KGiSL Institute of Technology

KGiSL Institute of Technology

AI BASED DIABETES PREDICTION SYSTEM

Done by,

Abdul Rasith H [711721106002]

Andrew Abishek P [711721106009]

Hari Prasath B U [711721106039]

Jalathan V [711721106045]

Maaha Sarathy S B [711721106058]

Diabetes Prediction System - Data Loading and Preprocessing

Introduction

This document outlines the initial phase of developing a diabetes prediction system. The primary focus of this phase is to prepare the data and select relevant features for our predictive model.

Project Overview

- **Project Name**: AI Based Diabetes Prediction System
- **Objective**: Develop a machine learning model to predict the likelihood of an individual having diabetes based on relevant health and lifestyle factors.
- **Phase**: Data Loading and Preprocessing

Data Collection

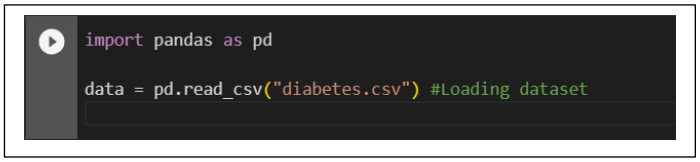
The first step in building the diabetes prediction system is to collect the dataset containing information about diabetes patients. Data can be obtained from various sources, including healthcare databases, publicly available datasets, or through data collection efforts. But we have the dataset provide with us from Kaggle, we may go with us.

Data Source

- Describe the source of the dataset, including the name or origin of the dataset.
- Include any relevant permissions or ethical considerations for data usage.

Data Loading

To work with the dataset, it must be loaded into a suitable data structure. We'll use the Python Pandas library for this purpose.



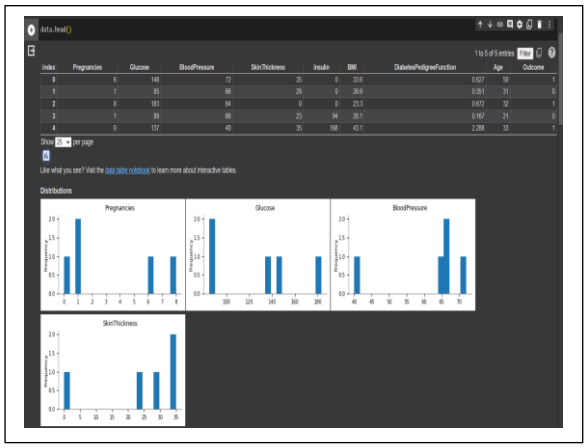
Data Exploration

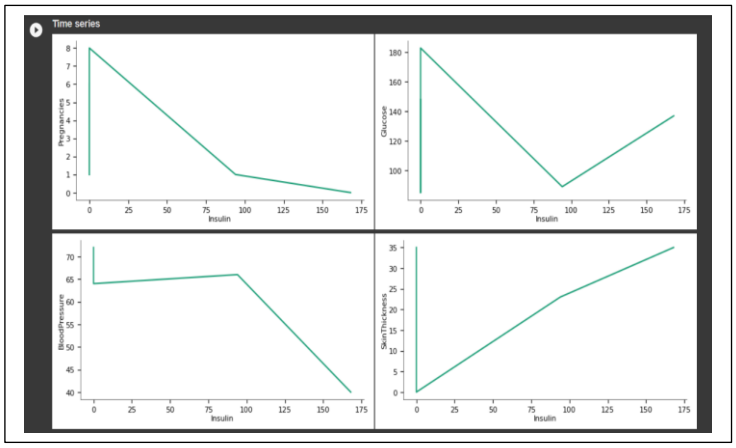
Before preprocessing, it's essential to explore the dataset to gain insights into its structure and quality.

