Project Requirements Document: Diabetes Prediction Model Development

Objective

Develop a machine learning model to predict whether a patient has a chronic condition (`Chronic_flag`), leveraging a dataset of diabetes patients. The project involves data exploration, preprocessing, feature engineering, and model building, with the following considerations and requirements:

Dataset Overview

Columns Description

Column Name	Description	Notes
Unique_Identifier	Unique identifier for each patient.	Will not be used in the model; serves only for tracking.
Gender	Patient's gender (e.g., Male, Female).	Needs to be encoded into numeric format (e.g., 1 for Male, 0 for Female).
Religion	Patient's religious affiliation.	Suggested to drop, as it is unlikely to significantly impact the model's performance and doesn't have high correlation with output column.
Nationality	Patient's nationality.	Contains – for missing values; requires discussion on whether to fill values, drop rows, or exclude column specially because it doesn't have high correlation with output column. Also contain data at uppercase with lower case solve it in future by transform to one look before loading into data or better make it as choices for user not entered text data.
Avg_HBA1C Results	Average HbA1c results, indicating blood sugar control.	Contains `0` values that are incorrect and should be handled via imputation or removal. Range of values should fit in: Normal Range: 4% to 5.6% Prediabetes: 5.7% to 6.4% Diabetes: 6.5% or higher
HBA1C test Compliance	Indicates the patient's adherence to taking HbA1c tests.	Binary; no special notes at this stage.
D_Of_Birth	Patient's date of birth.	Needs transformation to calculate `Age` for inclusion in the model.
Diagnosis_Type	Type of diabetes diagnosis (e.g., Type I, Type II).	Contains only `Type II` in the dataset; suggested to drop as it provides no variability.
Chronic_flag	Target variable indicating chronic condition (1 = Yes, 0 = No).	Imbalanced dataset; oversampling (e.g., SMOTE) will be needed to address bias.

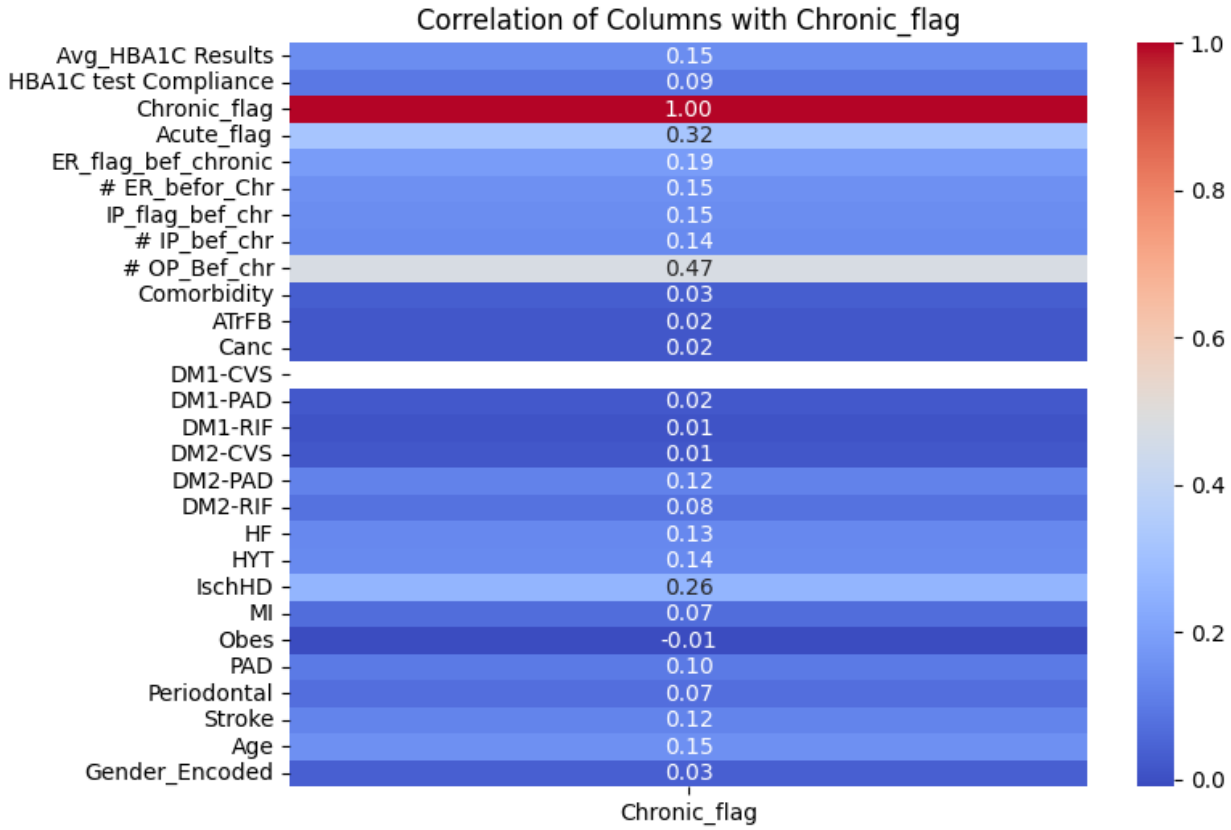
Acute_flag	Indicates acute conditions related to diabetes (1 = Yes, 0 = No).	Binary; needs scaling to match other features.
ER_flag_bef_chronic	Indicates whether the patient visited the ER before chronic diagnosis.	Binary; needs scaling.
# ER_befor_Chr	Number of ER visits before chronic diagnosis.	Requires outlier handling and potential scaling.
IP_flag_bef_chr	Indicates whether the patient had inpatient admissions before chronic diagnosis.	Binary; needs scaling.
# IP_bef_chr	Number of inpatient admissions before chronic diagnosis.	Requires outlier handling and potential scaling.
# OP_Bef_chr	Number of outpatient visits before chronic diagnosis.	Requires outlier handling and potential scaling.
Comorbidity	Other conditions present alongside diabetes.	Requires encoding if categorical.
ATrFB	Presence of atrial fibrillation.	Binary; needs scaling.
Canc	Presence of cancer.	Binary; needs scaling.
DM1-CVS / DM2-CVS	Cardiovascular conditions associated with diabetes (Type I/II).	Binary; needs scaling.
DM1-PAD / DM2-PAD	Peripheral Artery Disease associated with diabetes (Type I/II).	Binary; needs scaling.
DM1-RIF / DM2-RIF	Renal insufficiency associated with diabetes (Type I/II).	Binary; needs scaling.
HF	Presence of heart failure.	Binary; needs scaling.
HYT	Presence of hypertension.	Binary; needs scaling.
IschHD	Presence of ischemic heart disease.	Binary; needs scaling.
MI	Presence of myocardial infarction (heart attack).	Binary; needs scaling.
Obes	Presence of obesity.	Binary; needs scaling.
PAD	Presence of peripheral artery disease.	Binary; needs scaling.

