AI use in Business Environment

HEALTHCARE DIAGNOSIS USING AI

BUSINESS CASE STUDY

CARDIOVASCULAR DISEASE USE CASE

TABLE OF CONTENTS

1

CARDIOVASCULAR DISEASE

OVERVIEW

Cardiovascular disease (CVD) group of conditions affecting the heart or blood vessels.

Importance of Early Diagnosis: Early and accurate diagnosis is crucial for effective treatment and prevention of CVD.
Challenges in Diagnosis: Accurately diagnosing CVD can be challenging due to the disease's complexity and the variability of symptoms.
Consequences of Misdiagnosis: Misdiagnosis or delayed diagnosis can have serious consequences for patients, such as heart attack, stroke, or even death.

BUSINESS PROBLEM

Misdiagnosis or **delayed diagnosis** can have serious consequences for patients with cardiovascular disease leading to higher healthcare costs, decreased patient satisfaction, and poorer patient outcome.

2

KEY STATISTICS

CVD accounts **32%** of Global Deaths

Year : 2019
Figures : Billions

Source : World Bank, WHO

3

HEART DISEASE CLASSIFICATION

EXTRACTING **BUSINESS VALUES** FROM **RAW DATA**

Data Preprocessing

Cleaning & Transforming Data

Data Visualization

Exploring Data Graphically

Data Modeling

Training Classifying Models

Result

Assessing Model Performance

EXTRACTING **BUSINESS VALUES** FROM **RAW DATA**

Data Preprocessing

Cleaning & Transforming Data

Data Visualization

Exploring Data Graphically

Data Modeling

Training Classifying Models

Result

Assessing Model Performance

KNOWING THE DATA

Heart Disease Analysis of 303 persons which include their age, sex and all the basic health related experimental data(like cholesterol, fasting blood sugar etc.)or medical history to classify whether they are suffering from Heart Disease or not.

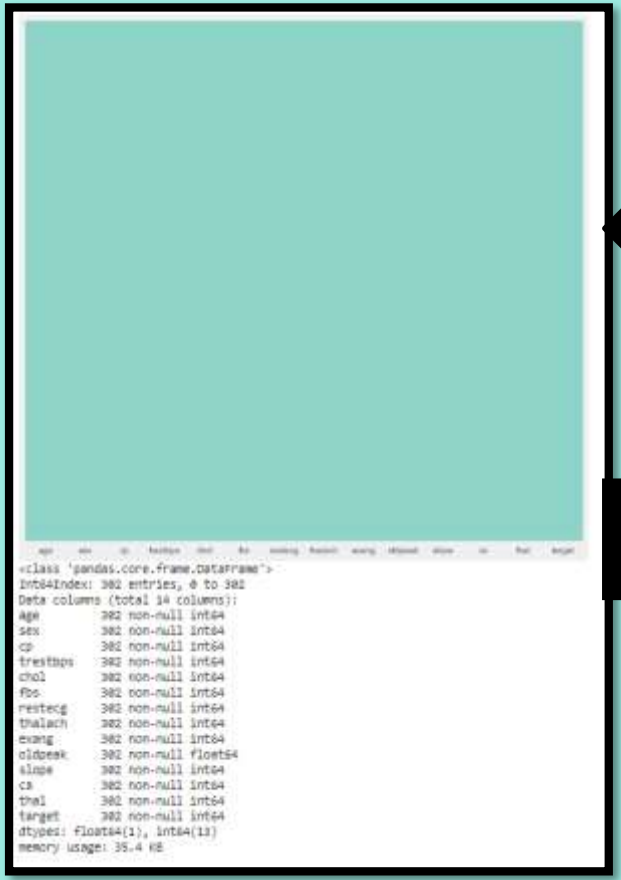
```
# getting the shape
data.shape
```

```
(303, 14)
```

```
# reading the head of the data
data.head()
```

