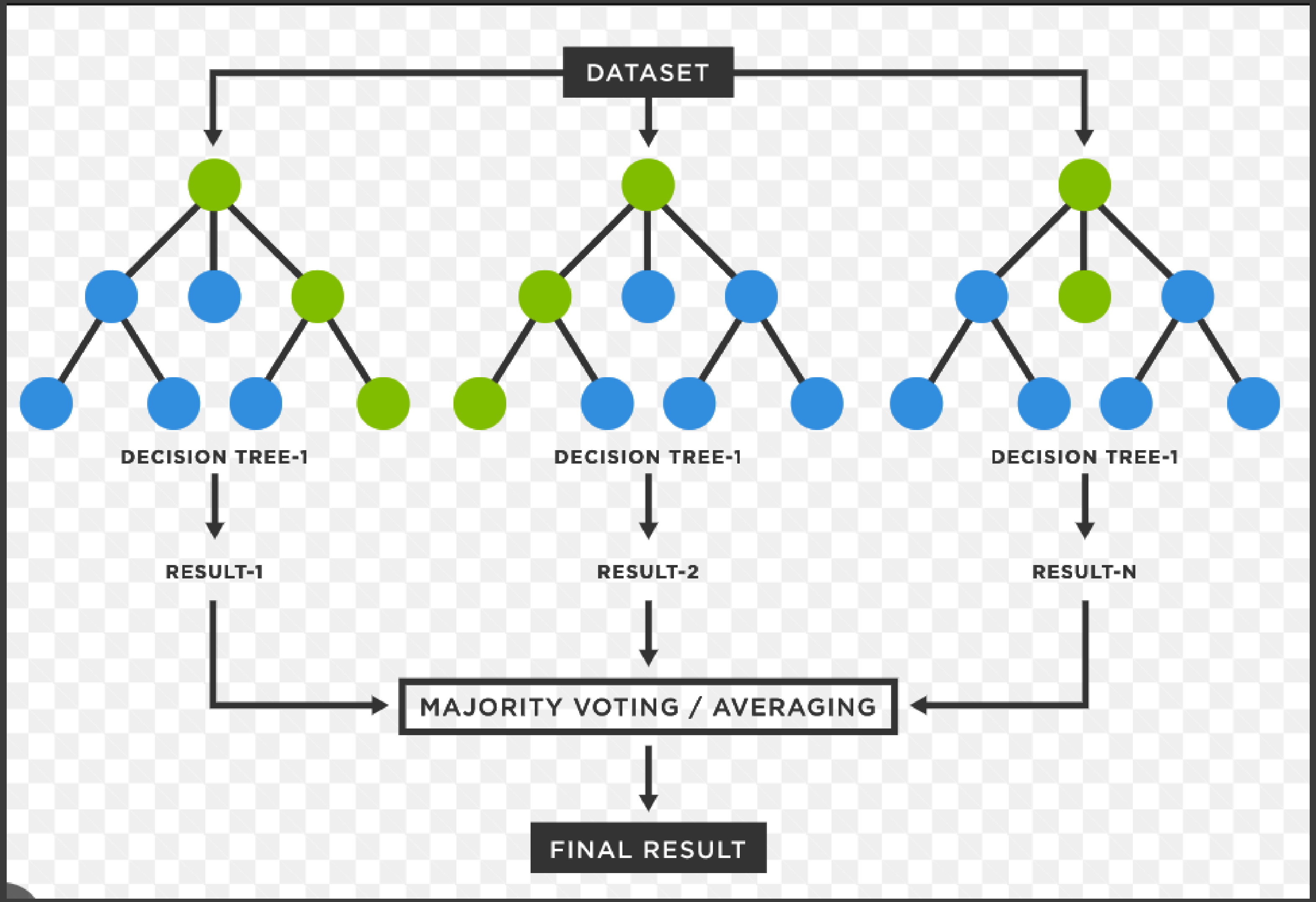
Logistic Regression:

- Logistic regression models a relationship between predictor variables and a categorical response variable.
- Logistic regression helps us estimate a probability of falling into a certain level of the categorical response given a set of predictors.



