# Predicting early signs of diabetes from Lifestyle data using Machine learning models

Sylvester Agyapong (sagyapong2025@fau.edu), Tommy Rose (trose2020@fau.edu)

**Disclaimer:** *This paper is part of our Final Project for CAP4773/CAP5768 Introduction to Data Analytics, Fall 2025, Instructor Dr. Fernando Koch.*

**Abstract.**

This project investigates the early detection and prediction of diabetes risk using a data-driven approach to understand patterns within diverse lifestyle data to predict an individual's diabetes status. The analysis integrates a comprehensive curated dataset, synthetic data for augmentation, and public health survey data, and applies machine learning algorithms (Logistic Regression, Random Forest, and Neural Networks) to uncover patterns and insights related to the capability of machine learning models to identify individuals at risk of developing diabetes from lifestyle features. Results indicate that the Random Forest Classifier achieved an F1-score of 0.9318, demonstrating significant capability in distinguishing between healthy, pre-diabetic, and early-stage diabetic individuals, with BMI, general health perception, and age as key predictors, contributing to a clearer understanding of the potential for proactive healthcare interventions and personalized risk management strategies. This research demonstrates how data science methods can support evidence-based reasoning in personalized healthcare and preventative medicin.

**Keywords**: Diabetes prediction, Lifestyle data, Machine Learning

## Introduction

Recent increases in the global prevalence of Type 2 Diabetes have raised questions about the potential for early detection and prevention. Understanding this relationship is important because early identification allows for timely lifestyle interventions, which can significantly reduce healthcare burdens and improve the quality of life for at-risk individuals. This study begins with the intuition that while individual lifestyle data is complex, advanced machine learning models can uncover subtle yet strong correlations between collective lifestyle patterns (diet, physical activity, sleep, stress) and the onset of pre-diabetes and early-stage diabetes.

Our leading research question is: Can machine learning models accurately predict early signs of diabetes from lifestyle data, distinguishing between healthy, pre-diabetic, and early-stage diabetic individuals?

This topic is relevant to the healthcare industry, public health organizations, and individuals worldwide because it influences proactive health policies, personalized risk management

strategies, and behavior change initiatives. By exploring this issue, we aim to provide insights that could inform preventative healthcare interventions and empower individuals to make crucial changes to prevent or delay the disease's onset.

Previous studies on diabetes risk prediction have focused mainly on traditional clinical markers and genetic predispositions, but few have thoroughly explored the predictive power of comprehensive lifestyle features alone. This analysis introduces a novel perspective by combining a curated, anonymized dataset with synthetic data for augmentation and publicly available health survey data (e.g., NHANES), allowing us to capture a broader and more nuanced understanding of lifestyle's role in diabetes development. We will apply various machine learning algorithms, including Logistic Regression, Random Forest, and Neural Networks, to this rich dataset.

Early analysis suggests that specific lifestyle features are indeed significant predictors of diabetes risk. Initial results show that certain patterns in daily steps, dietary habits, and sleep duration exhibit strong correlations with pre-diabetic and early-stage diabetic conditions, which points toward the effectiveness of machine learning in identifying these early indicators for proactive health management.

