AI- Based Diabetes Pricdiction System

1.INTRODUCTION

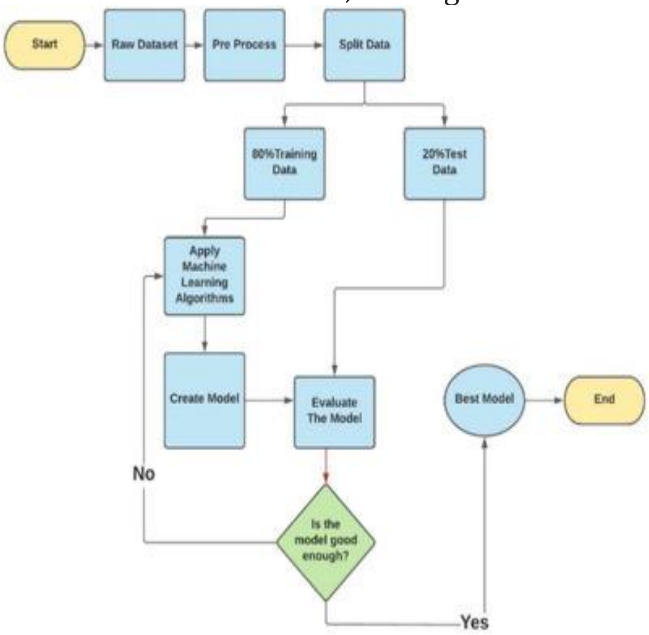
Diabetes is a chronic disease that directly affects the pancreas, and the body is incapable of producing insulin . Insulin is mainly responsible for maintaining the blood glucose level. Many factors, such as excessive body weight, physical inactivity, high blood pressure, and abnormal cholesterol level, can cause a person get affected by diabetes.

This private dataset has been obtained from female employees of Rownak Textile Mills Ltd, Dhaka, Bangladesh, referred to as the 'RTML dataset' in this paper. We have collected six features from 203 individuals, that is, pregnancy, glucose, blood pressure, skin thickness, BMI, age, and final outcome of diabetes.

- Another contribution of this work is to keep similarities with the feature of the Pima Indian dataset. The missing insulin feature of the RTML dataset was predicted using a semi-supervised technique.
- A website and an Android application have been developed with the finalized best-performed model of this research work to make instantaneous predictions with real-time data.

2.PROPOSED SYSTEM:-

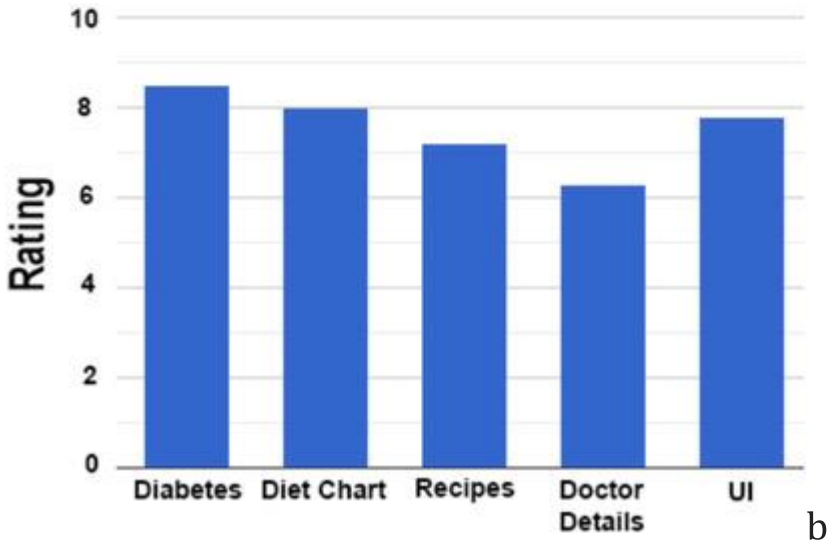
This section describes the working procedures and implementation of various machine learning techniques to design the proposed automatic diabetes prediction system. Figure 1 shows the different stages of this research work. First, the dataset was collected and preprocessed to remove the necessary discrepancies from the dataset, for example, replacing null instances with mean values, dealing with imbalanced class issues etc.



WORKING SEQUENCES OF DIABETES PRIDICTION SYSTEM

3.RESULTS AND DISCUSSION:-

This section presents the results and discussion of the proposed automatic diabetes prediction system. First, the performance of various machine learning techniques is discussed. Next, the implemented website framework and Android smartphone application are demonstrated.



