Fasting Plasma Glucose (FPG) Level Analysis Using Health and Lifestyle Biomarkers

1. Data Pre-processing :

To ensure data quality and consistency, several preprocessing steps were applied:

I. Handling Missing Values:

- Blank spaces in the dataset were replaced with 'NA' values for accurate detection of missing data.
- Rows with missing values in critical health-related columns (e.g., '*SB'P*, '*DBP'*, '*ALT*', '*FPG'*, *Smoking status*, and *Drinking status*) were removed to maintain data reliability.
- Mean imputation was used for missing values in '*HDL-c(mmol/L)'* and '*LDL(mmol/L)'*, grouped by BMI categories to ensure imputations were contextually appropriate.
- **KNN Imputation:** Missing values in the following key health indicators were imputed using K-Nearest Neighbours (KNN) with 20 neighbours to utilise the patterns in the dataset:

    - *Triglyceride(mmol/L)*
    - *Cholesterol(mmol/L)*
    - *AST(U/L)*
    - *BUN(mmol/L)*
    - *CCR(umol/L)*

  If 'K' is too small, it could make imputation sensitive to noise, while large 'K' might smooth the values. Given the dataset size (~50,000 rows), testing multiple K-values was computationally expensive, 20 provided a practical choice.

II. Feature Engineering:

- **Age Binning:** The '*Age (y)*' column was categorized into six meaningful age groups (Teen, Young Adult, Adult, Middle Age, Old, Senior Citizen) to facilitate analysis.

- **BMI Binning:** The '*BMI(kg/m2)*' column was divided into Underweight, Healthy, Overweight, and Obese categories to enhance interpretability.

III. Column Removal:

- Irrelevant or redundant columns ('*id'*, '*site'*, '*height(cm)'*, *weight(kg)*, '*Diabetes diagnosed during follow-up?')* were removed to streamline the dataset.
- The column '*censor of diabetes at follow-up'* was dropped due to irrelevance to the analysis objective.

IV. Renaming Columns:

- Columns with unclear or complex labels were renamed for better readability, e.g.,

    a) Gender(1, male; 2, female) → Gender

    b) smoking status(1,current smoker;2, ever smoker;3,never smoker) → Smoker).

V. Data Type Optimization:

- Categorical variables ('*Gender'*, '*Smoker*', '*Drinker'*, '*Family History (Diabetes)'*) were converted to category type to optimize memory usage and improve performance.

- Numerical categories were mapped to meaningful labels (e.g., 1 → Male, 2 → Female for Gender).

VI. Outlier Detection:

To identify potential outliers, the Interquartile Range (IQR) method was applied to numerical features. No outliers were detected in any of the features.

These cleaning steps ensure a well-structured dataset, minimizing inconsistencies while preserving valuable insights for further analysis.

2. Data Visualization :

I. Data Distribution:

- **Gender Distribution**: The dataset has more male observations than females.
- **Smoking & Drinking Status**: The dataset is dominated by non-smokers and current drinkers. No observation is a non-drinker (either current or past drinkers).
- **Age Group Distribution**: Most individuals fall into the adult category, followed by middle-aged individuals.
- **BMI Distribution**: The dataset is primarily composed of individuals with a healthy BMI, followed by those who are overweight, with around 9% of the data consisting of obese and underweight individuals.

