DATA SCIENCE

Business Required Document [BRD] – The Secret Sauce!

TRIBE : A

- Addanki Pallavi – Team Leader
- Anshuman Kumar Tiwari
- Bhargav Rohit
- Dhriti Dogra
- Jaswant Panigrahy
- Pawan Kumar Patel
- Polaki Siddharth
- Rajat Kumar
- Rohan Mule
- Rohit Kumar Gupta
- Shreyansh Malewar
- Vaishnavi Chauhan

Data Science

Data science is an essential part of any industry today, given the massive amounts of data that are produced. Data science is one of the most debated topics in the industries these days. Its popularity has grown over the years, and companies have started implementing data science techniques to grow their business and increase customer satisfaction.

"Data really powers everything that we do."

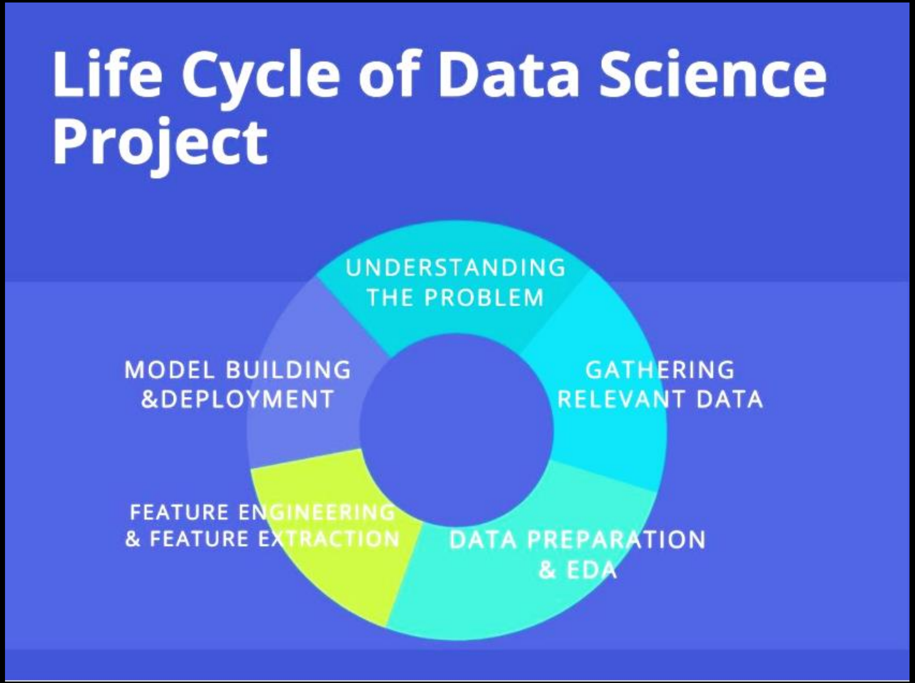
Data Science Life Cycle

A Data Science Life Cycle is nothing but a repetitive set of steps that you need to take to complete and deliver a project/product to your client. Although the data science projects and the teams involved in deploying and developing the model will be different, every data science life cycle will be slightly different in every other company. However, most of the data science projects happen to follow a somewhat similar process.

