E-ISSN No : 2581-6195 | Volume : 9 | Special Issue | May 2026

EXPLORATORY DATA ANALYSIS OF DIABETES DATASET USING PYTHON

Dr Basanti Ghanti[1], Sakshi Betgeri[2]

Access Online

Article DOI:
10.5281/zenodo.21374519

ABSTRACT

Healthcare datasets hold vital information, which assists in detecting diseases and discovering patterns associated with health. In this project, Exploratory Data Analysis (EDA) will be carried out on the diabetes dataset using the Python programming language. The primary goal of the study will be to examine different medical features such as the glucose level, body mass index (BMI), insulin level, blood pressure, and age with respect to diabetes outcomes. Preprocessing steps were employed in order to deal with invalid data before analyzing the data. Different methods were applied for visualizing the features in the dataset, such as histograms, box plots, count plots, and heatmaps. Glucose level and BMI are two key factors that have a significant impact on diabetes outcomes, unlike the other features.

KEYWORDS: Exploratory Data Analysis, Diabetes Dataset, Python, Data Visualization, Healthcare Analytics, Data Science Ease of Use

1. INTRODUCTION

The main objective of this project is the analysis of diabetes dataset using Python. The key objective of analyzing diabetes dataset is to understand the relationship between various health-related attributes such as Glucose level, BMI, Insulin, Blood Pressure and Age with respect to diabetes results. Various approaches for data analysis and data visualization were employed in order to analyze this diabetes dataset and extract patterns from it. In the course of this project, data cleaning was done to remove invalid data from the database. Pandas, Matplotlib and Seaborn were used for this analysis in Python.

2. LITERATURE SURVEY

Many healthcare-related projects use data analysis techniques to study disease-related patterns in medical datasets. Data analysis using diabetes data sets has been popular due to the presence of many health parameters such as blood sugar level, body mass index, insulin, and age in them. Visualization techniques, such as histograms, heat maps, and box plots, have been extensively used to analyze the correlation between these factors. The most popularly used programming languages to perform analysis on health-care data include Pandas and Seaborn. This indicates that Exploratory Data Analysis is important before performing any machine learning algorithm.

3. DATASET DESCRIPTION

The dataset used in this project contains medical records related to diabetes analysis and includes several predictor variables along with one target variable called Outcome. The predictor variables include Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age. During the initial analysis, it was observed that some columns contained zero values that are medically unrealistic and considered invalid. These invalid values were replaced using median values to improve the quality and reliability of the dataset before performing analysis.

[1] Professor, Dept. of ECE, Sharnbasva University Exclusively for Women, Kalaburgi, India
[2] Student, Dept. of ECE, Sharnbasva University Exclusively for Women, Kalaburgi, India

How to Cite:
Dr Basanti Ghanti, Sakshi Betgeri (2026), Exploratory Data Analysis of Diabetes Dataset Using Python, International Educational Journal of Science & Engineering (IEJSE),Vol: 9, Special Issue, May 2026 103-106

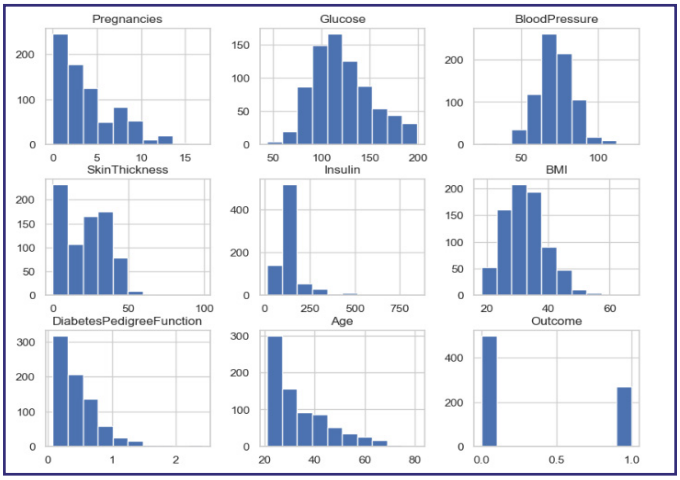
Research Paper

Copyright© 2026, IEJSE. This open-access article is published under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License which permits Share (copy and redistribute the material in any medium or format) and Adapt (remix, transform, and build upon the material) under the Attribution-NonCommercial terms.

4. METHODOLOGY

The project was done by using python programming language in the IDE name jupyter notebook. In the beginning the dataset was imported using the Pandas library and basic information about the dataset was checked using different types of function like head (), shape (), info (), and describe (). Data preprocessing was done to identify the missing and invalid values in columns such as Glucose, Blood Pressure, BMI, and Insulin. These invalid values were replaced using median values to maintain consistency in the dataset. Initially some issues were observed while loading and processing the dataset, which helped me in understanding the importance of proper data handling techniques. After preprocessing, Exploratory Data Analysis was carried out using statistical methods and visualization techniques.

