**Data Analysis**

**Data Preprocessing**

Effective preprocessing methods are essential for improving diabetes detection systems' accuracy. The main preprocessing procedures—Noise Identification and Handling, Zero-Value imputation, Outlier Analysis.

It was observed that the ML models performed well when few features namely - ‘Age’, ‘BMI’, and ‘SkinThickness’ were dropped from the input dataset. So, these features were dropped to obtain a better, higher accuracy with reliable results.

1. **Outlier Analysis**

When we plot a box-whisker plot, as in Figure 2, we can see that there are some outliers that need to be handled, which generally would be zero-values or missing values

1. **Noise Identification and Handling**

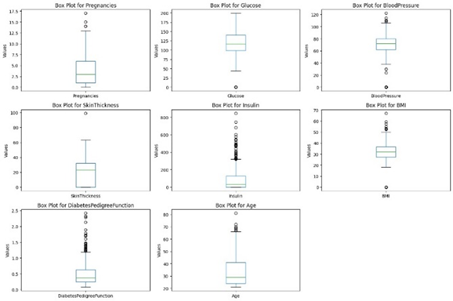
Likewise, some features like as 'BMI', 'SkinThickness', and 'Age' may not significantly contribute to the predictive task at hand or may introduce unnecessary noise into the modeling process. As a result, these attributes are strategically removed from the dataset using the

'drop' function in the pandas library.

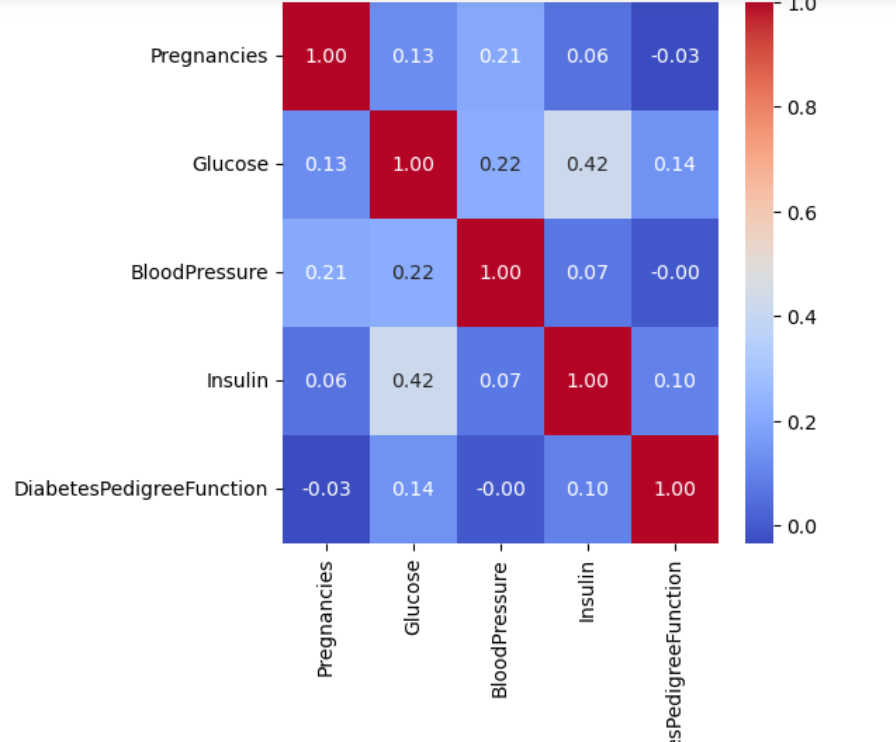
1. **Zero-Value imputation**

This dataset has no records having missing or null values, however it has some inconsistent values corresponding to features such as 'Insulin', 'SkinThickness', 'Glucose', and 'BloodPressure'. It has some zero-values instances which are practically erroneous. Instances where such values are zero are identified and replaced with the mean value corresponding to that feature calculated from non-zero instances.

