# Diabetes Prediction

August 29, 2021

[1]:
```
import pandas as pd
import numpy as np
```

[18]:
```
data=pd.read_csv("Diabetes Dataset.csv")
data.head(50)
```

[18]:

| | Unnamed: 0 | Pregnancies | Glucose | blood pressure | skin thickness | Insulin |
|---|---|---|---|---|---|---|
| 0 | NaN | 6.0 | 148.0 | 72.0 | 35.0 | 0.0 |
| 1 | NaN | 1.0 | 85.0 | 66.0 | 29.0 | 0.0 |
| 2 | NaN | 8.0 | 183.0 | 64.0 | 0.0 | 0.0 |
| 3 | NaN | 1.0 | 89.0 | 66.0 | 23.0 | 94.0 |
| 4 | NaN | 0.0 | 137.0 | 40.0 | 35.0 | 168.0 |
| 5 | NaN | 5.0 | 116.0 | 74.0 | 0.0 | 0.0 |
| 6 | NaN | 3.0 | 78.0 | 50.0 | 32.0 | 88.0 |
| 7 | NaN | 10.0 | 115.0 | 0.0 | 0.0 | 0.0 |
| 8 | NaN | 2.0 | 197.0 | 70.0 | 45.0 | 543.0 |
| 9 | NaN | 8.0 | 125.0 | 96.0 | 0.0 | 0.0 |
| 10 | NaN | 4.0 | 110.0 | 92.0 | 0.0 | 0.0 |
| 11 | NaN | 10.0 | 168.0 | 74.0 | 0.0 | 0.0 |
| 12 | NaN | 10.0 | 139.0 | 80.0 | 0.0 | 0.0 |
| 13 | NaN | 1.0 | 189.0 | 60.0 | 23.0 | 846.0 |
| 14 | NaN | 5.0 | 166.0 | 72.0 | 19.0 | 175.0 |
| 15 | NaN | 7.0 | 100.0 | 0.0 | 0.0 | 0.0 |
| 16 | NaN | 0.0 | 118.0 | 84.0 | 47.0 | 230.0 |
| 17 | NaN | 7.0 | 107.0 | 74.0 | 0.0 | 0.0 |
| 18 | NaN | 1.0 | 103.0 | 30.0 | 38.0 | 83.0 |
| 19 | NaN | 1.0 | 115.0 | 70.0 | 30.0 | 96.0 |
| 20 | NaN | 3.0 | 126.0 | 88.0 | 41.0 | 235.0 |
| 21 | NaN | 8.0 | 99.0 | 84.0 | 0.0 | 0.0 |
| 22 | NaN | 7.0 | 196.0 | 90.0 | 0.0 | 0.0 |
| 23 | NaN | 9.0 | 119.0 | 80.0 | 35.0 | 0.0 |
| 24 | NaN | 11.0 | 143.0 | 94.0 | 33.0 | 146.0 |
| 25 | NaN | 10.0 | 125.0 | 70.0 | 26.0 | 115.0 |
| 26 | NaN | 7.0 | 147.0 | 76.0 | 0.0 | 0.0 |
| 27 | NaN | 1.0 | 97.0 | 66.0 | 15.0 | 140.0 |
| 28 | NaN | 13.0 | 145.0 | 82.0 | 19.0 | 110.0 |
| 29 | NaN | 5.0 | 117.0 | 92.0 | 0.0 | 0.0 |
| 30 | NaN | 5.0 | 109.0 | 75.0 | 26.0 | 0.0 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 31 | NaN | 3.0 | 158.0 | 76.0 | 36.0 | 245.0 |
| 32 | NaN | 3.0 | 88.0 | 58.0 | 11.0 | 54.0 |
| 33 | NaN | 6.0 | 92.0 | 92.0 | 0.0 | 0.0 |
| 34 | NaN | 10.0 | 122.0 | 78.0 | 31.0 | 0.0 |
| 35 | NaN | 4.0 | 103.0 | 60.0 | 33.0 | 192.0 |
| 36 | NaN | 11.0 | 138.0 | 76.0 | 0.0 | 0.0 |
| 37 | NaN | 9.0 | 102.0 | 76.0 | 37.0 | 0.0 |
| 38 | NaN | 2.0 | 90.0 | 68.0 | 42.0 | 0.0 |
| 39 | NaN | 4.0 | 111.0 | 72.0 | 47.0 | 207.0 |
| 40 | NaN | 3.0 | 180.0 | 64.0 | 25.0 | 70.0 |
| 41 | NaN | 7.0 | 133.0 | 84.0 | 0.0 | 0.0 |
| 42 | NaN | 7.0 | 106.0 | 92.0 | 18.0 | 0.0 |
| 43 | NaN | 9.0 | 171.0 | 110.0 | 24.0 | 240.0 |
| 44 | NaN | 7.0 | 159.0 | 64.0 | 0.0 | 0.0 |
| 45 | NaN | 0.0 | 180.0 | 66.0 | 39.0 | 0.0 |
| 46 | NaN | 1.0 | 146.0 | 56.0 | 0.0 | 0.0 |
| 47 | NaN | 2.0 | 71.0 | 70.0 | 27.0 | 0.0 |
| 48 | NaN | 7.0 | 103.0 | 66.0 | 32.0 | 0.0 |
| 49 | NaN | 7.0 | 105.0 | 0.0 | 0.0 | 0.0 |