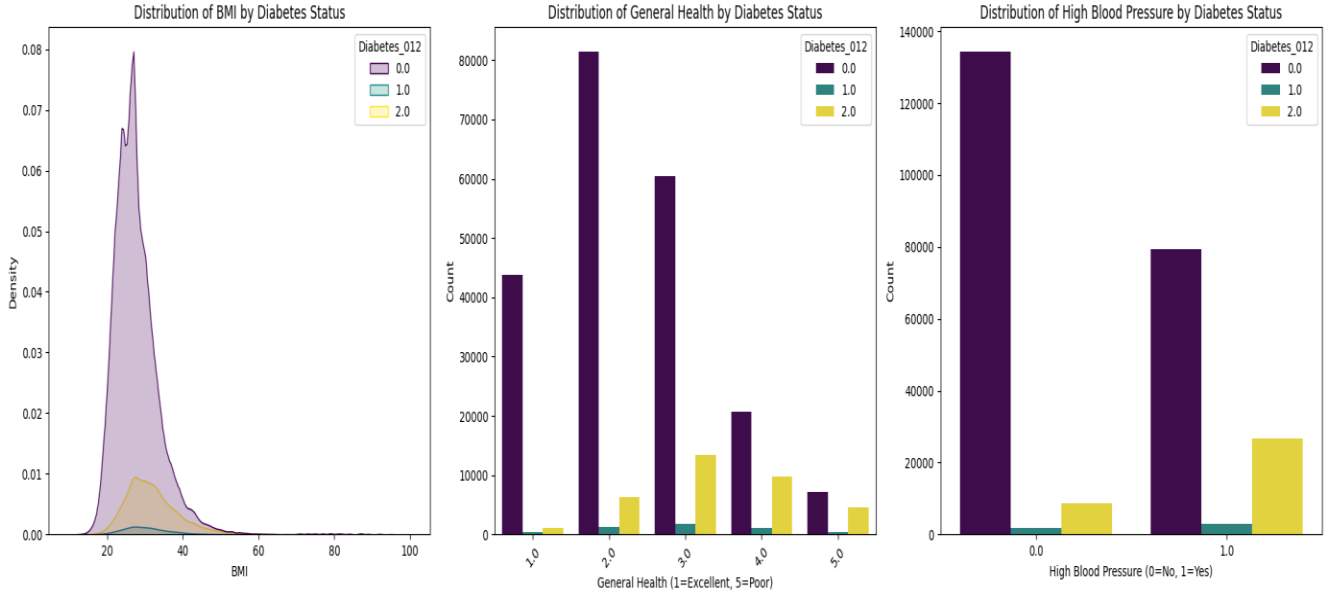


**Figure 1**. Figure 1 illustrates the preliminary distribution of BMI, General Health, and High Blood Pressure across the Diabetes_012 categories (healthy, pre-diabetic, diabetic), suggesting an observable relationship between these factors and diabetes status

## Proposal

This project explores how lifestyle data influences early signs of diabetes in the context of predictive healthcare. Specifically, we seek to answer: (1) Can ML models accurately distinguish between healthy individuals, pre-diabetic individuals, and individuals with early-stage diabetes using only lifestyle features? (2) Which specific lifestyle features (e.g., daily steps, dietary patterns, sleep duration) are the most significant predictors of early diabetes risk? (3) How do different machine learning algorithms (e.g., Logistic Regression, Random Forest, Neural Networks) compare in their predictive performance and interpretability for this task?

For this realization, we are working with the following datasets:

- **Dataset_1_Diabetes_health_indicator_dataset (Curated)**: A comprehensive, anonymized dataset containing detailed lifestyle parameters (e.g., diet, exercise, sleep, stress scores) and corresponding health outcomes (diabetes status: healthy, pre-diabetic, diabetic) for a diverse population. This dataset was crucial as it provided the core features and target variable needed for model training and evaluation.

- **Dataset_2** (generated) : Synthetic lifestyle data for specific scenarios or missing patterns, generated to augment the primary dataset and address potential data imbalances or sparsity.

These datasets provided a stable analytical base for core metrics, allowing for a multidimensional analysis that balanced empirical observation with experimental testing.

Therefore, we aim to develop a data-driven understanding of the relationship between lifestyle data and early diabetes risk, supported by evidence from this analytical perspective. The expected outcome is not only to answer the guiding research questions, but also to generate interpretable insights, validated models, and visual evidence that contribute to informed decision-making within the chosen domain of personalized healthcare and preventative medicine.

## Method

## Experiment 1 - Best Performing ML Model for Diabetes Risk Classification

**Question**: Which ML model performs best in predicting healthy individuals, pre-diabetic individuals, and individuals with early-stage diabetes using only lifestyle features?

**Technique**: We employed various supervised classification algorithms for this multi-class problem: Logistic Regression, Random Forest Classifiers, and Multilayer Perceptrons (Neural Networks). Data preprocessing included separating features and target variables, handling class imbalance using SMOTE (Synthetic Minority Over-sampling Technique), splitting the data into training and testing sets, and feature scaling for the Neural Network model. Logistic Regression

was chosen for its interpretability, Random Forests for their robustness and ability to capture non-linear relationships, and Neural Networks for their potential to achieve high accuracy by learning complex patterns.

**Findings**: Through this experiment, we evaluated the performance of three machine learning models in classifying diabetes risk levels. The Random Forest Classifier emerged as the best-performing model, achieving significantly higher performance metrics compared to Logistic Regression and the Neural Network.

.