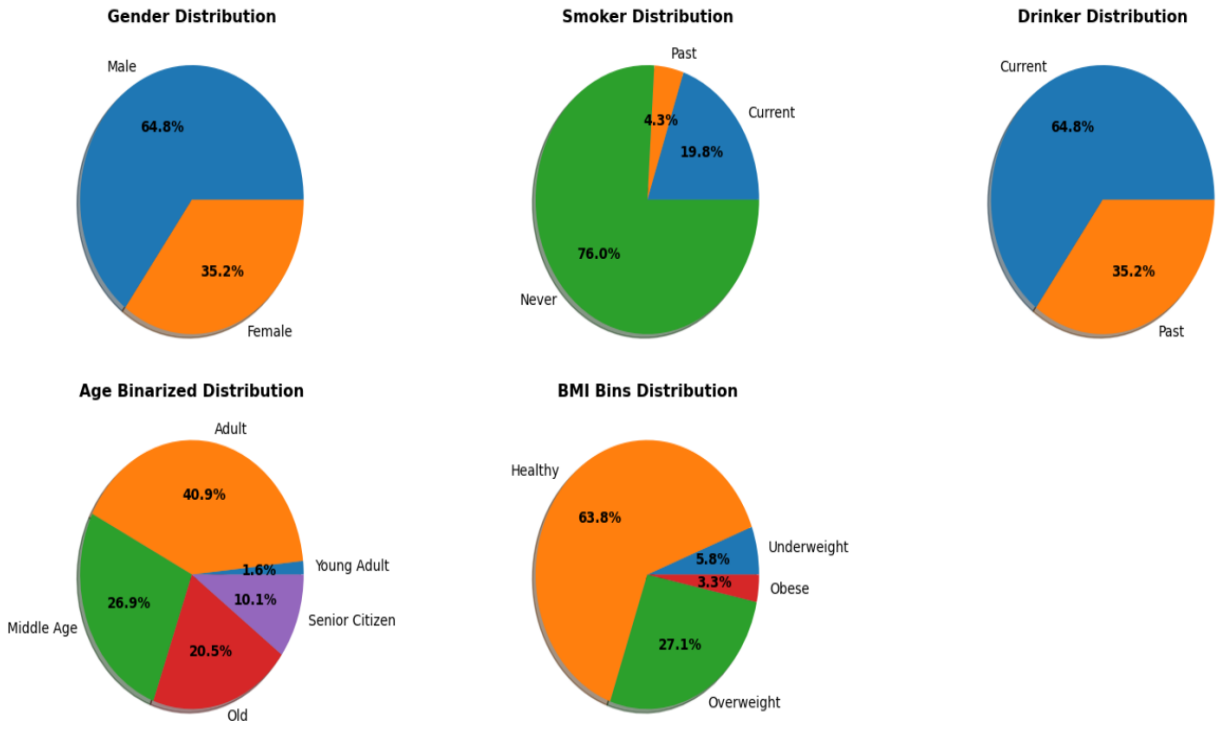


Fig 1. Distribution of Categorical Features in the dataset

II. Data Visualization:

To explore how **FPG levels** vary across different categorical variables, violin plots were generated for **Age Binarized, BMI Bins, Smoking Status, and Drinking Status**, with a distinction made between genders.