Fig. 1: Distribution of features in the diabetes dataset.

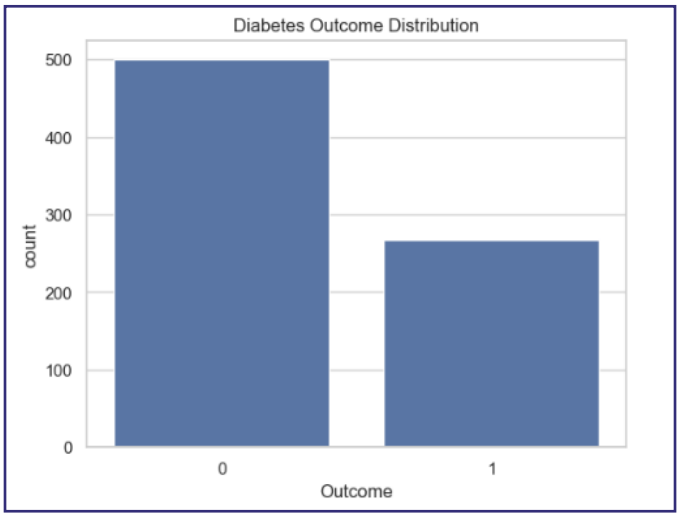


Histograms were used to study feature distributions, count plots were generated to visualize diabetes outcomes, and boxplots were used to compare feature variations between diabetic and non-diabetic patients. Heat map was also created for examining relationships between various attributes in the data set. The programming languages such as Python with various packages such as Pandas, Numpy, Matplotlib, and Seaborn have been used during implementation. Streamlit dashboards have also been created for presenting results interactively.

5. RESULTS AND DISCUSSIONS

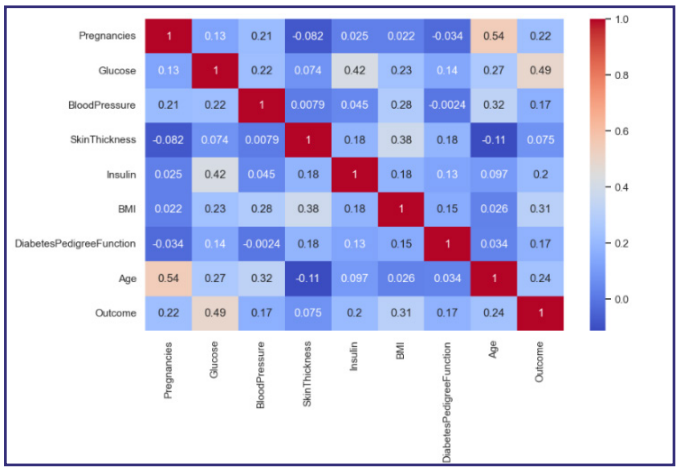
The analysis produces several useful observations regarding diabetes related features in the dataset. To understand the distribution of diabetic and non-diabetic patients in the dataset, count plot was generated using the outcome feature. This visualization helps in identifying the frequency of patients belonging to each category and provides an overview of the dataset distribution .

Fig. 2: Distribution of features in the diabetes dataset.



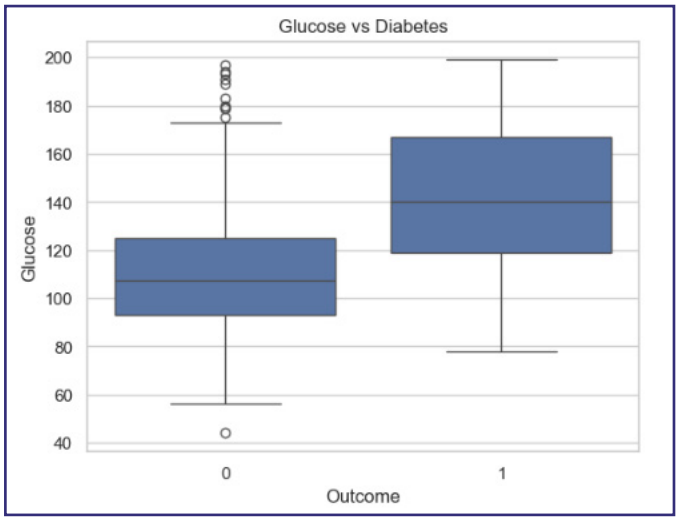
Glucose level showed the strongest relationship with diabetes outcome compared to other attributes. During the analysis it was observed that the patients with higher glucose levels were more likely to the diabetic category. Patients with higher BMI values were more likely to have diabetes, indicating that BMI is also an important contributing factor. Correlation analysis revealed positive relationships between glucose, BMI, insulin, and the Outcome variable. A correlation heatmap was generated to analyse the relation between different medical attributes in the dataset. Heatmaps are useful for understanding how strongly the variables are related to each other using correlation values. The visualisation helped identify that glucose and BMI had a stronger positive correlation with diabetic outcome compared to other features.