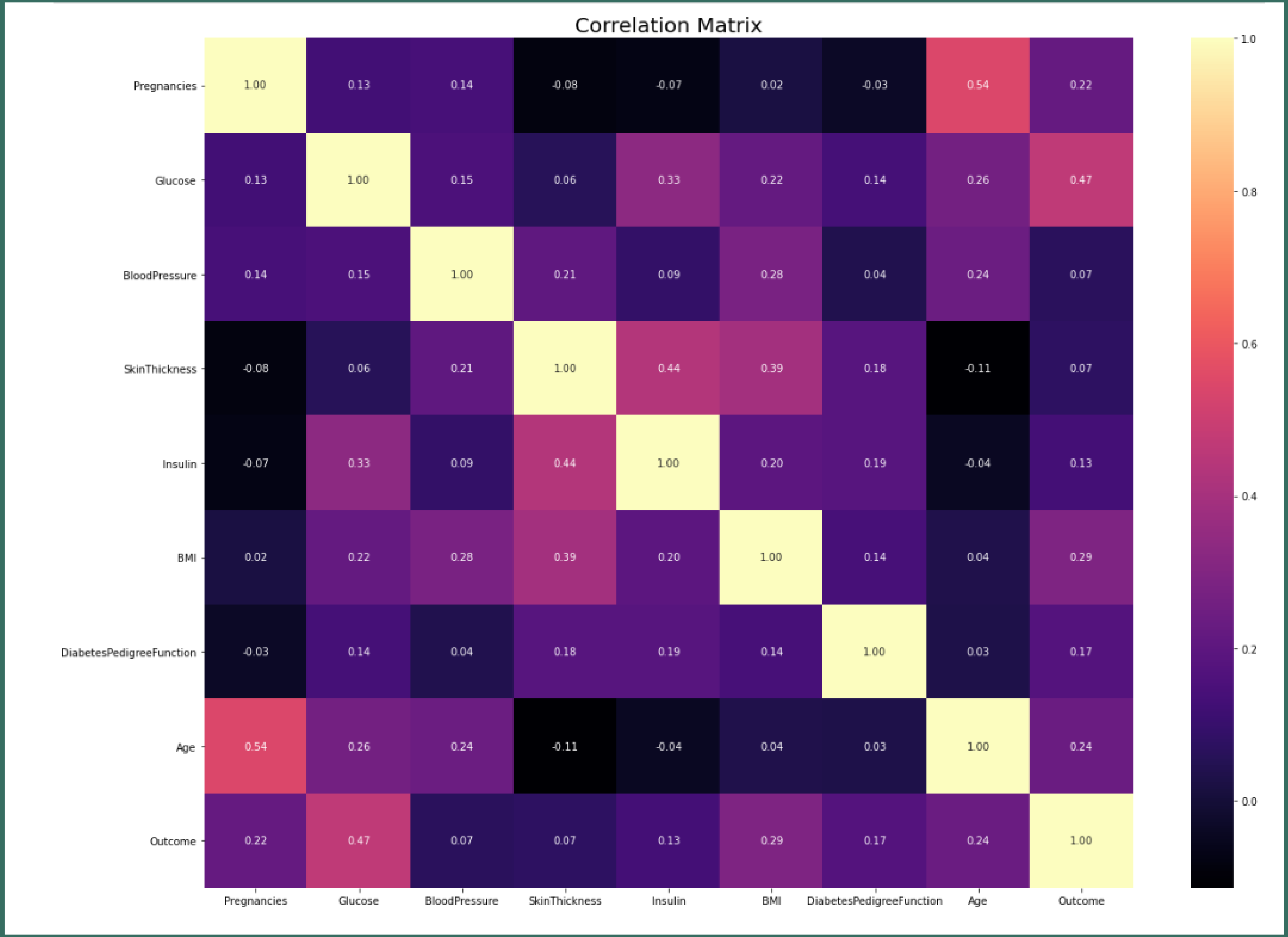


Random Forest classifier:



- Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique.
- It learning, which is a process of combining multiple classifiers to solve a complex problem, and at the end, the results are either made an average of all the classifiers or mode of all the classifiers. is based on the concept of ensemble .
- The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