Program:-

Input,

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

```
                                                         In [24]:
```

Input,

```
linkcode
df.head()
```

input,

```
# Now, we can look at where are missing values
df.isnull().sum()
```

output,

```
Pregnancies                    0
Glucose                        5
```

```
BloodPressure                35
SkinThickness               227
Insulin                     374
BMI                          11
DiabetesPedigreeFunction      0
Age                           0
Outcome                       0
dtype: int64
```

input,

```
# The missing values will be filled with the median values of each var
iable.
def median_target(var):
    temp = df[df[var].notnull()]
    temp = temp[[var, 'Outcome']].groupby(['Outcome'])[[var]].median()
.reset_index()
    return temp
```

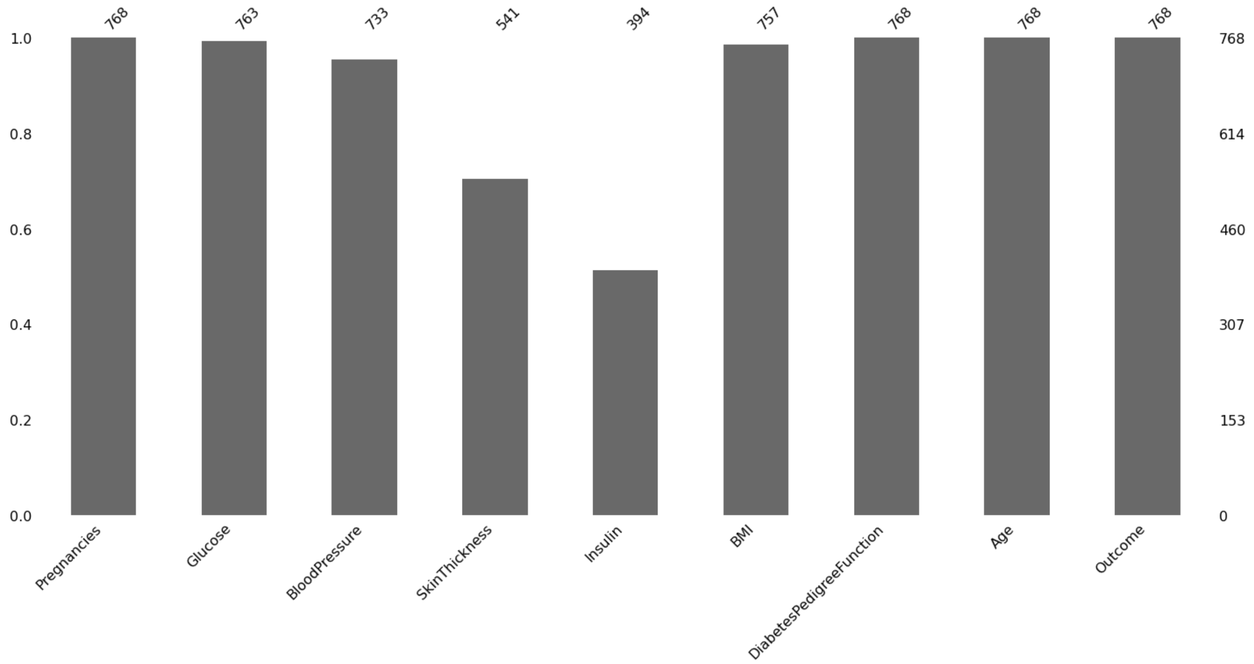
input,

```
# The missing values will be filled with the median values of each var
iable.
def median_target(var):
    temp = df[df[var].notnull()]
    temp = temp[[var, 'Outcome']].groupby(['Outcome'])[[var]].median()
.reset_index()
    return temp
```

input,

```
# Have been visualized using the missingno library for the visualizati
on of missing observations.
# Plotting
import missingno as msno
msno.bar(df);
```

output,



Input,

```
# The missing values will be filled with the median values of each var
iable.
def median_target(var):
    temp = df[df[var].notnull()]
    temp = temp[[var, 'Outcome']].groupby(['Outcome'])[[var]].median()
.reset_index()
    return temp
```

input,

```
# The values to be given for incomplete observations are given the med
ian value of people who are not sick and the median values of people w
ho are sick.
columns = df.columns
columns = columns.drop("Outcome")
for i in columns:
    median_target(i)
    df.loc[(df['Outcome'] == 0 ) & (df[i].isnull()), i] = median_targe
t(i)[i][0]
    df.loc[(df['Outcome'] == 1 ) & (df[i].isnull()), i] = median_targe
t(i)[i][1]
```

input,

```
df.head()
```

input,

```
# Missing values were filled.
```

```
df.isnull().sum()Out[30]
```

output,

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

conclusion:-

Diabetes can be a reason for reducing life expectancy and quality. Predicting this chronic disorder earlier can reduce the risk and complications of many diseases in the long run. In this paper, an automatic diabetes prediction system using various machine learning approaches has been proposed. The open-source Pima Indian and a private dataset of female Bangladeshi patients have been used in this work.

SMOTE and ADASYN preprocessing techniques have been applied to handle the issue of imbalanced class problems. This research paper reported different performance metrics, that is, precision, recall, accuracy, F1 score, and AUC for various machine learning and ensemble techniques.

The XGBoost classifier achieved the best performance with 81% accuracy and an F1 score and AUC of 0.81 and 0.84, respectively, with the ADASYN approach. Next, the domain adaptation technique has been applied to demonstrate the versatility of the proposed prediction system

. Finally, the best-performed XGBoost framework has been deployed into a website and smartphone application to predict diabetes instantly. There are some future scopes of this work, for example, we recommend getting additional private data with a larger cohort of patients to get better results. Another extension of this work is combining machine learning models with fuzzy logic techniques and applying optimization approaches.