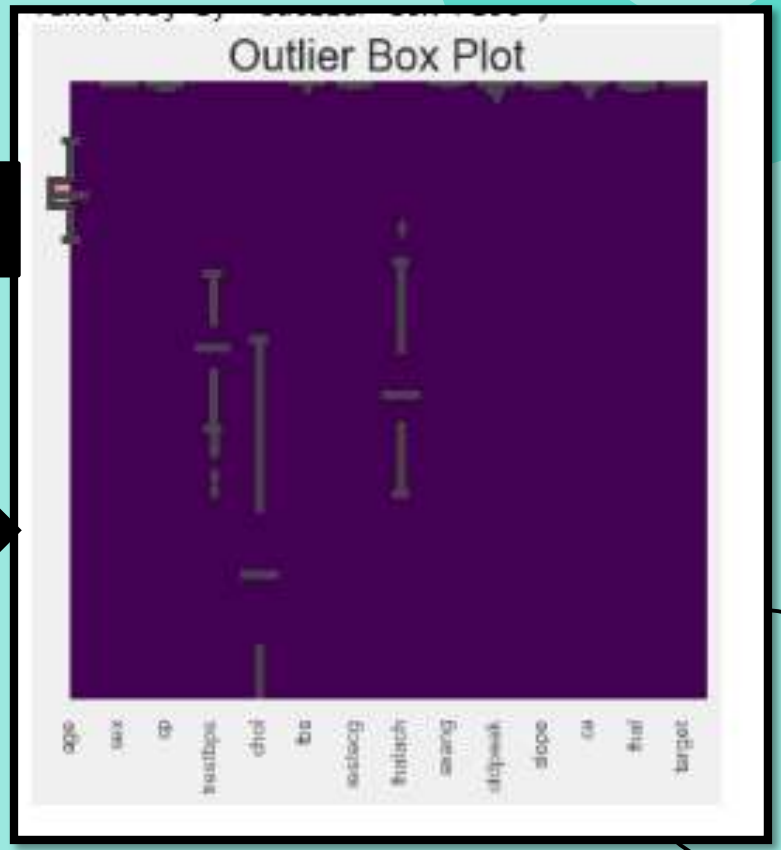
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

