KIET Group of Institutions, Ghaziabad

Computer Science

Internship Report

on

Artificial Intelligence

Summer Internship at Department of Computer Science

(Mini-Project)

August-September

(2021)

Submitted By:

Name of Student: Archit Rajesh Srivastava
Course/Branch(Sem-Sec.): B Tech – Computer Science (3A)
Class Roll No.: 38

INDEX

ACKNOWLEDGEMENT

I've got this golden opportunity to express my kind gratitude and sincere thanks to my Head of Institution, KIET Group of Institutions of Engineering and Technology, and Head of Department of Computer Science, project coordinator and my mentors for their kind support and necessary counselling in the preparation of this project report. I'm also indebted to each and every person responsible for the making up of this project directly or indirectly.

I must also acknowledge our deep debt of gratitude to each one of my colleagues who led this project to come out in the way it is. It's my hard work and untiring sincere efforts and mutual cooperation to bring out the project work. Last but not the least, I would like to thank my parents for their sound counselling and cheerful support. They have always inspired us and kept our spirit up.

Name of Student: Archit Rajesh Srivastava
Course and Branch: B Tech; Computer Science
Semester: Third
University Roll No: 2000290120039

CERTIFICATE

This is to certify that the internship project report entitled "**Diabetes prediction using classification techniques - DIBPredict**" submitted by Mr. **Archit Rajesh Srivastava** in the Department of **Computer Science** of KIET Group of Institutions, Ghaziabad, affiliated to Dr. A. P. J. Abdul Kalam Technical University, Lucknow, Uttar Pradesh, India, is a record of candidate summer internship. **He has successfully completed his internship** under my supervision and guidance and is worthy of consideration for the same.

Signature of Supervisor:

Supervisor's Name:

Date:

ABSTRACT

Diabetes is an increasingly growing health issue due to our inactive lifestyle. If it is detected in time then through proper medical treatment, adverse effects can be prevented. To help in early detection, technology can be used very reliably and efficiently. Using machine learning we have built a predictive model that can predict whether the patient is diabetes positive or not.

In this project, the objective is to predict whether the person has Diabetes or not based on various parameters like Glucose level, Insulin, Age, BMI. We will use the PIMA India Databases from the UCI Machine learning repository. We will develop this project in six steps which follows data gathering to model deployment.

All the standard libraries like NumPy, pandas, matplotlib and seaborn were used. We use NumPy for linear algebra operations, pandas for using data frames, matplotlib and seaborn for plotting graphs. The dataset is imported using the pandas command read_csv().

INTRODUCTION TO PROJECT

Project:

In this project, the objective is to predict whether the person has Diabetes Or not based on various features like Glucose level, Insulin, Age, and BMI.

Contributors:

Archit Rajesh Srivastava (CS Branch, 2000290120039)

Adrika Tripathi (CS Branch, 2000290120015)

Prashant Kumar Pal (CSE Branch, 2000290100102)

Dataset used:

The dataset used in this project is **Pima Indians Diabetes Dataset** downloaded using Kaggle This original dataset has been provided by the **National Institute of Diabetes and Digestive and Kidney Diseases**. The link to the dataset and code is provided in the GitHub repository link below.

This dataset is used to predict whether a patient is likely to get diabetes based on the input parameters like Age, Glucose, Blood pressure, Insulin, BMI, etc. Each row in the data provides relevant information about the patient. It is to be noted that all patients here are females minimum 21 years old belonging to Pima Indian heritage.

Features of Dataset used:

The dataset contains 768 individuals' data with 9 features set. The detailed description of all the features are as follows:

- *Pregnancies:* indicates the number of pregnancies
- *Glucose:* indicates the plasma glucose concentration
- *Blood Pressure:* indicates diastolic blood pressure in mm/Hg
- Skin Thickness: indicates triceps skinfold thickness in mm
- Insulin*:* indicates insulin in U/mL
- *BMI:* indicates the body mass index in kg/m2
- *Diabetes Pedigree Function:* indicates the function which scores likelihood of diabetes based on family history
- *Age:* indicates the age of the person
- *Outcome:* indicates if the patient had a diabetes or not (1 = yes, 0 = no)

Link to GitHub repository:

https://github.com/archit1203/DiabetesPrediction

DETAILS OF TASKS ASSINGNED

MOOC COURSE:

We were required to complete 30 hrs. of MOOC. The certificate was submitted titled "**Introduction to Artificial Intelligence (AI)**". The course was instructed by Rav Ahuja.

Literature Review Report:

A summary of research papers was required to be submitted that consists of at least 5 papers. The titles of the research papers are as follows:

1. **Machine learning in cybersecurity: A review** - Anand Handa | Ashu Sharma | Sandeep K. Shukla
2. **Artificial intelligence in liver disease** - Hye Won Lee| Joseph JY Sung | Sang Hoon Ahn
3. **Insightful artificial intelligence -** Marta Halina
4. **Minds and Computers: An Introduction to the Philosophy of Artificial Intelligence** - Matt Carter
5. **Next generation technologies for smart healthcare: challenges, vision, model, trends and future directions -** Shreshth Tuli | Shikhar Tuli | Gurleen Wander | Praneet Wander | Sukhpal Singh Gill | Schahram Dustdar | Rizos Sakellariou | Omer Rana8

MINI PROJECT:

We built a mini project on prediction of diabetes using various machine learning techniques. The project was built on Jupyter notebook using Python language.