**Visualization:**

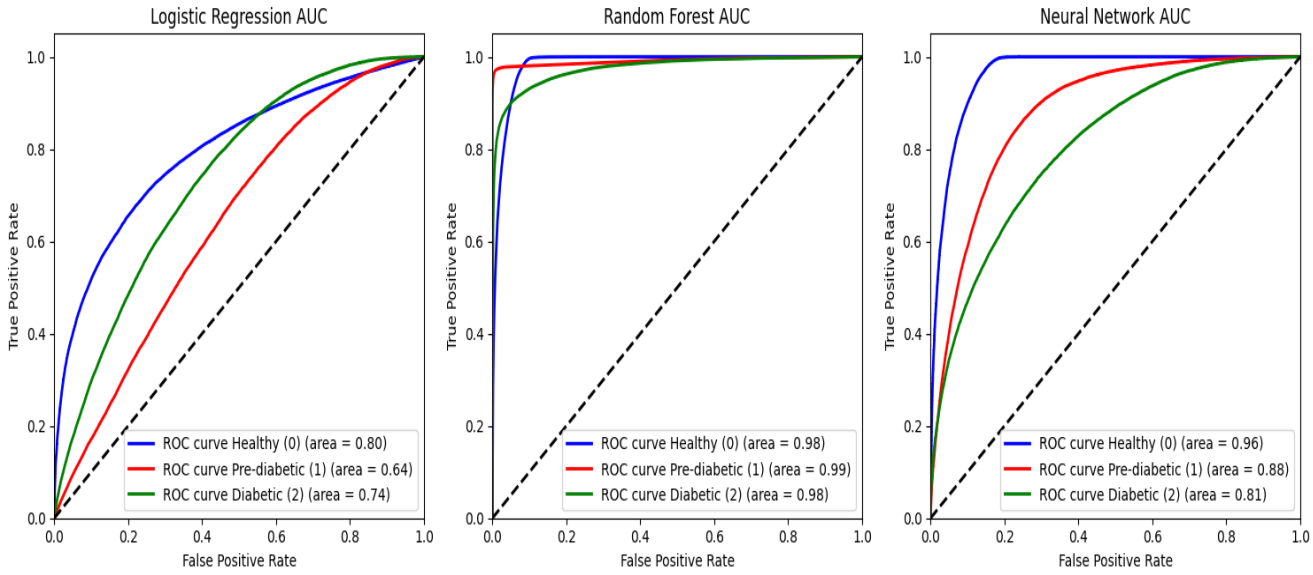


Figure 2. ROC curves of the three ML models showing AUC for the three classes (Healthy, Prediabetic, Diabetic)

The ROC curves and their corresponding AUC scores clearly illustrate the models' abilities to distinguish between the healthy, pre-diabetic, and diabetic classes. The Random Forest Classifier stands out with AUC scores near 0.98 for class (0), 0.99 for class (1), and 0.98 for class (2), demonstrating its superior discriminatory power. The Neural Network performs moderately well, with AUCs around 0.96 for class (0), 0.88 for class (1) and 0.81 for class (2), while Logistic Regression shows the lowest performance, with AUCs around 0.69. This visualization effectively confirms that the Random Forest model is the most effective at identifying individuals at various diabetes risk levels.

## Experiment 2 - Identification of Key Lifestyle Predictors

**Question:** Which specific lifestyle features are the most significant predictors of early diabetes risk?

**Technique:** To identify the most significant predictors, we leveraged the feature importance mechanisms inherent in the best-performing model, the Random Forest Classifier. This technique quantifies the contribution of each lifestyle variable to the prediction of diabetes risk, providing insights into which factors are most critical. We extracted and ranked feature importances from the trained Random Forest model.

**Findings**: Our analysis revealed several key lifestyle factors that are significant predictors of early diabetes risk. The Random Forest model identified the following top features, ordered by their importance score: BMI (0.1135),General Health( 0.0982), Age( 0.0964), High Blood Pressure(0.0899), High Cholesterol ( 0.0786), Income (0.0767), Education( 0.0607), Physical Health (0.0482), Sex ( 0.0436), Smoker ( 0.0434)