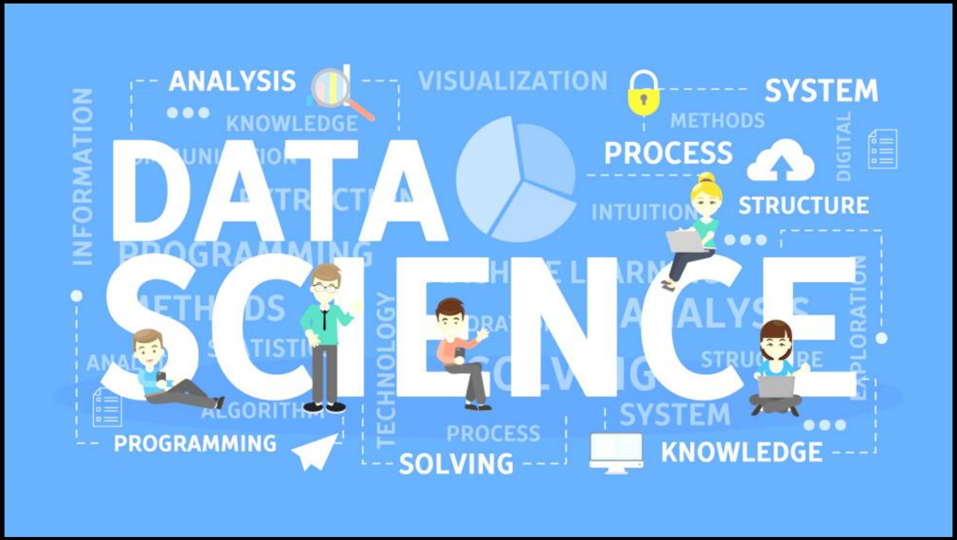
In order to start and complete a data science-based project, we need to understand the various roles and responsibilities of the people involved in building, developing the project. Let us take a look at those employees who are involved in a typical data science project:

Who Are Involved in The Projects:

1. Business Analyst
2. Data Analyst
3. Data Scientists
4. Data Engineer
5. Data Architect
6. Machine Learning Engineer


Life Cycle of Data Science Project
UNDERSTANDING THE PROBLEM
GATHERING RELEVANT DATA
DATA PREPARATION & EDA
FEATURE ENGINEERING & FEATURE EXTRACTION
MODEL BUILDING &DEPLOYMENT

Summary Statement

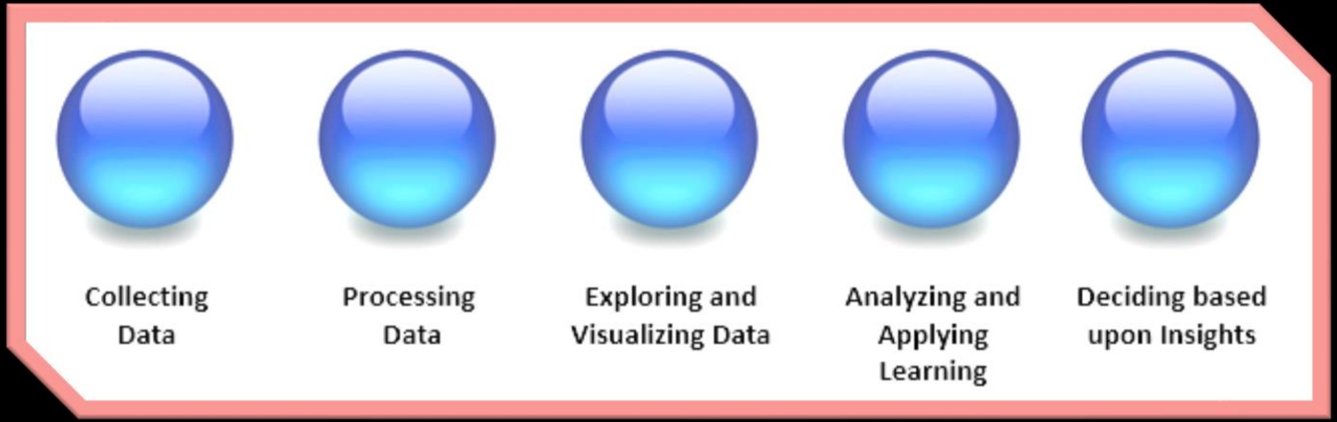


Project Requirement will combine some data set in a particular domain with some computational, statistical, or analytical technique to produce an interesting result, e.g. a model, a predictive model, a software system, or a visualization. The project will necessarily require the construction of software for data ingestion, wrangling, computation and analyses, and the production of a report or application. The final requirements include a demonstration of the software product created by the team as well as a short paper describing the system and it's execution in terms of the data science pipeline.

