# Diabetes Prediction

## Based on Random Forest Algorithm

T.E. mini-project report submitted in partial fulfilment of the

requirements of the degree of

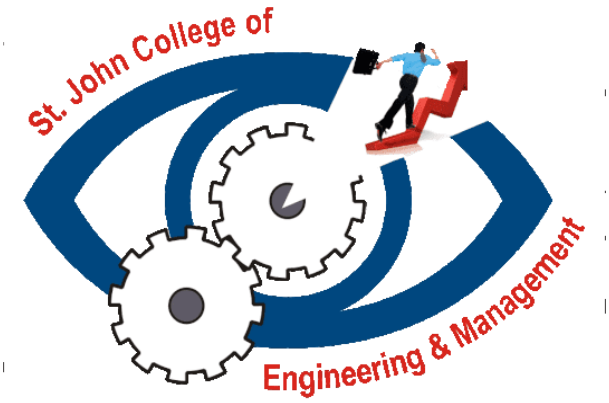
**Bachelor of Engineering**
**in**
**Information Technology**

by


| | | |
|---|---|---|
| **Dodamani Ankita Mallikarjun** | **EU1184012** | **(19)** |
| **Kuveskar Amit Ravindra** | **EU1164028** | **(32)** |
| **Wadile Pranali Ratilal** | **EU1154017** | **(76)** |


Under the guidance of

**Ms. Linda John**
(Assistant Professor)





**Department of Information Technology**

**St. John College of Engineering and Management, Palghar**

**University of Mumbai**


2020–2021

# CERTIFICATE

This is to certify that the T.E. mini-project entitled "**Diabetes Prediction based on Random Forest Algorithm**" is a bonafide work of

| | | |
|---|---|---|
| **Dodamani Ankita Mallikarjun** | **EU1184012** | **(19)** |
| **Kuveskar Amit Ravindra** | **EU1164028** | **(32)** |
| **Wadile Pranali Ratilal** | **EU1154017** | **(76)** |

submitted to University of Mumbai in partial fulfilment of the requirement for the award of the degree of "**Bachelors of Engineering in Information Technology**" during the academic year 2020-2021.

**Ms. Linda John**

Guide

**Mrs Anita Chaudhari**

Head of Department

**Dr. G.V. Mulgund**

Principal

# T.E. Mini-Project Report Approval

This mini-project synopsis entitled ***Diabetes Prediction based on Random Forest Algorithm*** by ***Ankita Mallikarjun Dodamani, Amit Ravindra Kuveskar, Pranali Ratilal Wadile*** is approved for the degree of ***Bachelors of Engineering in Information Technology*** from ***University of Mumbai***.

**Examiners**

1.----------

2.----------

Date:

Place:

# Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

---

Signature
Ankita Mallikarjun Dodamani (EU1184012)

---

Signature
Amit Ravindra Kuveskar (EU1164028)

---

Signature
Pranali Ratilal Wadile (EU1154017)

Date:

# ABSTRACT

Diabetes is taken into account together of the deadliest and chronic disease that causes a rise in glucose. Random Forest algorithms are often used for each classification and regression tasks and also it is a type of ensemble learning method.

A paper published in the Journal of the Association of Physicians of India in 2019 has argued the importance of screening of every pregnant woman for high blood glucose even if no symptoms are exhibited. India has an estimated 62 million people with Type 2 diabetes mellitus (DM); this number is expected to go up to 79.4 million by 2025.

Random Forest is a supervised learning algorithm. The “forest” it builds, is an ensemble of Decision Trees. In simple words: Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. One big advantage of random forest is, that it can be used for both classification and regression problems & current machine learning systems Random forest is also a very handy algorithm because the default hyperparameters it uses often produce a good prediction result.