1. **Normalization**

Normalization methods that are used in this study to change feature values to a particular range. As we have numerical features corresponding to different feature, we normalized it to avoid any unwanted bias over any feature

**Co-linearity analysis of the features**

****

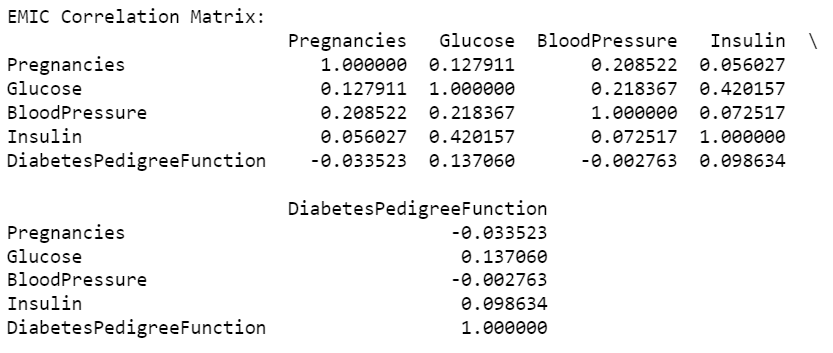
**Correlation Matrix**

**Inference:**

* **Pregnancies and Glucose:** The correlation coefficient between 'Pregnancies' and 'Glucose' is 0.127911, indicating a weak positive correlation. This suggests that there is a slight tendency for women with higher glucose levels to have slightly more pregnancies, though the correlation is not very strong.
* **Pregnancies and BloodPressure:** The correlation coefficient between 'Pregnancies' and 'BloodPressure' is 0.208522, indicating a weak positive correlation. This suggests a slight tendency for women with higher blood pressure levels to have slightly more pregnancies, but again, the correlation is not strong.
* **Pregnancies and Insulin:** The correlation coefficient between 'Pregnancies' and 'Insulin' is 0.056027, indicating a very weak positive correlation. There doesn't seem to be a significant relationship between the number of pregnancies and insulin levels.
* **Pregnancies and DiabetesPedigreeFunction:** The correlation coefficient between 'Pregnancies' and 'DiabetesPedigreeFunction' is -0.033523, indicating a very weak negative correlation. This suggests that there is little to no linear relationship between the number of pregnancies and the diabetes pedigree function.
* **Pregnancies and Outcome:** The correlation coefficient between 'Pregnancies' and 'Outcome' is 0.221898, indicating a weak positive correlation. This suggests that there may be a slight tendency for women with more pregnancies to have a higher likelihood of a positive diabetes outcome.
* **Glucose and BloodPressure:** The correlation coefficient between 'Glucose' and 'BloodPressure' is 0.218367, indicating a weak positive correlation. This suggests that there is a slight tendency for individuals with higher glucose levels to also have slightly higher blood pressure.
* **Glucose and Insulin:** The correlation coefficient between 'Glucose' and 'Insulin' is 0.420157, indicating a moderate positive correlation. This suggests that there is a moderate linear relationship between glucose levels and insulin levels, which is expected due to the role of insulin in glucose metabolism.
* **Glucose and DiabetesPedigreeFunction:** The correlation coefficient between 'Glucose' and 'DiabetesPedigreeFunction' is 0.137060, indicating a weak positive correlation. This suggests that there may be a slight tendency for individuals with higher glucose levels to have slightly higher diabetes pedigree function values.
* **Glucose and Outcome:** The correlation coefficient between 'Glucose' and 'Outcome' is 0.492928, indicating a moderate positive correlation. This suggests that there is a moderate linear relationship between glucose levels and the likelihood of a positive diabetes outcome.

**EMIC Correlation**

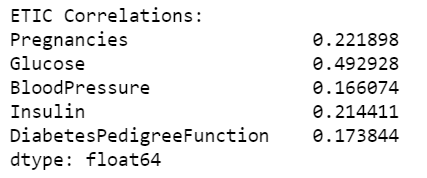
The EMIC correlation matrix helps to identify the strengths and directions of linear relationships between exogenous variables. While some correlations are present, multicollinearity does not appear to be severe among the variables analyzed. This analysis provides valuable insights for understanding the interplay between different factors in the dataset and can inform subsequent modeling and analysis approaches.