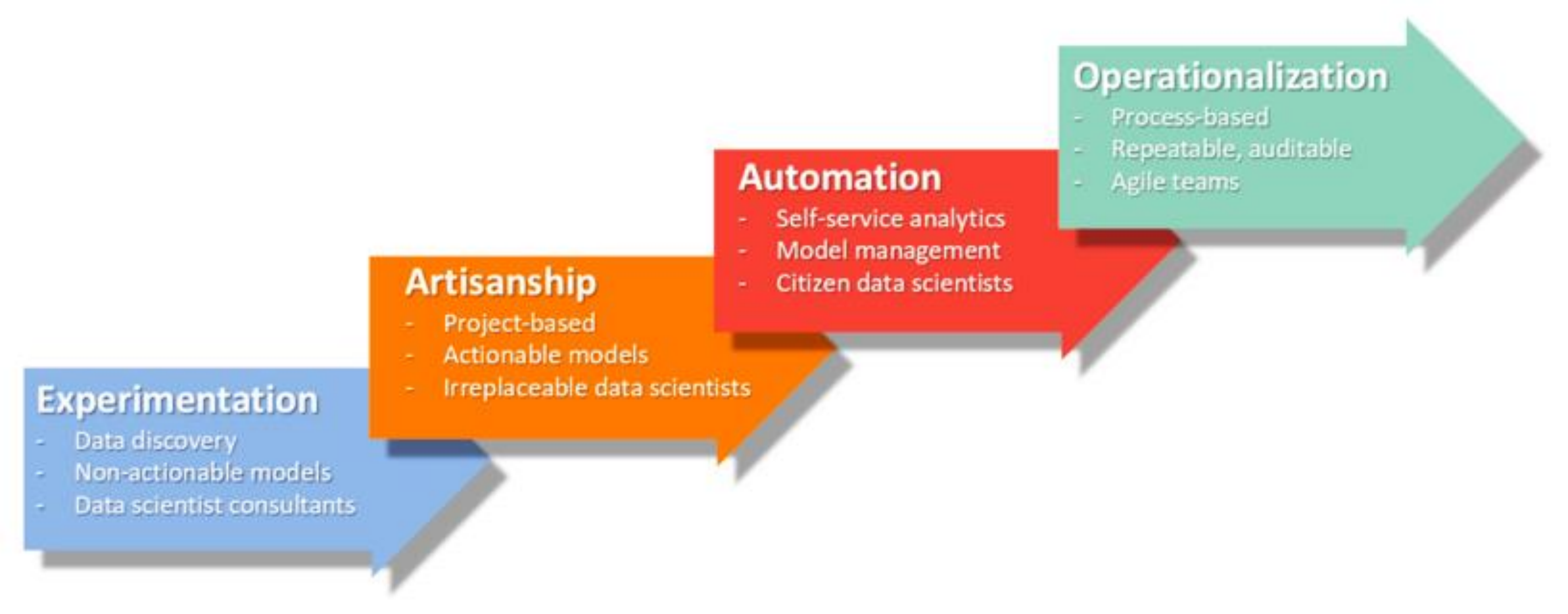
Project Objectives

- ✓ Understand the Business
- ✓ Get the Data
- ✓ Explore and Clean the Data
- ✓ Enrich the Dataset
- ✓ Build Helpful Visualizations
- ✓ Get Predictive
- ✓ Iterate, Iterate, Iterate

- ✓ Define The Business Impact
- ✓ Connect Business Objectives to The Data
- ✓ Clarify The Objective


Collecting Data
Processing Data
Exploring and Visualizing Data
Analyzing and Applying Learning
Deciding based upon Insights

Needy-Need!