**Keywords:** *Diabetes , Random Forest Algorithm*

# TABLE OF CONTENTS

**Page No**

# LIST OF FIGURES

# Abbreviations

| Figure No | Figure Name |
|---|---|
| NCD | Non-Communicable Disease |
| DM | Diabetes Mellitus |
| GDM | Gestational Diabetes Mellitus |
| BMI | Body Mass Index |
| BI | Business Intelligence |
| EHR | Electronic Health Records |

# CHAPTER 1

# INTRODUCTION

### 1.1 Motivation:

Diabetes mellitus, commonly known as diabetes, is a metabolic disease that causes high blood sugar. It is a Non-Communicable Disease (NCD) that occurs either when the pancreas does not produce enough insulin (a hormone that regulates blood sugar, or glucose), or when the body cannot effectively use the insulin, it produces. A paper published in the Journal of the Association of Physicians of India in 2019 has argued the importance of screening of every pregnant woman for high blood glucose even if no symptoms are exhibited. India has an estimated 62 million people with Type 2 diabetes mellitus (DM); this number is expected to go up to 79.4 million by 2025. In parallel with the increase in diabetes prevalence, there seems to be an increasing prevalence of gestational DM (GDM), that is, diabetes diagnosed during pregnancy. Healthcare resources are insufficient. There is inadequate awareness among public. This results in a large population being hesitant to access healthcare system for diseases with not so "obvious" implications like GDM. In predictive Analysis, firstly the dataset of all the patients containing their details such as no. of pregnancies, glucose levels, blood pressure, BMI, insulin, age etc. and outcome is considered, out of which 'outcome' becomes our target class which tells us accurately whether the patient shall have diabetes or not. Using Random Forest Algorithm, Trained data is used to predict in test data accurately whether the outcome is 1 (Positive Results) or 0 (Negative Results).

### 1.2 Problem System

To develop a system that can accurately and quickly predict diabetes in patients using details like Glucose levels, Blood Pressure, BMI, Insulin etc. which are present in generic medical reports. This system can be used to generate health analysis reports especially for Pregnant Women.

## 1.3 Objective:

The Objective of the project are as follows;

- To predict diabetes in pregnant women in most efficient way using only the most basic relevant information which is freely available in medical reports.
- To judge whether the results will be positive or negative for diabetes.
- To inform patients right before they are at Risk/prone to/or ought to have diabetes in near future.

## 1.4 Scope:

1. The Tool can very well work for large datasets and can accurately predict score over 93% of results from the given user data. This score can be improved by cleaning data more efficiently and pre-processing the data which in our case was considered to be an 'Ideal dataset'. In real time, there might be several scenarios where there may be inconsistent data, empty rows, redundant or outdated which may affect the results and predictions.
2. The Tool works on test, train and score method to generate the predictions. specific entries such as patient ID is the only tuple to identify the patient individually. In The current scenario, patient is selected at random from the test data, it can be improved to predict for a single individual or a batch of individuals Out of the whole test dataset.
3. The BI tool representation is in form of pie charts only. It visualizes the overall data of the whole system. It can be improved to generate for a single individual or a batch of individuals in other forms of representation with more detailed visualization.

# CHAPTER 2

# LITERATURE SURVEY

### 2.1 Risk prediction of type II diabetes based on random forest model

In this paper Authors have proposed a type II diabetes prediction model based on random forest which aims at analyzing some readily available indicators (age, weight, waist, hip, etc.) effects on diabetes and discovering some rules on given data. The method can significantly reduce the risk of disease through digging out a clear and understandable model for type II diabetes from a medical database. Random forest algorithm uses multiple decision trees to train the samples, and integrates weight of each tree to get the final results.

**Conclusion:** These indicators are relatively easy to obtain (they can be measured at home), so we can greatly reduce the cost of diagnosis.

**Gap:** Expand other indicators to predict the risk of disease and update the perspective of data mining are the future direction of the prediction

### 2.2 Classification of Diabetes using Random Forest with Feature Selection Algorithm

In this paper Authors have predicted the level of occurrence of diabetes. They predict the level of occurrence of diabetes using Random Forest, a Machine Learning Algorithm. Using the patient's Electronic Health Records (EHR) we can build accurate models that predict the presence of diabetes.

**Conclusion:** There is no remedy for diabetes mellitus perception can lessen the repine word complications and subdue the charge. Decision Tree accuracy 75.2, Bagging with Decision Tree but rather accuracy 81.3, Random Forest accuracy 85.6, Random Forest with Feature Selection accuracy 92.02.

**Gap:** To obtain stronger precision by gathering patient information from hospitals, clinics, and by using more choice supporting technologies with true parameters and potent classifications, forward to increasing precision.

