A Data-Driven Population-Based Targeted Intervention for Diabetes Prevention and Management: A Multidimensional Approach to Enhance Community Health Outcomes

Krushi Patel[1] Shrinidhi Rajesh[2]

[1]Department of Health Informatics, DePaul University, Chicago, IL, USA
[2]Department of Applied Statistics, DePaul University, Chicago, IL, USA

Abstract

This paper presents a population based targeted intervention approach for diabetes prevention and management, focusing on addressing the diverse needs of individuals across the United States. Our intervention program is tailored to specific demographics, including individuals aged 55-69 with lower educational backgrounds, compromised general health, limited income, elevated BMI, and high blood pressure. Leveraging advanced data analysis techniques and CDC population health competencies, our program encompasses a range of community-based initiatives. These include health education workshops, support groups, affordable fitness classes, financial assistance programs, and health screening camps. By integrating multidimensional approaches and collaborating with healthcare providers, our interventions aim to mitigate diabetes risk factors and enhance disease management. Through evidence-based practices and continuous monitoring, our program strives to achieve sustainable improvements in health outcomes and reduce health disparities. Our commitment to community engagement and evidence-based decision-making underscores our dedication to creating a healthier future for all individuals affected by diabetes. Additionally, we conducted a literature search on machine learning in healthcare, recent advancements in diabetes management, and the effectiveness of CDC interventions, informing our approach. It is important to note that this intervention is a continuous process, and this paper represents just the beginning of our ongoing efforts to combat diabetes and improve population health.

1 Introduction

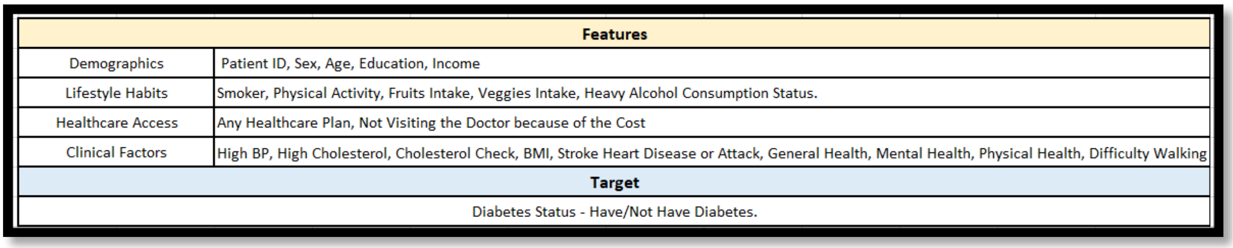
1.1 Background of the Study

Diabetes, a prevalent metabolic disorder, poses a significant public health concern globally. With its incidence steadily increasing, particularly in the United States, there is a pressing need for proactive measures to curb its impact and improve outcomes for those affected. Diabetes, characterized by abnormal levels of blood sugar, affects millions globally and poses serious health risks if left unmanaged. It is a chronic condition that requires lifelong management to prevent complications such as heart disease, kidney failure, and blindness. Lifestyle factors also play a crucial role in the development and progression of diabetes. Understanding these factors and their interactions is essential for effective prevention and management strategies.

The project is focused on one clearly defined goal, Developing Targeted Population-Based Interventions. Our goal is to utilize advanced data analysis and machine learning techniques to develop precise interventions aimed at preventing diabetes within the population of United States. By analyzing factors such as demographics, lifestyle choices, and healthcare access, we aim to tailor community-based interventions to the unique needs and characteristics of at-risk populations. Our project embarks on a mission of paramount importance: to combat the diabetes epidemic through data-driven insights and evidence-based interventions for diabetes prevention, aligning with the CDC Hi 5 principle.

1.2 Dataset

Central to our efforts, we are utilizing the Diabetes Health Indicators Dataset, a rich collection curated by the Centers for Disease Control and Prevention (CDC) downloaded through the UCI repository library (ucimlrepo) in Python [1].

