FPT Education

FPT UNIVERSITY

Diabete Prediction from Health Data

Duong Tan Hung Thinh - Pham Huynh Quy An - Tran Nguyen Phuoc Nhan
Supervisor: Mr. Nguyen Quoc Trung

DSP391m

Spring Semester - 2025

- Ho Chi Minh City, 12 Jan 2025 -

Contents

I. Data Collection and Cleaning

1. Data Tasks:

The process of analyzing the Hóc Môn General Hospital diabetes dataset begins with defining the objective, understanding the dataset, and identifying potential data quality issues that may impact analysis. The goal is to ensure clean and reliable data for effective decision-making and predictive modeling.

Before starting data preprocessing and analysis, it is essential to establish clear objectives. In the context of the Hóc Môn dataset, the objectives may include:

- Assessing the prevalence of diabetes among patients visiting the hospital over the past two months, categorized by demographics such as age, gender, and BMI.
- Identifying key risk factors contributing to diabetes, including blood glucose levels, HbA1c, blood pressure, family history, and lifestyle factors.
- Detecting patterns and trends in diabetes-related health indicators to support early diagnosis and preventive healthcare strategies.

2. Data Collection Process

The dataset is collected from Hóc Môn General Hospital, which records patient visit data related to diabetes diagnosis and risk factors. The data is sourced from hospital medical records and clinical assessments over the past two months.

The data collection methods include:

- Electronic Medical Records (EMR): Patient information, including demographic details, medical history, and test results (e.g., blood glucose, HbA1c, blood pressure), is recorded by healthcare professionals.
- Clinical Assessments: Doctors and nurses conduct physical examinations and laboratory tests to assess diabetes risk factors during patient visits.
- Patient Questionnaires: Some data is gathered through self-reported lifestyle surveys, capturing dietary habits, physical activity levels, and family history of diabetes.

3. Data Sources and Acquisition

The Hóc Môn General Hospital Diabetes Dataset is derived from patient records collected over the past two months. It includes medical and demographic data from individuals who visited the hospital for diabetes-related consultations and screenings. The dataset captures essential health indicators such as blood glucose levels, HbA1c, blood pressure, BMI, and lifestyle factors, offering valuable insights into diabetes prevalence and risk factors within the local population.

The dataset contains structured information on age, gender, medical history, and family history of diabetes, enabling a comprehensive analysis of trends and patterns. It is typically available in formats like CSV, Excel, and SQL database files, ensuring ease of access and manipulation for research and predictive modeling. Data fields are standardized to maintain consistency and facilitate integration with other health datasets.

While the dataset provides a detailed snapshot of diabetes cases in Hóc Môn, there are challenges associated with data completeness and potential biases. Some records may have missing values due to incomplete patient information, and self-reported lifestyle factors may introduce inaccuracies. Additionally, since the dataset is based on hospital visits, it may not fully represent individuals who have undiagnosed diabetes or those who do not seek medical care.

To address these limitations, data preprocessing techniques such as missing value imputation and outlier detection can be applied. Moreover, further data collection efforts, including community health screenings and follow-up surveys, could enhance the dataset's representativeness and reliability for diabetes research and prevention strategies.

4. Data Cleaning Strategies

To ensure a high-quality dataset for diabetes analysis and prediction, the following data cleaning and preprocessing steps will be applied to the Hóc Môn General Hospital Diabetes Dataset:

- Create a Diabetes Column: A new binary column, "Diabetes", will be created based on the diagnosis column, where patients diagnosed with diabetes will be assigned 1 (Diabetic) and non-diabetic patients will be assigned 0 (Non-Diabetic).
- Remove Low-Sample Features: Any feature (column) with fewer than 1,000 valid samples will be removed from the dataset to eliminate underrepresented or incomplete attributes that may introduce noise.
- Standardize Gender Encoding: The Gender column will be converted to numerical values for consistency, where 0 represents Female and 1 represents Male.
- Impute Missing Values with Predictive Modeling: Missing values will be filled using machine learning models (Linear Regression, Random Forest, and XGBoost). Only predictions with an R^2 score ≥ 0.95 will be accepted to ensure high accuracy in imputed values. Otherwise, missing values will remain unchanged.
- Statistical Visualization & Analysis:
    - Distribution plots will be used to visualize feature distributions before and after missing value imputation.
    - Correlation analysis will help identify relationships between features.
    - Comparative distribution analysis will be conducted to evaluate how filling missing values affects the overall dataset structure.