We were assigned a task to take the dataset of diabetes patient and analyze its various attributes and by selecting suitable attributes and with the help of various machine learning algorithms We had to develop a model and train it with this dataset so that it can predict whether a patient is diabetic or not on its own. Here, we used various algorithms like Logistic regression, Decision tree regression, Random Forest and KNN and ran the code on Jupyter Notebook and use various ML libraries like Pandas, seaborne, NumPy, matplotlib to generate better result and to achieve a good accuracy

TECHNICAL LEARNINGS

During the MOOC, I learned about what AI is, its applications and how it is transforming our lives. Basic concepts of AI and how it learns from the data. I also learnt about issues and concerns surrounding AI, including - ethical considerations, bias, jobs, etc. - their impact on society. At last, I learnt about the future with AI, as well as hear from experts about their advice to learn and start a career in AI. And also demonstrate AI in action by utilizing Computer Vision to classify images.

Also, we were introduced to many IDEs like Jupyter notebook and Spyder, PyCharm, etc and multiple libraries also. With a huge dataset, we ran the code on college's supercomputer PARAM SHAVAK, which is based on Linux OS and it was also something new.

To sum up things, following were the things that We learned during this project:

- Working on different IDEs.
- Basic knowledge of Python programming.
- Using different libraries on Python.
- Basics of Artificial intelligence.
- Different algorithms and their working principle.
- Working remotely on a Linux based computer.
- Searching internet and communities to find things and fix up errors in codes.
- Working with team.
- How to differentiate between different algorithms and where to use them.
- Accuracy, Precision and recall values along with confusion matrix.
- Reading and evaluating research papers and how to write them.

PARAM SHAVAK SUPERCOMPUTER

The PARAM SHAVAK SUPER COMPUTING FACILITY For AI AND ML at the Centre of Excellence consist of dedicated and hardworking students along with highly skilled faculties of the computer science department.

It aims to provide practical knowledge and familiarity with how the industry is adopting and utilizing AI and ML. The students are provided dedicated cabins which with all the necessary equipment including systems installed with the latest version of Linux.

For high-performance computation, they have been provided Param Shavak desktop supercomputer by CDAC. The Param Shavak has dual intel Xenon Gold 6132, 14 cores server grade processors with each core having a minimum 2.6GHz clock speed, 96GB (expandable) DDR4 2666MHz RAM in a balanced configuration, two 1GbE networks ports, 2x16 PCI-E Gen3 slots for GPU/Co-processors with NVIDIA P5000 GPUs. For deployment of their projects, students are provided NVIDIA Jetson devices. They have also been provided with access to the latest contents published by top publishers, including O'Reilly and Wiley.

PROJECT

In [60]:

```
#Importing Libraries
```

In [61]:

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
```

In [62]:

```
#Importing Dataset
```

In [64]:

```
#Extracting data
dataset=pd.read_csv('diabetes.csv')
dataset.head()
```

Out[64]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunctio
0	6	148	72	35	0	33.6	0.62
1	1	85	66	29	0	26.6	0.35
2	8	183	64	0	0	23.3	0.67
3	1	89	66	23	94	28.1	0.16
4	0	137	40	35	168	43.1	2.28

In [65]:

```
#Our dataset dimesnions
dataset.shape
```

Out[65]:

```
(768, 9)
```

In [66]:

```
dataset.describe()
```

Out[66]:

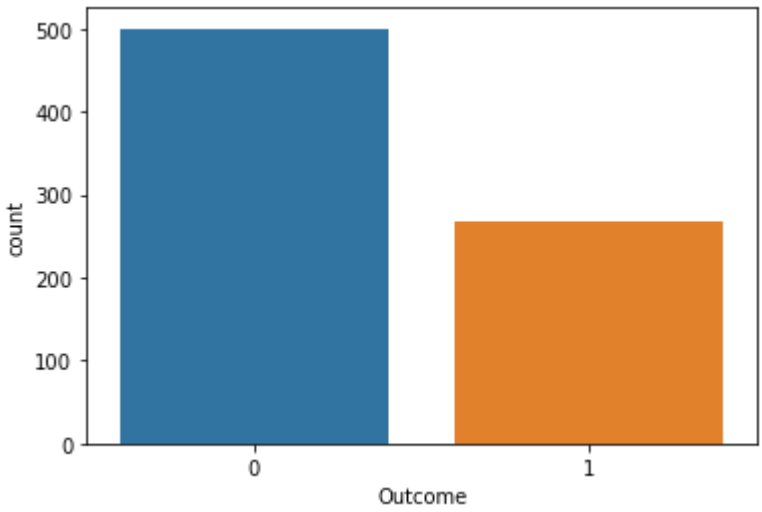
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	Diabete
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	

In [67]:

```
#Countng values of outcomes having 0 or 1, 0 means non diabetic and 1 means diabetic
sns.countplot(x='Outcome',data=dataset)
```

Out[67]:

```
<AxesSubplot:xlabel='Outcome', ylabel='count'>
```



In [68]:

```
dataset['Outcome'].value_counts()
```

Out[68]:

```
0    500
1    268
Name: Outcome, dtype: int64
```

In [69]:

```
dataset.groupby('Outcome').mean()
```

Out[69]:

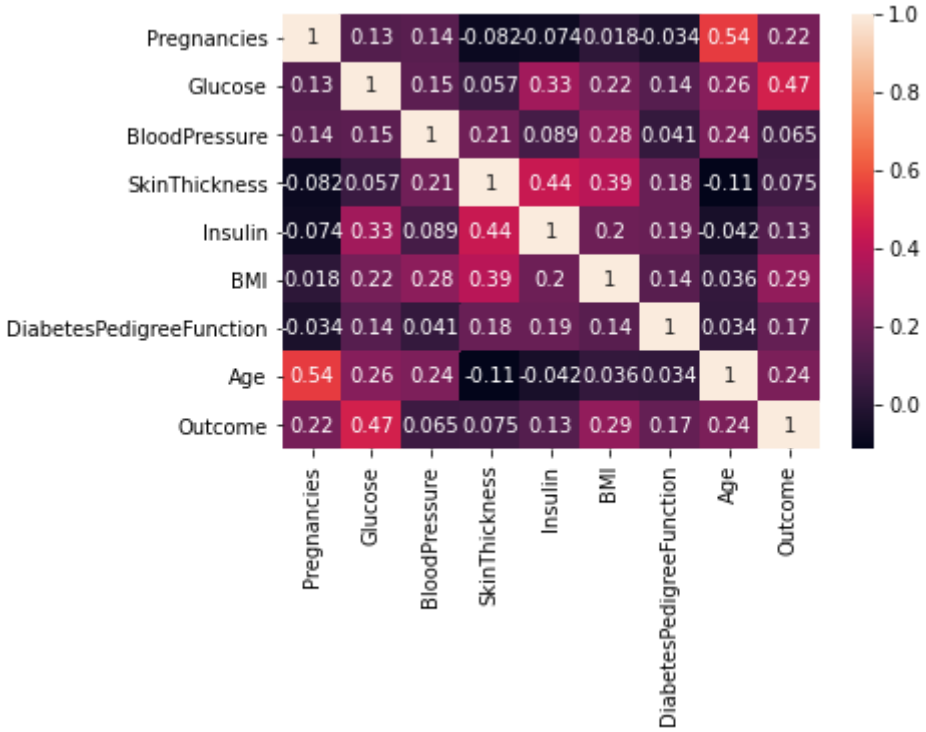
Outcome	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	Diab
0	3.298000	109.980000	68.184000	19.664000	68.792000	30.304200	
1	4.865672	141.257463	70.824627	22.164179	100.335821	35.142537	

In [70]:

```
#Correlation matrix to show correlation between two variables, 0.x means x% similar
corr_mat=dataset.corr()
sns.heatmap(corr_mat, annot=True)
```

Out[70]:

<AxesSubplot:>



In [71]:

```
#Ex: correlatiom between Glucose and Outcome is 47% that means output depends majorly on Gl
```

In [72]:

```
#Check if any null or empty data is present in dataset
dataset.isna().sum()
```

Out[72]:

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

In [73]:

```
#Feature matrix - Taking all our independent columns into single array and dependent values
x=dataset.iloc[:,:-1].values  #Independent matrix
y=dataset.iloc[:,-1].values
```

In [74]:

```
x.shape
```

Out[74]:

```
(768, 8)
```

In [75]:

```
x[0] #refering to column 1 in dataset i.e pregnancies
```

Out[75]:

```
array([  6.   , 148.   ,  72.   ,  35.   ,   0.   ,  33.6  ,   0.627,
        50.   ])
```

In [76]:

```
y
```

Out[76]:

```
array([1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 0,
       1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1,
       0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0,
       1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0,
       1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1,
       1, 1, 0, 0, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 1,
       1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,
       1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1,
       0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 0, 1,
       1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 1, 1, 1, 1, 0, 1, 1,
       1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0,
       1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 1, 0, 0,
       1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 1, 0,
       0, 1, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 0,
       1, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 1, 1, 1, 0, 0, 1, 0, 1, 0,
       0, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,
       0, 0, 0, 1, 1, 1, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0,
       0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 1, 0,
       0, 1, 0, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 1,
       0, 1, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,
       1, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0,
       1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0,
       1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 1, 1, 0, 0, 1, 1, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,
       0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0,
       0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 1, 0,
       0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0,
       1, 0, 0, 0, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1,
       0, 1, 1, 1, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1,
       0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0,
       0, 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0,
       0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0,
       1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0],
      dtype=int64)
```

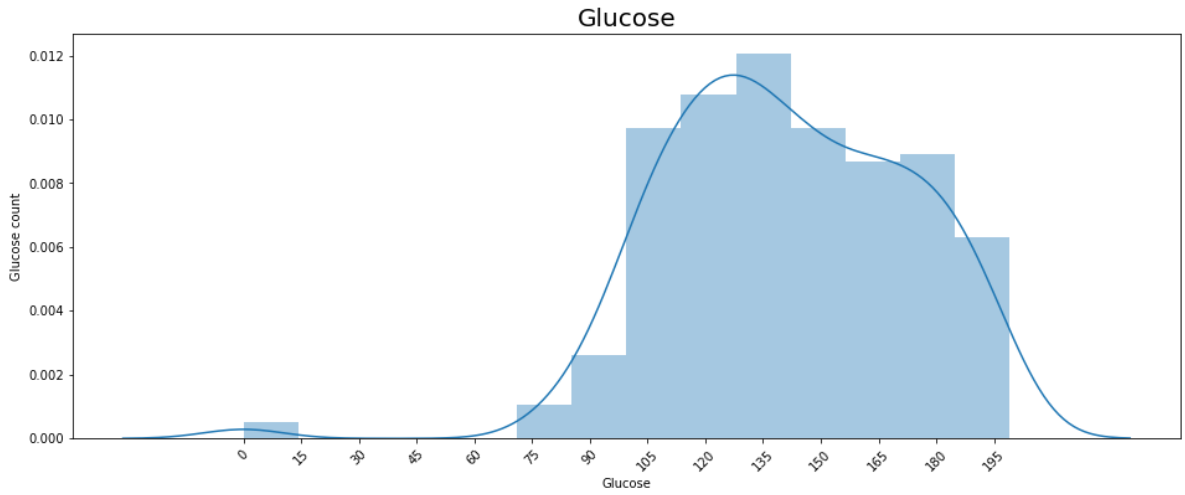
In [77]:

```
#glucose for diabetic
fig = plt.figure(figsize =(16,6))

sns.distplot(dataset["Glucose"][dataset["Outcome"] == 1])
plt.xticks([i for i in range(0,201,15)],rotation = 45)
plt.ylabel("Glucose count")
plt.title("Glucose",fontsize = 20)
```

```
C:\Users\aa\anaconda3\lib\site-packages\seaborn\distributions.py:2557: Futur
eWarning: `distplot` is a deprecated function and will be removed in a futur
e version. Please adapt your code to use either `displot` (a figure-level fu
nction with similar flexibility) or `histplot` (an axes-level function for h
istograms).
  warnings.warn(msg, FutureWarning)
```

Out[77]:

```
Text(0.5, 1.0, 'Glucose')
```



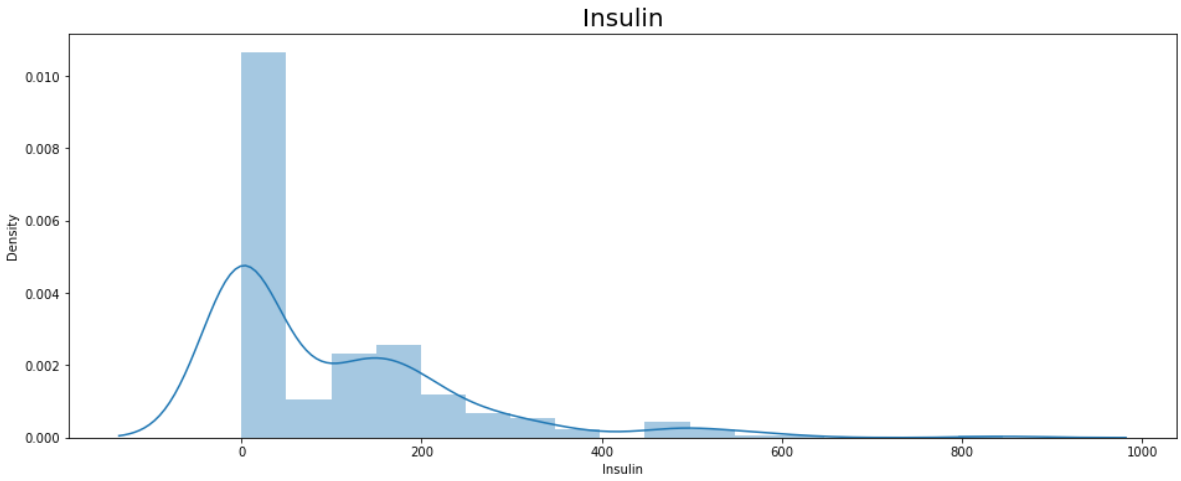
In [78]:

```
#insulin for diabetic

fig = plt.figure(figsize = (16,6))

sns.distplot(dataset["Insulin"][dataset["Outcome"]==1])
plt.xticks()
plt.title("Insulin",fontsize = 20)
```

```
C:\Users\aa\anaconda3\lib\site-packages\seaborn\distributions.py:2557: Futur
eWarning: `distplot` is a deprecated function and will be removed in a futur
e version. Please adapt your code to use either `displot` (a figure-level fu
nction with similar flexibility) or `histplot` (an axes-level function for h
istograms).
  warnings.warn(msg, FutureWarning)
```

Out[78]:

```
Text(0.5, 1.0, 'Insulin')
```



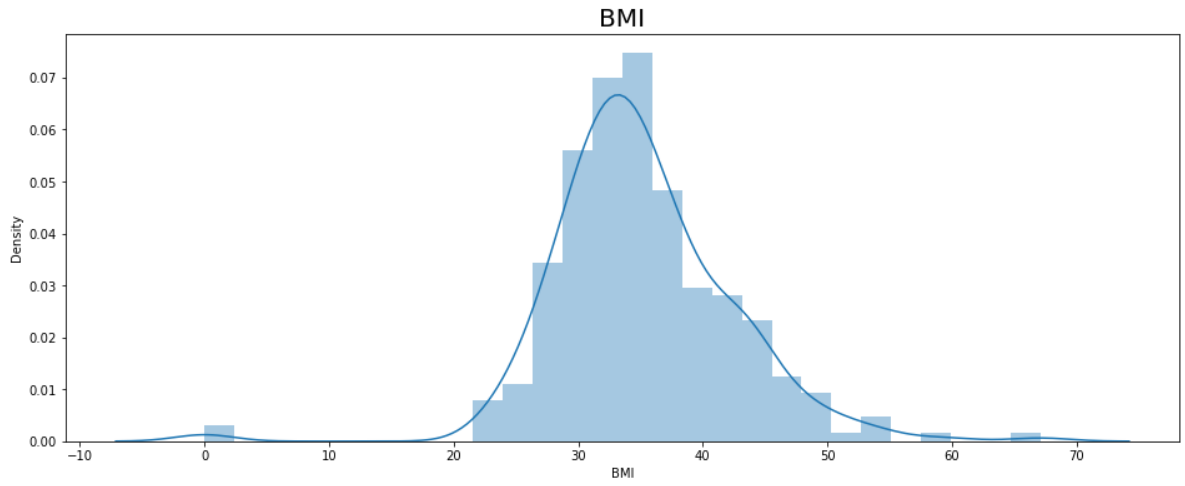
In [79]:

```
#BMI for diabetic
fig = plt.figure(figsize =(16,6))

sns.distplot(dataset["BMI"][dataset["Outcome"]==1])
plt.xticks()
plt.title("BMI",fontsize = 20)
```

```
C:\Users\aa\anaconda3\lib\site-packages\seaborn\distributions.py:2557: Futur
eWarning: `distplot` is a deprecated function and will be removed in a futur
e version. Please adapt your code to use either `displot` (a figure-level fu
nction with similar flexibility) or `histplot` (an axes-level function for h
istograms).
  warnings.warn(msg, FutureWarning)
```

Out[79]:

```
Text(0.5, 1.0, 'BMI')
```



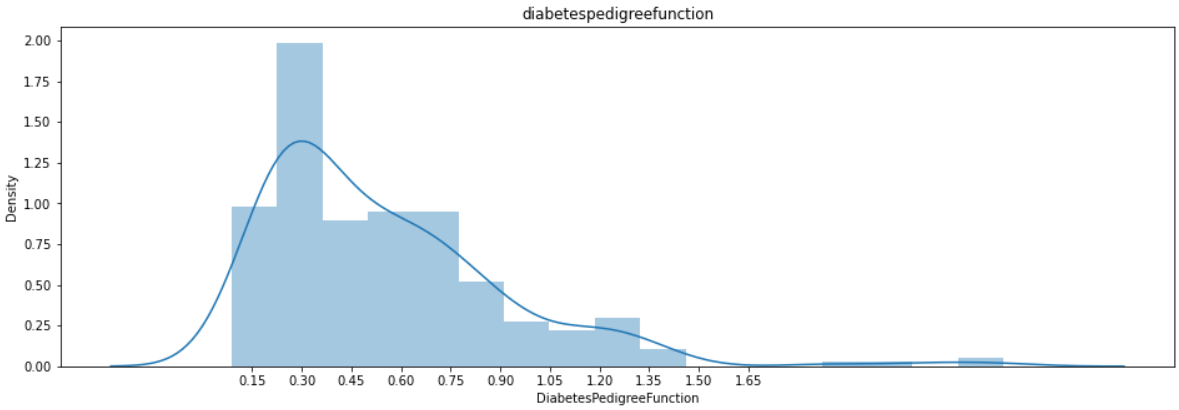
In [80]:

```
#diabeticpedigreefunction for diabetic
fig = plt.figure(figsize = (16,5))
sns.distplot(dataset["DiabetesPedigreeFunction"][dataset["Outcome"] == 1])
plt.xticks([i*0.15 for i in range(1,12)])
plt.title("diabetespedigreefunction")
```

```
C:\Users\aa\anaconda3\lib\site-packages\seaborn\distributions.py:2557: Futur
eWarning: `distplot` is a deprecated function and will be removed in a futur
e version. Please adapt your code to use either `displot` (a figure-level fu
nction with similar flexibility) or `histplot` (an axes-level function for h
istograms).
  warnings.warn(msg, FutureWarning)
```

Out[80]:

```
Text(0.5, 1.0, 'diabetespedigreefunction')
```



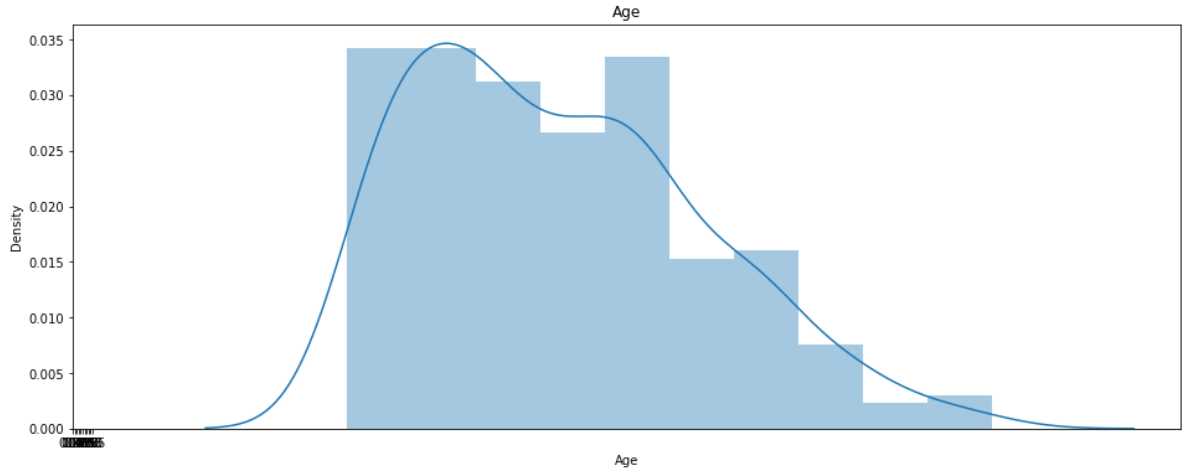
In [81]:

```
#Age for diabetic
fig = plt.figure(figsize = (16,6))

sns.distplot(dataset["Age"][dataset["Outcome"] == 1])
plt.xticks([i*0.15 for i in range(1,12)])
plt.title("Age")
```

```
C:\Users\aa\anaconda3\lib\site-packages\seaborn\distributions.py:2557: Futur
eWarning: `distplot` is a deprecated function and will be removed in a futur
e version. Please adapt your code to use either `displot` (a figure-level fu
nction with similar flexibility) or `histplot` (an axes-level function for h
istograms).
  warnings.warn(msg, FutureWarning)
```

Out[81]:

```
Text(0.5, 1.0, 'Age')
```



In [82]:

```
#Removing unnessary columns
x = dataset.drop(["Pregnancies","BloodPressure","SkinThickness","Outcome"],axis = 1)
y = dataset.iloc[:,-1]
```

In [117]:

```
#splitting dataset into training set and test set

from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.20, random_state=0)
#test_size 0.2 means for testing data 20% and training data 80%
```

In [118]:

```
x_train.shape #80% of original dataset (769,9) after removing unnecceasry data
```

Out[118]:

```
(614, 5)
```

In [119]:

```
x_test.shape #20% of original dataset (769,9) after removing unnecceasry data
```

Out[119]:

```
(154, 5)
```

In [120]:

```
#Feature Scaling - To standardize the independent features present in the data in a fixed r
#If feature scaling is not done, then a machine learning algorithm tends to weigh greater v
#values as the lower values, regardless of the unit of the values.
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
x_train = sc.fit_transform(x_train)
x_test = sc.transform(x_test)
```

In [121]:

```
x_train
```

Out[121]:

```
array([[ 0.91569367,  0.3736349 ,  0.37852648,  0.67740401,  1.69955804],
       [-0.75182191, -0.69965674, -0.50667229, -0.07049698, -0.96569189],
       [ 1.38763205,  5.09271083,  2.54094063, -0.11855487, -0.88240283],
       ...,
       [-0.84620959, -0.69965674, -0.94927168, -0.95656442, -1.04898095],
       [-1.12937261, -0.69965674, -0.26640405, -0.50001442,  0.11706589],
       [ 0.47521786, -0.69965674, -4.07275877,  0.52121586,  2.94889395]])
```

In [122]:

```
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier(n_neighbors =25, metric = 'minkowski')
#n_neighbors is 25 bcoz for x_train we got 614 which is near to 25^2
#metric means on what factor choosing so as its KNN so our metric is minkowski i.e., distan
knn.fit(x_train, y_train)
```

Out[122]:

```
KNeighborsClassifier(n_neighbors=25)
```

In [123]:

```
#Predicting the data
knn_y_pred = knn.predict(x_test)
```

In [124]:

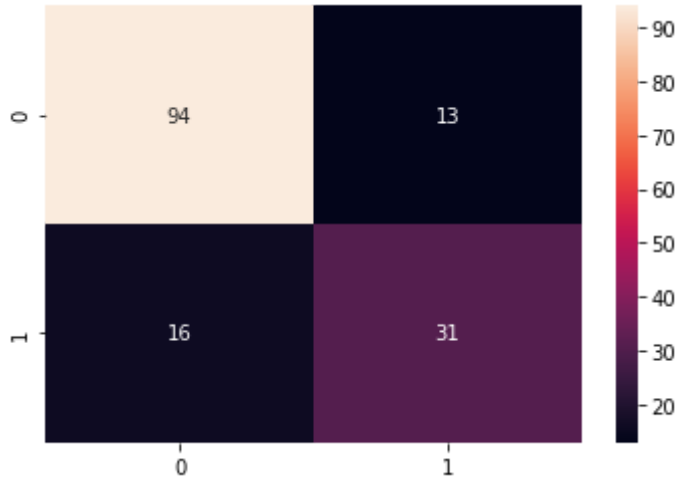
```
#Predicting the data
knn_y_pred = knn.predict(x_test)
```

In [125]:

```
# Confusion matrix - To check how many are correct or wrong
from sklearn.metrics import confusion_matrix
knn_cm = confusion_matrix(y_test, knn_y_pred)
sns.heatmap(knn_cm, annot=True)
```

Out[125]:

```
<AxesSubplot:>
```



In [126]:

```
# The above heatmap says 0,0 means true negative and 1,1 means true positive
# and 0,1 means even person is negative but showing result positive
# and 1,0 means person is positive but shows negative so its danger so we need to accurate
```

In [127]:

```
print("Correct:",sum(knn_y_pred==y_test))
print("Incorrect : ",sum(knn_y_pred != y_test))
print("Accuracy:",sum(knn_y_pred ==y_test)/len(knn_y_pred))
```

```
Correct: 125
Incorrect :  29
Accuracy: 0.8116883116883117
```

In [128]:

```
#Verfying accuracy using inbuilt methods
from sklearn.metrics import accuracy_score
accuracy_score(y_test,knn_y_pred)
```

Out[128]:

```
0.8116883116883117
```

In [129]:

```
from sklearn.svm import SVC
svc=SVC(kernel="linear",random_state=0)
svc.fit(x_train,y_train)
```

Out[129]:

```
SVC(kernel='linear', random_state=0)
```

In [130]:

```
svc_y_pred = svc.predict(x_test)
```

In [131]:

```
svc_cm = confusion_matrix(y_test,svc_y_pred)
print(svc_cm)
```

```
[[96 11]
 [18 29]]
```

In [132]:

```
print("Correct:",sum(svc_y_pred == y_test))
print("Incorrect : ",sum(svc_y_pred != y_test))
print("Accuracy:",sum(svc_y_pred ==y_test)/len(knn_y_pred))
```

```
Correct: 125
Incorrect :  29
Accuracy: 0.8116883116883117
```

In [133]:

```
import pickle
pickle.dump(svc, open('classifier.pkl', 'wb'))
```

In [134]:

```
pickle.dump(sc, open('sc.pkl', 'wb'))
```

In [ ]:

```

```

OUTCOME OF INTERNSHIP

The internship was a good learning experience for me. I learned about various machine learning algorithms and classification, regression and its implementation. We were also taught about Python and using it via Anaconda and Jupyter notebook. Various datatypes like list, tuple and dictionary. We also learnt about Data Analysis & visualization –using NumPy, panda matplotlib, SciPy etc.

The project we built will give us confidence in our knowledge of various machine learning techniques, python IDE and whatever we learned will help us build more such AI projects.

Some highlights of what we learned in the internship:

- Introduction to Artificial Intelligence
- Types of AI
- Ethical question about AI
- Python
    - Data Types
    - Functions in Python
    - Loops

- Data analysis and Exploration
    - Data Manipulation using NumPy, Pandas and Matplotlib.
    - Preprocessing of machine learning

- "Machine learning" a field of AI & its application
    - Implementation of various machine learning algorithms
    - Evaluation of various machine learning algorithm
- Working on PARAM SHAVAK supercomputer
- Project on Diabetes Prediction

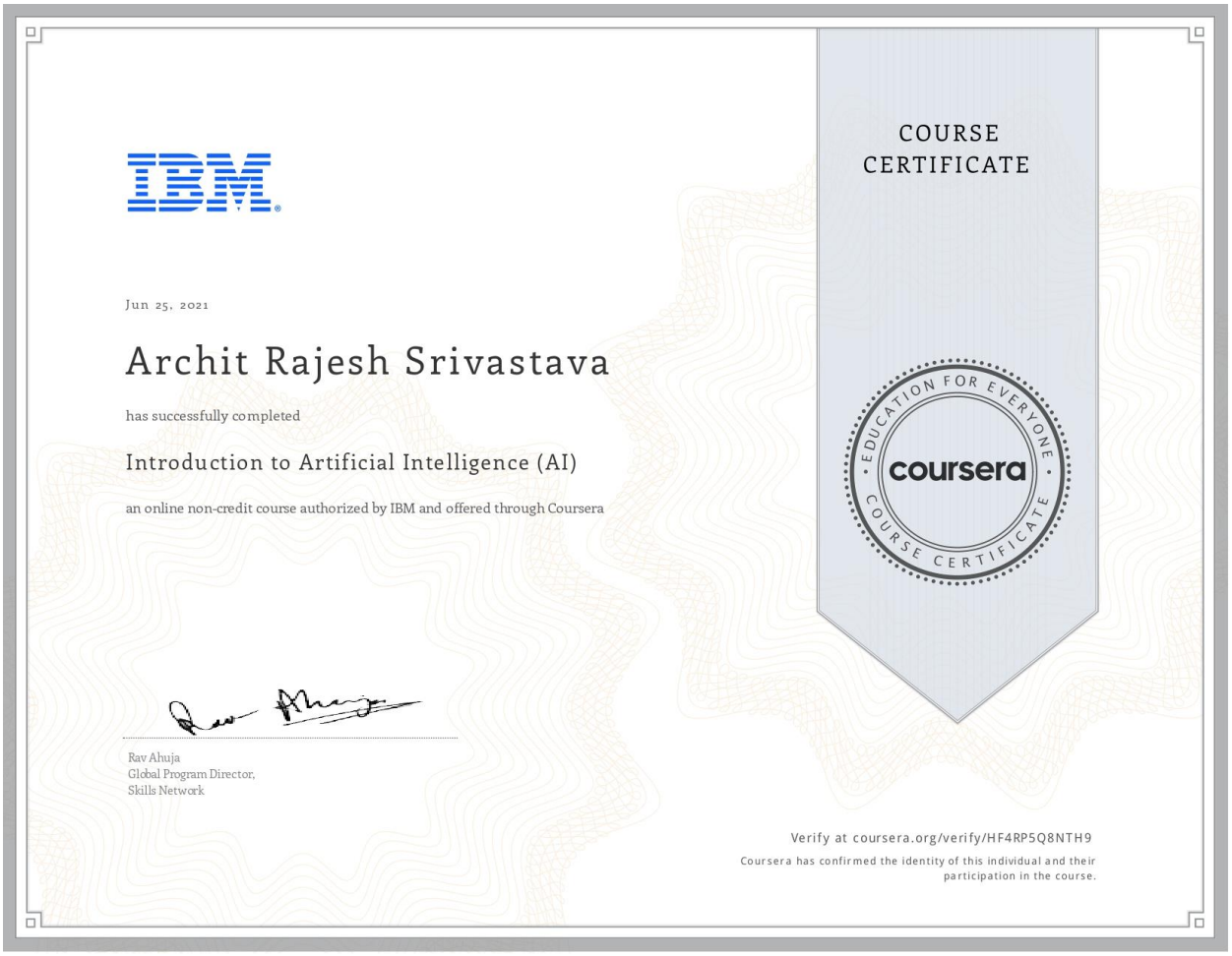
FUTURE SCOPE

We build the model using some Machine Learning algorithms to predict the whether the person is diabetic or not, we can use some other algorithms such as Naive Bayes, SVM (Support Vector Machine), ID3 ,etc to make our model more efficient.

We can setup a module for visitor's query, where visitors can post their queries to doctors and doctors can send reply to those queries.

We can also add a treatment module, where a patients can upload their diagnostic reports and doctors can see these reports and then they will suggest them proper treatment details.

MOOC Certificate



https://coursera.org/share/8582396c3e3c68a21a183b0d74f4dbcd

LITERATURE REVIEW REPORT

Autonomous Vehicles and Embedded Artificial Intelligence:

The Challenges of Framing Machine Driving Decisions

Applied Artificial Intelligence - 2021

Martin Cunneen | Martin Mullins | Finbarr Murphy

- The paper talks about the challenges and risks in the domain on autonomous vehicles.
- AVs can help mitigate any errors associated with human driving decisions. But they also create new potential risk factors and governing challenges that need to be considered.
- There are other advantages of AVs as well like enhancing the mobility of those who lack it.
- AV risk mitigation cannot be undertaken by governments alone, but must rather be a multi-stakeholder phenomenon.
- There have been concerns about the safety of this technology. But statistics show it is safer.
- AV offers a driving experience that is unaffected by human-like issues like intoxication, distraction, fatigue, and poor behavioral decisions such as speeding.
- Although humans are statistically proficient drivers, human drivers make human errors.

Artificial intelligence in Internet of things

The Institution of Engineering and Technology Journals - 2018

Ashish Ghosh | Debasrita Chakraborty | Anwesha Law

The Internet's functionality is evolving from the Internet of computers (IoC) to the "Internet of Things (IoT)." Furthermore, cyber -physical systems (CPSs) are arising from the integration of numerous elements such as infrastructure, embedded devices, smart objects, humans, and physical environments. The writers' goal is to create a massive "Internet of Everything in a Smart Cyber Physical Earth." The combination of IoT and CPS with "data science" could be the next "smart revolution." The issue then becomes how to handle the massive amounts of data generated with the current computing capability.