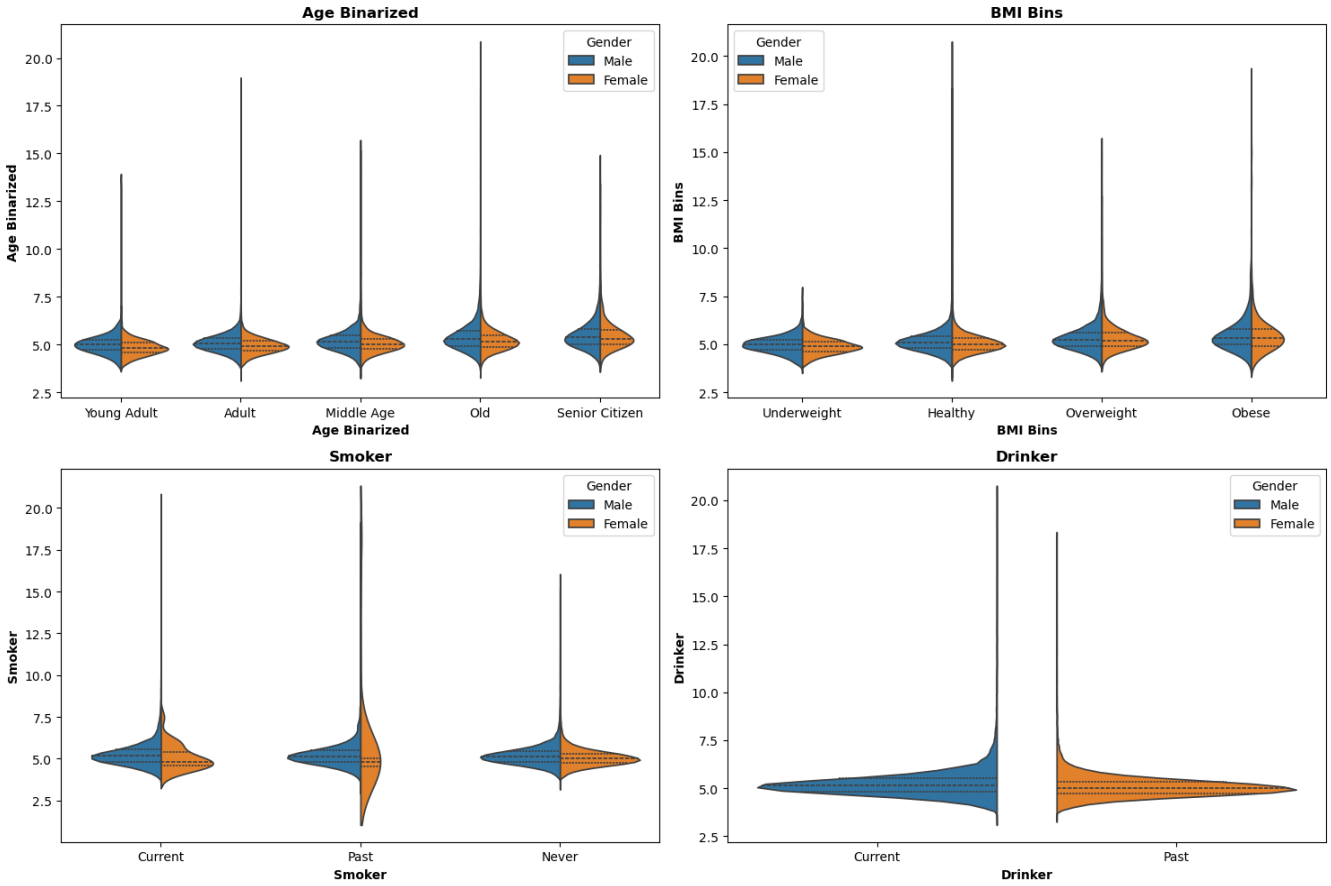


Fig 2. Comparison of Categorical Features in the dataset

Key Observations:

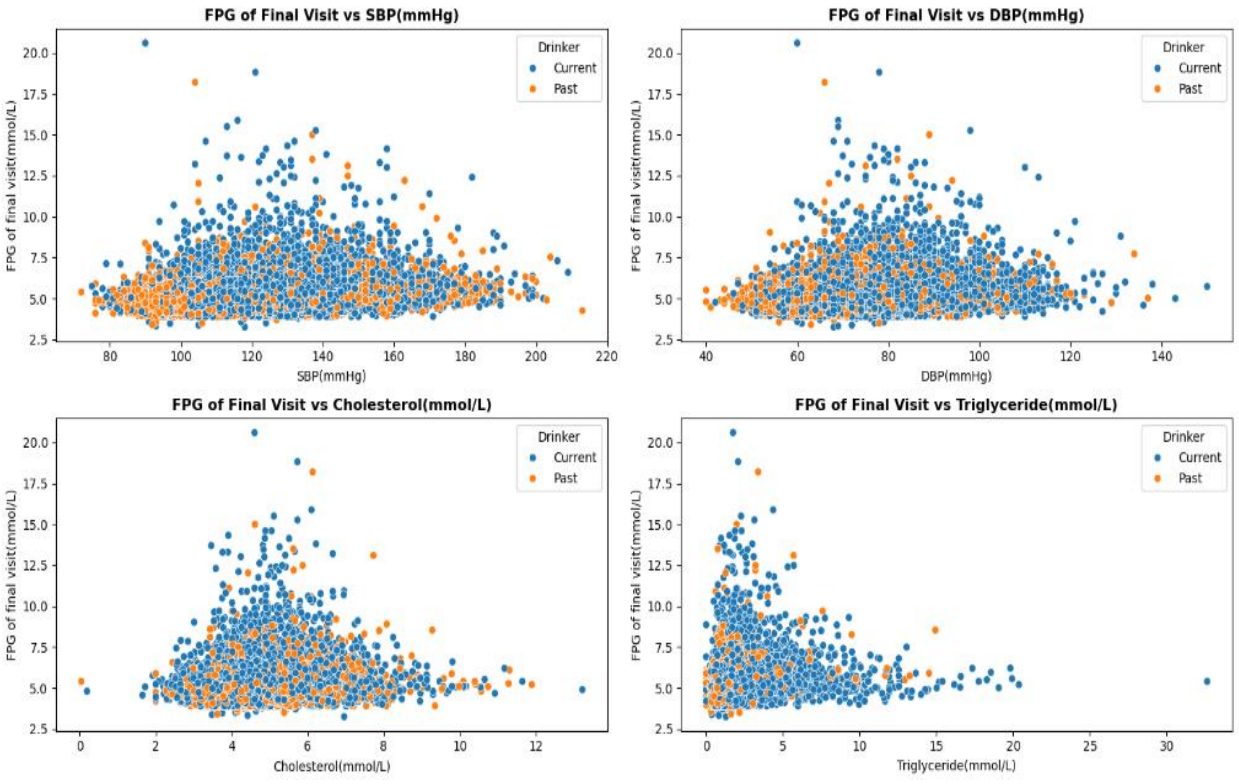
- **<u>Age vs. FPG</u>:** The median FPG in the final visit remains consistent across all age groups, with males showing a slightly higher median value than females. However, further statistical analysis is required to determine if this difference is significant. Notably, individuals aged **35-45** exhibit the highest FPG values, reaching up to **20 mmol/L**, which suggests the need for further investigation. Examining potential contributing factors such as **working hours, exercise habits, and diet** in this age group could provide insights into the observed trend.