* **Pregnancies and Glucose (0.127911):** There is a weak positive correlation between the number of pregnancies and glucose levels. While there is some tendency for women with more pregnancies to have slightly higher glucose levels, the correlation is not substantial.
* **Pregnancies and BloodPressure (0.208522):** A weak positive correlation exists between the number of pregnancies and blood pressure levels. This suggests a slight tendency for women with more pregnancies to have slightly higher blood pressure, though the correlation is not strong.
* **Pregnancies and Insulin (0.056027):** There is a very weak positive correlation between the number of pregnancies and insulin levels. The correlation coefficient indicates that there is little to no linear relationship between these variables.
* **Pregnancies and DiabetesPedigreeFunction (-0.033523):** A very weak negative correlation is observed between the number of pregnancies and the diabetes pedigree function. This suggests that there is minimal linear relationship between the two variables.
* **Glucose and BloodPressure (0.218367):** A weak positive correlation exists between glucose levels and blood pressure. Individuals with higher glucose levels tend to have slightly higher blood pressure, although the correlation is not strong.
* **Glucose and Insulin (0.420157):** A moderate positive correlation is observed between glucose levels and insulin levels. This correlation indicates a moderate linear relationship between these two variables, which aligns with the role of insulin in glucose metabolism.
* **Glucose and DiabetesPedigreeFunction (0.137060):** There is a weak positive correlation between glucose levels and the diabetes pedigree function. Individuals with higher glucose levels may exhibit slightly higher diabetes pedigree function values, though the correlation is not strong.
* **BloodPressure and Insulin (0.072517):** A very weak positive correlation exists between blood pressure and insulin levels. The correlation coefficient suggests that there is minimal linear relationship between these variables.
* **BloodPressure and DiabetesPedigreeFunction (-0.002763):** There is virtually no correlation between blood pressure and the diabetes pedigree function, as indicated by the very weak correlation coefficient.
* **Insulin and DiabetesPedigreeFunction (0.098634):** A very weak positive correlation is observed between insulin levels and the diabetes pedigree function. The correlation coefficient suggests minimal linear relationship between these variables.

**ETIC Correlation Analysis**

The ETIC (Endogenous True Indicator Criterion) correlations provide insights into the relationships between the endogenous (dependent) variable and the exogenous (independent) variables in a dataset. In the context of the provided correlations from the PIMA Indian Diabetes dataset, the ETIC analysis reveals the following correlations:

* **Pregnancies (0.221898):** There is a weak positive correlation between the number of pregnancies and the diabetes outcome. This suggests that women with a higher number of pregnancies may have a slightly higher likelihood of a positive diabetes outcome.
* **Glucose (0.492928):** Glucose levels exhibit a moderate positive correlation with the diabetes outcome. Individuals with higher glucose levels tend to have a higher likelihood of a positive diabetes outcome, indicating the significant impact of glucose levels on diabetes.
* **BloodPressure (0.166074):** There is a weak positive correlation between blood pressure levels and the diabetes outcome. Individuals with higher blood pressure levels may have a slightly higher likelihood of a positive diabetes outcome, though the correlation is not strong.
* **Insulin (0.214411):** Insulin levels demonstrate a weak positive correlation with the diabetes outcome. Individuals with higher insulin levels may have a slightly higher likelihood of a positive diabetes outcome, though the correlation is not substantial.
* **DiabetesPedigreeFunction (0.173844):** The diabetes pedigree function exhibits a weak positive correlation with the diabetes outcome. Individuals with higher diabetes pedigree function values may have a slightly higher likelihood of a positive diabetes outcome, though the correlation is not strong.

These correlations help elucidate the relationship between each independent variable and the diabetes outcome. Variables with higher correlations indicate a stronger influence on the diabetes outcome. In this analysis, glucose levels demonstrate the most substantial correlation with the diabetes outcome, followed by pregnancies, insulin levels, diabetes pedigree function, and blood pressure, respectively.