MISSING VALUES AND OUTLIERS

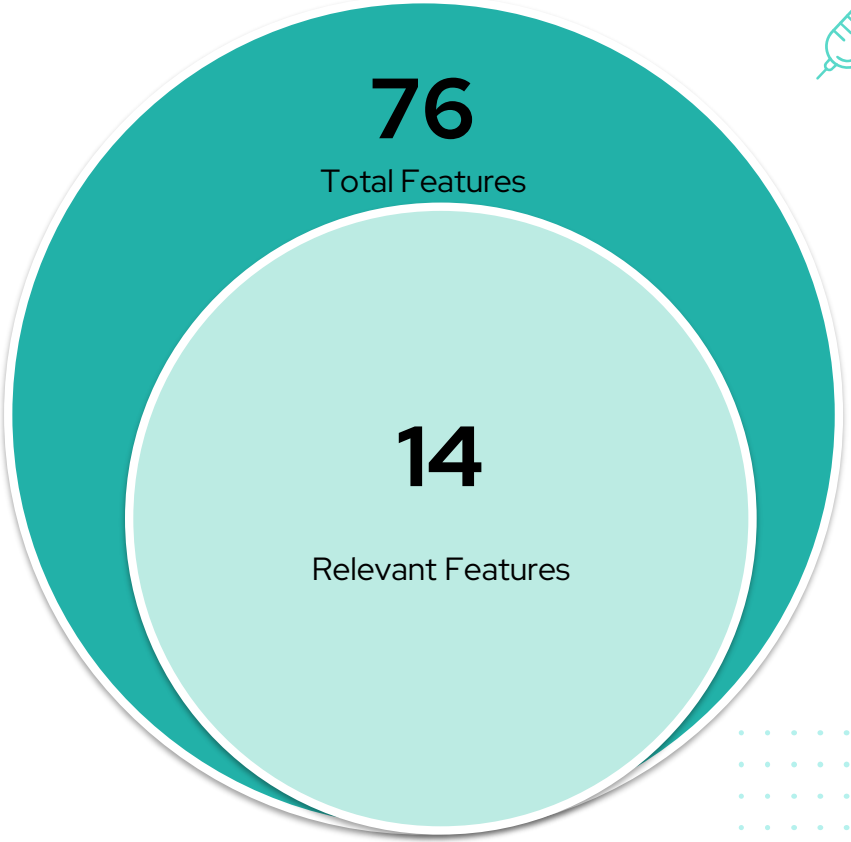


No Missing Values

No Significant Outliers



Feature Selection



FEATURE DICTIONARY

column name : Description

age : The person's age in years

sex : The person's sex (1 = male, 0 = female)

cp : The chest pain experienced (Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic)

trestbps : The person's resting blood pressure (mm Hg on admission to the hospital)

chol : The person's cholesterol measurement in mg/dl

fbs : The person's fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)

restecg : Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)

thalach : The person's maximum heart rate achieved

exang : Exercise induced angina (1 = yes; 0 = no)

oldpeak : ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot. See more here)

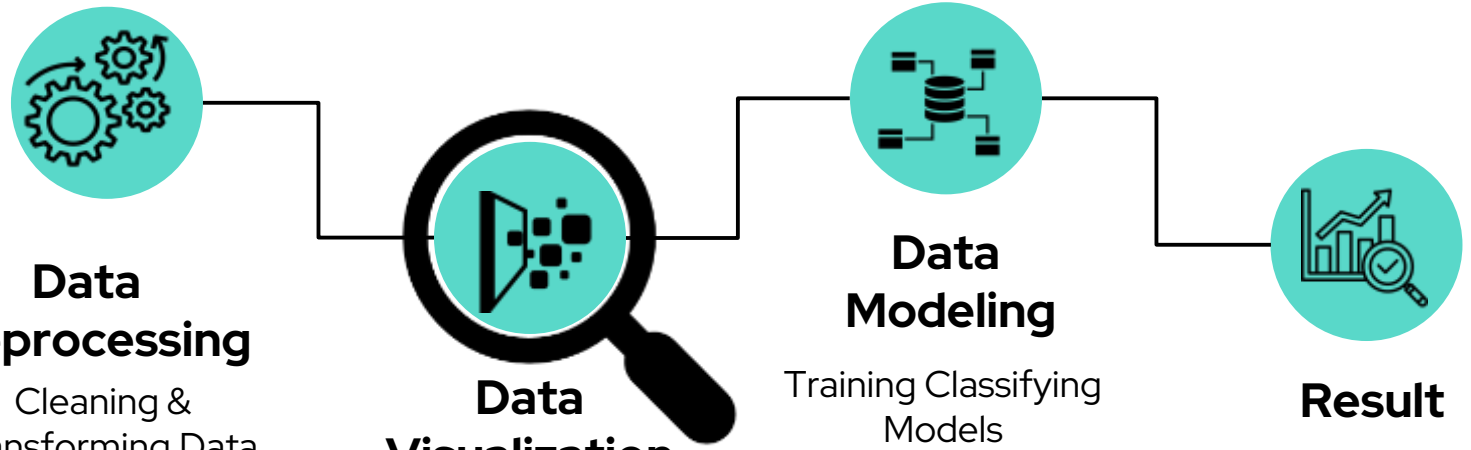
slope : the slope of the peak exercise ST segment (Value 1: upsloping, Value 2: flat, Value 3: downsloping)

ca : The number of major vessels (0-3)

thal : A blood disorder called thalassemia (3 = normal; 6 = fixed defect; 7 = reversable defect)

target : Heart disease (0 = no, 1 = yes)