- **<u>BMI vs. FPG</u>:** The median FPG remains nearly consistent across different BMI categories, suggesting that **BMI might not be a primary factor influencing FPG levels**. However, a **small subset of individuals with a healthy BMI** exhibits a tendency toward higher FPG values. The violin plot indicates that while most individuals have FPG levels concentrated around a lower range, a few cases extend to unusually high values with low probability. This tail extension suggests the presence of **outliers or specific subgroups** that may require further analysis.

- **<u>Smoking vs. FPG</u>:** Males who are **current or past smokers** exhibit a slightly higher median FPG compared to females in the same category. Additionally, the spread of FPG values among **female past smokers** appears wider, suggesting **greater variability in FPG levels** in this group. This variation could indicate that some females with a smoking history experience elevated or significantly reduced FPG, warranting further investigation.

Scatter Plot Analysis: FPG of Final Visit vs Clinical Biomarkers :

This section visualizes the relationship between **Fasting Plasma Glucose (FPG) at the final visit** and various clinical biomarkers, with data points color-coded based on **drinking status**. The scatter plots provide insights into potential correlations and patterns among the features.



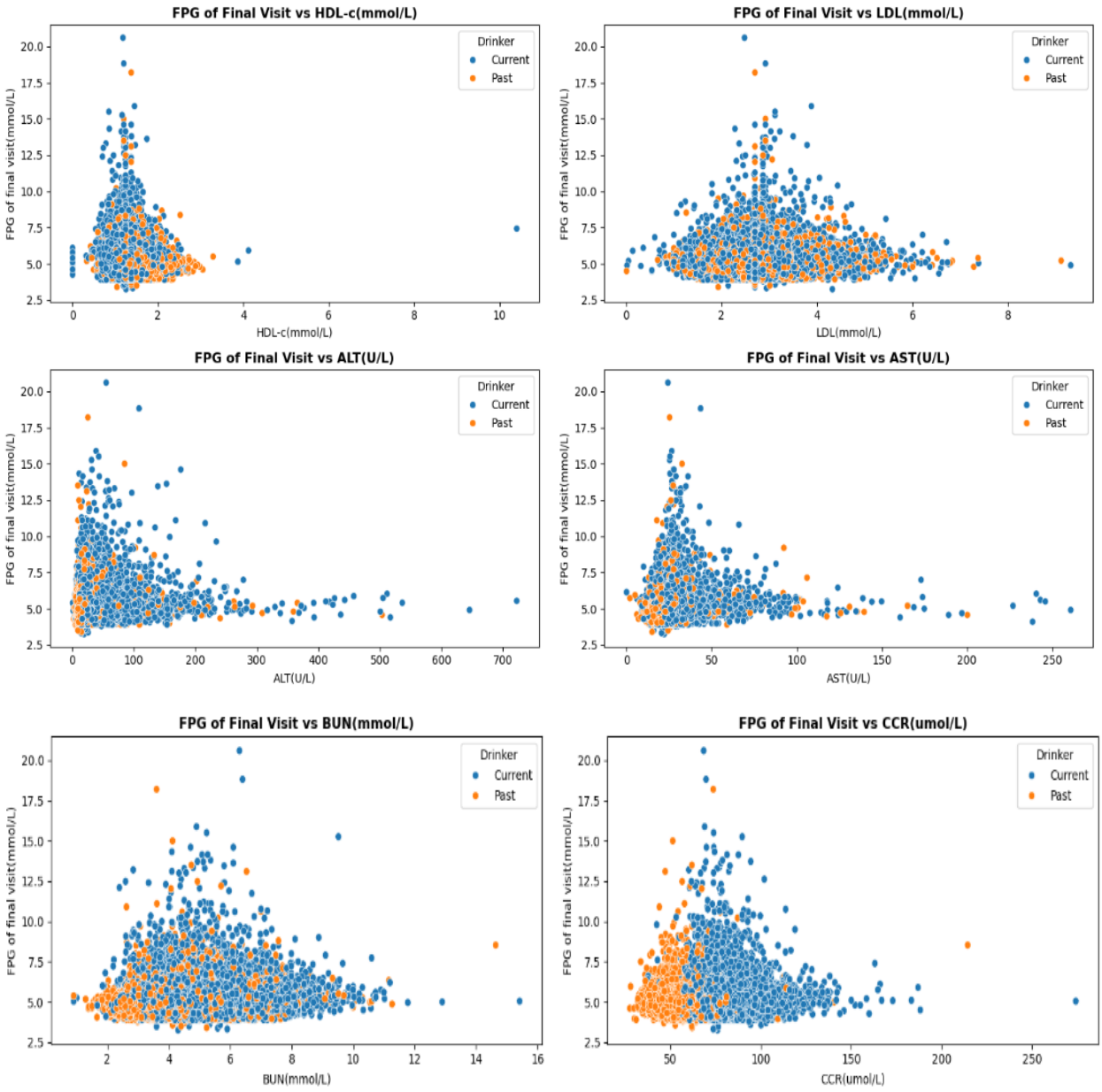
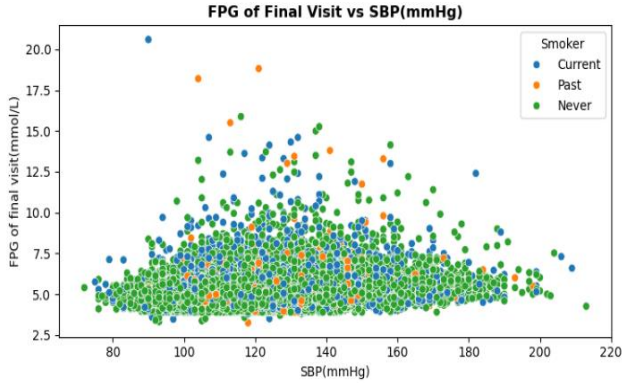
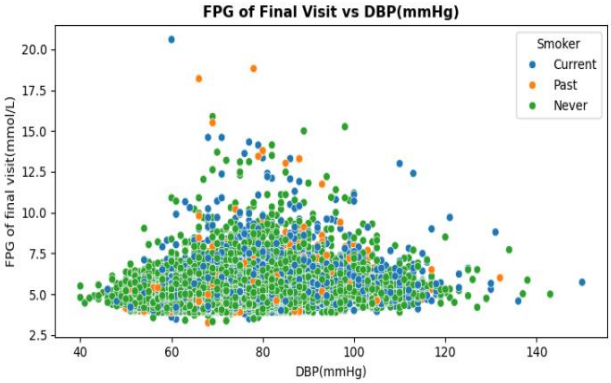
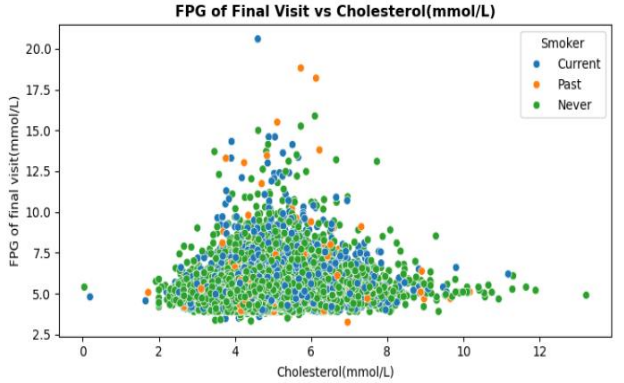