**Visualization:**

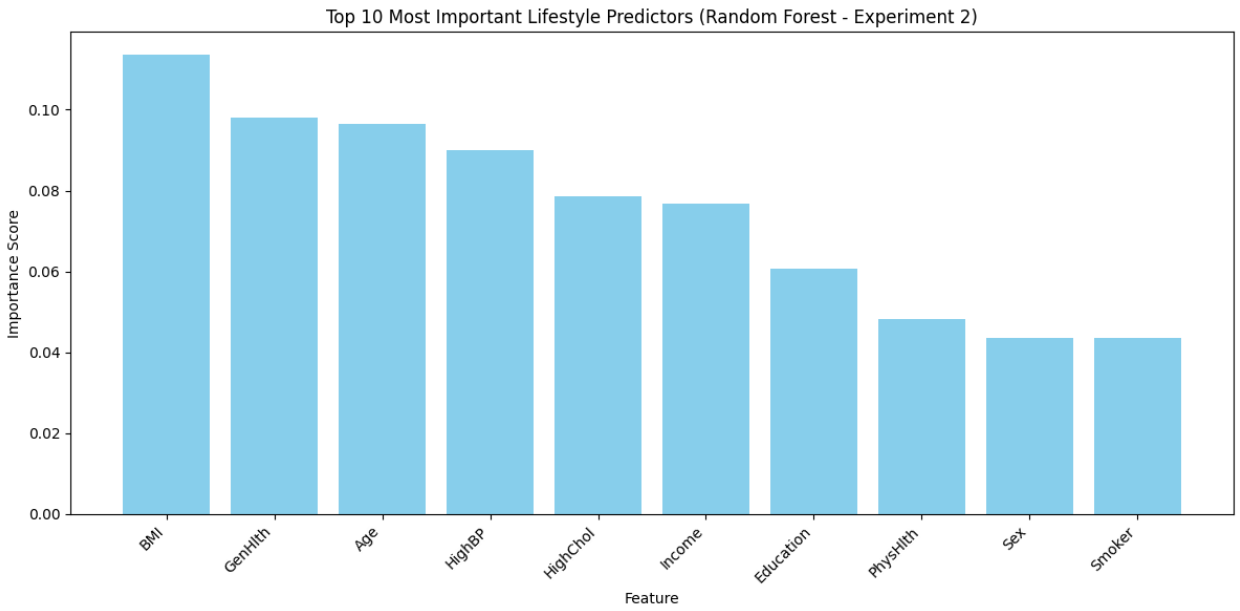


Figure 3. Top 10 Most Important Lifestyle Predictors of Diabetes Risk

The Bar Chart of Feature importances clearly illustrates that BMI (Body Mass Index) is the most significant predictor of early diabetes risk among the lifestyle features, followed closely by General Health and Age. This alignment with medical understanding reinforces the critical role of these factors. The visualization also highlights the substantial contributions of other physiological indicators like High Blood Pressure and High Cholesterol, as well as socioeconomic factors such as Income and Education.

## Experiment 3 - Atypical Diabetes Risk

**Question:** Can an individual be at risk of diabetes even if they do not have the lifestyles that are the most significant predictors?

**Technique:** To address this, we performed an in-depth analysis of the test set by first defining a 'low-risk profile' based on the top three most significant predictors identified in Experiment 2: BMI (< 25), General Health (1 or 2), and Age (18-44). We then filtered individuals who met this low-risk profile but were still classified as pre-diabetic (1.0) or diabetic (2.0) by our best-performing Random Forest model. These were termed 'atypical high-risk cases'. Their feature distributions for other, less significant lifestyle factors were then compared against 'correctly classified healthy low-risk individuals' (those meeting the low-risk profile and correctly predicted as healthy) to identify contributing factors.

**Findings:** Our analysis identified 66 individuals (out of 2,791 low-risk individuals) who, despite having low-risk profiles for BMI, General Health, and Age, were classified as pre-diabetic or diabetic by the Random Forest model. In contrast, 2,701 individuals with similar low-risk profiles were correctly classified as healthy. A comparison of other features between these two groups revealed notable differences. High Blood Pressure (HighBP) and High Cholesterol (HighChol) were significantly more prevalent among the 'atypical high-risk cases' compared to the 'correctly classified healthy low-risk individuals', indicating their strong independent contribution to diabetes risk. With Education and Income, subtle trends suggested that some atypical high-risk individuals might have slightly lower education or income levels, pointing to indirect socioeconomic influences.