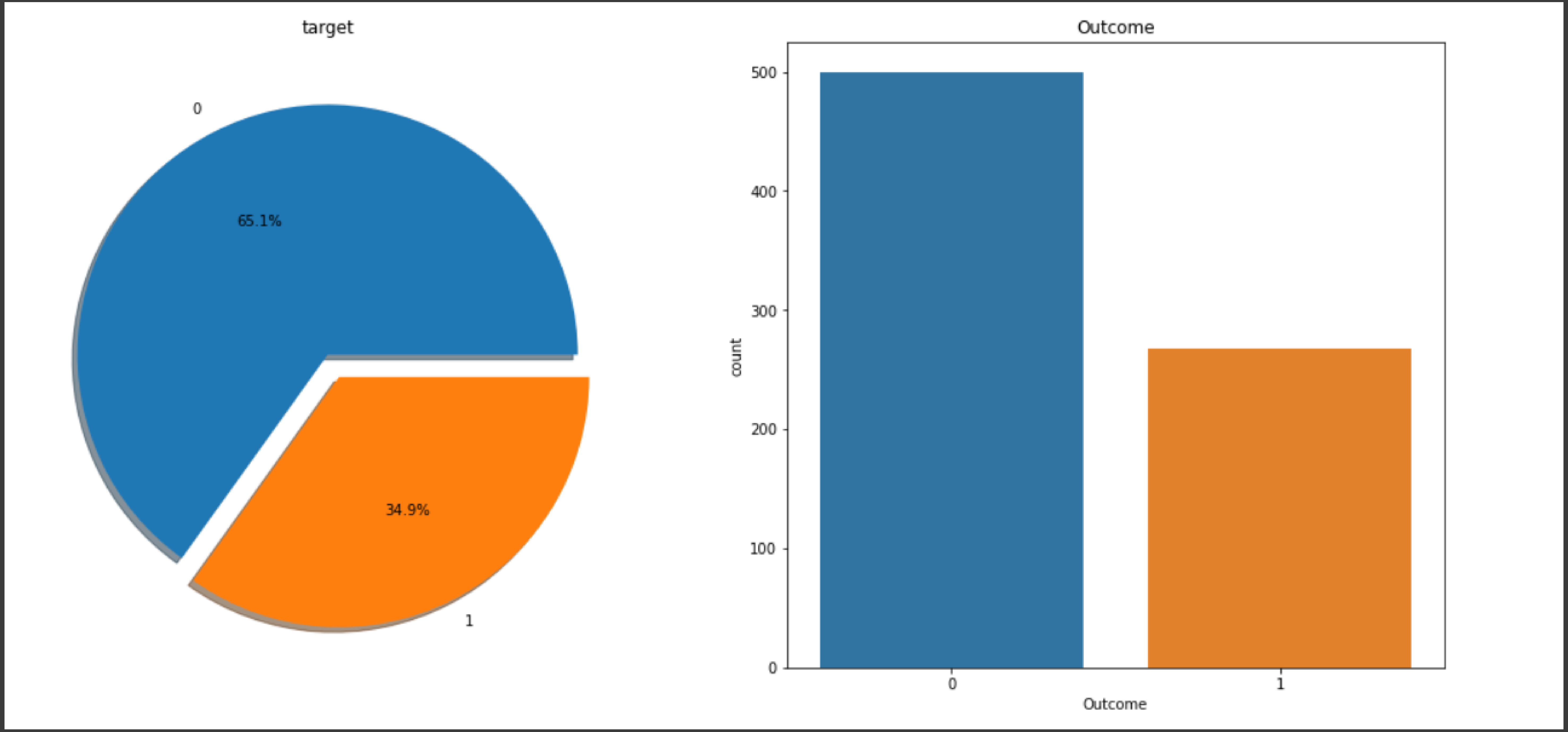
Dataset:



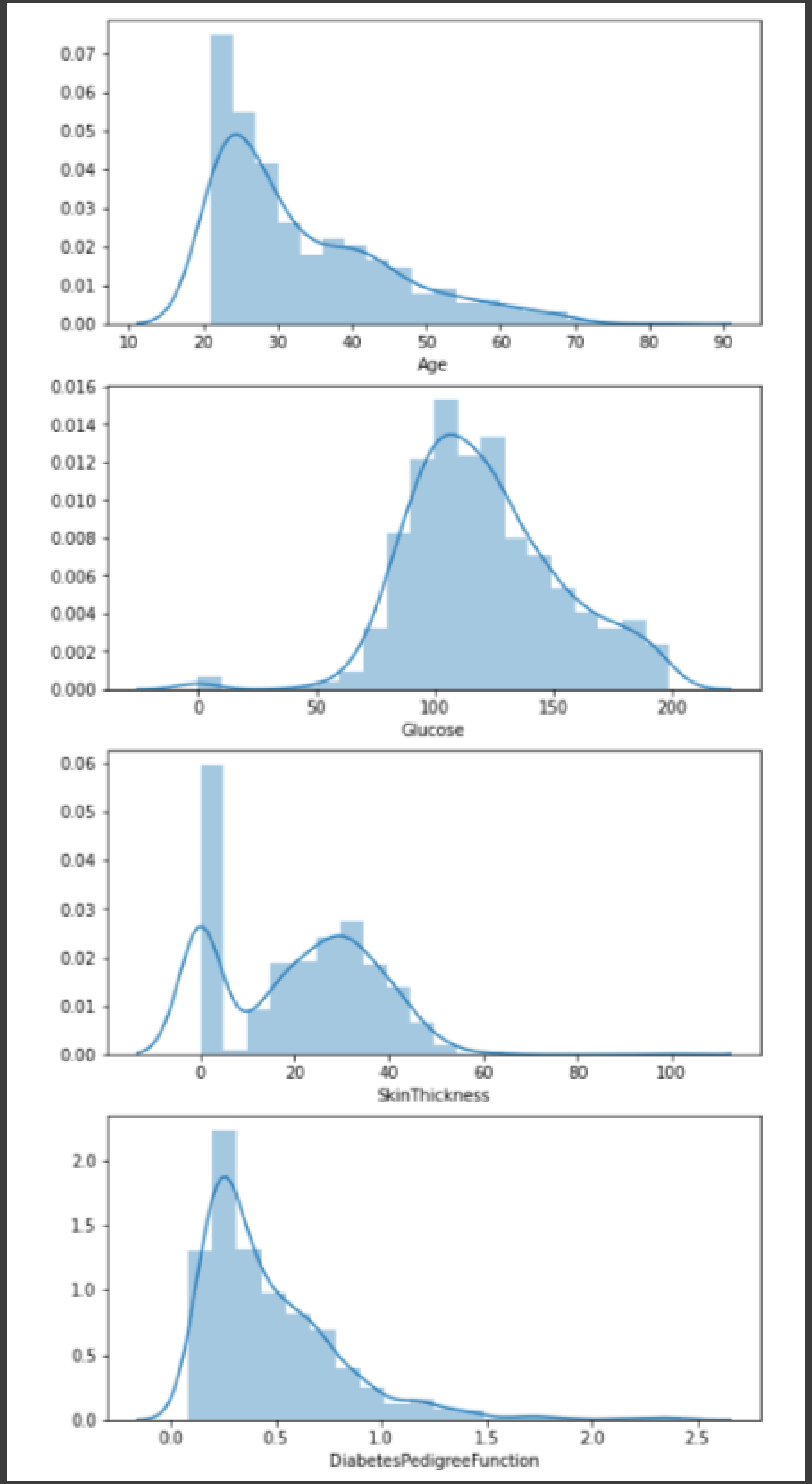
Correlation Matrix

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.00	0.13	0.14	-0.08	-0.07	0.02	-0.03	0.54	0.22
Glucose	0.13	1.00	0.15	0.06	0.33	0.22	0.14	0.26	0.47
BloodPressure	0.14	0.15	1.00	0.21	0.09	0.28	0.04	0.24	0.07
SkinThickness	-0.08	0.06	0.21	1.00	0.44	0.39	0.18	-0.11	0.07
Insulin	-0.07	0.33	0.09	0.44	1.00	0.20	0.19	-0.04	0.13
BMI	0.02	0.22	0.28	0.39	0.20	1.00	0.14	0.04	0.29
DiabetesPedigreeFunction	-0.03	0.14	0.04	0.18	0.19	0.14	1.00	0.03	0.17
Age	0.54	0.26	0.24	-0.11	-0.04	0.04	0.03	1.00	0.24
Outcome	0.22	0.47	0.07	0.07	0.13	0.29	0.17	0.24	1.00