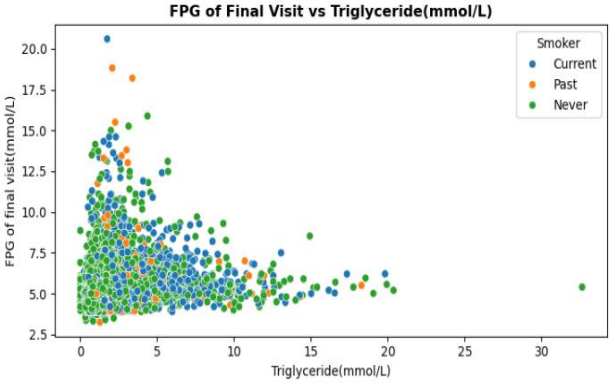
Fig 3. Scatter Plots of FPG at Final Visit vs. Various Health Indicators by Drinking Status

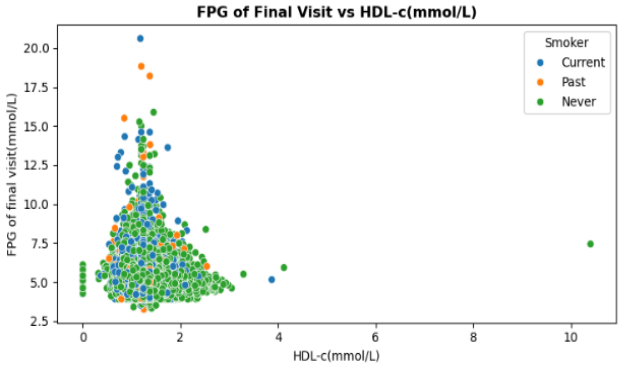
Key Observations: No significant relationship is observed between **FPG at the final visit** and the other features. However, some **current drinkers** exhibit **elevated FPG levels** across multiple features (Triglycerides, HDL, LDL, SBP, DBP, etc.). Despite this, the trend is not strong enough to establish a definitive pattern.