Operationalization
- Process-based
- Repeatable, auditable
- Agile teams

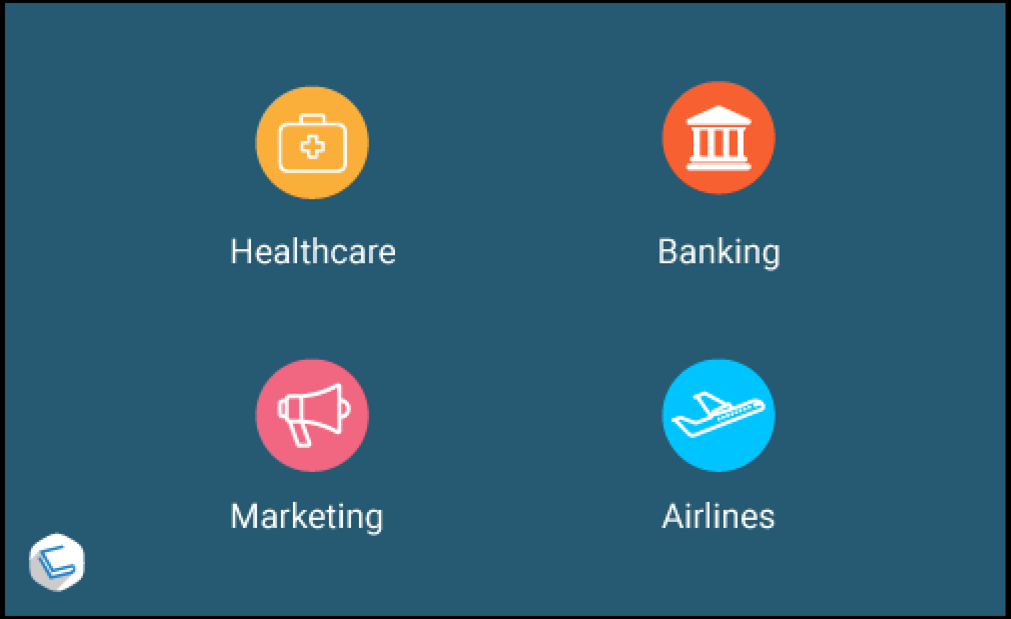
Automation
- Self-service analytics
- Model management
- Citizen data scientists

Artisanship
- Project-based
- Actionable models
- Irreplaceable data scientists

Experimentation
- Data discovery
- Non-actionable models
- Data scientist consultants

Data Science Projects helps organizations identify and refine target audiences by combining existing data with other data points for developing useful insights.

SCOPE



When we are working on a data science project, the project in a professional environment, it is likely to be an integral part of a more extensive setup. There may be people or teams that might get affected by the project or maybe are part of your team. A nicely laid out scope gives us a grip on the problem outlines and facilitates communications with stakeholders.

There are 4 Parts of a Data Project Scope as follows:

1. Context
   1. Needs (the project is trying to fulfill)
   2. Vision (of the achievement)
   3. Outcome

The STD

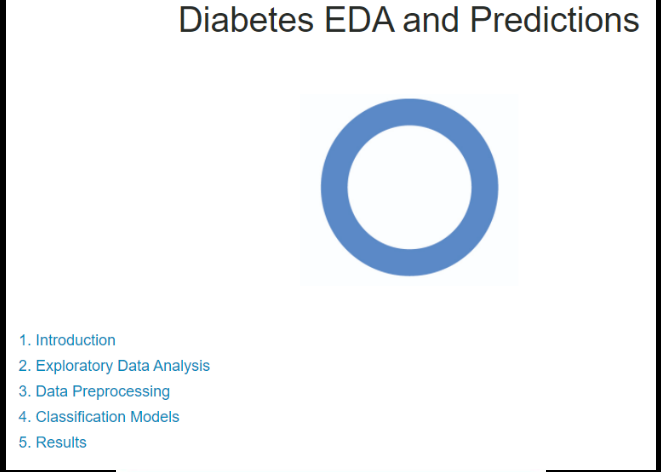
Each project is broken down into many small deliverable tasks that will given a short-time estimation. Deliverables are grouped into cycles. These small tasks will be pulled by team members until completion, trying to finish all tasks until the cycle ends. This helps to ensure that all stakeholders are aware of what is required and when it will be required.