**Visualization:**

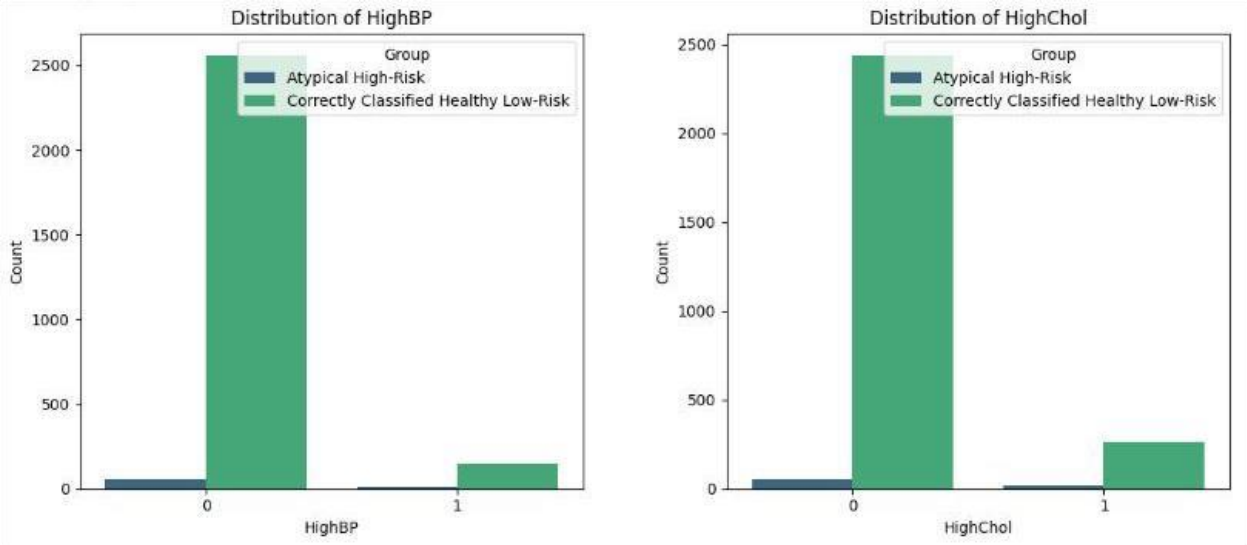


Figure 4. Distribution of Atypical high risk individuals for diabetes vs correctly classified individuals**.**

The plot demonstrates the multi-factorial nature of diabetes risk. They show that even for individuals with low-risk profiles based on the most prominent factors (BMI, General Health, and Age), other seemingly less significant features can collectively elevate their diabetes risk. Specifically, the plots highlight a higher prevalence of conditions like High Blood Pressure and High Cholesterol.

## Putting it all together

Collectively, the three experiments indicate that machine learning models can effectively predict early diabetes risk using lifestyle data, offering a nuanced understanding beyond primary risk factors**.** The findings suggest that the multi-factorial nature of diabetes risk holds under different analytical perspectives, confirming the critical role of comprehensive data in predictive healthcare.The individual experiments complement each other in the following ways:

Experiment 1 and Experiment 3 provided the foundational evidence that ML models, particularly the Random Forest Classifier, can accurately distinguish between healthy, pre-diabetic, and diabetic individuals based on lifestyle features. Experiment 1's confusion matrix visually confirmed the high predictive power, while Experiment 3 quantified this performance through F1-scores, establishing Random Forest as the superior model for this task. This established the 'what' – that prediction is possible and which model is best suited.

Experiment 2 then delved into the 'why,' identifying the most significant features driving these predictions. While BMI, General Health, and Age were the top predictors, the Random Forest's robust performance in Experiment 1 and 3 implicitly indicated its ability to capture the combined influence of these and other, less prominent, features.

The Experiment 3 built upon these findings by showing that even individuals with low-risk profiles based on the top 3 predictors can still be at risk due to the cumulative effect of other, individually less significant, factors (like HighBP, HighChol, MentHlth, PhysHlth, Education, and Income). This final experiment confirmed the multi-factorial nature of diabetes risk and showcased the Random Forest model's ability to identify these complex, non-obvious interactions. This answered the 'how' and 'who else' aspects of the research question, providing a comprehensive understanding of nuanced risk.

Based on the combined evidence, we can conclude that machine learning models, particularly Random Forest, can accurately predict early diabetes risk by distinguishing between healthy, pre-diabetic, and early-stage diabetic individuals using only lifestyle data. The experiments collectively support the hypothesis that a multi-factorial approach, leveraging various lifestyle features beyond just the primary indicators, is crucial for effective early detection and personalized preventative strategies in diabetes management.