### 2.3 Random Forest Algorithm for the Prediction of Diabetes

In this paper Authors have used to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by using machine learning technique which provides advance support for predicting the accuracy rate of diabetes.

**Conclusion**: Systematic efforts area unit created in coming up with a system that finally ends up among the prediction of illness like genetic defect. Throughout this work Random Forest algorithms area unit studied and evaluated on varied measures.

# CHAPTER 3

# PROPOSED SYSTEM

## 3.1 Block Diagram of the Proposed System

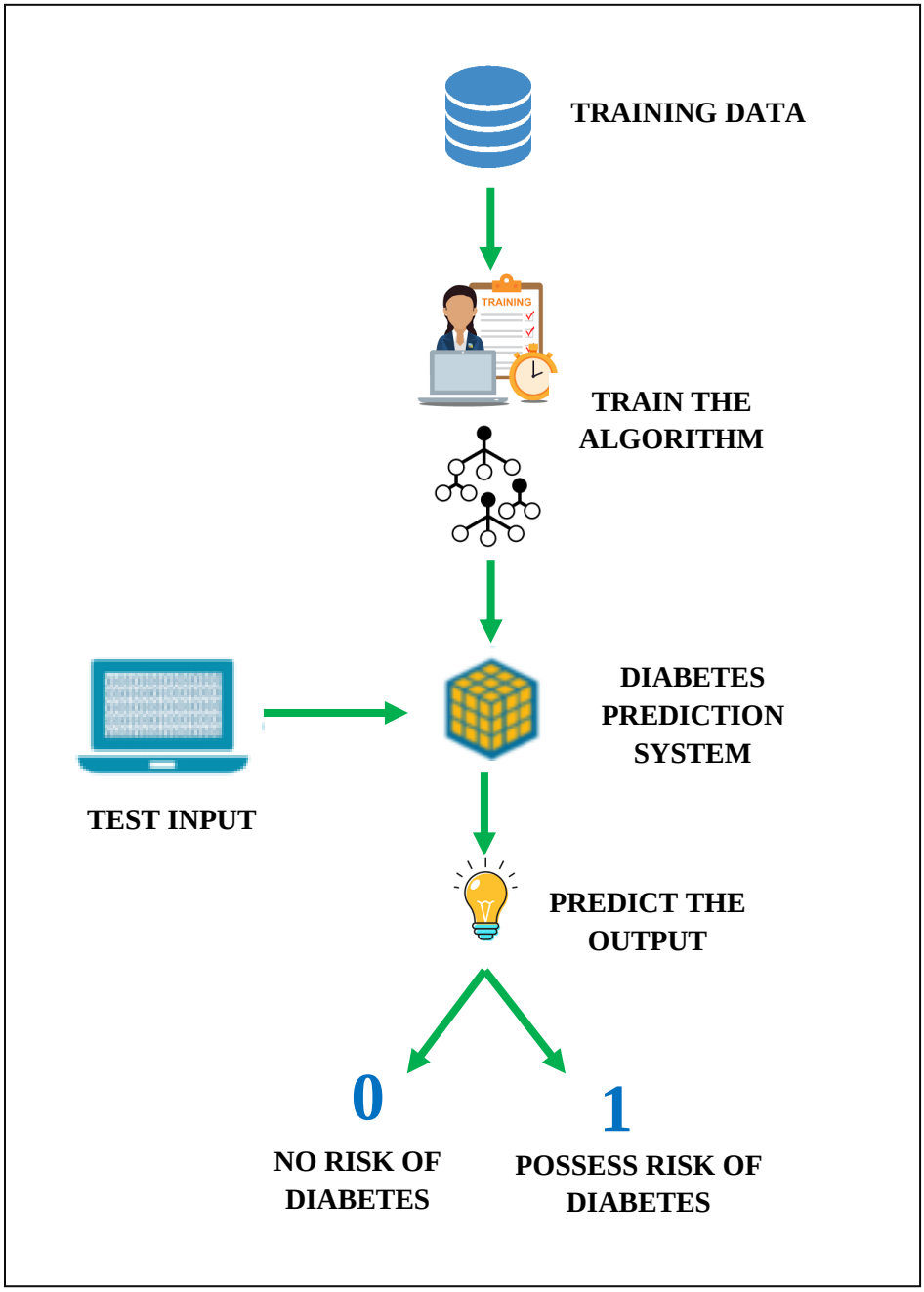



*Figure 3.1: Block Diagram of the System*

## 3.2 Layout of Proposed Work

**3.2.1 User Input:** In this step an input dataset is to be given which becomes the test data for our system. This file can be an excel spreadsheet in .csv or .xlsx or .ods format.