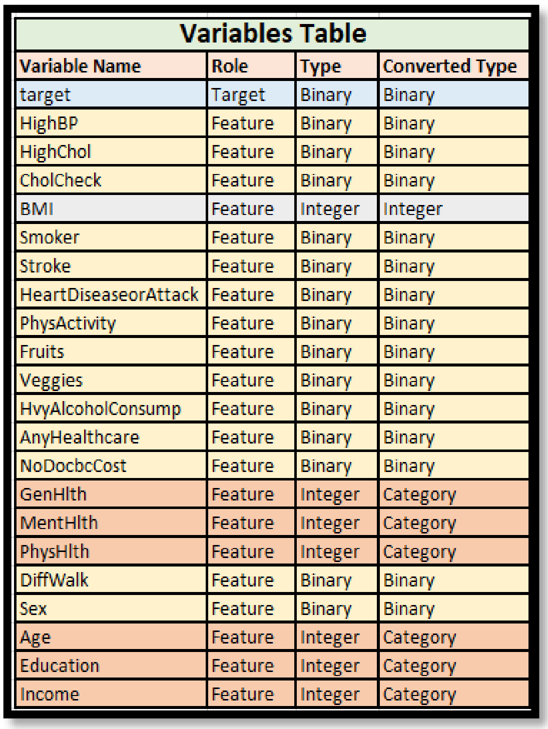


Features	
Demographics	Patient ID, Sex, Age, Education, Income
Lifestyle Habits	Smoker, Physical Activity, Fruits Intake, Veggies Intake, Heavy Alcohol Consumption Status.
Healthcare Access	Any Healthcare Plan, Not Visiting the Doctor because of the Cost
Clinical Factors	High BP, High Cholesterol, Cholesterol Check, BMI, Stroke Heart Disease or Attack, General Health, Mental Health, Physical Health, Difficulty Walking
Target	
Diabetes Status - Have/Not Have Diabetes.	

Table 1: Variable in the Diabetes Dataset

This extensive dataset encapsulates a diverse array of information, ranging from demographic profiles to laboratory findings and survey responses, meticulously collected from individuals spanning the breadth of the United States [Table 1]. These features are collected based on few of the demographics, lab results and through patient survey. The Diabetes Health Indicators Dataset is characterized by its tabular and multivariate nature, reflecting the intricate interplay between various factors influencing diabetes prevalence. Positioned within the domain of health and medicine, this dataset is inherently tailored to support classification tasks, enabling the categorization of individuals based on their diabetes status. Comprising both categorical and integer features [Table 2], the dataset boasts a substantial volume, with 253,680 instances and 21 distinct features. At its core, the creation of this dataset is driven by a fundamental objective to unravel the complex nexus between lifestyle choices and diabetes within the American populace.

The target variable encapsulates individual's diabetes statuses, delineating between those afflicted by diabetes, those exhibiting pre-diabetic conditions, and individuals classified as healthy. The target variable is one-hot encoded, with healthy and the prediabetic people as 0 and the people with diabetes as 1. This pivotal classification serves as the cornerstone for subsequent analyses and interventions aimed at addressing diabetes prevalence and its associated risk factors [2].



Variables Table

Variable Name	Role	Type	Converted Type
target	Target	Binary	Binary
HighBP	Feature	Binary	Binary
HighChol	Feature	Binary	Binary
CholCheck	Feature	Binary	Binary
BMI	Feature	Integer	Integer
Smoker	Feature	Binary	Binary
Stroke	Feature	Binary	Binary
HeartDiseaseorAttack	Feature	Binary	Binary
PhysActivity	Feature	Binary	Binary
Fruits	Feature	Binary	Binary
Veggies	Feature	Binary	Binary
HvyAlcoholConsump	Feature	Binary	Binary
AnyHealthcare	Feature	Binary	Binary
NoDocbcCost	Feature	Binary	Binary
GenHlth	Feature	Integer	Category
MentHlth	Feature	Integer	Category
PhysHlth	Feature	Integer	Category
DiffWalk	Feature	Binary	Binary
Sex	Feature	Binary	Binary
Age	Feature	Integer	Category
Education	Feature	Integer	Category
Income	Feature	Integer	Category

Table 2: Type of Variables

1.3 Literature Reviews

American Diabetes Association Professional Practice Committee [3] provides critical insights into effective diabetes management. Emphasizing early detection, lifestyle modifications, and personalized care plans, these guidelines serve as a cornerstone for our intervention approach. By understanding these standards, our project aims to optimize health outcomes for individuals at risk of diabetes, leveraging the latest evidence-based practices.

Dunya Tomic et.al.[4] proposed an expanding spectrum of health risks associated with diabetes. Traditionally, diabetes management focused on well-known complications like cardiovascular diseases and neuropathy. However, this study highlights newer complications such as increased cancer risk, cognitive impairment, liver diseases, and susceptibility to infections among diabetic individuals. Understanding and addressing these emerging complications are essential components of our project's comprehensive approach to healthcare. By incorporating the understanding from this study, our project aims to optimize health outcomes for individuals at risk of diabetes.