Schedule

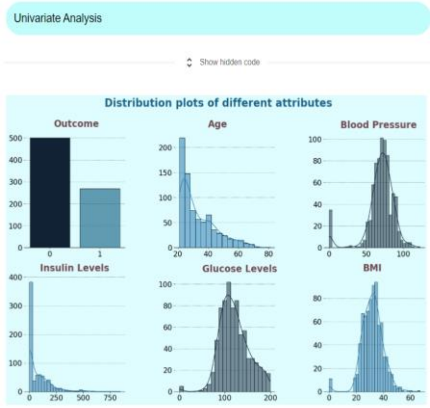
Timeline

Deadline

Assumptions

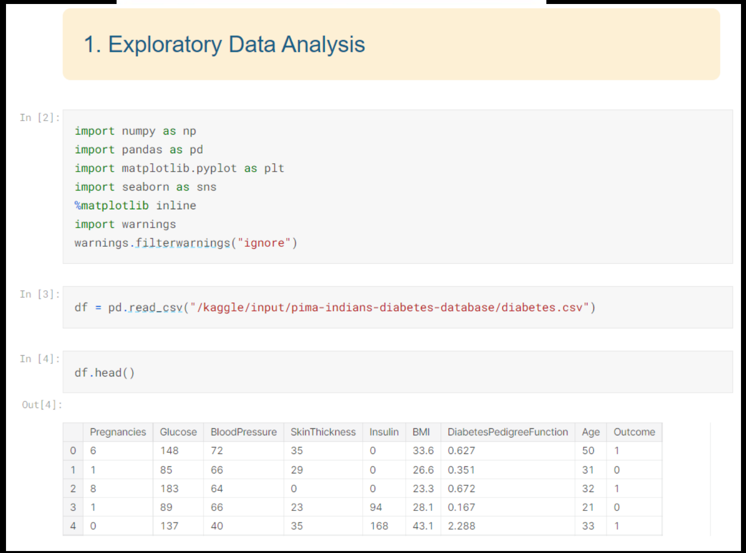


Diabetes EDA and Predictions



Univariate Analysis

Show hidden code



1. Exploratory Data Analysis

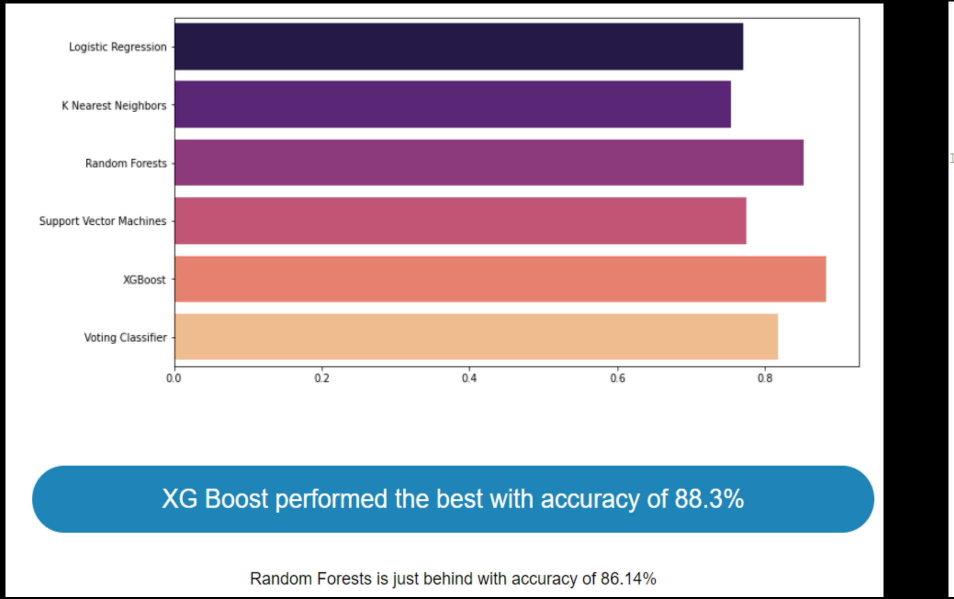
```
In [2]:
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
import warnings
warnings.filterwarnings("ignore")
```

```
In [3]:
df = pd.read_csv("/kaggle/input/pima-indians-diabetes-database/diabetes.csv")
```

```
In [4]:
df.head()
```