By following this structured data preprocessing approach, the dataset will be clean, standardized, and optimized for diabetes prediction models.

5. Data Preprocessing Steps

Data preprocessing involves transforming raw medical records into a structured format suitable for analysis and machine learning models. The first step is encoding categorical variables, such as Gender, into numerical values to ensure efficient processing. Label encoding will be used to convert gender into binary values (0 for Female, 1 for Male), while other categorical features, if present, may be transformed using one-hot encoding when necessary. Additionally, feature selection techniques will be applied to remove irrelevant or redundant variables, improving both computational efficiency and model accuracy.

Another critical step is normalization and scaling of continuous variables. Key features such as blood glucose levels, BMI, and HbA1c will be standardized using Min-Max scaling or Z-score normalization. This ensures all numerical attributes are on a comparable scale, preventing certain variables from disproportionately influencing the model's learning process.

Finally, feature engineering will be applied to create new variables that enhance predictive capabilities. For instance, BMI values may be categorized into weight classes, and a new diabetes risk score may be derived based on multiple risk factors like age, blood pressure, and family history. These preprocessing techniques refine the dataset, making it more structured and effective for diabetes prediction and analysis.

6. Exploratory Data Analysis Overview

Exploratory Data Analysis (EDA) helps uncover patterns, trends, and anomalies in the Hóc Môn General Hospital Diabetes Dataset. The process begins with computing summary statistics, including mean, median, standard deviation, and frequency distributions, to understand the overall characteristics and distribution of variables such as blood glucose levels, HbA1c, BMI, and blood pressure.

To gain deeper insights, visualization techniques will be applied:

- Histograms to examine feature distributions.
- Box plots to detect outliers in key medical indicators.
- Scatter plots to analyze relationships between continuous variables.
- Heatmaps of correlation coefficients to identify potential multicollinearity, ensuring that highly correlated features do not negatively impact predictive modeling.

Additionally, segmentation analysis will be conducted to compare diabetes prevalence across different demographic groups. By stratifying the dataset based on age, gender, and BMI categories, we can explore disparities in diabetes risk and related health behaviors. These insights can guide preventive healthcare strategies and improve predictive model performance.

7. EDA Techniques

Exploratory Data Analysis (EDA) involves univariate, bivariate, and multivariate analyses to better understand the Hóc Môn General Hospital Diabetes Dataset and prepare it for predictive modeling.

- Univariate Analysis: This step focuses on individual variables by computing mean, variance, skewness, and distribution to identify potential anomalies and outliers. Histograms and box plots will be used to examine the spread of key features such as blood glucose, HbA1c, and BMI, helping detect irregularities that may require corrective actions.
- Bivariate Analysis: Relationships between two variables will be explored using scatter plots, correlation matrices, and statistical tests. Chi-square tests will assess categorical relationships (e.g., gender and diabetes prevalence), while correlation analysis will help determine strong predictors of diabetes, such as BMI, blood pressure, and physical inactivity.
- Multivariate Analysis: Advanced techniques like Principal Component Analysis (PCA) and clustering will be applied to identify hidden patterns in high-dimensional data. PCA helps with dimensionality reduction, improving model efficiency, while clustering techniques can segment patients into risk groups, providing deeper insights into diabetes patterns.

By combining these EDA techniques, we can extract meaningful patterns, refine feature selection, and improve diabetes prediction models.

8. Data Visualization Tools

Effective data visualization tools help communicate insights from the Hóc Môn General Hospital Diabetes Dataset, enabling better decision-making and analysis.