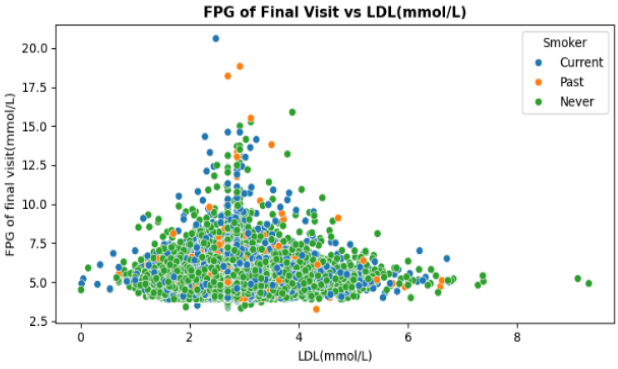

FPG of Final Visit vs SBP(mmHg)
FPG of final visit(mmol/L)
SBP(mmHg)
Smoker
Current
Past
Never

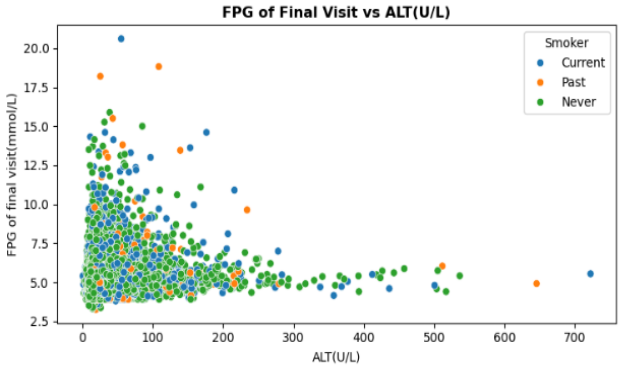

FPG of Final Visit vs DBP(mmHg)
FPG of final visit(mmol/L)
DBP(mmHg)
Smoker
Current
Past
Never

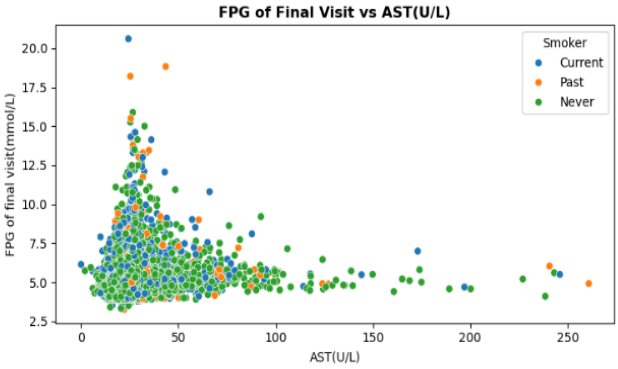

FPG of Final Visit vs Cholesterol(mmol/L)
FPG of final visit(mmol/L)
Cholesterol(mmol/L)
Smoker
Current
Past
Never


FPG of Final Visit vs Triglyceride(mmol/L)
FPG of final visit(mmol/L)
Triglyceride(mmol/L)
Smoker
Current
Past
Never


FPG of Final Visit vs HDL-c(mmol/L)
FPG of final visit(mmol/L)
HDL-c(mmol/L)
Smoker
Current
Past
Never


FPG of Final Visit vs LDL(mmol/L)
FPG of final visit(mmol/L)
LDL(mmol/L)
Smoker
Current
Past
Never


FPG of Final Visit vs ALT(U/L)
FPG of final visit(mmol/L)
ALT(U/L)
Smoker
Current
Past
Never


FPG of Final Visit vs AST(U/L)
FPG of final visit(mmol/L)
AST(U/L)
Smoker
Current
Past
Never

