AI BASED DIABETES PREDICTION SYSTEM

(PHASE 3)

**Introduction:**

In this phase 3 we are going to implement the programming codes for predicting diabetes system using Artificial Intelligence by using modules such as numpy pandas,matlplotlib in python programming language.

Here we collect the patients biological datas such as Blood Glucose level,Blood Insulin level,Blood pressure,Skin thickness etc.

**Code:**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

dataset = pd.read\_csv('diabetes.csv')

dataset.head()

dataset.shape

dataset.info()

dataset.describe().T

dataset.isnull().sum()

sns.countplot(x = 'Outcome',data = dataset)

sns.pairplot(data = dataset, hue = 'Outcome')

plt.show()

sns.heatmap(dataset.corr(), annot = True)

plt.show()

dataset\_new = dataset

dataset\_new[["Glucose", "BloodPressure", "SkinThickness", "Insulin", "BMI"]] = dataset\_new[["Glucose", "BloodPressure", "SkinThickness", "Insulin", "BMI"]].replace(0, np.NaN)

dataset\_new.isnull().sum()

dataset\_new["Glucose"].fillna(dataset\_new["Glucose"].mean(), inplace = True)

dataset\_new["BloodPressure"].fillna(dataset\_new["BloodPressure"].mean(), inplace = True)

dataset\_new["SkinThickness"].fillna(dataset\_new["SkinThickness"].mean(), inplace = True)

dataset\_new["Insulin"].fillna(dataset\_new["Insulin"].mean(), inplace = True)

dataset\_new["BMI"].fillna(dataset\_new["BMI"].mean(), inplace = True)

dataset\_new.isnull().sum()

The above program is written in Python language and utilizes several libraries for data manipulation and visualization:

**import pandas as pd:** This imports the Pandas library, which is commonly used for data manipulation and analysis.

**import numpy as np:** This imports the NumPy library, which provides support for working with arrays and matrices.

**import matplotlib.pyplot as plt:** This imports the Matplotlib library, which is used for creating visualizations like plots and charts.

**import seaborn as sns:** This imports the Seaborn library, which is built on top of Matplotlib and provides a high-level interface for creating attractive and informative statistical graphics.

**dataset** = **pd.read\_csv('diabetes.csv'):** This line reads a CSV file named 'diabetes.csv' using Pandas and stores it in a variable called `dataset`.

**dataset.head():** This displays the first few rows of the dataset to provide an overview.

**dataset.shape:** This returns the dimensions (number of rows and columns) of the dataset.

**dataset.info()**: This provides information about the dataset, including the data types and non-null counts for each column.

**dataset.describe().T:** This generates summary statistics for numerical columns and transposes the result for better readability.

**dataset.isnull().sum():** This counts the number of missing values in each column of the dataset.

**sns.countplot(x='Outcome', data=dataset):** This creates a count plot to visualize the distribution of the 'Outcome' column.

**sns.pairplot(data=dataset, hue='Outcome'):** This creates a pair plot to visualize relationships between variables, with different colors for different outcomes.

**plt.show():** This displays the visualizations created using Matplotlib and Seaborn.

**sns.heatmap(dataset.corr(), annot=True):** This generates a heatmap to visualize the correlation between different numerical variables.

**plt.show():** This displays the correlation heatmap.

The next section of code performs **data preprocessing:**

- `dataset\_new = dataset`: This creates a copy of the original dataset.

- `dataset\_new[["Glucose", "BloodPressure", "SkinThickness", "Insulin", "BMI"]] = dataset\_new[["Glucose", "BloodPressure", "SkinThickness", "Insulin", "BMI"]].replace(0, np.NaN)`: This replaces zero values in specific columns with NaN (Not a Number).

- `dataset\_new.isnull().sum()`: This counts the missing values after the replacement.

- The following lines use `fillna()` to replace the NaN values with the mean of each respective column.

**dataset\_new.isnull().sum():** This checks for missing values again after the replacement.

**Conclusion:**

Overall, this program is designed to read a diabetes dataset from a CSV file, perform exploratory data analysis, and preprocess the data by handling missing values. The visualizations and summary statistics provide insights into the dataset's characteristics.

Hence using the above the diabetes can be predicted by Artificial Intelligence