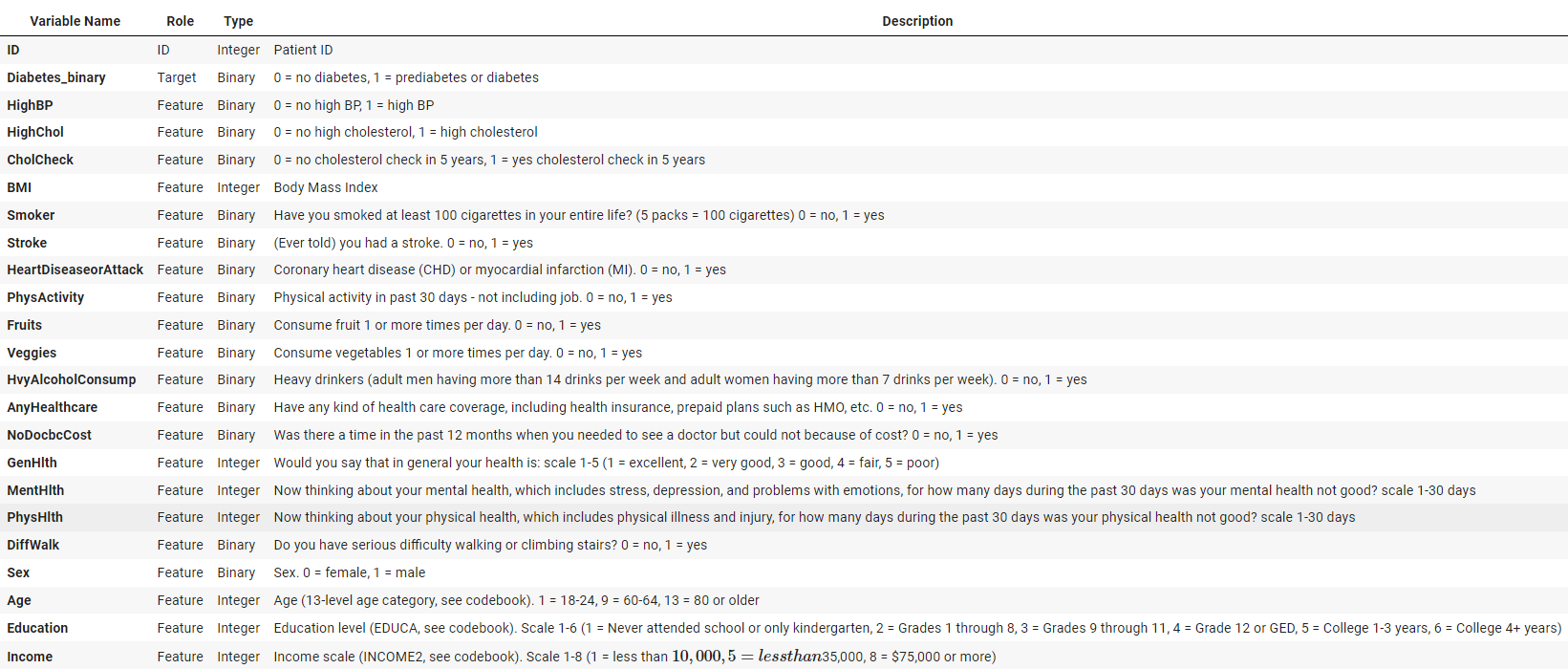
CDC Diabetes Health Indicators Project Report

# Introduction and Data Understanding

The goal of this project is to predict diabetes occurrence using a dataset of health indicators provided by the CDC. The main objective is to identify key factors that influence diabetes prevalence, thereby developing a predictive model to aid in early detection and prevention. By leveraging machine learning techniques, we aim to create a robust model that can accurately predict diabetes risk and provide actionable insights for public health interventions.

The dataset comprises various health indicators, including Body Mass Index (BMI), physical activity levels, cholesterol levels, and other demographic and lifestyle factors. The target variable in this dataset is diabetes\_binary, a binary indicator signifying whether an individual has diabetes (1) or not (0).

To understand the dataset comprehensively, we conducted thorough data exploration and cleaning. This involved imputation methods, engineering new features like BMI categories, and normalizing continuous variables to ensure uniformity in model training. We also performed bivariate analysis to explore relationships between the target variable and other features, identifying significant correlations that guided our feature selection process. This foundational work set the stage for building effective predictive models and uncovering crucial health indicators related to diabetes.



###### **Imputation Techniques:**

* *Continuous Variables:* Applied mean/median imputation using the SimpleImputer from sklearn.impute for variables like BMI and Age.
* *Categorical Variables:* Used mode imputation for categorical features like Sex and Education to replace missing values with the most frequent category.