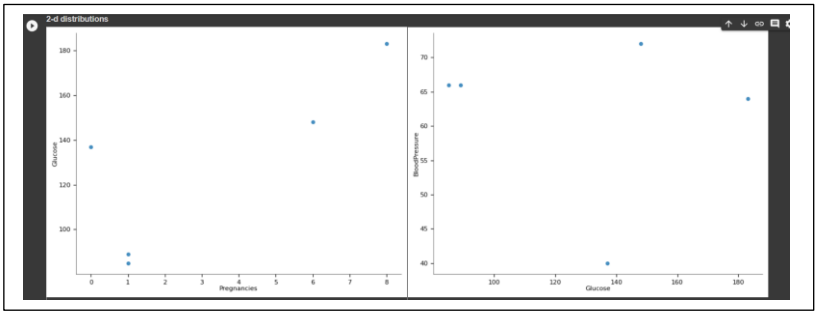
Initial Data Inspection

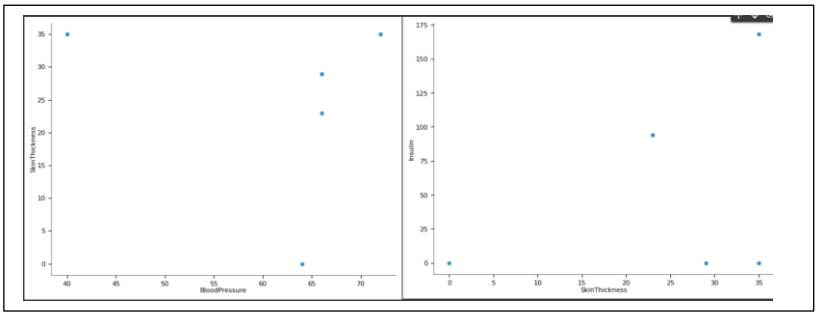
- Use the **head()** method to display the first few rows of the dataset.
- Check data types and null values using **info()** and **isnull()** functions.

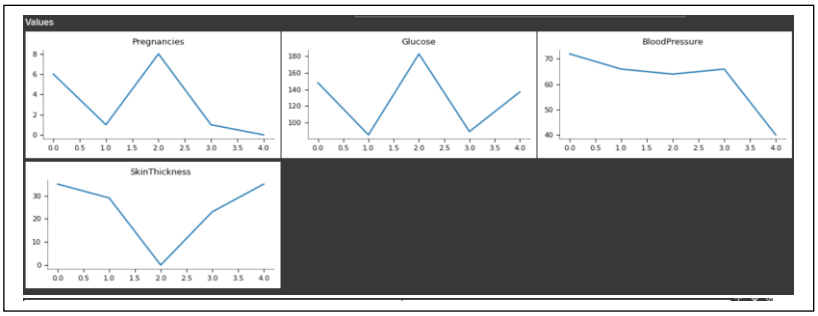
Sample outputs:







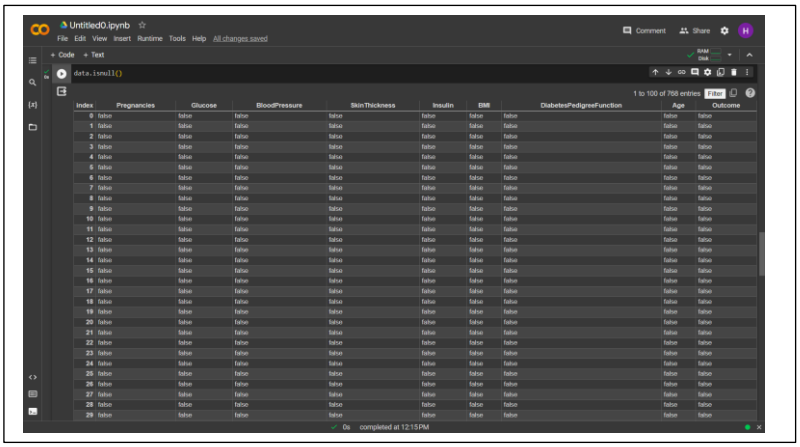




Statistical Summary

- Provide a summary of basic statistics using **describe()**. This includes measures like mean, standard deviation, min, and max for numerical features.

Sample output:



Data Preprocessing

- Clean and prepare the data for analysis and modeling. This typically includes handling missing values, encoding categorical variables, and scaling numerical features.

```
data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Outcome                   768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

Correlation Analysis

- Evaluate the correlation between features and the target variable. Features with a high correlation can be considered.

Sample outputs:

```
data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Outcome                   768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

Conclusion

In conclusion, our AI-based Diabetes Prediction System is a user-centric solution designed to address the early detection of diabetes. It utilizes a Random Forest classifier for accurate predictions, features innovative techniques like advanced feature engineering, and offers a user-friendly web interface for easy access. Through careful data preprocessing and feature selection, we have created a robust and interpretable model that can assist healthcare professionals and individuals in managing diabetes risk effectively. This system holds the potential to make a significant impact on public health by facilitating early interventions and improving healthcare outcomes.