**3.2.2 Test & Train Data:** Here the Training data is used by the diabetes prediction system and trained to predict outcomes. If a .csv file is used then it will be converted into machine readable format by python libraries 'pandas' and 'numpy' to extract the numeric data present in the dataset.

**3.2.3 Diabetes Prediction System:** Here our main structure of system resides which used python language to execute the code and display results. The 'Jupyter' is an improved 'iPython' notebook service which serves as an predefined environment to run and execute the code and display results simultaneously.

**3.2.4 Classification Algorithms:** In this step, Random Forest algorithm coded in python uses the train data to predict the outcome of the testing data. It predicts whether patient is a diabetes 'POSITIVE' or 'NEGATIVE'.

**3.2.5 Prediction:** Outcome of the classification algorithm gives 0 for predicting Negative results and 1 for predicting Positive results. Out of 400 tested data, random forest accurately predicts over 95% of the test data i.e. 380 rows accurately for diabetic predictions successfully.

## 3.3 Implementation

In [1]:
```
import pandas as pd
import seaborn as sns
import numpy as np
import matplotlib.pyplot as plt
from sklearn.metrics import roc_auc_score
%matplotlib inline
```

In [2]:
```
df=pd.read_csv('diabetes.csv')
df.head()
```

Out[2]:

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | A |
|---|---|---|---|---|---|---|---|---|
| 0 | 2 | 138 | 62 | 35 | 0 | 33.6 | 0.127 | |
| 1 | 0 | 84 | 82 | 31 | 125 | 38.2 | 0.233 | |
| 2 | 0 | 145 | 0 | 0 | 0 | 44.2 | 0.630 | |
| 3 | 0 | 135 | 68 | 42 | 250 | 42.3 | 0.365 | |
| 4 | 1 | 139 | 62 | 41 | 480 | 40.7 | 0.536 | |

In [3]:
```
#Lets Describe the data
df.describe()
```

Out[3]:

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | Diabete |
|---|---|---|---|---|---|---|---|
| count | 2000.000000 | 2000.000000 | 2000.000000 | 2000.000000 | 2000.000000 | 2000.000000 | |
| mean | 3.703500 | 121.182500 | 69.145500 | 20.935000 | 80.254000 | 32.193000 | |
| std | 3.306063 | 32.068636 | 19.188315 | 16.103243 | 111.180534 | 8.149901 | |
| min | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | |
| 25% | 1.000000 | 99.000000 | 63.500000 | 0.000000 | 0.000000 | 27.375000 | |
| 50% | 3.000000 | 117.000000 | 72.000000 | 23.000000 | 40.000000 | 32.300000 | |
| 75% | 6.000000 | 141.000000 | 80.000000 | 32.000000 | 130.000000 | 36.800000 | |
| max | 17.000000 | 199.000000 | 122.000000 | 110.000000 | 744.000000 | 80.600000 | |