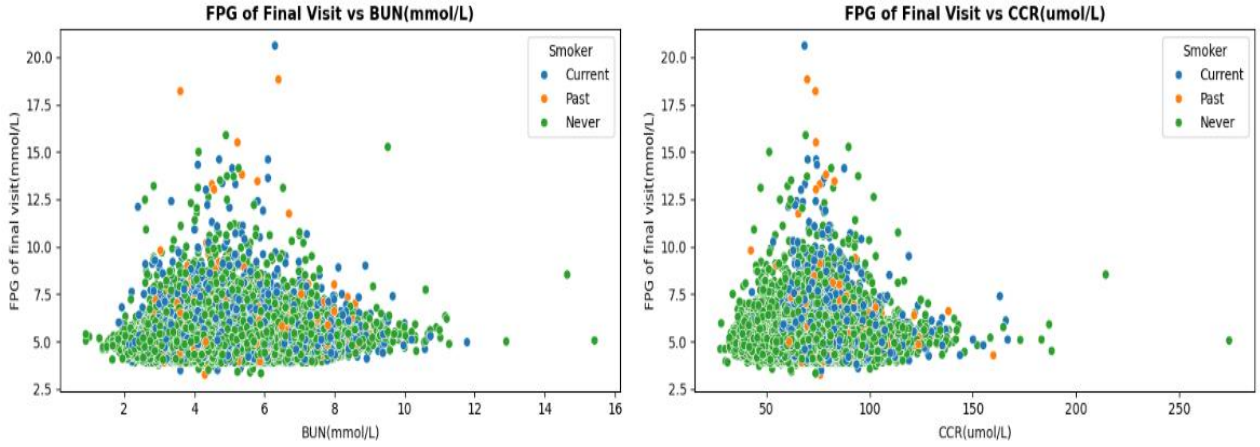
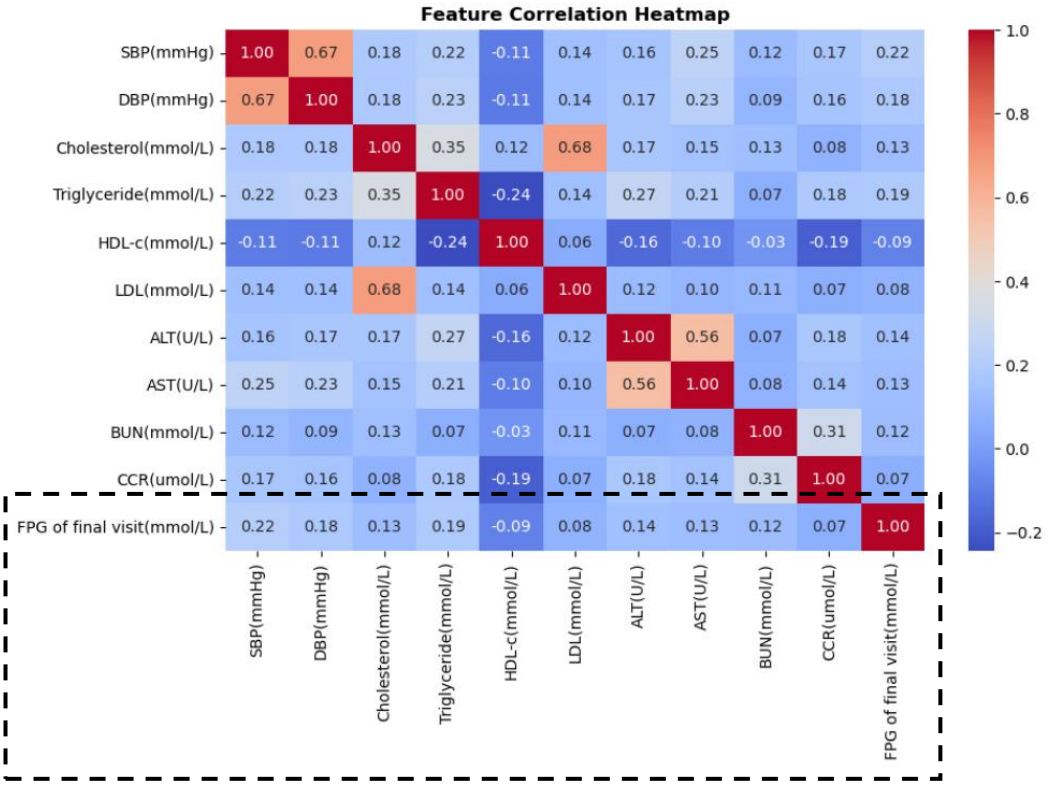


Fig 4. Scatter Plots of FPG at Final Visit vs. Various Health Indicators by Smoking Status

Key Observations: There is no significant relationship observed between **FPG at the final visit** and the other features across different drinking habits.

Exploring Relationships Between Biological Markers and FPG:

Since no clear trend was observed between the biological markers and FPG at the final visit, a correlation matrix was examined to explore potential relationships.



Observation : The correlation coefficients between FPG at the final visit and other biological markers are all low (< 0.25), indicating weak linear relationships. This suggests that no single marker strongly influences FPG in isolation, and other factors or non-linear interactions might play a role.