Classification of Customer Reviews Using Machine Learning Algorithms

Applied Artificial Intelligence - 2021

-Behrooz Noori

- This paper talks about categorization of customer review using Machine Learning Classification algorithms.
- Two approaches of sentiment analysis: Semantic Orientation and Machine Learning.
- The data was collected from a travel website called TripAdvisor.com which is a travel website
- After the preprocessing steps the author applies different Machine Learning methods.
- F-measure is used for performance metric.
- The highest accuracy of review classification was achieved by Decision Tree and C4.5

Comparative Investigation of Learning Algorithms for Image Classification with Small Dataset

Applied Artificial Intelligence - 2021

Imran Iqbal | Gbenga Abiodun Odesanmi | Jianxiang Wang | Li Liu

- The paper is an experimental study on different Learning optimization algorithms for Image classification.
- An optimization algorithm is the method by which the weights and biases are updated for a neural network.
- The author discusses different Learning algorithms like SGD, AdaGrad, RMSprop, AdaDelta, Adam etc.
- The Algorithm is tested for different optimizations and the comparison is done on training time, accuracy and cost.
- AdaGrad was found to be performing much better at learning rate 0.0008, mini-batch size 32, number of layers 50, and number of epoch 70

Automatic Diagnosis of Attention Deficit Hyperactivity Disorder Using Machine Learning

Applied Artificial Intelligence - 2021

Tianhua Chen | Grigoris Antoniou | Marios Adamou | Ilias Tachmazidis | Pan Su

- This paper discusses about diagnosing ADHD patients through the use of Machine Learning.
- ADHD is a social health issue which causes negative consequences in academics and career if went untreated.
- The researchers collected anonymized data from National Health Service of ADHD patients which contained information like demographic and age.
- K-fold Cross validation is used with k= (number of training examples) for training the model onto the data.
- Different Machine Learning Algorithms were tested like Logistic Regression, Decision Tree, KNN, Random Forest.
- Decision Tree was found to be performing the best giving more than 80% accuracy.

DAILY LOG

Name of Student	Archit Rajesh Srivastava
Roll No.	2000290120039
Name of Course	Mini Project on Artificial Intelligence
Date of Commencement of Training:	06-09-2021
Date of Completion of Training:	24-09-2021

S.No.	Please specify the learning of the day	Date
1	Enrolled in coursera MOOC	01/09/2021
2	Started watching content of Week 1	02/09/2021
3	Completed Quiz of Week 1	03/09/2021
4	Started watching content of Week 2	04/09/2021
5	Completed Quiz of Week 2	05/09/2021
6	Started watching content of Week 3 / Orientation Class for Internship	06/09/2021
7	Completed Quiz of Week 3 / Learnt about Python and ML techniques	07/09/2021
8	Started watching content of Week 4 / Learnt about Python and ML techniques	08/09/2021
9	Completed Quiz of Week 4 / Learnt about Python and ML techniques	09/09/2021
10	Learnt about Python and ML techniques	10/09/2021

11	Learnt about Python and ML techniques	11/09/2021
12	Read some articles on My Loft	12/09/2021
13	Learnt about Python and ML techniques	13/09/2021
14	Learnt about Param Shavak Supercomputer	14/09/2021
15	Read some articles on My Loft	15/09/2021
16	Read some articles on My Loft	16/09/2021
17	Started on the Project	17/09/2021
18	Did research on database and techniques to be used	18/09/2021
19	Started on the code	19/09/2021
20	Did the coding	20/09/2021
21	Started the Report writing	21/09/2021