EXTRACTING **BUSINESS VALUES** FROM **RAW DATA**



Data Preprocessing

Cleaning & Transforming Data

Data Visualization

Exploring Data Graphically

Data Modeling

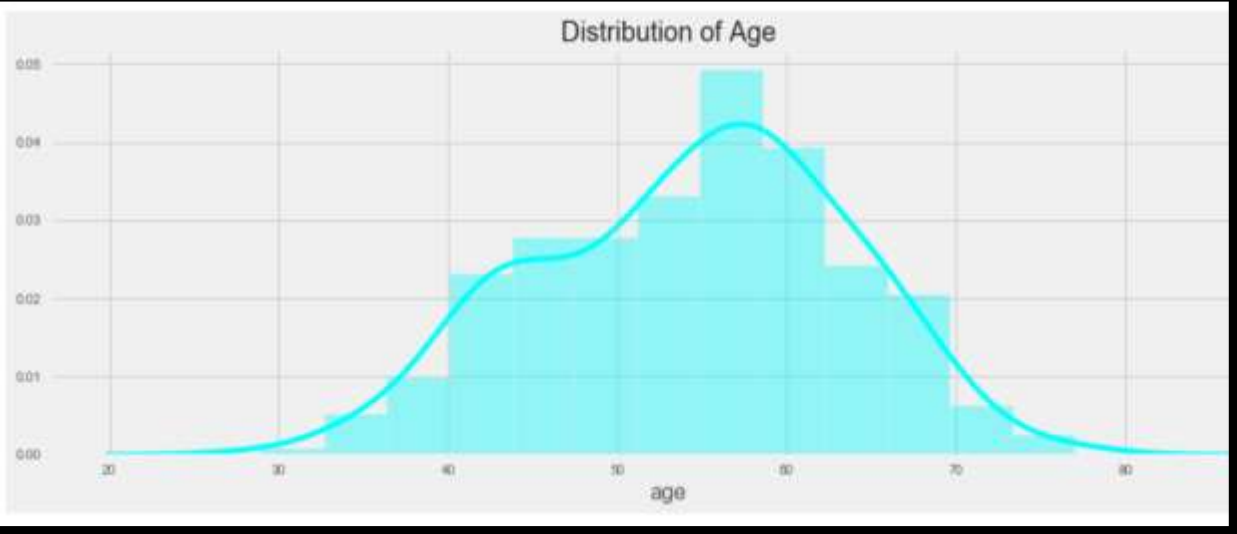
