**AI-Based Diabetes Prediction System**

au211521243058:Guru Venkatesh MJ

****

**Phase 3: Development Part 1 - Building a Diabetes Prediction System**

In Phase 3 of developing your diabetes prediction system, Building a diabetes prediction system involves several key steps, including loading and preprocessing the dataset, selecting relevant features, and preparing the data for model training. Here is a step-by-step process on how to begin this project:

**Step 1: Data Collection:**

* Begin by collecting a dataset that contains relevant information for predicting diabetes. Common datasets for diabetes prediction include the Diabetes dataset from scikit-learn or publicly available medical datasets.

Dataset Link:[**https://www.kaggle.com/datasets/mathchi/diabetes-data-set**](https://www.kaggle.com/datasets/mathchi/diabetes-data-set)



**Step 2: Data Loading:**

* We load the dataset into our data analysis environment, using Python with the Pandas library, enabling us to access and manipulate the data effectively

import pandas as pd

import numpy as np

from sklearn.preprocessing import StandardScaler

#from sklearn.linear\_model import LogisticRegression

from sklearn.tree import DecisionTreeClassifier

from sklearn.svm import SVC

from sklearn.naive\_bayes import BernoulliNB

from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import accuracy\_score, confusion\_matrix

import matplotlib.pyplot as plt

import seaborn as sns

data = pd.read\_csv("/kaggle/input/diabetes-data-set/diabetes.csv")

data.head()

|  | **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 |

**Step 3: Data Exploration:**

* To gain a deep understanding of our dataset, we initiate data exploration. This process includes:  
  1. Checking for missing values to ensure data quality.  
  2. Obtaining summary statistics to grasp the dataset's characteristics.  
  3. Evaluating data types, distribution, and patterns.

data.describe()

| Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | Age | Outcome |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| count | 768.000000 | 768.000000 | 768.000000 | 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 | 0.471876 | 33.240885 | 0.348958 |
| std | 3.369578 | 31.972618 | 19.355807 | 15.952218 | 115.244002 | 7.884160 | 0.331329 | 11.760232 | 0.476951 |
| min | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.078000 | 21.000000 | 0.000000 |
| 25% | 1.000000 | 99.000000 | 62.000000 | 0.000000 | 0.000000 | 27.300000 | 0.243750 | 24.000000 | 0.000000 |
| 50% | 3.000000 | 117.000000 | 72.000000 | 23.000000 | 30.500000 | 32.000000 | 0.372500 | 29.000000 | 0.000000 |
| 75% | 6.000000 | 140.250000 | 80.000000 | 32.000000 | 127.250000 | 36.600000 | 0.626250 | 41.000000 | 1.000000 |
| max | 17.000000 | 199.000000 | 122.000000 | 99.000000 | 846.000000 | 67.100000 | 2.420000 | 81.000000 | 1.000000 |

*data['BMI'] = data['BMI'].replace(0,data['BMI'].mean())*

data['BloodPressure'] = data['BloodPressure'].replace(0,data['BloodPressure'].mean())

data['Glucose'] = data['Glucose'].replace(0,data['Glucose'].mean())

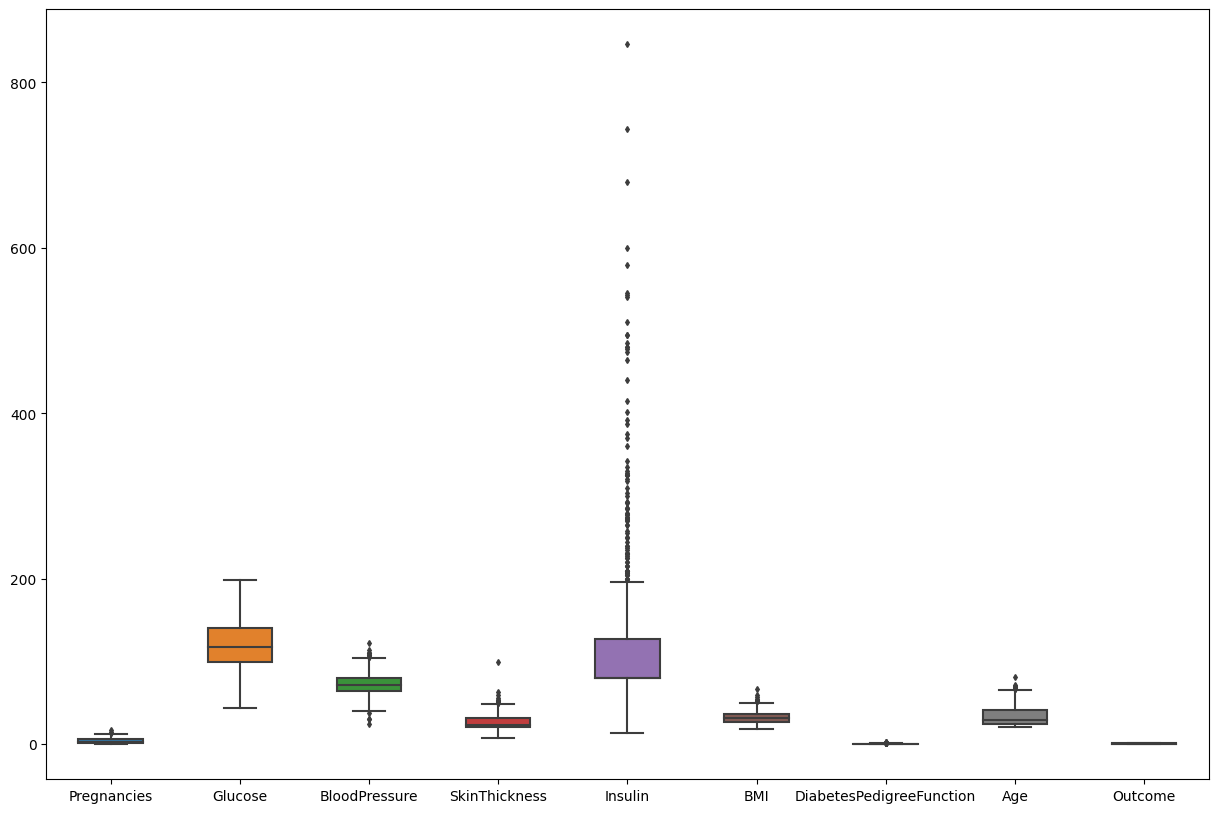
data['Insulin'] = data['Insulin'].replace(0,data['Insulin'].mean())