In [4]:
```
#information of  dataset
df.info()
```

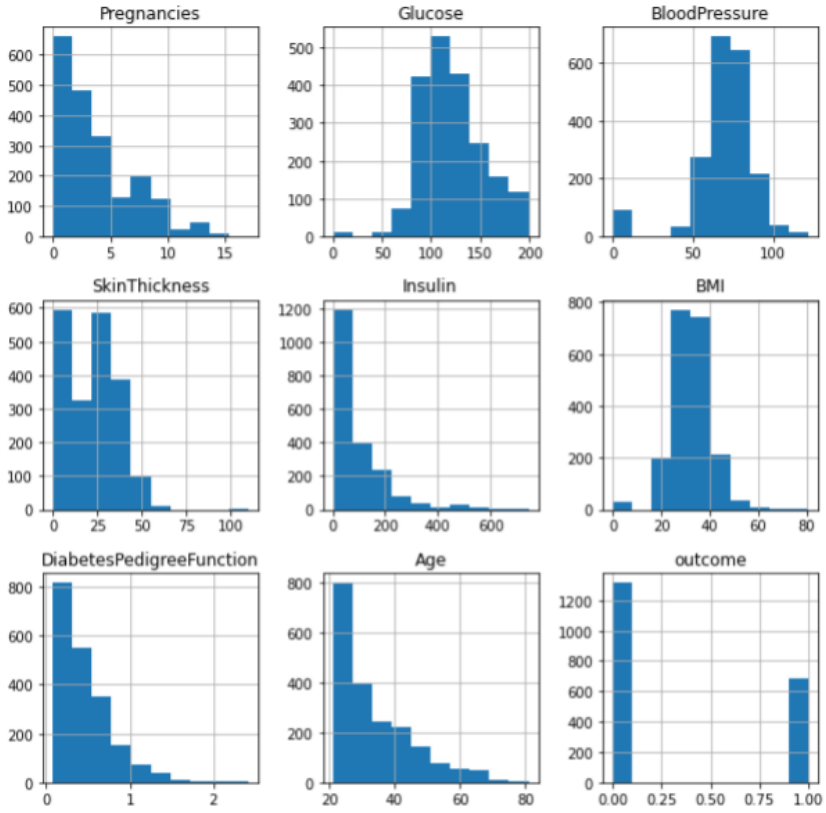
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Pregnancies               2000 non-null   int64  
 1   Glucose                   2000 non-null   int64  
 2   BloodPressure             2000 non-null   int64  
 3   SkinThickness             2000 non-null   int64  
 4   Insulin                   2000 non-null   int64  
 5   BMI                       2000 non-null   float64
 6   DiabetesPedigreeFunction  2000 non-null   float64
 7   Age                       2000 non-null   int64  
 8   outcome                   2000 non-null   int64  
dtypes: float64(2), int64(7)
memory usage: 140.8 KB
```

In [5]:
```
#any null values
#not neccessary in above information we can see
df.isnull().values.any()
```

Out[5]:
```
False
```

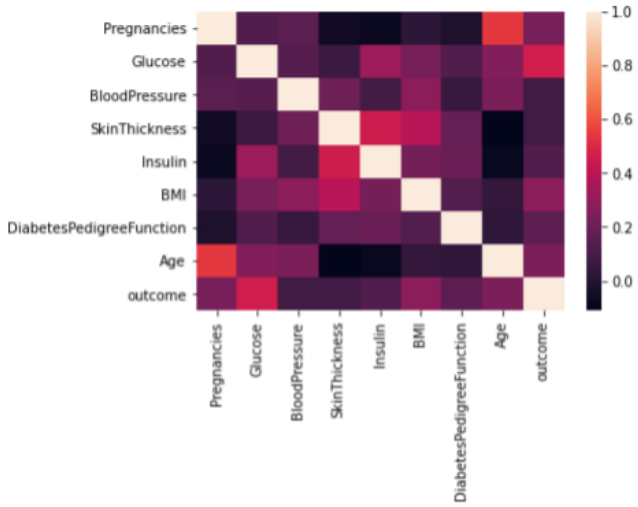


In [6]:
```
#histogram
df.hist(bins=10,figsize=(10,10))
plt.show()
```



In [7]:
```
#correlation
sns.heatmap(df.corr())
#we can analyse skin thinkness,insulin, Pregnencies and age are full independent