- Static Visualizations: Tools like Matplotlib and Seaborn will be used to create histograms, box plots, scatter plots, and heatmaps, allowing researchers to explore trends, distributions, and correlations within the dataset.
- Interactive Dashboards: Power BI and Tableau can be used to develop dynamic dashboards that enable users to filter, drill down, and explore data in real-time, making it easier to identify key diabetes risk factors.
- Geospatial Analysis: If location data is available, choropleth maps can be used to visualize regional diabetes prevalence, helping policymakers focus healthcare interventions in high-risk areas.
- Advanced Interactive Visualizations: Techniques such as drill-down charts, animations, and interactive scatter plots can enhance engagement, providing a more detailed and dynamic understanding of diabetes trends.

By leveraging these visualization techniques, we can enhance data-driven decision-making, improve predictive model interpretation, and raise public health awareness regarding diabetes.

II. Exploratory Data Analysis

1. Overview

The dataset consists of multiple medical parameters and laboratory test results of patients recorded at HocMon General Hospital in the past two months. It includes key clinical attributes such as:

- Demographics: Gender, Age
- Diabetes Status: Presence of diabetes (binary classification)
- Blood and Biochemical Markers: Glucose levels, Hemoglobin A1c (HBA), Cholesterol levels, Kidney function tests (Creatinine, Urea, EGFR), Liver function tests (ALT, Bilirubin, Albumin)
- Lipid Profile: HDL, LDL, Triglycerides
- Electrolytes: Sodium, Chloride
- Other Tests: Microalbuminuria, Acid Phosphatase, Gamma-Glutamyl Transferase (GGT)

Initial Observations:

- Missing Values: Some lab test values are missing for certain patients, requiring data imputation or removal.
- Class Imbalance: The dataset includes both diabetic and non-diabetic patients, but further analysis is needed to determine the distribution.
- Age Distribution: Patients range from 50 to 69 years old, with a predominance of middle-aged and elderly individuals.
- Comorbidities: Many patients have hypertension (I10), cardiovascular diseases (e.g., myocardial infarction, arrhythmia), and metabolic disorders (e.g., hyperkalemia).

3. Key Insights from Data Distribution

- Glucose Levels: Several patients exhibit high fasting glucose (GLUD) and postprandial glucose levels, indicating potential diabetes or impaired glucose tolerance.
- Hemoglobin A1c (HBA): This long-term diabetes indicator will be analyzed for correlation with glucose levels.
- Lipid Profile: High LDL and low HDL levels are common among patients with diabetes and hypertension.
- Kidney Function: Creatinine and EGFR values vary across patients, with some cases indicating potential kidney dysfunction.
- Liver Function: Elevated ALT and bilirubin levels in some diabetic patients suggest possible liver complications.

2. Dataset

The dataset consists of 5,233 patient records with various clinical and demographic features, including laboratory test results and indicators commonly associated with diabetes diagnosis. These features vary in completeness, with some having nearly full observations while others contain significant amounts of missing data. The presence of missing values can pose challenges in statistical analysis and machine learning models, making it essential to assess the quality of each feature before proceeding with imputation or modeling.