data['SkinThickness'] = data['SkinThickness'].replace(0,data['SkinThickness'].mean())

fig, ax = plt.subplots(figsize=(15,10))

sns.boxplot(data=data, width= 0.5,ax=ax, fliersize=3)

OUT:

****

**Step 4: Feature Selection:**

* We select relevant features, the variables that are strongly associated with diabetes. Feature selection is a critical step, and it typically involves domain knowledge and statistical analysis.

data.isnull().sum()

OUT:

Pregnancies 0

Glucose 0

BloodPressure 0

SkinThickness 0

Insulin 0

BMI 0

DiabetesPedigreeFunction 0

Age 0

Outcome 0

dtype: int64

**Step 5: Target Variable:**

* We identify the target variable, the key element we aim to predict in our model. In this case, our target variable is a binary indicator that represents the presence or absence of diabetes.

X = data.drop(columns = ['Outcome'])

y = data['Outcome']

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X,y,test\_size=0.25,random\_state=0)

X\_train.shape, X\_test.shape

OUT: ((576, 8), (192, 8))

**Step 6: Data Preprocessing:**

* To prepare our data for modeling, we perform necessary data preprocessing tasks, including:  
   1. Handling missing values by imputing them (for example, replacing them with the mean).  
  2. Encoding categorical variables if needed.  
  3. Standardizing or normalizing features to ensure that all variables are on a common scale.

import pickle  
def scaler\_standard(X\_train, X\_test):  
 #scaling the data  
 scaler = StandardScaler()  
 X\_train\_scaled = scaler.fit\_transform(X\_train)  
 X\_test\_scaled = scaler.transform(X\_test)  
 file = open('standardScalar.pkl','wb')  
 pickle.dump(scaler,file)  
 file.close()  
 return X\_train\_scaled, X\_test\_scaled  
X\_train\_scaled, X\_test\_scaled = scaler\_standard(X\_train, X\_test)

X\_train\_scaled

OUT:

array([[ 1.50755225, -1.09947934, -0.89942504, ..., -1.45561965,

-0.98325882, -0.04863985],

[-0.82986389, -0.1331471 , -1.23618124, ..., 0.09272955,

-0.62493647, -0.88246592],

[-1.12204091, -1.03283573, 0.61597784, ..., -0.03629955,

0.39884168, -0.5489355 ],

...,

[ 0.04666716, -0.93287033, -0.64685789, ..., -1.14021518,

-0.96519215, -1.04923114],

[ 2.09190629, -1.23276654, 0.11084355, ..., -0.36604058,

-0.5075031 , 0.11812536],

[ 0.33884418, 0.46664532, 0.78435594, ..., -0.09470985,

0.51627505, 2.953134 ]])

**Conclusion:**By the end of Phase 3, we have successfully loaded and preprocessed our dataset, chosen relevant features, and defined our target variable. We are now well-prepared to move on to the subsequent phases, which involve data splitting, model selection, training, and evaluation. Our ultimate goal is to develop an accurate diabetes prediction system to benefit individuals' health and well-being.

This phase is a crucial foundation for the rest of the project, as the quality of data and feature selection significantly impact the performance of the diabetes prediction model.