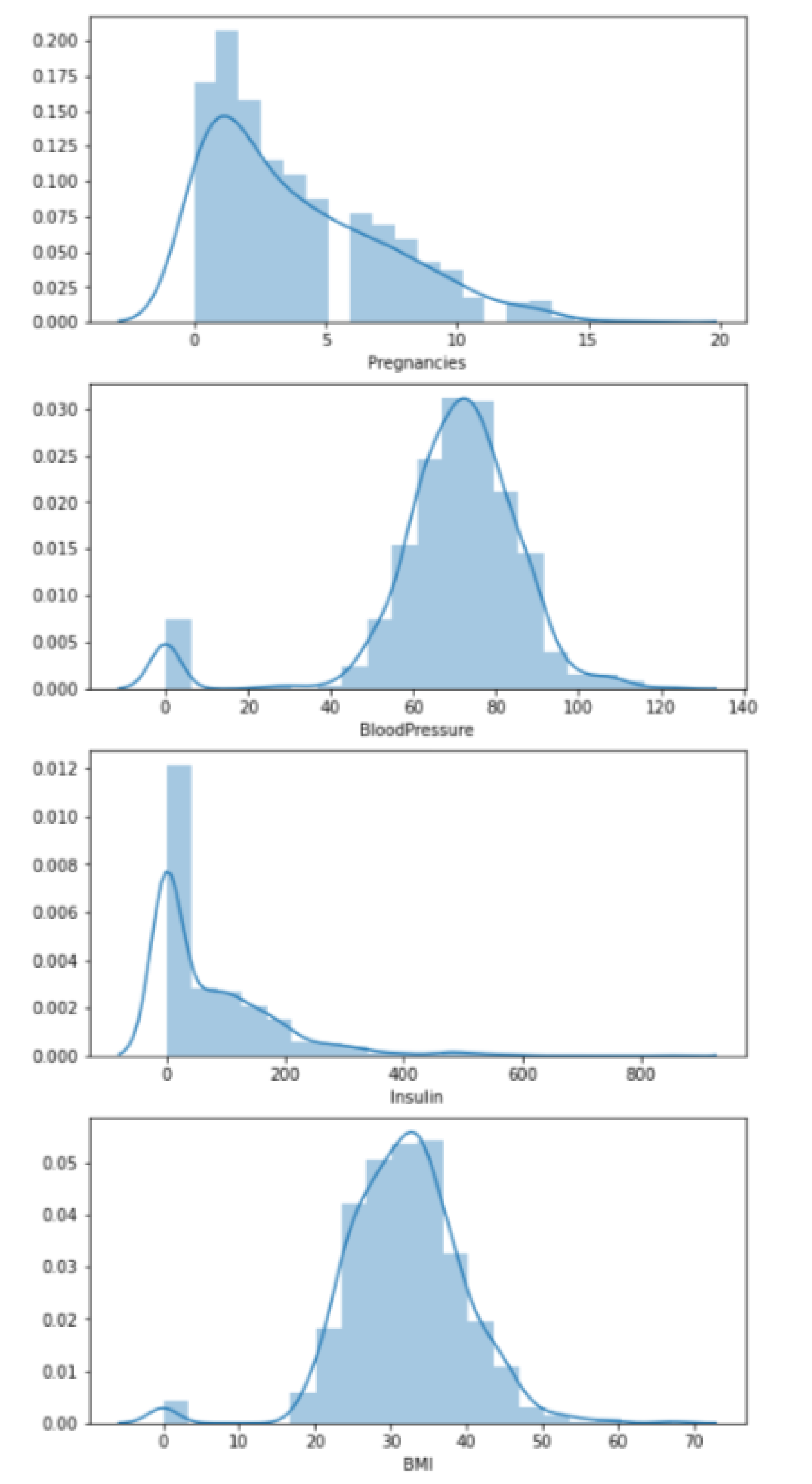
Distribution of Outcome Variable:



Histogram:



0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
10 20 30 40 50 60 70 80 90
Age
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
0 50 100 150 200
Glucose
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0 20 40 60 80 100
SkinThickness
2.0
1.5
1.0
0.5
0.0
0.0 0.5 1.0 1.5 2.0 2.5
DiabetesPedigreeFunction



0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0 5 10 15 20
Pregnancies
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0 20 40 60 80 100 120 140
BloodPressure
0.012
0.010
0.008
0.006
0.004
0.002
0.000
0 200 400 600 800
Insulin
0.05
0.04
0.03
0.02
0.01
0.00
0 10 20 30 40 50 60 70
BMI

Conclusion

- As per the main objective of the project is to classify and identify Diabetes Patients Using ML algorithms is being discussed throughout the project.

- we build the model using some machine learning algorithms such as logistic regression, decision tree, Random Forest and Gradient Boosting, these all are supervised machine learning algorithm in machine learning.

- As part of the future scope, we hope to try out different algorithms to optimize the feature output process, increase the feature similarity of data to improve the model's representation capability.

Reference

[1] Diabetes, World Health Organization (WHO)

[2] Mujumdar, A., & Vaidehi, V. (2019). Diabetes prediction using machine learning algorithms. Procedia Computer Science, 165, 292-299.

[3] Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access, 8, 76516-76531.

[4] Joshi, T. N., & Chawan, P. P. M. (2018). Diabetes prediction using machine learning techniques. Ijera, 8(1), 9-13.

[5] Sarwar, M. A., Kamal, N., Hamid, W., & Shah, M. A. (2018, September). Prediction of diabetes using machine learning algorithms in healthcare. In 2018 24th international conference on automation and computing (ICAC) (pp. 1-6). IEEE.

[6] Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. Procedia computer science, 132, 1578-1585.

THANK

YOU!!