## Conclusion

This study set out to understand whether machine learning models can accurately predict early signs of diabetes from lifestyle data, distinguishing between healthy, pre-diabetic, and early-stage diabetic individuals**.** Through a combination of data preprocessing (SMOTE for class imbalance), model training (Logistic Regression, Random Forest, Neural Network), and detailed analysis of

performance and feature importances, we found that yes, ML models, especially Random Forest, can achieve high accuracy in this prediction task**.** These results provide evidence that a data-driven approach leveraging lifestyle indicators offers significant potential for proactive healthcare interventions**.** The project demonstrated the strong ability of Machine Learning models in making predictions on diabetes from lifestyle data. The Random Forest Classifier achieved an F1-score of 0.9318, clearly showing its capability to distinguish between different diabetes risk levels using only lifestyle features. Beyond traditional markers, the analysis highlighted BMI, General Health, and Age as primary predictors, while also uncovering the nuanced contributions of factors like HighBP, HighChol, Mental/Physical Health, and socioeconomic indicators. The atypical cases analysis powerfully illustrated that individuals with seemingly low-risk profiles based on top predictors can still be at high risk due to the cumulative effect of other factors. The Random Forest model's ability to identify these complex interactions is a key strength. The project successfully compared Logistic Regression, Random Forest, and Neural Networks, offering a clear trade-off between interpretability and performance, with Random Forest providing the best balance.

This project, while comprehensive, faced several limitations. The BRFSS2015 dataset, being based on self-reported data, introduces potential recall bias, and although synthetic data from SMOTE balanced classes, it did not introduce new real-world variation. A key challenge was the inherent lack of interpretability in Neural Networks, making it difficult to fully explain their predictions in a medical context. Furthermore, the cross-sectional nature of the dataset limited our ability to infer causality or track changes in diabetes risk over time, highlighting the need for longitudinal data to gain deeper insights into disease progression. Finally, the model's generalizability might be restricted to populations similar to those in the BRFSS2015 dataset, necessitating further validation with diverse datasets. Looking ahead, future research could focus on incorporating longitudinal data for modeling risk progression and evaluating the impact of interventions over time. Exploring advanced interpretability techniques such as SHAP or LIME would help to better understand the complex decision-making processes of the Random Forest and Neural Network models. Integrating clinical data, like blood glucose levels or HbA1c, with lifestyle information could lead to more robust predictive models. Developing and testing personalized intervention strategies based on identified risk factors, especially for atypical high-risk individuals, is another promising avenue. Lastly, investigating the potential of real-time wearable device data could significantly enhance early detection and intervention strategies in preventative healthcare.

## Appendix 1 - Research Assets

This appendix consolidates all **research assets** developed or used in this project, including datasets, analysis notebooks, and supporting materials. Each item listed below should be accessible through a public or shareable link, ensuring transparency and reproducibility of the research process.

**Modelling**

| Notebook Title | Description / Purpose | Link |
|---|---|---|
| CAP 4773 Final Project | Complete notebook containing all stages of data processing, exploration, modeling, and visualization for this research project. | https://colab.research.google.com/drive/1PDY8i2jON1XopHvikUadHnueBnqccAAM?usp=sharing |

**Datasets**

| Dataset Title | Type | Description / Purpose | Link |
|---|---|---|---|
| Dataset_1 – Diabetes_health_indicator_dataset | Curated | Integrated dataset combining [https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset] and [ https://www.kaggle.com/code/vedatgul/feature-engineering-with-diabetes-dataset/notebook] for the training our models | [https://drive.google.com/file/d/1kU-CeWeiaW4XS7sfYaaSMkUbwKrL4-OL/view?usp=sharing] and [https://drive.google.com/file/d/19m1u5_7Lg72gl6Lk7F_Gz-fEqG-iRqHM/view?usp=sharing] |
| Dataset_3 – Generated diabetes dataset | Generated | Synthetic dataset simulating lifestyle factors that may contribute to diabetes for model validation and testing. | [https://drive.google.com/file/d/1nPH00KR2KEPTldD0tfJ65zdzu0cERZTM/view?usp=sharing] |

In Dataset_1 – Diabetes_health_indicator_dataset, we have several lifestyles factors as well as with some associated health factors. Factors like BMI, high blood pressure, general health, smokers status serve as the features which is used for training our models with the diabetes status(0-No diabetes, 1-prediabetes, 2-diabetes) as target.

Dataset_3 – Generated diabetes dataset has some generated lifestyle factors which is combined with our main dataset to deal with imbalances since we were missing some data in the primary dataset to allow us to train a balanced ML model to predict diabetes.