**Data Interpretation**

The PIMA Indian Diabetes dataset is derived from medical examinations conducted on PIMA Native American women, aiming to understand factors contributing to diabetes prevalence within this population. The dataset is widely used in research to develop predictive models for diabetes risk assessment and management strategies.

**Variables:**

The dataset comprises several health-related variables:

**Pregnancies:** Number of pregnancies the individual has had.

**Glucose:** Plasma glucose concentration after a 2-hour oral glucose tolerance test.

**Blood Pressure**: Diastolic blood pressure measurement in mm Hg.

**Insulin:** 2-hour serum insulin measurement in mu U/ml.

**Diabetes Pedigree Function:** Diabetes pedigree function providing information about diabetes mellitus history in relatives.

**Age:** Age of the individual.

**Outcome:** Binary variable indicating whether the individual has diabetes (1) or not (0).

**Diabetes Risk Factors:** Analysis of the dataset can help identify significant risk factors associated with diabetes development. Variables like glucose levels, insulin levels, and family history (diabetes pedigree function) may contribute to higher diabetes prevalence.

**Predictive Modeling:** The dataset can be used to develop predictive models to identify individuals at high risk of developing diabetes. Machine learning algorithms can utilize variables such as glucose, insulin, and age to predict diabetes outcomes.

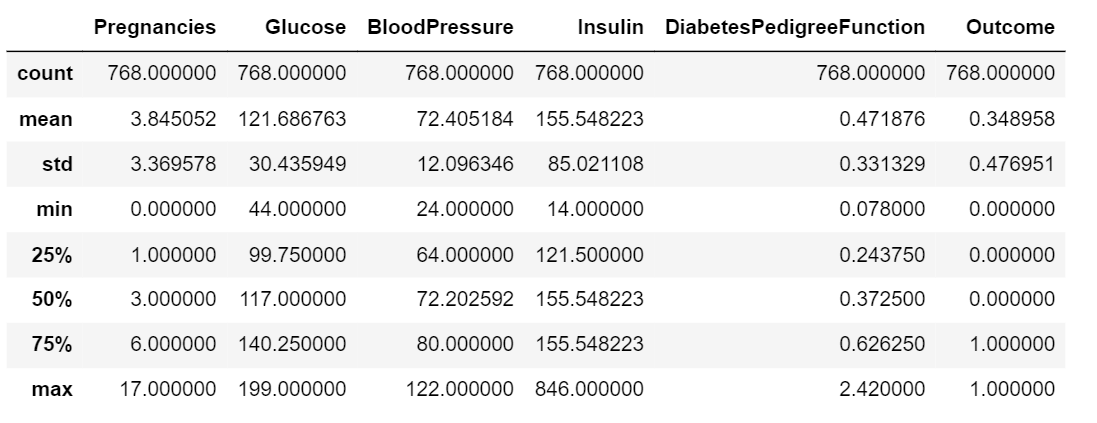
**Clinical Insights:** Insights gained from the dataset can inform clinical practice by highlighting the importance of regular monitoring of glucose levels, blood pressure, and insulin levels in high-risk populations such as PIMA Native American women.

**Preventive Strategies:** Understanding the dataset can guide the development of preventive strategies, including lifestyle modifications, dietary interventions, and early screening programs aimed at reducing diabetes incidence among at-risk populations.

**Research and Policy Implications:** Findings derived from the dataset may inform public health policies and interventions targeted at reducing diabetes prevalence and improving healthcare outcomes among vulnerable populations.

In summary, interpreting the PIMA Indian Diabetes dataset involves recognizing the multifaceted aspects of diabetes risk factors, predictive modeling, clinical insights, preventive strategies, and their implications for research and policy. Through thorough analysis and interpretation, valuable insights can be gleaned to address the challenges posed by diabetes within the PIMA database.

