**AI Based Diabetes Prediction System**

**Team Members:**

* Thajul Niyas -720921104110

## Phase-1 Document Submission

## Project : Diabetes Prediction

## A person using a pen to check the blood sugar level

## Abstract:

## Diabetes mellitus, a chronic metabolic disorder, affects millions of individuals worldwide and poses significant healthcare challenges. Early detection and management of diabetes can mitigate its complications and improve the quality of life for those affected. Machine learning, with its capability to analyze vast amounts of medical data, has emerged as a valuable tool in diabetes prediction. In this study, we propose a machine learning-based approach for diabetes prediction using a comprehensive dataset of clinical and demographic features.

## Our methodology involves data preprocessing, feature selection, and the development of predictive models. We explore various machine learning algorithms, including logistic regression, decision trees, support vector machines, and neural networks, to identify the most accurate and robust model for diabetes prediction. Feature importance analysis is conducted to understand the factors contributing to the prediction, aiding in clinical interpretation.

## Objective:

## The objective is to build an AI-powered diabetes prediction system that uses machine learning algorithms to analyze medical data and predict the likelihood of an individual developing diabetes. The system aims to provide early risk assessment and personalized preventive measures, allowing individuals to take proactive actions to manage their health.

## Phase-1 : *Data Preprocessing and Feature Engineering*

## Data Collection:

## Data Collection is the process of gathering, measuring and recording information or data on a specific topic or subjects of interest. It’s a fundamental step in research ,analysis and more. We need a dataset containing medical features such as glucose levels, blood pressure, BMI, etc., along with information about whether the individual has diabetes or not

## Dataset link:  <https://www.kaggle.com/datasets/mathchi/diabetes-data-set>

## Input:

import warnings

warnings.filterwarnings('ignore')

import os

for dirname, \_, filenames **in** os.walk('/kaggle/input'):

for filename **in** filenames:

print(os.path.join(dirname, filename))

import seaborn as sns

import matplotlib.pyplot as plt

import plotly.express as px

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

df.head(10)

|  | Pregnancies | Glucose | Blood Pressure | Skin Thickness | Insulin | Diabetes pedigree function | BMI | Age | outcome |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 1 | 148 | 72 | 35 | 0 | 0.627 | 33.6 | 50 | 1 |
| 1 | 6 | 85 | 66 | 29 | 0 | 0.351 | 26.6 | 31 | 0 |
| 2 | 1 | 183 | 64 | 0 | 0 | 0.672 | 23.3 | 32 | 1 |
| 3 | 8 | 89 | 66 | 23 | 94 | 0.167 | 28.1 | 21 | 0 |
| 4 | 1 | 137 | 40 | 35 | 168 | 2.288 | 43.1 | 33 | 1 |
| 5 | 0 | 116 | 74 | 0 | 0 | 0.201 | 25.6 | 30 | 0 |
| 6 | 5 | 78 | 50 | 32 | 88 | 0.248 | 31.0 | 26 | 1 |
| 7 | 3 | 115 | 0 | 0 | 0 | 0.134 | 35.3 | 29 | 0 |
| 8 | 10 | 197 | 70 | 45 | 543 | 0.158 | 30.5 | 53 | 1 |
| 9 | 2 | 125 | 96 | 0 | 0 | 0.232 | 0.0 | 54 | 0 |

## Output:

## Data Preprocessing:

## Data preprocessing Is a crucial step in my machine learning project. It involves cleaning , transforming , and organizing raw data into a format that is suitable for analysis or model training. Proper data Preprocessing can significantly improve the quality of results and the performance of machine learning models. The medical data needs to be cleaned, normalized, and prepared for training machine learning models .Here are some common data preprocessing tasks:

## 1.Data Cleaning:

## Handling missing data’s

## Handling Outliers

## Noise Reduction

## Code:

df.shape

(768, 9)

In [15]:

df=df.drop\_duplicates()

df.shape

(768, 9)

df.isnull().sum()

Pregnancies 0

Glucose 0

BloodPressure 0

SkinThickness 0

Insulin 0

BMI 0

DiabetesPedigreeFunction 0

Age 0

Outcome 0

df.columns

Index(['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',

'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome'],

dtype='object')

print("No. of Zero Values in Blood Pressure ", df[df['BloodPressure']==0].shape[0])

No. of Zero Values in Blood Pressure 35

## 2.Data Transformation:

## Data encoding

## Feature Scaling

## Feature Engineering

## 3.Data Reduction and sampling:

## Down sampling

## Up sampling

## Feature Selection:

## Feature selection is a crucial step in data analytics, particularly in machine learning and statistical analysis, it involves choosing a subset of the most relevant and informative features from the original dataset while discarding irrelevant or redundant ones. Feature selection offers several benefits, including improved model performance, reduced overlifting, and reduced computational complexity.  We will select relevant features that can impact diabetes risk prediction

## Model-Selection:

## Model selection is a critical step in data analysis , machine learning, and statistical modelling. It involves choosing the most appropriate model or algorithm for a given dataset and problem. The goal is to select a model that effectively captures the underlying patterns in the data, Providers good predictive performance, and is interpretable for making informed decisions.  We can experiment with various machine learning algorithms like Logistic Regression, Random Forest, and Gradient Boosting.

## Logistic regression

* from sklearn.metrics import accuracy\_score
* print("Train Accuracy of Logistic Regression: ", lr.score(X\_train, y\_train)\*100)
* print("Accuracy (Test) Score of Logistic Regression: ", lr.score(X\_test, y\_test)\*100)
* print("Accuracy Score of Logistic Regression: ", accuracy\_score(y\_test, lr\_pred)\*100)
* Train Accuracy of Logistic Regression: 77.36156351791531
* Accuracy (Test) Score of Logistic Regression: 77.27272727272727
* Accuracy Score of Logistic Regression: 77.27272727272727

## Evaluation:

## ****Evaluation in data analysis refers to the process of assessing the performance, quality, and validity of analytical results, models, or insights derived from data.it is a crucial step to determine the effectiveness and reliability of data driven decisions or predictions. Evaluation techniques and criteria may vary depending on the specific type of analysis.**** We will evaluate the model's performance using metrics like accuracy, precision, recall, F1-score, and ROC-AUC.

## Area under curve

* auc= roc\_auc\_score(y\_test, lr\_pred)
* print("ROC AUC SCORE of logistic Regression is ", auc)
* ROC AUC SCORE of logistic Regression is 0.7327726532826913

**Iterative Improvement:**

**We will fine-tune the model parameters and explore techniques like feature engineering to enhance prediction accuracy. Iterative improvement in data analytics refers to the process of continuously refining and enhancing analytical models, algorithms, and processes by repeatedly analyzing data. This approach is commonly used in various data-driven fields, such as machine learning, business intelligence and data science to achieve better results and insights overtime**