| | BMI | DiabetesPedigreeFunction | Age | Outcome |
|---|---|---|---|---|
| 0 | 33.6 | 0.627 | 50.0 | 1.0 |
| 1 | 26.6 | 0.351 | 31.0 | 0.0 |
| 2 | 23.3 | 0.672 | 32.0 | 1.0 |
| 3 | 28.1 | 0.167 | 21.0 | 0.0 |
| 4 | 43.1 | 2.288 | 33.0 | 1.0 |
| 5 | 25.6 | 0.201 | 30.0 | 0.0 |
| 6 | 31.0 | 0.248 | 26.0 | 1.0 |
| 7 | 35.3 | 0.134 | 29.0 | 0.0 |
| 8 | 30.5 | 0.158 | 53.0 | 1.0 |
| 9 | 0.0 | 0.232 | 54.0 | 1.0 |
| 10 | 37.6 | 0.191 | 30.0 | 0.0 |
| 11 | 38.0 | 0.537 | 34.0 | 1.0 |
| 12 | 27.1 | 1.441 | 57.0 | 0.0 |
| 13 | 30.1 | 0.398 | 59.0 | 1.0 |
| 14 | 25.8 | 0.587 | 51.0 | 1.0 |
| 15 | 30.0 | 0.484 | 32.0 | 1.0 |
| 16 | 45.8 | 0.551 | 31.0 | 1.0 |
| 17 | 29.6 | 0.254 | 31.0 | 1.0 |
| 18 | 43.3 | 0.183 | 33.0 | 0.0 |
| 19 | 34.6 | 0.529 | 32.0 | 1.0 |
| 20 | 39.3 | 0.704 | 27.0 | 0.0 |
| 21 | 35.4 | 0.388 | 50.0 | 0.0 |
| 22 | 39.8 | 0.451 | 41.0 | 1.0 |
| 23 | 29.0 | 0.263 | 29.0 | 1.0 |
| 24 | 36.6 | 0.254 | 51.0 | 1.0 |
| 25 | 31.1 | 0.205 | 41.0 | 1.0 |

```
 26  39.4                       0.257  43.0      1.0
 27  23.2                       0.487  22.0      0.0
 28  22.2                       0.245  57.0      0.0
 29  34.1                       0.337  38.0      0.0
 30  36.0                       0.546  60.0      0.0
 31  31.6                       0.851  28.0      1.0
 32  24.8                       0.267  22.0      0.0
 33  19.9                       0.188  28.0      0.0
 34  27.6                       0.512  45.0      0.0
 35  24.0                       0.966  33.0      0.0
 36  33.2                       0.420  35.0      0.0
 37  32.9                       0.665  46.0      1.0
 38  38.2                       0.503  27.0      1.0
 39  37.1                       1.390  56.0      1.0
 40  34.0                       0.271  26.0      0.0
 41  40.2                       0.696  37.0      0.0
 42  22.7                       0.235  48.0      0.0
 43  45.4                       0.721  54.0      1.0
 44  27.4                       0.294  40.0      0.0
 45  42.0                       1.893  25.0      1.0
 46  29.7                       0.564  29.0      0.0
 47  28.0                       0.586  22.0      0.0
 48  39.1                       0.344  31.0      1.0
 49   0.0                       0.305  24.0      0.0
```

[21]:
```
data=data.drop(columns=['Unnamed: 0'])
```

[23]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 769 entries, 0 to 768
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    float64
 1   Glucose                   768 non-null    float64
 2   blood pressure            768 non-null    float64
 3   skin thickness            768 non-null    float64
 4   Insulin                   768 non-null    float64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    float64
 8   Outcome                   768 non-null    float64
dtypes: float64(9)
memory usage: 54.2 KB
```

[25]:
```
data.isnull().sum()
```

[25]:
```
Pregnancies                 1
Glucose                     1
blood pressure              1
skin thickness              1
Insulin                     1
BMI                         1
DiabetesPedigreeFunction    1
Age                         1
Outcome                     1
dtype: int64
```

[29]:
```
data=data.drop([768])
```

[30]:
```
data.isnull().sum()
```

[30]:
```
Pregnancies                 0
Glucose                     0
blood pressure              0
skin thickness              0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

[34]:
```
x=data.iloc[:,:-1]
y=data.iloc[:,-1]
print(y)
```

```
0      1.0
1      0.0
2      1.0
3      0.0
4      1.0
      ...
763    0.0
764    0.0
765    0.0
766    1.0
767    0.0
Name: Outcome, Length: 768, dtype: float64
```

[46]:
```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train,␣
 ↪y_test=train_test_split(x,y,test_size=50,random_state=0)
```

[47]:
```
from sklearn.ensemble import RandomForestClassifier
classifier = RandomForestClassifier(n_estimators=6, criterion='entropy',␣
↪random_state=0)
classifier.fit(x_train, y_train)
y_pred = classifier.predict(x_test)
```

[48]:
```
from sklearn.metrics import accuracy_score
accuracy=round(accuracy_score(y_pred, y_test),2)*100
print("Random Forest Algorithm Accuracy is: ",accuracy)
```

```
Random Forest Algorithm Accuracy is:  78.0
```

[49]:
```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score,r2_score,classification_report
lr=LogisticRegression(solver='lbfgs',max_iter=1000)
lr.fit(x_train,y_train)
y_pred=lr.predict(x_test)
accuracy1=round(accuracy_score(y_pred, y_test),2)*100
print("Logistic Regression Algorithm Accuracy is: ",accuracy1)
```

```
Logistic Regression Algorithm Accuracy is:  86.0
```

[52]:
```
from sklearn.neighbors import KNeighborsClassifier
knn=KNeighborsClassifier(n_neighbors=2)
knn.fit(x_train,y_train)
y_pred=knn.predict(x_test)
accuracy2=round(accuracy_score(y_pred, y_test),2)*100
print("KNN Algorithm Accuracy is: ",accuracy2)
```

```
KNN Algorithm Accuracy is:  76.0
```

[ ]:
```

```