###### **Feature Engineering:**

* *New Feature Creation:*
  + BMI Categories**:** Transformed the continuous BMI values into categorical bins (e.g., Underweight, Normal weight, Overweight, Obese) using pd.cut() to create more meaningful insights.
  + Age Binning: Grouped ages into predefined bins (e.g., 0-18, 19-35, 36-50, 51-65, 66+) to reduce variability and improve model performance by creating a new categorical variable Age\_Binned.
* *Bivariate Analysis:*
  + Techniques Used:
    - Cross-tabulation: Used pd.crosstab() to explore the relationship between diabetes\_binary and categorical features such as HighBP, HighChol, and PhysActivity.
    - Correlation Matrix: Computed correlation coefficients using df.corr() to understand linear relationships between diabetes\_binary and continuous features like BMI, Age, and Income.
    - Visualizations: Utilized bar plots, point plots, and violin plots with seaborn and matplotlib to visually represent associations between the target variable and other features.

###### **Data Balancing:**

* *Techniques Applied:*
  + **SMOTE (Synthetic Minority Over-sampling Technique):** Used imblearn.over\_sampling.SMOTE to synthetically generate new samples for the minority class (diabetes\_binary = 1) by interpolating between existing samples.
  + **NearMiss:** Applied imblearn.under\_sampling.NearMiss to reduce the majority class (diabetes\_binary = 0) by selecting the closest majority samples to the minority samples.
  + **SMOTETomek:** Combined oversampling and undersampling techniques using imblearn.combine.SMOTETomek to achieve a balanced dataset by removing Tomek links (pairs of samples from opposite classes that are closest to each other).
  + **Result Comparison:** Compared the performance of simple SMOTE and SMOTETomek using cross-validation to determine that SMOTE provided better results during modeling.

###### **Normalization:**

* **Technique:** Standardized continuous features using StandardScaler from sklearn.preprocessing to scale features to have zero mean and unit variance, ensuring that all features contribute equally to the model.

# Exploratory Data Analysis (EDA)

Found strong correlations between diabetes and factors like BMI, physical inactivity, and cholesterol levels using df.corr() and heatmaps.

**Key Observations:**

1. **BMI and General Health:**
   * **Correlation:** High BMI is significantly associated with poor general health. Individuals with higher BMI scores tend to report worse overall health conditions.
   * **Age Factor:** BMI increases with age, indicating that older populations are more prone to obesity-related health issues, which in turn raises the risk of diabetes.
2. **Income and Age:**
   * **Socioeconomic Influence:** Older individuals generally have lower income levels. This suggests that socioeconomic factors may influence health outcomes, including diabetes prevalence, as lower income can restrict access to healthcare and healthy lifestyle choices.
3. **Physical Inactivity:**
   * **Impact:** Physical inactivity is a strong predictor of diabetes. Individuals who do not engage in regular physical activity have higher rates of diabetes, highlighting the importance of promoting active lifestyles for diabetes prevention.
4. **High Blood Pressure and Cholesterol:**
   * **Health Indicators:** High blood pressure and high cholesterol levels are prevalent among diabetic individuals. These conditions often coexist with diabetes, indicating the need for integrated healthcare approaches to manage these interrelated health issues.
5. **Smoking Habits:**
   * **Risk Factor:** Smoking is more common among those with diabetes, suggesting that smoking cessation programs could be beneficial in diabetes management and prevention.
6. **Dietary Habits:**
   * **Nutrition:** Although not explicitly detailed in the dataset, it is inferred that poor dietary habits contribute to higher BMI and diabetes risk. Further research into dietary patterns could provide more insights.
7. **Mental Health:**
   * **Correlation:** There is a slight association between poor general health and mental health issues. Mental health conditions like stress and depression may exacerbate physical health problems, including diabetes.
8. **Healthcare Access:**
   * **Access and Outcomes:** Limited access to healthcare services is more common among lower-income and older individuals, potentially contributing to higher diabetes rates in these groups. Ensuring equitable access to healthcare can improve diabetes management and outcomes.

# Model Training and Evaluation

We trained several models, including Logistic Regression, Decision Trees, Random Forests, and Gradient Boosting. Each model was trained using the selected features, and predictions were made on the test data. Our primary evaluation metrics included:

* **AUC Score**: This metric measures the area under the ROC curve, indicating the model's ability to distinguish between classes.
* **Accuracy**: The proportion of correctly classified instances out of the total instances.
* **Precision**: The ratio of true positive predictions to the total predicted positives, reflecting the model's accuracy in identifying positive instances.
* **Recall**: The ratio of true positive predictions to all actual positives, indicating the model's ability to capture all positive instances.
* **F-measure**: The harmonic mean of precision and recall, providing a balanced measure of the model's performance.

### Best Performing Model

After comparing the performance of all models, we found that the **Random Forest** model emerged as the best performer, particularly in terms of the AUC score. The Random Forest model demonstrated superior ability to distinguish between positive and negative instances, achieving the highest AUC score among all models. This indicates that Random Forest had the best overall performance in identifying diabetes health indicators.

### Hyperparameter Tuning

To further enhance the performance of the Random Forest model, we conducted hyperparameter tuning using grid search. This involved systematically varying key parameters and evaluating the model's performance to identify the optimal settings. The tuning process significantly improved the model's accuracy, precision, recall, and F-measure, confirming that the Random Forest model was not only the best but also well-calibrated for our dataset.