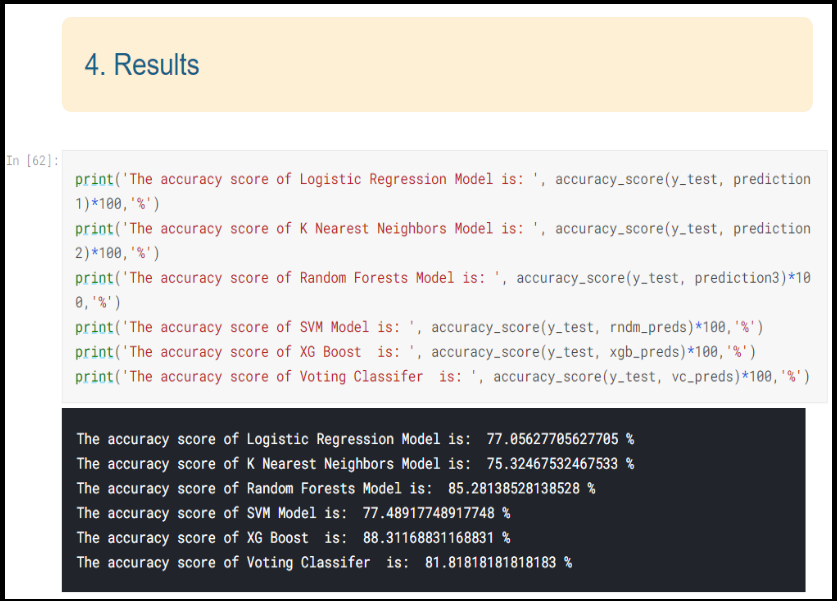
Out[4]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1



XG Boost performed the best with accuracy of 88.3%

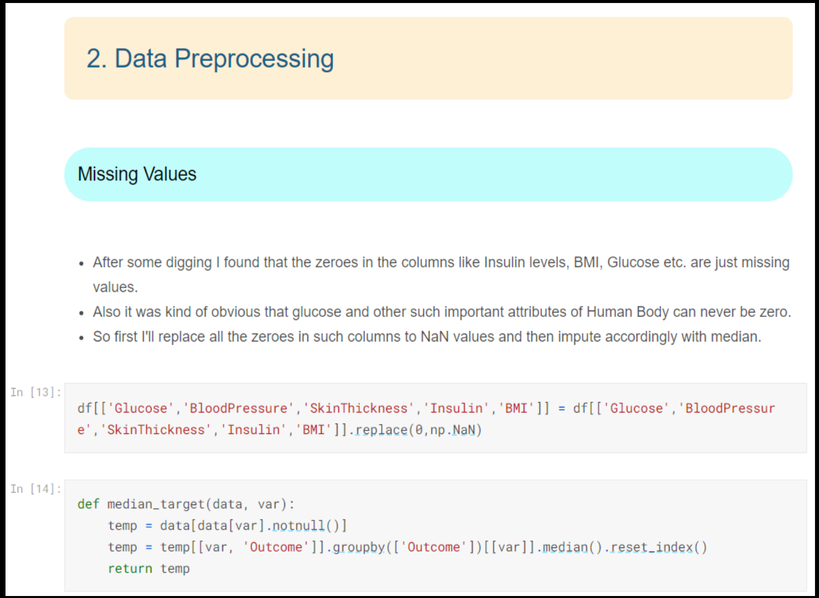
Random Forests is just behind with accuracy of 86.14%



4. Results

```
In [62]:
print('The accuracy score of Logistic Regression Model is: ', accuracy_score(y_test, prediction1)*100,'%')
print('The accuracy score of K Nearest Neighbors Model is: ', accuracy_score(y_test, prediction2)*100,'%')
print('The accuracy score of Random Forests Model is: ', accuracy_score(y_test, prediction3)*100,'%')
print('The accuracy score of SVM Model is: ', accuracy_score(y_test, rndm_preds)*100,'%')
print('The accuracy score of XG Boost  is: ', accuracy_score(y_test, xgb_preds)*100,'%')
print('The accuracy score of Voting Classifer  is: ', accuracy_score(y_test, vc_preds)*100,'%')
```

```
The accuracy score of Logistic Regression Model is:  77.05627705627705 %
The accuracy score of K Nearest Neighbors Model is:  75.32467532467533 %
The accuracy score of Random Forests Model is:  85.28138528138528 %
The accuracy score of SVM Model is:  77.48917748917748 %
The accuracy score of XG Boost  is:  88.31168831168831 %
The accuracy score of Voting Classifer  is:  81.81818181818183 %
```