#age and pregncies has negative correlation
```

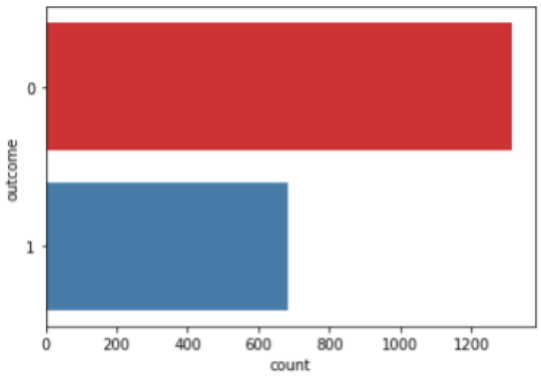
Out[7]: <AxesSubplot:>





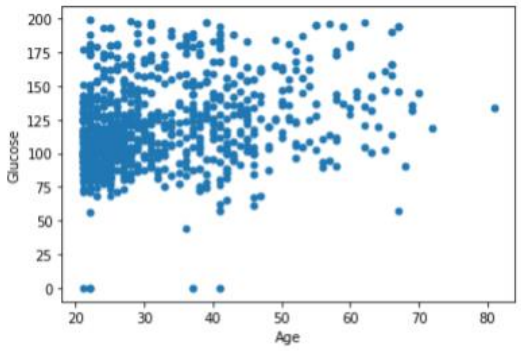

In [8]:
```
#Lets count total outcome in each target 0 1
#0 means no diabetes
#1 means patient with diabetes
sns.countplot(y=df['outcome'],palette='Set1')
```

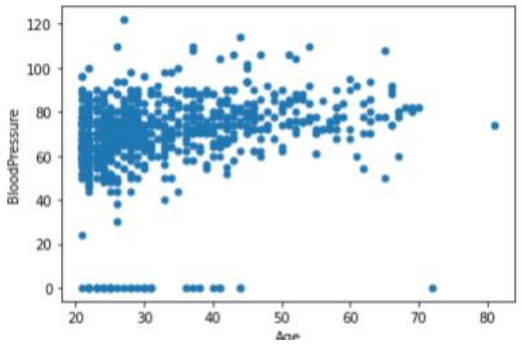
Out[8]: <AxesSubplot:xlabel='count', ylabel='outcome'>



In [9]:
```
#scatter plot to unterstant
df.plot(kind='scatter',x='Age',y='Glucose')
df.plot(kind='scatter',x='Age',y='BloodPressure')
```

Out[9]: <AxesSubplot:xlabel='Age', ylabel='BloodPressure'>




In [10]:
```
X=df.drop(columns=['outcome'])
y=df['outcome']
```

In [11]:
```
# Train Test split
from sklearn.model_selection import train_test_split
train_X,test_X,train_y,test_y=train_test_split(X,y,test_size=0.2)
```

In [12]:
```
train_X.shape,test_X.shape,train_y.shape,test_y.shape
```

Out[12]: ((1600, 8), (400, 8), (1600,), (400,))

In [13]:
```
#Random forest
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score