Training Classifying Models

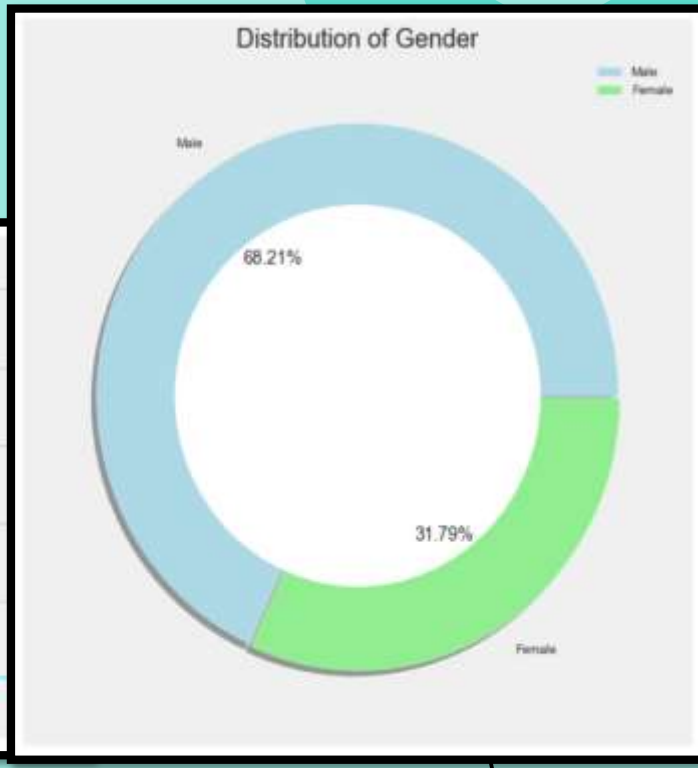
Result

Assessing Model Performance

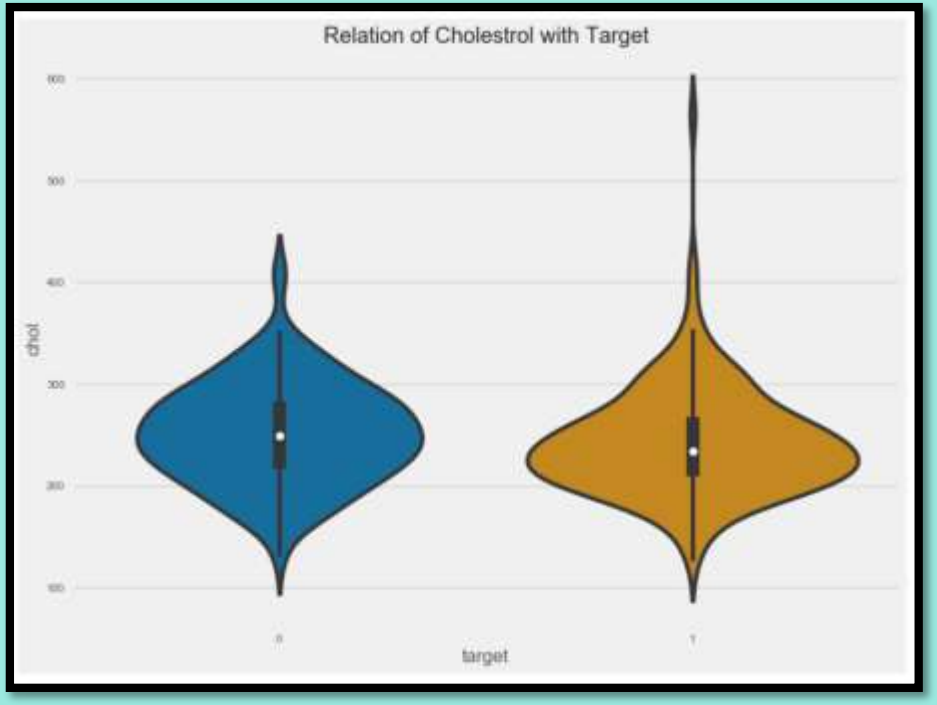
Correlations amongst the different Features