**Descriptive Statistics**

****

* **Pregnancies:** The average number of pregnancies among the women in the dataset is approximately 3.85, with a standard deviation of around 3.37. The range of pregnancies varies from 0 to 17.
* **Glucose:** The mean glucose level is approximately 121.69 mg/dL, with a standard deviation of 30.44 mg/dL. Glucose levels range from 44 mg/dL to 199 mg/dL.
* **BloodPressure:** The average blood pressure reading is approximately 72.41 mm Hg, with a standard deviation of 12.10 mm Hg. Blood pressure readings range from 24 mm Hg to 122 mm Hg.
* **Insulin:** The mean insulin level is approximately 155.55 mu U/ml, with a standard deviation of 85.02 mu U/ml. Insulin levels range from 14 mu U/ml to 846 mu U/ml.
* **DiabetesPedigreeFunction:** The average diabetes pedigree function value is approximately 0.47, with a standard deviation of 0.33. Values for the diabetes pedigree function range from 0.078 to 2.42.
* **Outcome:** The outcome variable indicates whether a patient has diabetes (1) or not (0). The mean outcome value is approximately 0.35, suggesting that around 35% of the individuals in the dataset have diabetes.

These statistics provide a comprehensive overview of the distribution, central tendency, and variability of the key variables in the dataset. Understanding these statistics is crucial for identifying patterns, outliers, and potential relationships between variables, which can inform further analysis and modeling efforts aimed at understanding and predicting diabetes outcomes in the PIMA Indian population

**Descriptive Dimensions**

**Temporal Dimension:**

Analysis of temporal trends in diabetes prevalence over time if the dataset includes timestamps or temporal information.

Understanding how diabetes risk factors and outcomes may have changed over different time periods, if applicable.

**Geographical Dimension:**

Examining potential geographical variations in diabetes prevalence and risk factors among PIMA Indian women across different regions or communities.

Exploring the influence of environmental factors, access to healthcare, and lifestyle differences on diabetes outcomes.

**Demographic Dimension:**

Investigating the impact of demographic factors such as age, ethnicity, socioeconomic status, and education level on diabetes risk and outcomes.

Understanding how diabetes prevalence and risk factors may vary across different demographic groups within the PIMA Indian population.

**Clinical Dimension:**

Analyzing clinical characteristics and medical history of individuals in the dataset, including previous diagnoses, treatments, and comorbidities.

Identifying patterns and associations between clinical variables and diabetes outcomes, such as the relationship between insulin levels and diabetes incidence.

**Behavioral Dimension:**

Exploring lifestyle factors and health behaviors that may influence diabetes risk, such as diet, physical activity, smoking, and alcohol consumption.

Understanding the role of behavioral interventions and lifestyle modifications in diabetes prevention and management strategies.

**Genetic Dimension:**

Investigating genetic predispositions and family history of diabetes among PIMA Indian women, including the influence of genetic markers and hereditary factors on diabetes risk.

Exploring gene-environment interactions and epigenetic mechanisms underlying diabetes susceptibility.

**Environmental Dimension:**

Examining environmental factors and contextual variables that may contribute to diabetes risk, including access to healthcare, food environment, socioeconomic disparities, and urbanization.

Understanding how environmental determinants interact with individual characteristics to shape diabetes outcomes within the PIMA Indian population.

**Micro-Interlocutor Analysis**

Micro-interlocutor analysis, also known as micro-level analysis, involves examining specific aspects or interactions within a dataset to gain deeper insights into underlying patterns, relationships, and phenomena. In the context of the PIMA Indian Diabetes dataset, micro-interlocutor analysis may encompass several key areas:

**Feature Importance Analysis:**

Assessing the importance of each feature (e.g., pregnancies, glucose levels, blood pressure, insulin levels, diabetes pedigree function) in predicting diabetes outcomes.

Utilizing techniques such as feature importance scores from machine learning algorithms (e.g., random forests, gradient boosting) to identify the most influential variables.

**Correlation Analysis:**

Exploring correlations between pairs of variables to uncover potential relationships and dependencies.

Analyzing correlation matrices to identify pairs of variables that exhibit strong positive or negative correlations, which may indicate underlying associations or multicollinearity issues.

**Outlier Detection:**

Identifying outliers or anomalous data points within the dataset that deviate significantly from the majority of observations.

Investigating the potential causes of outliers and assessing their impact on analysis and model performance.