Periodontal	Presence of gum disease.	Binary; needs scaling.
Stroke	History of stroke.	Binary; needs scaling.

Notes and Actionable Points

1. Correlation:

- The correlation heatmap shows no strong correlation between `Chronic_flag` and most other columns, including `Avg_HBA1C Results`, which seems unexpected. Overall, there are no highly correlated variables with `Chronic_flag`.:



2. **Missing Values**:
   - **Nationality**: Contains `-` for missing values. Discuss with the data team to decide on:
     - Obtaining correct values if available.
     - Dropping rows with missing values.
     - Excluding the column entirely.
   - **Avg_HBA1C Results**: Contains `0` values, which are invalid. Potential solutions:
     - Use ML-based imputation methods based on other patient attributes.
     - Drop rows with `0` values if correct values are unavailable.
   - **Religion**: Likely not impactful for prediction; recommend removing the column.

3. **Imbalanced Target Variable**:
    - `Chronic_flag` has a majority of `0` and a minority of `1`. Address imbalance using oversampling methods like **SMOTE** to avoid bias.
4. **Feature Engineering**:
    - **Age**: Calculate from `D_Of_Birth` and include it as a feature.
    - **Binary Columns**: Normalize to a common range (e.g., 0–1) to ensure consistency.
5. **Redundant Features**:
    - **Diagnosis_Type**: Contains only one value (`Type II`) across all rows. Exclude from modeling as it offers no predictive power.
6. **Data Encoding**:
    - Categorical columns (`Gender`, `Comorbidity`, etc.) must be encoded into numeric representations.
7. **Scaling**:
    - Ensure all numerical data is scaled (e.g., using **StandardScaler**) to align feature magnitudes.
8. **Outlier Detection**:
    - Columns like `# ER_befor_Chr`, `# IP_bef_chr`, and `# OP_Bef_chr` may contain outliers. Apply robust scaling or capping methods.