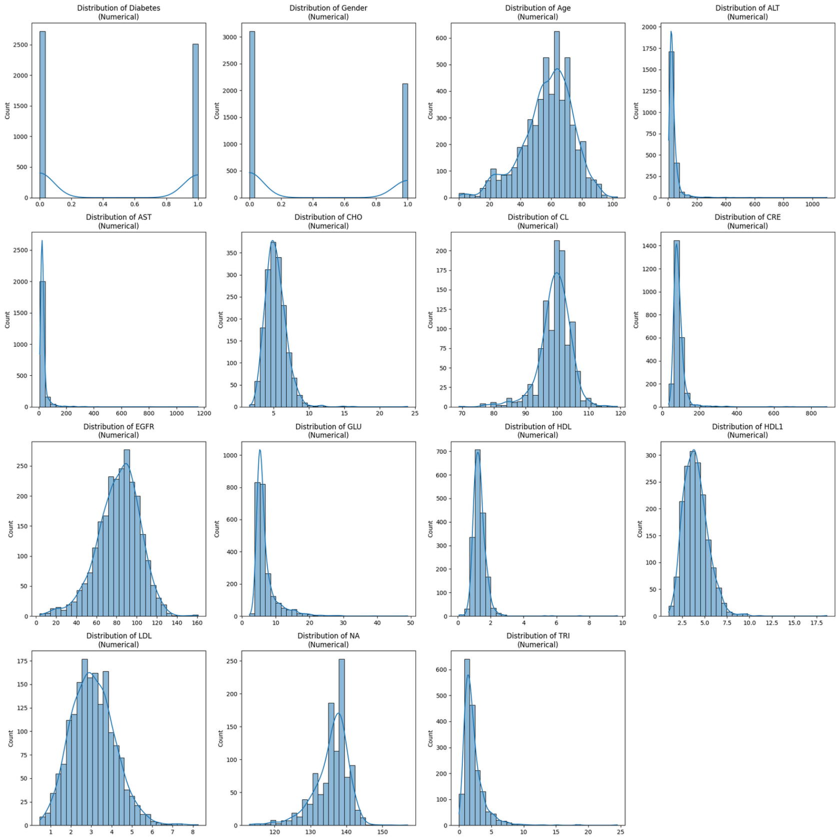
Initially, the dataset contained 33 columns, representing different biomarkers, metabolic indicators, and health metrics. However, many of these features had a very low number of recorded samples. For example, ALB (38 samples), AMY (20 samples), CKMB (17 samples), and ETHA (5 samples) had extremely sparse data, making them unreliable for meaningful analysis. Such a high percentage of missing values can lead to biased imputations and instability in model predictions. Retaining features with inadequate data might also introduce noise rather than valuable insights, ultimately affecting the robustness of predictive models.

To improve data quality and ensure that only statistically significant and well-represented features are included, a filtering threshold was applied, removing any feature with fewer than 1,000 valid samples. After this step, the dataset was refined to 15 columns, keeping only the most informative and well-populated features. The retained features include ALT, AST, CHO, CRE, GLU, HDL, LDL, and NA, which are commonly used in diabetes assessment and general health evaluations. This reduction in dimensionality enhances the dataset's reliability, reduces computational complexity, and ensures that subsequent analyses and machine learning models are trained on high-quality data.

By eliminating sparsely populated features, the dataset becomes more robust, improving the accuracy of imputation strategies and model predictions. Furthermore, focusing on well-represented clinical indicators allows for a more meaningful exploration of diabetes-related patterns, enabling better insights into potential risk factors and disease progression.

