PHASE 3 – data analysis

AI BASED DIABETES PREDICTION SYSTEM WITH MACHINE LEARNING USING PYTHON

SHRINIDHI V 2021504041

VIBIKSHA J 2021504053

VINODHINI R 2021504054

CHARUMATHI 2021504508

Phase 3 marks a critical juncture in our project, encompassing data loading, preprocessing, and comprehensive analysis. This stage is where raw data transforms into valuable insights, setting the stage for model development.



**TABLE OF CONTENTS**

1. INTRODUCTION
2. DATA PREPARATION
3. EPLORATORY DATA ANALYSIS
4. STATISTICAL ANALYSIS
5. OUTCOME SEPERATION
6. DATA DISTRIBUTION ANALYSIS
7. DIABETES PERCENTAGE ANALYSIS
8. DEPENDENCY ANALYSIS
9. DATA AUGMENTATION
10. ADDITIONAL DATA ANALYSIS
11. CORRELATION ANALYSIS
12. OBSERVATION ANS COMPARISONS
13. CONCLUSION
14. APPENDIX

**INTRODUCTION**

**Project Overview**: In this data analysis journey, we embark on a project of immense importance predicting diabetes with precision. Diabetes is a widespread global health concern, and early detection is pivotal for effective intervention. Our mission is to delve into the dataset, extract valuable insights, and construct predictive models to assist in early diabetes diagnosis.

**DATA PREPARATION**

**Library Import**: The engine that drives our analysis is a suite of Python libraries. We've harnessed the capabilities of Pandas for versatile data handling, NumPy for numerical operations, and Matplotlib and Seaborn for compelling data visualization. These tools form the bedrock of our data analysis.

**Google Drive Connection**: Recognizing the need for seamless data management and collaborative work environments, we've integrated our project with Google Drive. This connection streamlines data storage and accessibility, promoting a collaborative spirit among our team.

**Dataset Loading**: Our journey commences with the dataset. We load this reservoir of information, gaining access to the raw data that holds the potential to unlock valuable insights and patterns.

**EXPLORATORY DATA ANALYSIS (EDA)**

**Displaying the First and Last 10 Data Entries**: The data's initial and final ten entries offer our first glimpse into the dataset's structure. They provide a bird's eye view, helping us understand its organization and initial trends.

**Feature Overview**: In our dataset, each feature carries a unique potential for influencing diabetes prediction. A comprehensive feature overview:

1.***Pregnancies***: The number of pregnancies often correlates with diabetes risk. This feature might reveal the importance of reproductive history in predicting diabetes.

2.***Glucose***: As a primary indicator, glucose levels are central to diabetes prediction. Elevated glucose levels are a strong predictor.

3.***Blood Pressure***: High blood pressure can be both a cause and a consequence of diabetes. Understanding this feature is essential.

4.***Skin Thickness***: Skin thickness can affect insulin absorption, influencing blood sugar regulation. It may reveal the role of subcutaneous fat in diabetes risk.

5.***Insulin***: The insulin feature directly pertains to diabetes. Variations in insulin levels can indicate diabetes risk.

6.***BMI***: Body Mass Index (BMI) is a key marker of obesity, a significant risk factor for diabetes.

7.***Diabetes***: Pedigree Function: This function quantifies genetic influence on diabetes risk. Genetic predisposition plays a vital role in diabetes prediction.

8.***Age***: Age is a well established risk factor. As individuals age, the likelihood of diabetes increases.

9.***Outcome***: The outcome variable is our lodestar, indicating the presence (1) or absence (0) of diabetes. It serves as our North Star, guiding predictive model development.

**STATISTICAL ANALYSIS**

**Descriptive Statistics**: Our entry into statistical analysis provides a thorough examination of the dataset's core characteristics. Descriptive statistics, encompassing mean, median, standard deviation, minimum, and maximum values, paint a comprehensive picture of each feature.

**Grouping Data by Mean and Standard Deviation**: Grouping data based on mean and standard deviation reveals essential insights. Mean values act as beacons, illuminating central tendencies, while standard deviation measures variability. These statistics serve as foundations for identifying patterns.

**Data Visualization with Plots**: Data visualization breathes life into the numbers. Through histograms, box plots, and scatterplots, we reveal the essence of feature distributions, central tendencies, and relationships.