clf=RandomForestClassifier()
clf.fit(train_X,train_y)
y_pred=clf.predict(test_X)
accuracy_score(test_y,y_pred)
acc,roc=[],[]
ac=accuracy_score(test_y,y_pred)
acc.append(ac)
try:
    rc=roc_auc_score(test_y,y_pred)
except ValueError:
    pass
#rc=roc_auc_score(test_y,y_pred)
#rc=accuracy_score(test_y,y_pred)
roc.append(rc)
print("Accuracy {0} ROC {1}".format(ac,rc))

#display predicted
pd.DataFrame(data={'Actual':test_y,'Predicted':y_pred}).head(400)
```

```
Accuracy 0.985 ROC 0.9819545989758756
```

Out[13]:

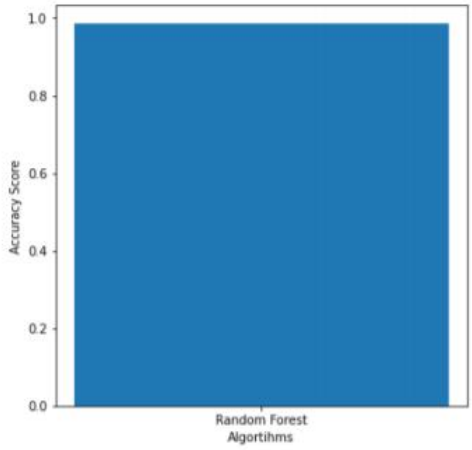
| | Actual | Predicted |
|---|---|---|
| **722** | 1 | 1 |
| **1812** | 0 | 0 |
| **142** | 0 | 0 |
| **531** | 0 | 0 |
| **707** | 0 | 0 |
| **...** | ... | ... |
| **1215** | 0 | 0 |
| **172** | 0 | 0 |
| **1069** | 1 | 1 |
| **266** | 1 | 1 |
| **595** | 1 | 1 |

400 rows × 2 columns

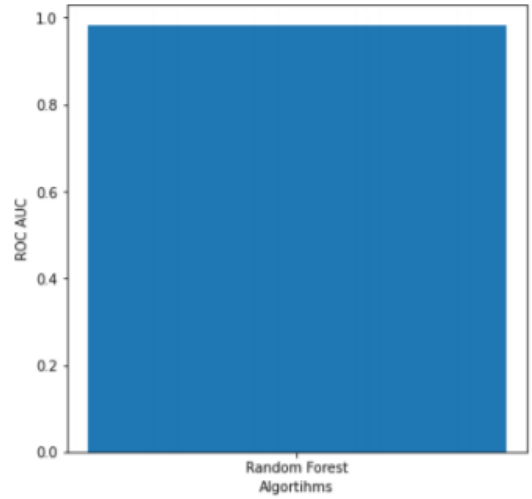
In [14]:
```
#Lets plot the bar graph

ax=plt.figure(figsize=(6,6))
plt.bar(['Random Forest',],acc,label='Accuracy')
plt.ylabel('Accuracy Score')
plt.xlabel('Algortihms')
plt.show()