3. Distribution

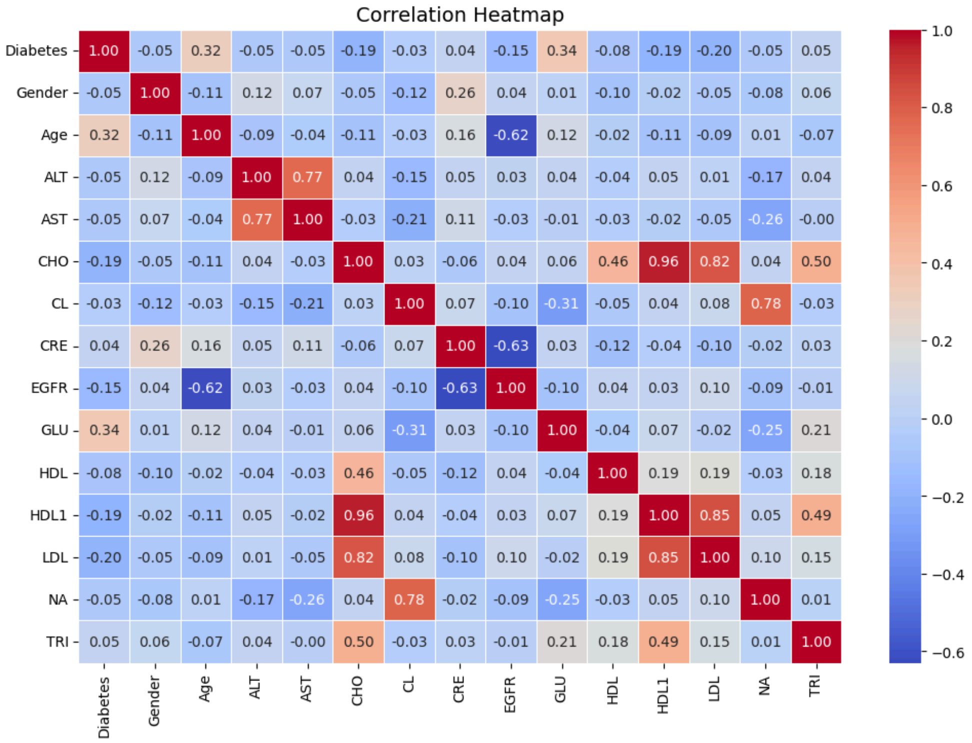
- The dataset contains a mix of categorical (e.g., Diabetes, Gender) and continuous numerical features. Many features show a right-skewed distribution, which suggests the presence of extreme values (potential outliers). Some features exhibit bimodal or multimodal distributions, which could indicate the presence of subgroups in the data.
- Diabetes (Categorical, Binary): The distribution is highly imbalanced, with roughly equal counts of 0 and 1. This suggests a balanced dataset for classification.
- Gender (Categorical, Binary): Similar to Diabetes, the data is split into two categories with a near-even distribution.
- Age: The age distribution is approximately normal, centered around middle age (40–60 years). A slight skew suggests a few younger and older individuals.
- ALT & AST (Liver Enzymes): Highly right-skewed, with a small number of extreme values. A transformation like log scaling or robust scaling might help.
- CHO, HDL, LDL, TRI (Cholesterol and Lipid Profiles): These values appear to follow a log-normal distribution, with some right-skewness. Some features (HDL, LDL) have long tails, which might be caused by outliers.
- CL, NA (Electrolytes): CL (Chloride) and NA (Sodium) have distributions that are nearly normal, but with minor skewness. There are potential outliers at the extreme ends.
- CRE & EGFR (Kidney Function): EGFR appears normally distributed, whereas CRE has a right-skewed distribution. The kidney function variables should be examined further to check for relationships with diabetes.
- GLU (Blood Glucose): The distribution is right-skewed, which is expected since diabetes is linked to high glucose levels. There might be a mix of diabetic and non-diabetic groups.



Features Distribution

3. Linear Distribution

- The correlation heatmap provides an overview of relationships between different features in the dataset. The color intensity indicates the strength of correlation, with red representing strong positive correlations and blue representing strong negative correlations.
- AST and ALT have a very high correlation (0.77), indicating they may be closely related biological markers.
- HDL1 and HDL (0.96) show an almost perfect correlation, suggesting they may represent similar or overlapping measures.
- LDL and HDL1 (0.85), as well as LDL and CHO (0.82), show strong positive correlations.
- GLU shows a moderate positive correlation with Diabetes (0.34), which is expected as glucose levels are a key indicator of diabetes.
- Age has a noticeable positive correlation with Diabetes (0.32), possibly indicating that older individuals in the dataset are more likely to have diabetes.
- EGFR has a notable negative correlation with Age (-0.62), which aligns with medical knowledge that kidney function (estimated glomerular filtration rate) declines with age. HDL has a negative correlation with Diabetes (-0.8), which suggests lower HDL cholesterol might be associated with diabetes.
- Some features, such as HDL1 and HDL, as well as AST and ALT, are highly correlated, which might introduce redundancy in modeling.
- Many features have very weak correlations with Diabetes, which may indicate they have little impact on predicting diabetes in this dataset.
- This correlation heatmap effectively highlights relationships between different health indicators. Feature engineering, non-linear analysis, and dimensionality reduction techniques should be considered before using these features in predictive modeling.