Demographics of Patients

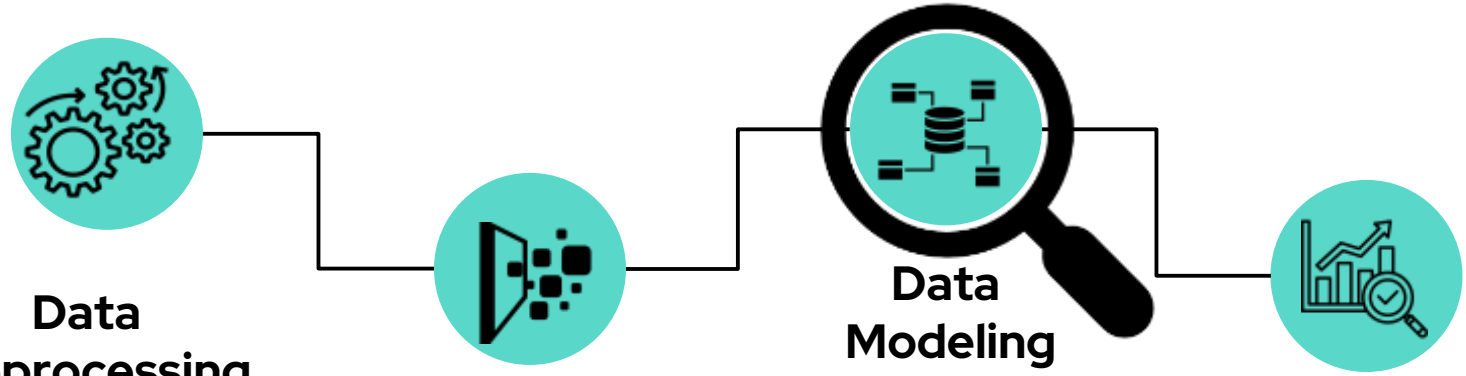




Cholesterol vs Heart Disease



EXTRACTING **BUSINESS VALUES** FROM **RAW DATA**



Data Preprocessing

Cleaning & Transforming Data

Data Visualization

Exploring Data Graphically

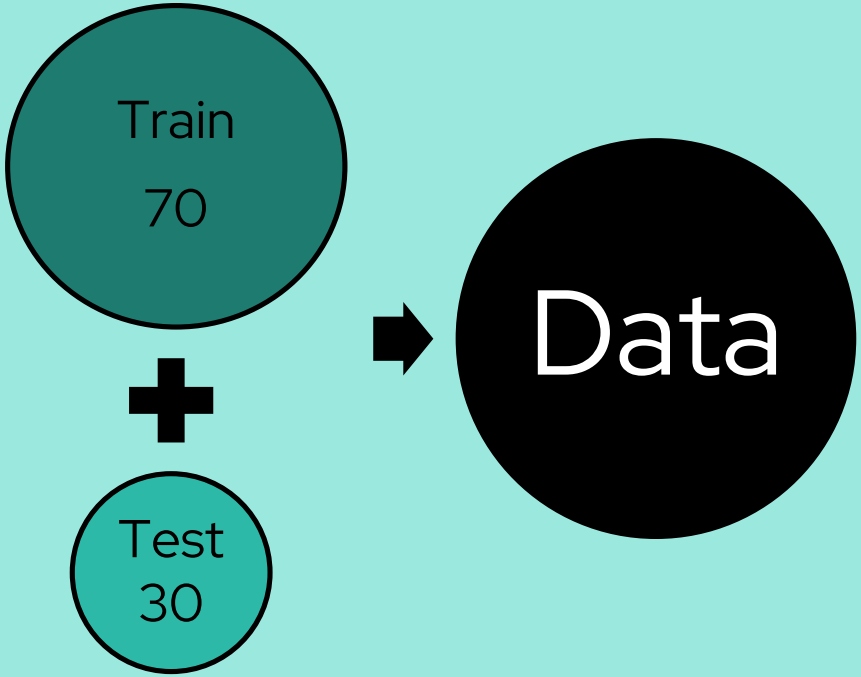
Data Modeling

Training Classifying Models

Result

Assessing Model Performance

The Optimal Train-Test Split Ratio



70:30 Split Ratio strikes a balance between overfitting and underfitting, allowing the model to generalize well to new data.

Sample Code

```
# splitting the sets into training and test sets

from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.3, random_state = 0)
#test_size=0.3 means data set divided into 70% and 30%
# getting the shapes
print("Shape of x_train :", x_train.shape)
print("Shape of x_test :", x_test.shape)
print("Shape of y_train :", y_train.shape)
print("Shape of y_test :", y_test.shape)
```

```
Shape of x_train : (211, 19)
Shape of x_test : (91, 19)
Shape of y_train : (211,)
Shape of y_test : (91,)
```