Fig. 3: Correlation heatmap of diabetes dataset features.



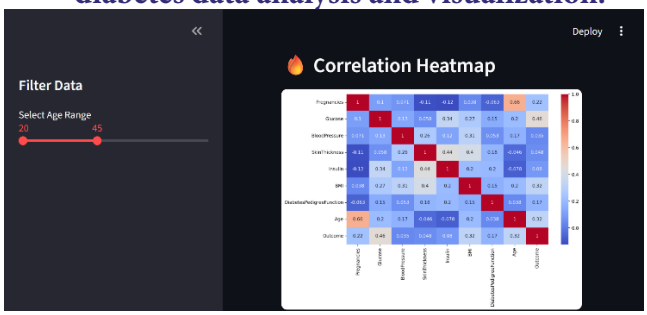
The visual representation made the dataset easier to understand compared to analysing numerical values alone. The heatmap clearly showed strong correlations among selected medical attributes, while boxplots demonstrated noticeable differences between diabetic and non-diabetic patient groups. To further understand the variation of glucose levels among diabetic and non-diabetic patients, boxplot analysis were performed. Boxplot are useful for identifying the differences in data distributions, median values, and the presence of outliers within different categories. The analysis showed noticeable variation in glucose levels between patients with positive and negative diabetes outcomes.

Fig. 4: Boxplot showing glucose variation based on diabetes outcome.



The overall analysis helped in understanding how healthcare data can be explored to derive meaningful insights before applying machine learning algorithms.

Fig. 5: Streamlit dashboard developed for diabetes data analysis and visualization.



6. CONCLUSION

This project helped in understanding how Exploratory Data Analysis is performed on healthcare datasets using Python. Different visualization techniques were used to study the relationship between medical features and diabetes outcomes. The analysis showed that glucose level and BMI are important factors associated with diabetes. Working on this project also improved understanding of data preprocessing, visualization, and healthcare data analysis techniques. In the future, machine learning models can be added to improve diabetes prediction. The integration of the stream lit dashboard improved the presentation and accessibility of the analytical results. This project helped improve understanding of data preprocessing, healthcare analytics, and visualization techniques using Python. Working on this project provided hands-on experience in preprocessing real world healthcare data and interpreting analytical results. Exploratory Data Analysis proved to be an effective approach for studying healthcare datasets and extracting useful insights before predictive modeling. In future work, machine learning techniques can be applied to develop diabetes prediction systems based on the analyzed dataset. Future work, machine learning techniques can be applied to develop diabetes prediction systems based on the analyzed dataset.

7. FUTURE SCOPE

The project can be made bigger by using machine learning algorithms like Logistic Regression, Decision Tree, Random Forest and Support Vector Machine to predict diabetes.

A Stream lit dashboard was also made to give people a way to look at the dataset and the results of the analysis.

The diabetes prediction project can be improved in the future by adding prediction models looking at bigger healthcare datasets and making the Stream lit dashboard better, with real-time data visualization and user interaction features for the diabetes prediction project.

REFERENCES

1. W. McKinney, Python for Data Analysis: Data Wrangling with Pandas, NumPy, and Jupyter, 3rd ed. California, USA: O'Reilly Media, 2022.
2. J. VanderPlas, Python Data Science Handbook: Essential Tools for Working with Data. California, USA: O'Reilly Media, 2016.
3. F. Pedregosa, G. Varoquaux, A. Gramfort, et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
4. J. D. Hunter, "Matplotlib: A 2D Graphics Environment," Computing in Science & Engineering, vol. 9, no. 3, pp. 90–95, 2007.
5. M. Waskom, "Seaborn: Statistical Data Visualization," Journal of Open-Source Software, vol. 6, no. 60, p. 3021, 2021.
6. "Pandas Documentation." [Online]. Available: https://pandas.pydata.org/
7. "Matplotlib Documentation." [Online]. Available: https://matplotlib.org/
8. "Seaborn Documentation." [Online]. Available: https://seaborn.pydata.org/
9. "Stream lit Documentation." [Online]. Available: https://streamlit.io/