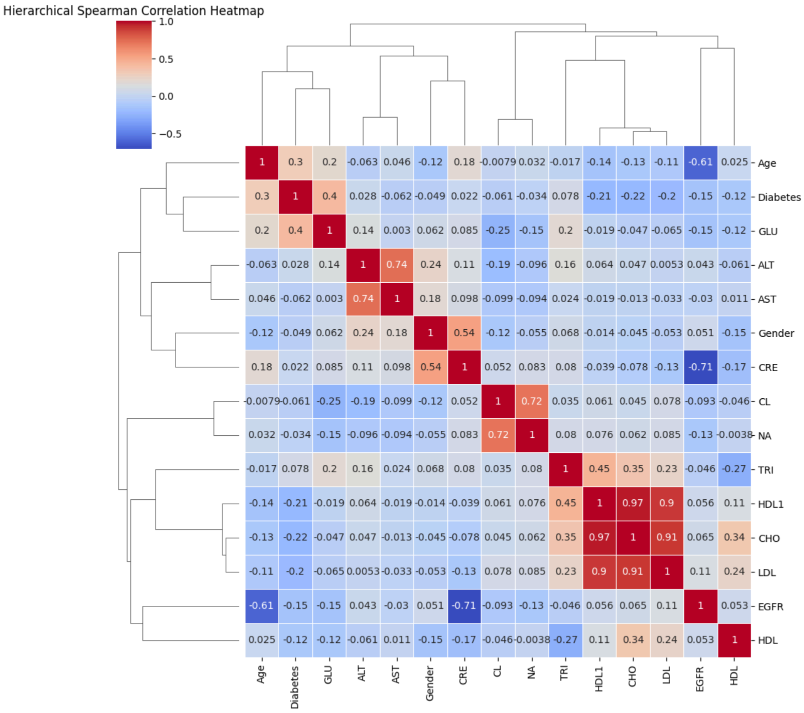


	Diabetes	Gender	Age	ALT	AST	CHO	CL	CRE	EGFR	GLU	HDL	HDL1	LDL	NA	TRI
Diabetes	1.00	-0.05	0.32	-0.05	-0.05	-0.19	-0.03	0.04	-0.15	0.34	-0.08	-0.19	-0.20	-0.05	0.05
Gender	-0.05	1.00	-0.11	0.12	0.07	-0.05	-0.12	0.26	0.04	0.01	-0.10	-0.02	-0.05	-0.08	0.06
Age	0.32	-0.11	1.00	-0.09	-0.04	-0.11	-0.03	0.16	-0.62	0.12	-0.02	-0.11	-0.09	0.01	-0.07
ALT	-0.05	0.12	-0.09	1.00	0.77	0.04	-0.15	0.05	0.03	0.04	-0.04	0.05	0.01	-0.17	0.04
AST	-0.05	0.07	-0.04	0.77	1.00	-0.03	-0.21	0.11	-0.03	-0.01	-0.03	-0.02	-0.05	-0.26	-0.00
CHO	-0.19	-0.05	-0.11	0.04	-0.03	1.00	0.03	-0.06	0.04	0.06	0.46	0.96	0.82	0.04	0.50
CL	-0.03	-0.12	-0.03	-0.15	-0.21	0.03	1.00	0.07	-0.10	-0.31	-0.05	0.04	0.08	0.78	-0.03
CRE	0.04	0.26	0.16	0.05	0.11	-0.06	0.07	1.00	-0.63	0.03	-0.12	-0.04	-0.10	-0.02	0.03
EGFR	-0.15	0.04	-0.62	0.03	-0.03	0.04	-0.10	-0.63	1.00	-0.10	0.04	0.03	0.10	-0.09	-0.01
GLU	0.34	0.01	0.12	0.04	-0.01	0.06	-0.31	0.03	-0.10	1.00	-0.04	0.07	-0.02	-0.25	0.21
HDL	-0.08	-0.10	-0.02	-0.04	-0.03	0.46	-0.05	-0.12	0.04	-0.04	1.00	0.19	0.19	-0.03	0.18
HDL1	-0.19	-0.02	-0.11	0.05	-0.02	0.96	0.04	-0.04	0.03	0.07	0.19	1.00	0.85	0.05	0.49
LDL	-0.20	-0.05	-0.09	0.01	-0.05	0.82	0.08	-0.10	0.10	-0.02	0.19	0.85	1.00	0.10	0.15
NA	-0.05	-0.08	0.01	-0.17	-0.26	0.04	0.78	-0.02	-0.09	-0.25	-0.03	0.05	0.10	1.00	0.01
TRI	0.05	0.06	-0.07	0.04	-0.00	0.50	-0.03	0.03	-0.01	0.21	0.18	0.49	0.15	0.01	1.00