MODELS IMPLEMENTED

Logistics Regression

- Linear model for Binary Classification
- Simple and Interpretable

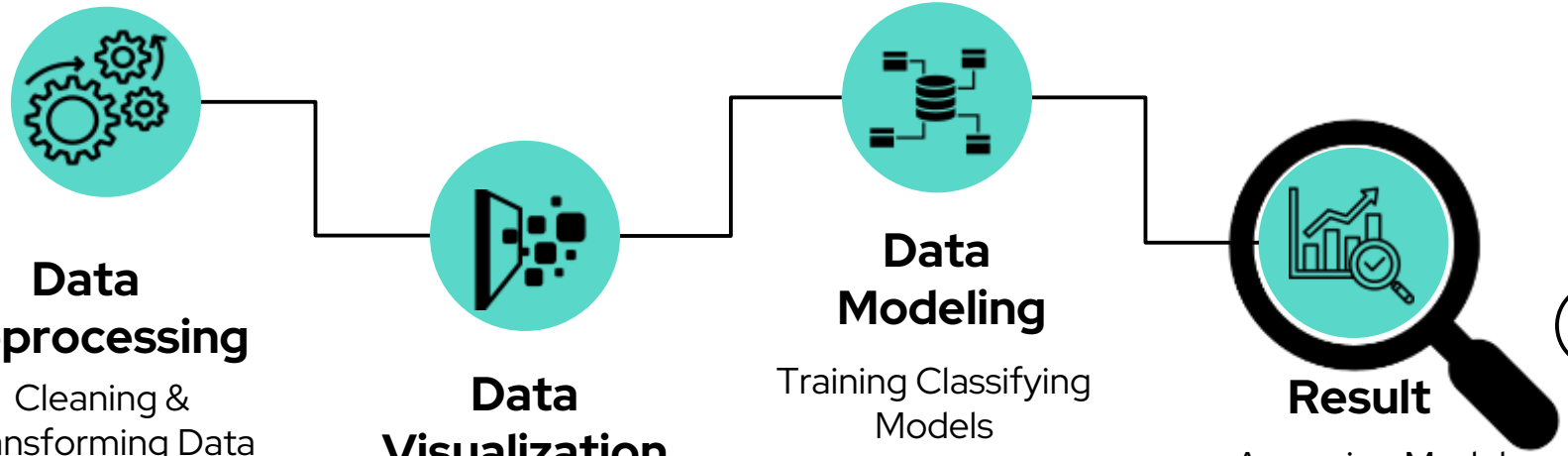
Decision Tree

- Recursive split of data based on informative features
- Handles complex datasets

Random Forest

- Aggregates multiple decision tress
- Handles complex datasets

EXTRACTING **BUSINESS VALUES** FROM **RAW DATA**



Data Preprocessing

Cleaning & Transforming Data

Data Visualization

Exploring Data Graphically

Data Modeling

Training Classifying Models

Result

Assessing Model Performance

Comparison of the models (1/2)

LOGISTIC REGRESSION

```
TRAINING ACCURACY : 0.8909952606635071
TESTING ACCURACY : 0.8241758241758241
              precision    recall  f1-score

           0       0.87      0.76      0.81
           1       0.79      0.89      0.84
```

DECISION TREE CLASSIFIER

```
TRAINING ACCURACY : 0.943127962085308
TESTING ACCURACY : 0.7582417582417582
              precision    recall  f1-score

           0       0.77      0.73      0.75
           1       0.75      0.78      0.77
```

RANDOM FOREST CLASSIFIER

```
TRAINING ACCURACY : 0.9478672985781991
TESTING ACCURACY : 0.7692307692307693
              precision    recall  f1-score

           0       0.79      0.73      0.76
           1       0.76      0.80      0.78
```

Comparison of the models (2/2)

- **Logistic regression** achieved the **highest testing** accuracy of 82.4% and **highest F1 score.**

- **Decision tree** had a testing accuracy of 75.8%, which is lower than both logistic regression and random forest, and may also be overfitting the training data.

- **Random forest** achieved **the highest training accuracy** of 94.8% but had a **lower testing accuracy** of 76.9% compared to logistic regression, which suggests that it may be overfitting the training data.

LOGISTIC REGRESSION

highest balanced accuracy on the test and train sets

4

BUSINESS TAKEAWAYS

POTENTIAL BENEFITS

- Accurate Diagnosis
- Improved Patient outcomes
- Cost savings
- Reduced Healthcare Resources Costs
- Improved Patient Satisfaction

LIMITATIONS AND FUTURE SCOPE

- Potential for scaling
- Potential for further research and development

THANK YOU

QUESTION?