To capture potential hidden relationships, new features were engineered, including ratios and interaction terms. These derived features aim to explore relationships that may not be evident from individual variables alone. A correlation heatmap was computed to check their association with FPG at the final visit.

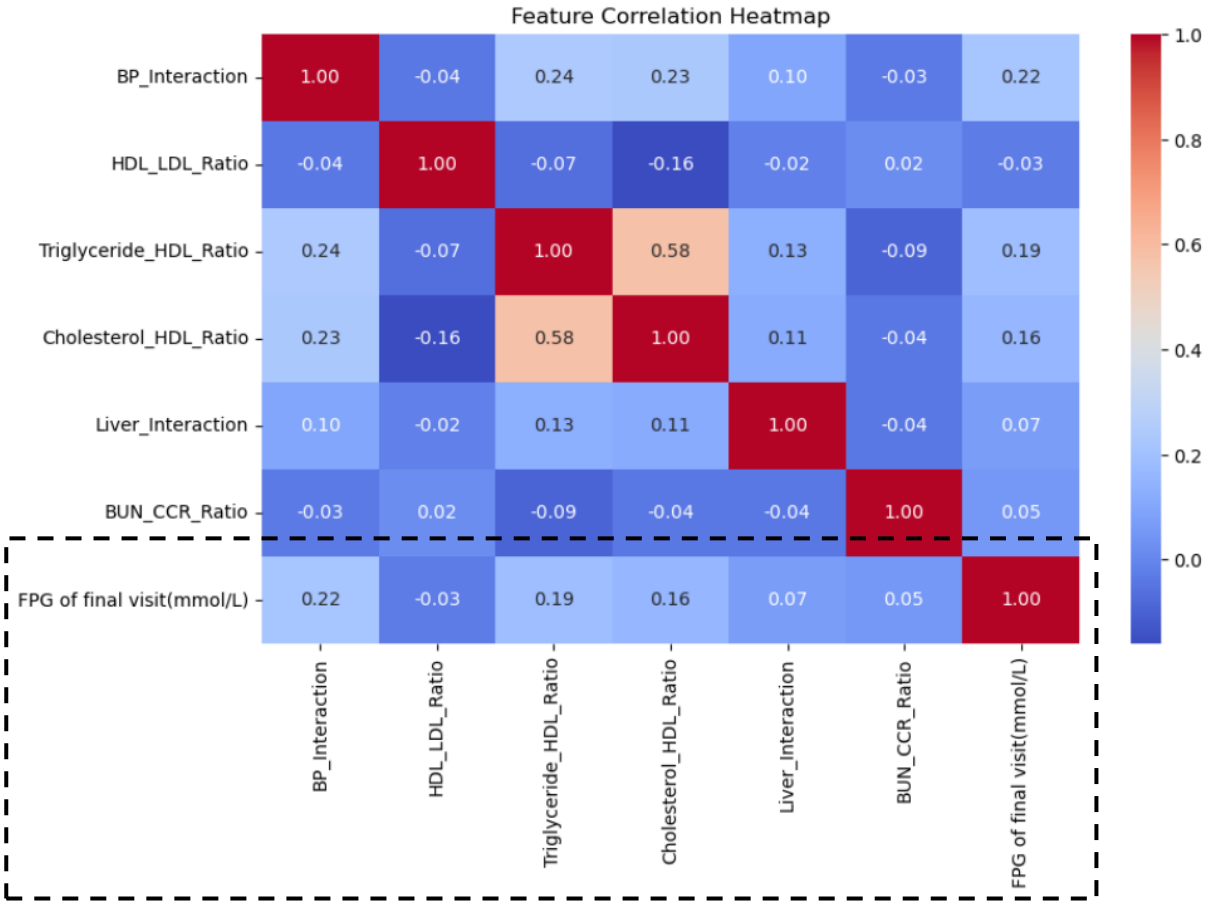


Fig 5. Heatmap of Correlation matrix between the features

Observation : The heatmap shows that the correlations remain relatively weak, indicating that more complex, non-linear modelling approaches may be required to uncover deeper patterns in the data or other factors might influence the FPG.

Conclusion

The project offered valuable insights into the limitations of using traditional biomarkers and behavioural variables to predict Fasting Plasma Glucose (FPG) levels. Despite rigorous preprocessing, visualization, and modelling efforts, weak correlations and minimal target variability restricted predictive performance. This highlighted the complexity of metabolic health patterns and emphasized the necessity of domain-specific knowledge, richer feature sets, and potentially longitudinal data for more accurate modelling in healthcare analytics.