Features Linear Correlation

4. Nonlinear Distribution

- While the given correlation matrix provides insights into the linear relationships between features, it does not capture nonlinear dependencies that may exist within the dataset. Here are key points to consider:
- The correlation coefficients in the table are based on Pearson correlation, which measures only linear relationships. nIf variables have a nonlinear but strong relationship, Pearson correlation may indicate a weak or even zero correlation. This limitation suggests the need for alternative techniques to detect hidden nonlinear patterns.
- Diabetes vs. Glucose (GLU): The correlation between Diabetes and GLU is 0.401, suggesting a moderate positive linear relationship. However, in reality, blood glucose levels and diabetes might follow a more complex relationship—e.g., a threshold effect, where a sharp increase in diabetes risk occurs after a certain glucose level. Nonlinear approaches (e.g., polynomial regression, decision trees, or mutual information) could provide deeper insights.
- Age vs. EGFR (-0.611): This strong negative correlation suggests that kidney function declines with age. However, this decline is likely not perfectly linear—kidney function may drop more steeply at older ages. A nonlinear regression or a spline model could capture this trend more accurately.
- CHO, HDL1, LDL, and TRI Relationships: Cholesterol (CHO), HDL (HDL1), and LDL (LDL) are highly correlated (0.965 and 0.905). However, lipid metabolism and cardiovascular risk may involve nonlinear interactions. Cholesterol's effect on diabetes or heart disease could be better analyzed using interaction terms or clustering methods.
- While Pearson correlation suggests some meaningful linear relationships (e.g., GLU and Diabetes or Age and EGFR), it misses nonlinear dependencies. Using mutual information, Spearman correlation, or machine learning models can help uncover deeper patterns in the data.