**Model Evaluation:**

Evaluating the performance of predictive models trained on the dataset to assess their accuracy, precision, recall, and other relevant metrics.

Conducting cross-validation, hyperparameter tuning, and model selection to optimize model performance and generalization to unseen data.

**Subgroup Analysis:**

Conducting subgroup analysis based on demographic characteristics (e.g., age, ethnicity), clinical profiles, or other relevant variables to identify differential patterns or associations.

Assessing whether certain subgroups exhibit distinct risk factors or outcomes compared to the overall population.

**Temporal Analysis:**

Examining temporal trends and changes in diabetes prevalence, risk factors, and outcomes over time if the dataset includes longitudinal data or timestamps.

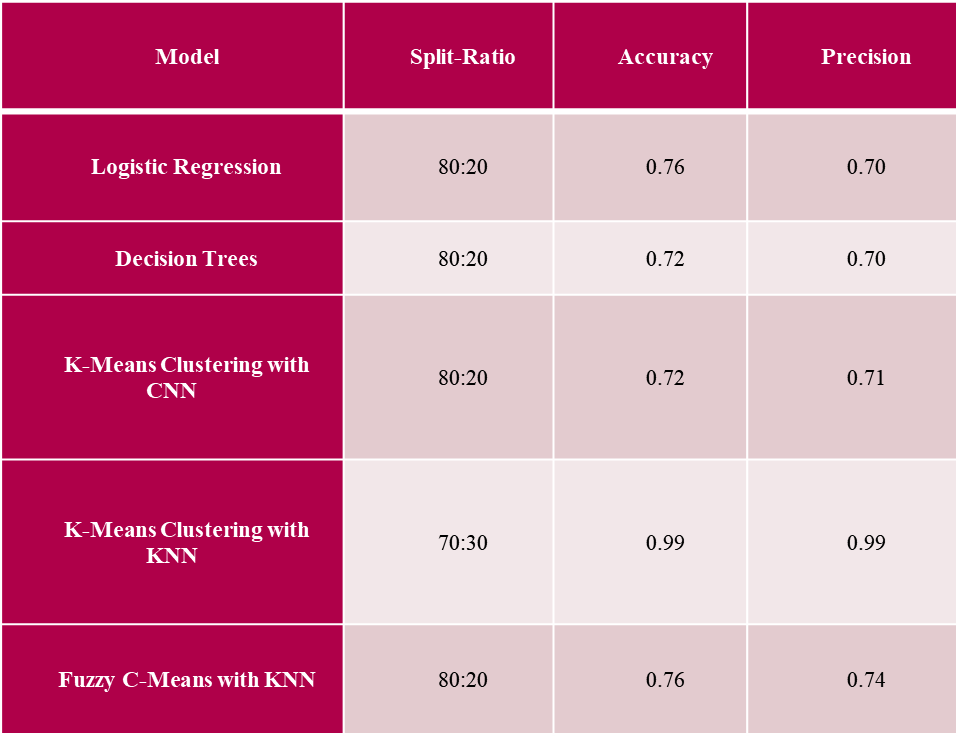
Investigating how temporal factors may influence diabetes incidence and progression within the PIMA Indian population.

**Dimensionality Reduction:**

Applying dimensionality reduction techniques such as principal component analysis (PCA) or t-distributed stochastic neighbor embedding (t-SNE) to visualize high-dimensional data and uncover underlying structures.

Reducing the dimensionality of the dataset while preserving as much relevant information as possible for subsequent analysis and modeling.

**Results**



Algorithms like Logistic Regression and Fuzzy C-Means with KNN have performed almost similarly with an accuracy of 76%. We have deployed Logistic Regression for the chat-bot integration as it involves simple mathematical functions that can be run over any light-weight server.

We have chosen Logistic Regression in lieu of its light-weight nature and simple classifying function. As we have numerical attributes, it would be better to analyze and handle the underlying patterns and give us a binary classifier. Also, we have deployed the model over a server that is connected with a chat-bot running using Telegram API implemented using NLTK framework.