ax=plt.figure(figsize=(6,6))
plt.bar(['Random Forest'],roc,label='ROC AUC')
plt.ylabel('ROC AUC')
plt.xlabel('Algortihms')
plt.show()
```





## 3.4 Visualization

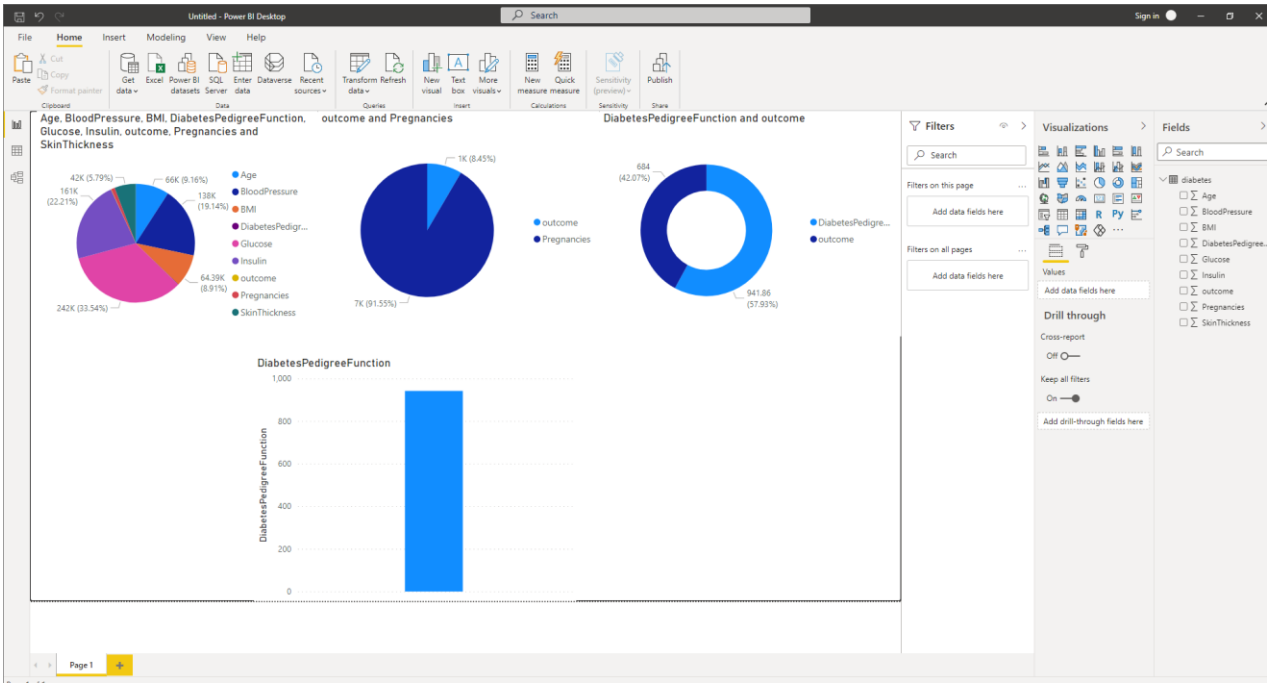



*Figure 3.4: Visualization using Power BI*

From the above Figure, we see the different data visualizations using Power BI. From the figure we can understand relation and visualize the whole data in form of pie chart. The Pie charts shows pregnancies and outcome show out of how many pregnancies would be diabetic and how many would be not. The Donut chart gives outcome and DiabetesPedigreeFunction grouped together and the last bar plot shows the overall number of pregnancies versus DiabetesPedigreeFunction.

# CHAPTER 4

# CONCLUSION AND FUTURE WORK

### 4.1 Limitations

Limitation was that we did not directly measure medication adherences, our data was mainly based on patient information.

### 4.2 Conclusion

GDM not only influences immediate maternal (preeclampsia, stillbirths, macrosomia, and need for cesarean section) and neonatal outcomes (hypoglycemia, respiratory distress), but also increases the risk of future Type 2 diabetes in mother as well as the baby. Higher glucose transfer to the foetus, when the mother has high blood sugar, stimulates the foetal pancreatic cells to start secreting insulin earlier and in higher quantities. Once initiated, it becomes self perpetuating. In addition, when the maternal glucose reading is high the amniotic fluid becomes glucose enriched, and after 20 weeks, when the foetus begins to swallow the amniotic fluid, which further stimulates production of insulin Prevention at the earliest stage of development of the foetus is essential to prevent children from becoming predisposed to diabetes or other non-communicable diseases (NCD) in future. In this project, Prediction of Diabetes using data mining classification techniques have been studied and implemented. To achieve this study, classification techniques such as Random Forest and Naïve Bayes are used for prediction and the dataset is visualized & represented using PowerBI tool..

### 4.3 Future Scope

This system just predicts diabetes disease using certain parameters specified. However, this system doesn't give recommendations regarding which medicine should be consumed. Also, we can increase the accuracy by trying out different algorithms like Support Vector Machine, Naïve Bayes, k-NN etc. The algorithm could then be compared and the one that gives best result could be used.

# CHAPTER 5

# REFERENCES

## References:

[1] K.Koteswara Chari, M.Chinna babu, Sarangarm Kodati Classification of Diabetes using Random Forest with Feature Selection Algorithm" International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-1, November 2019

[2] Weifeng Xu, Jianxin Zhang, Qiang Zhang*, Xiaopeng Wei "Risk prediction of type II diabetes based on random forest model" Key Lab of Advanced Design and Intelligent Computing, Ministry of Education Dalian University, Dalian, P. R. China

[3] K.VijiyaKumar, B.Lavany , I.Nirmala , S.Sofia Caroline "Random Forest Algorithm for the Prediction of Diabetes" Manakula Vinayagar Institute of Technology, Puducherry

# ACKNOWLEDGEMENT

We would like to express our deepest gratitude to all those who provided us the possibility to complete this report. We have taken efforts in this project, but it would not have been possible without the kind support and help of many individuals. We would like to extend our sincere thanks to all of them.

We take this opportunity to express our profound gratitude and deep regards to our teacher **Ms. Linda John** for her exemplary guidance, monitoring and constant encouragement throughout the course of this project. The blessing, help and guidance given by her, time to time shall carry us a long way in the journey of life on which we are about to embark.

We also take this opportunity to express a deep sense of gratitude to our Head of Information Technology Department, **Mrs. Anita Chaudhari** along with all the Teaching & Non-Teaching Staff for their cordial support, valuable information and guidance, which helped us in completing this task through various stages.

We express our sincere gratitude to our respected principal **Dr. G. V. Mulgund** for encouragement and facilities provided to us. We would also like to acknowledge the patience that our ever-beloved parents have shown during our efforts and the encouragement we have received from them.

Last but not least we would also like to thank to our Friends and Family who directly and indirectly helped and supported us during the whole course of the Project.

Ankita Mallikarjun Dodamani

Amit Ravindra Kuveskar

Pranali Ratilal Wadile