## Model Deployment

We successfully deployed a user-friendly Diabetes Prediction and Classification application using Streamlit. This deployment involved loading a pre-trained Random Forest classifier model and creating an intuitive web interface for users to input their health data and receive real-time predictions.

#### **Key Steps in Deployment**

1. **Model Preparation**:
   * We started by ensuring that the pre-trained Random Forest classifier model was correctly placed and accessible.
   * The model was saved as a pickle file, which allowed for easy loading and integration into the Streamlit application.
2. **User Interface Design**:
   * Using Streamlit, we designed an interface to collect essential health indicators, including demographic details, lifestyle habits, and medical history.
   * The user interface was built with various interactive widgets, making it easy for users to input their data.
   * The design was guided by results from dimensionality reduction, ensuring that the most significant features were captured effectively.
3. **Input Data Processing**:
   * User inputs were carefully converted into binary and numerical values that the model could understand.
   * Additional health metrics, such as BMI and Health Score, were calculated based on the input data.
   * This processing step ensured that the input data matched the features used to train the Random Forest model.
4. **Prediction Generation**:
   * The processed input data was fed into the pre-trained model via the pickle file.
   * The model predicted the probabilities of the user being Diabetic, Pre-Diabetic, or Non-Diabetic.
   * These probabilities were presented in percentages, providing a clear and concise assessment of the user’s health status.
5. **Result Display**:
   * The display was enhanced to effectively communicate the user's health status.
   * Detailed explanations of BMI and Health Score were provided alongside the prediction probabilities.
   * This information helped users understand their health metrics and the implications of their predicted diabetes risk.
6. **Deployment**:
   * Streamlit’s intuitive framework was utilized to deploy the application locally.
   * The deployment process was streamlined, resulting in an accessible and user-friendly platform for predicting diabetes risk.
   * The application provided real-time predictions, allowing users to assess their health status and take preventive measures effectively.

By leveraging Streamlit, we developed a robust and accessible web application for diabetes risk prediction. The deployment process was carefully planned and executed, ensuring that the application was both user-friendly and accurate. The Random Forest model’s integration, combined with a well-designed interface, allowed for real-time health assessments.

## Limitations

1. **Data Limitations**:
   * **Quality and Completeness**: The dataset used may have contained missing or inaccurate data, potentially affecting the model's accuracy and reliability.
   * **Generalizability**: The data was limited to a specific population, which might not be representative of the broader population. This limits the model’s applicability to other demographic groups or regions.
2. **Model Limitations**:
   * **Model Complexity**: While Random Forest is robust, it is also complex, making it difficult to interpret the importance of individual features.
3. **Technical Limitations**:
   * **Local Deployment**: The application was deployed locally, limiting its accessibility to a wider audience. Users need to have the necessary environment and dependencies installed.
4. **User Experience**:
   * **Input Data Collection**: Users need to manually input their health data, which can be time-consuming and prone to errors.
   * **Interpretation of Results**: While efforts were made to explain the results, some users might still find it challenging to understand the implications of their health metrics and predicted risk levels.

## Conclusions

1. **Model Performance**:
   * The Random Forest model demonstrated strong performance in predicting diabetes risk, with high accuracy and reliability. This underscores its potential as a valuable tool for early diabetes detection.
2. **User Engagement**:
   * The deployment of the application using Streamlit provided an interactive and user-friendly platform. Users could easily input their health data and receive real-time predictions, enhancing engagement and usability.
3. **Health Insights**:
   * The application not only predicted diabetes risk but also provided users with valuable health insights such as BMI and Health Score, aiding in a comprehensive understanding of their health status.

## Key Recommendations

* **Improve Lifestyle Factors:** Focus on reducing physical inactivity and managing BMI.
* **Community Engagement:**
  + **Education:** Educate communities on the importance of lifestyle factors in diabetes prevention.
  + **Collaboration:** Collaborate with healthcare providers to implement preventive measures and promote healthy living.
* **Further Research:**
  + **Additional Indicators:** Investigate the impact of dietary habits, genetic factors, sleep patterns, mental health, socioeconomic status, environmental factors, and physical measurements on diabetes prediction.

## Next Steps

1. **Improving Model Interpretability and Performance**:
   * **Regular Model Updates**: Periodically retrain the model with new data to maintain its accuracy and relevance. This ensures that the model adapts to any changes in population health trends or new medical insights.
2. **Technical Enhancements for Broader Access**:
   * **Cloud-Based Deployment**: Deploy the application on a cloud platform to enhance accessibility and scalability. This will allow users to access the tool from anywhere and handle a larger number of concurrent users.
   * **Integration with Healthcare Systems**: Integrate the application with electronic health records (EHRs) and other healthcare systems. This will streamline data input and improve the accuracy of predictions by leveraging comprehensive health data.
3. **User Experience Improvements**:
   * **Automated Data Entry**: Reduce the manual data entry burden by integrating with wearable devices and health apps that can automatically input relevant health data.
   * **Educational Resources**: Provide users with educational resources and guidelines on how to interpret their results and take preventive measures. This will empower users to make informed health decisions.
   * **Feedback and Iteration**: Implement a feedback mechanism within the application to gather user inputs and continuously improve the tool based on user experiences and suggestions.