Features Nonlinear Correlation

5. Filling Missing Values

- Handling missing values in medical datasets, such as those related to diabetes, is crucial for ensuring accurate analyses and predictions. This approach integrates machine learning models, feature correlation analysis, and statistical adjustments to impute missing values while maintaining the integrity of the dataset. By first addressing skewed distributions, selecting the most relevant features, and leveraging advanced imputation techniques, this method ensures that missing data is replaced in a way that reflects the true characteristics of the original dataset.
- Many real-world datasets, particularly those in healthcare, contain skewed distributions that can distort model predictions. Skewness is common in variables such as glucose (GLU), triglycerides (TRI), and cholesterol (CHO). If not corrected, highly skewed features may bias machine learning models, leading to inaccurate imputations. To address this, several transformation techniques are applied before imputation. Log transformation is used for positively skewed variables to reduce extreme values, while Box-Cox transformation stabilizes variance for positive-only data. Yeo-Johnson transformation is applied to datasets containing zero or negative values, ensuring that all numerical features follow a more normal distribution. Quantile transformation further enhances normalization by mapping the data to a Gaussian-like distribution, improving model performance.
- Feature selection plays a critical role in ensuring high-quality imputations. Instead of using all available variables, only the top four most correlated features are selected for imputing each missing variable. These features are identified through the Spearman correlation matrix, which captures both linear and nonlinear relationships. By focusing on the most relevant predictors, this method prevents overfitting and ensures that the imputation process reflects real-world dependencies within the dataset. To avoid data leakage, missing values in these selected features are replaced with their mode (most frequent value) before model training.
- The imputation process utilizes multiple machine learning models to predict missing values based on the most relevant features. Four models are tested: Random Forest Regressor, XGBoost Regressor, Multi-Layer Perceptron (MLP) Regressor, and K-Nearest Neighbors (KNN) Regressor. Each model is trained on the observed (non-missing) data and evaluated using the R^2 score to determine its predictive accuracy. To optimize model performance, hyperparameters are adjusted dynamically, increasing the number of estimators from 100 to 2000 in increments of 100. If a model achieves an R^2 score of 0.95 or higher, training stops early, selecting the best-performing model for each target variable.
- Once missing values are predicted, further statistical adjustments are applied to ensure that the imputed values match the distribution of the original dataset. Kernel Density Estimation (KDE) is used to generate synthetic samples that preserve the natural variability of the target variable. Additionally, histogram matching with quantile transformation ensures that the imputed values maintain the same statistical properties as the observed data. This prevents extreme or unrealistic values that could skew downstream analyses.

- After imputation, the previously applied transformations are reversed to restore the data to its original scale. This step ensures that variables such as glucose and cholesterol remain interpretable while retaining their improved statistical properties. The final imputed dataset is saved as "imputed_data.csv", ready for further analysis. By combining statistical transformations, feature selection, machine learning-based imputation, and distribution-based adjustments, this method provides a robust and statistically sound approach to handling missing data in complex datasets.

6. Conclusion

With the imputed dataset ready, the next steps involve validating the imputation quality and further refining the dataset for downstream analysis. The following steps will ensure the integrity and usability of the dataset:

- Perform distributional comparison between original (non-missing) and imputed values using histograms and KDE plots. Conduct statistical tests such as the Kolmogorov-Smirnov test or the Mann-Whitney U test to check if imputed values follow the same distribution as the observed data.
- Investigate whether new meaningful features can be derived from existing ones (e.g., ratios like ALT/AST for liver function). Apply Principal Component Analysis (PCA) or t-SNE to explore underlying patterns and reduce dimensionality if necessary.
- Train machine learning models on the imputed dataset to predict diabetes outcomes. Compare the performance of models with and without the imputed dataset to measure the impact of imputation. Explore advanced models like deep learning or ensemble methods for improved accuracy.
- Evaluate whether the dataset is generalizable to different populations, particularly for Vietnamese diabetes patients in this case. Consider external validation using a separate dataset or real-world clinical data.
- If the dataset is intended for real-world applications, integrate it into a dashboard or API for medical professionals. Ensure ethical and regulatory compliance when using imputed medical data in decision-making processes.

Handling missing data effectively is crucial for maintaining the reliability of medical datasets, especially in diabetes research. This approach leverages feature selection, machine learning-based imputation, and statistical adjustments to ensure that missing values are replaced in a way that preserves the dataset’s integrity. The incorporation of skewness correction, nonparametric transformations, and KDE-based sampling enhances the accuracy of imputations while preventing bias.

Furthermore, nonlinear correlation analysis using Spearman correlation, hierarchical clustering, and LOESS smoothing provides a deeper understanding of complex dependencies in the dataset. This allows for more informed feature selection and better predictive modeling.

Moving forward, validating the imputed dataset, performing predictive analysis, and testing generalizability will be essential. If successful, this dataset can be used for developing robust diabetes prediction models that aid in early diagnosis and treatment recommendations. By integrating advanced data imputation techniques, this approach ensures that real-world healthcare data remains useful even in the presence of missing values, ultimately contributing to more accurate and actionable medical insights.