**OUTCOME SEPARATION**

**Creating a Separate Variable for Outcome**: We take the pivotal step of isolating the outcome variable. By designating it as the target variable for predictive models, we establish clarity and readiness for the model training phase.

**DATA DISTRIBUTION ANALYSIS**

**Histograms (Histplot):** Histograms are the storytellers of feature distributions. They unveil the unique shape, skewness, and central tendencies of each feature's distribution. These visual representations provide the foundation for deeper analysis.

**Distribution Plots (Distplot):** Distribution plots offer an in depth exploration of feature distributions. They reveal the subtleties of skewness, kurtosis, and data spread, allowing us to develop a richer narrative of the data.

**DIABETES PERCENTAGE ANALYSIS**

**Plotting the Percentage of People with Diabetes**: Visualizing the proportion of individuals with diabetes in our dataset is critical. The plotted chart provides an at a glance understanding of the balance between the two classes—individuals with diabetes and those without. This chart is a key reference for class balance.

**DEPENDENCY ANALYSIS**

**Scatterplot Analysis**: Scatterplots are the key to unlocking feature relationships. They allow us to identify patterns, both direct and indirect, between two features. These patterns inform our understanding of how specific feature pairs interact and potentially influence diabetes risk.

**Pairplot Analysis**: The pair plot takes us into the world of comprehensive feature dependencies. It simultaneously examines multiple feature relationships, offering a holistic view of the dataset's intricate interconnections. The pair plot is our Rosetta Stone for deciphering complex relationships.

**DATA AUGMENTATION**

**Augmentation for Increased Data**: Data augmentation injects vitality into our dataset. It introduces variation, diversifying the data and increasing its volume. Data augmentation is an essential step in preparing the dataset for robust model training.

**Reading Augmented Dataset:** With the augmented dataset in our toolkit, we set out to explore the impact of data augmentation on our analysis. By repeating earlier analyses on this enriched dataset, we uncover hidden insights and assess the value of augmentation.

**ADDITIONAL DATA ANALYSIS**

**Repeating Earlier Analyses on Augmented Dataset**: Our foray into the augmented dataset involves revisiting the same analyses that we performed on the original dataset. This comparative approach allows us to evaluate the efficacy of data augmentation. Differences and similarities in our observations shed light on the influence of augmented data on our analysis.

**CORRELATION ANALYSIS**

**Correlation Matrix**: Understanding feature relationships is a crucial part of our data analysis. The correlation matrix quantifies these relationships, revealing strong and weak correlations, both positive and negative. Features with high correlations may significantly impact each other and have implications for model development.

**Heatmap Visualization**: The correlation matrix, when visualized as a heatmap, transforms into a vivid representation of feature relationships. This heatmap simplifies the identification of significant correlations, making it an indispensable tool for feature selection and model development.

**OBSERVATIONS AND COMPARISONS**

**Observations from the Data Analysis**: Our analysis has uncovered a wealth of valuable insights, ranging from feature distributions to dependencies and correlations. These observations provide the compass for our journey into the model development phase. We now possess a nuanced understanding of the dataset's dynamics.

**Comparison of Original and Augmented Datasets**: Comparing the original and augmented datasets allows us to assess the impact of data augmentation. By identifying differences or similarities in our findings, we gauge the value of data augmentation in enriching our dataset.

**CONCLUSION**

**Summary of Key Findings**: Our data analysis journey has been fruitful, yielding numerous insights that form the bedrock of our diabetes prediction project. These insights encompass the distribution of features, intricate dependencies, and significant correlations. They provide the map for our future steps.

**Next Steps**: With data analysis complete, we set our sights on the horizon. Our next steps involve the development and evaluation of predictive models. Equipped with the knowledge distilled from this analysis, we seek to construct robust models for diabetes prediction.

**APPENDIX**

**Code Snippets**: This section serves as a reference point for the technical aspects of our analysis. It contains relevant code snippets, providing transparency in the data analysis process.

**Additional Visualizations**: Supplementary visualizations, charts, and graphs that enhance our understanding of the dataset and feature relationships are found here. These visual aids offer further clarity and context.

Link for the google colab to access the data we have analysed so far.

<https://colab.research.google.com/drive/14bcNlyr0ehZ0pmKKCaHHFAHUi9ZW2qzP?usp=sharing>