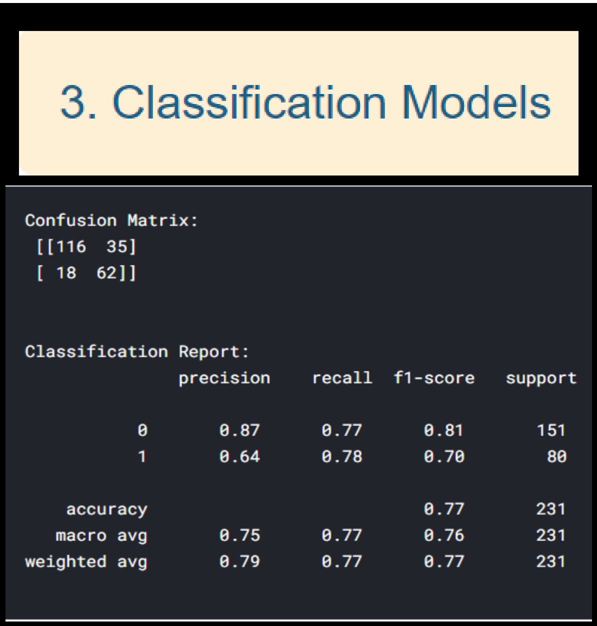
2. Data Preprocessing

Missing Values

- After some digging I found that the zeroes in the columns like Insulin levels, BMI, Glucose etc. are just missing values.
- Also it was kind of obvious that glucose and other such important attributes of Human Body can never be zero.
- So first I'll replace all the zeroes in such columns to NaN values and then impute accordingly with median.

```
In [13]:
df[['Glucose','BloodPressure','SkinThickness','Insulin','BMI']] = df[['Glucose','BloodPressure','SkinThickness','Insulin','BMI']].replace(0,np.NaN)
```

```
In [14]:
def median_target(data, var):
    temp = data[data[var].notnull()]
    temp = temp[[var, 'Outcome']].groupby(['Outcome'])[[var]].median().reset_index()
    return temp
```



3. Classification Models

```
Confusion Matrix:
 [[116  35]
 [ 18  62]]


Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.77      0.81       151
           1       0.64      0.78      0.70        80

    accuracy                           0.77       231
   macro avg       0.75      0.77      0.76       231
weighted avg       0.79      0.77      0.77       231
```

Deal with Data

Although it contains figures, statistics, and facts, **unstructured data** is usually text-heavy or configured in a way that's difficult to analyze.

This is how we will work with the given data –

Working Together : Specializes in cleaning, sorting, or analyzing unstructured data that will automatically pasre, sort and analyze unstructured data in relatively new areas of development

Evaluate the Value of your Data, and Clean your Records : Not all unstructured data is worth analyzing, or even worth keeping. If the data are coming from a source that won't yield much value for your organization, you should consider deleting it.

Take a Random Sample and Create a "Dictionary" : Analyzing the entire text file of your data manually is a virtually impossible task or at least an incredibly time-intensive one. Instead, it's better to take a random sample or stratified sample from the collection, and use that to build a "dictionary" that you can use to find similar patterns in the rest of the data.

Clean the Entire Dataset : By using the framework you created from a random sample, you should be able to write a script that allows you to clean your entire dataset. Ideally, you'll be able to classify and segment those data so you can analyze it easily in the future.

Analyze it : Assuming your data is properly structured and easy to digest, you can analyze those data and start making decisions based on the insights you gain. Once structured, you can treat your data like any other structured dataset you come across.

AT THE END, ALL THAT MATTERS IS:

Process
People
Technology
Product
Materials and Supplies
Facilities
Machinery and Equipment
Others as necessary (Depending on the Organization)

THANK YOU!

References -

ReQtest

kaggle

KDnuggets

GeeksforGeeks

Towards
Data Science

Google
Images