Hafsa Habehh et.al.[5] proposes the incorporation of machine learning (ML) into healthcare has shown promising advancements in patient outcomes prediction, diagnosis, and treatment from various sources. Esteva et al. (2017) and Rajpurkar et al. (2017) exemplify ML's impact in enhancing diagnostic accuracy, particularly in medical imaging, with applications like skin cancer detection and pneumonia identification from X-rays. Leveraging these ML-driven insights, our project aims to develop tailored ML based interventions for individuals at risk of diabetes to optimize health outcomes.

In the paper by DeSalvo et.al [6], the transformation of public health strategies is analyzed from historical models to the modern Public Health 3.0 framework. Initially centered on infectious disease control and basic hygiene (Public Health 1.0), followed by an emphasis on chronic disease management and workforce readiness (Public Health 2.0), the shift to Public Health 3.0 signifies a holistic approach. This approach designates public health officials as Chief Health Strategists, fostering collaboration across sectors to address social determinants impacting health outcomes. Challenges highlighted include inadequate funding relative to medical care and the need for robust data strategies. Recommendations advocate for strengthened accreditation standards, improved data accessibility, and innovative funding mechanisms. This review underscores the ongoing need for integrated public health systems that proactively address community health through comprehensive strategies, resonating with efforts in diabetes prevention and management initiatives.

In their recent publication, Khalifa and Albadawy [7] underscore how AI is reshaping diabetes care. Highlighting AI's potential, the authors detail its application in predictive risk assessments, personalized treatment optimization, and precise diagnostics through advanced data analysis. This technological integration aims to enhance patient outcomes by tailoring interventions based on individual health profiles. However, challenges such as data security, algorithmic bias, and interdisciplinary collaboration must be addressed for effective implementation. Khalifa and Albadawy advocate for ongoing research and regulatory frameworks to harness AI's full potential in diabetes management, aligning with our efforts to integrate innovative technologies into comprehensive healthcare strategies.

The Carney et al [8], underscores the pivotal role of data modernization in enhancing chronic disease management in the US. With six out of ten adults affected by chronic diseases, including diabetes and heart disease, the initiative addresses critical challenges such as outdated infrastructures and slow data processing that were exacerbated during the COVID-19 pandemic. By integrating public health data with electronic health records and leveraging cloud technologies, the CDC aims to improve data accuracy, accessibility, and the integration of social determinants of health (SDOH) to promote equity. Carney et al. highlight the initiative's successes in enhancing data precision and informing evidence-based decision-making, paving the way for more effective public health interventions and reduced healthcare costs. This aligns with our project's goal of integrating comprehensive health data to optimize diabetes prevention and management strategies, ensuring they are informed by real-time insights and contribute to improved health outcomes. The full review of these 6 literatures can be found on Github for download [10].

2 Methodology

2.1 Data Analysis Process

2.1.1 Data Preprocessing:

The data was loaded into Jupyter Lab, an interactive development environment for running Python code, to initiate the data analysis process. After loading the dataset, it was analyzed for Missing Values to avoid misinterpretations. The dataset is huge and the outliers were not evident, thus all data points were included. Scaling and Normalization were not required for this data set. The data set was pre encoded using one-hot and label. The feature 'GenHlth' was transformed, the scaling for initially 5-1(poor to excellent). This was restructured as 1-5(poor to excellent). Feature selection was performed using two methods: Random Forest Classifier (RFC) and Low Threshold (LT). RFC yielded 8 features, while LT yielded 15 features with a threshold of 0.1. With time as the constraint, feature engineering was not performed, however, with the interventions and data collected from the participants requires us to perform FE in the future to come up with a more refined analysis.

2.1.2 Exploratory Data Analysis (EDA):

EDA was conducted both before and after feature selection to explore relationships between features and the target variable [Figure 1]. This iterative process helped refine insights and validate initial findings. Initial exploration involved calculating descriptive statistics (mean, median, standard deviation, count and proportion). Selected features were visualized to uncover patterns and relationships with the target variable. Techniques such as scatter plots, heatmaps, and bar charts were employed to identify potential risk factors for diabetes. These plots and descriptiveness helped understand the relationship between the target and the features. Visualizations, such as histograms, bar graphs and box plots, were also used to understand the distributions and identify potential outliers or trends to summarize the dataset's characteristics.

A correlation plot was generated to assess relationships between numerical features. This step helped identify multicollinearity and understand how variables relate to each other. Dimensionality Reduction and Pattern Recognition: These techniques are required once we get more data from the participants utilizing the intervention techniques, for now, we did not utilize these techniques.

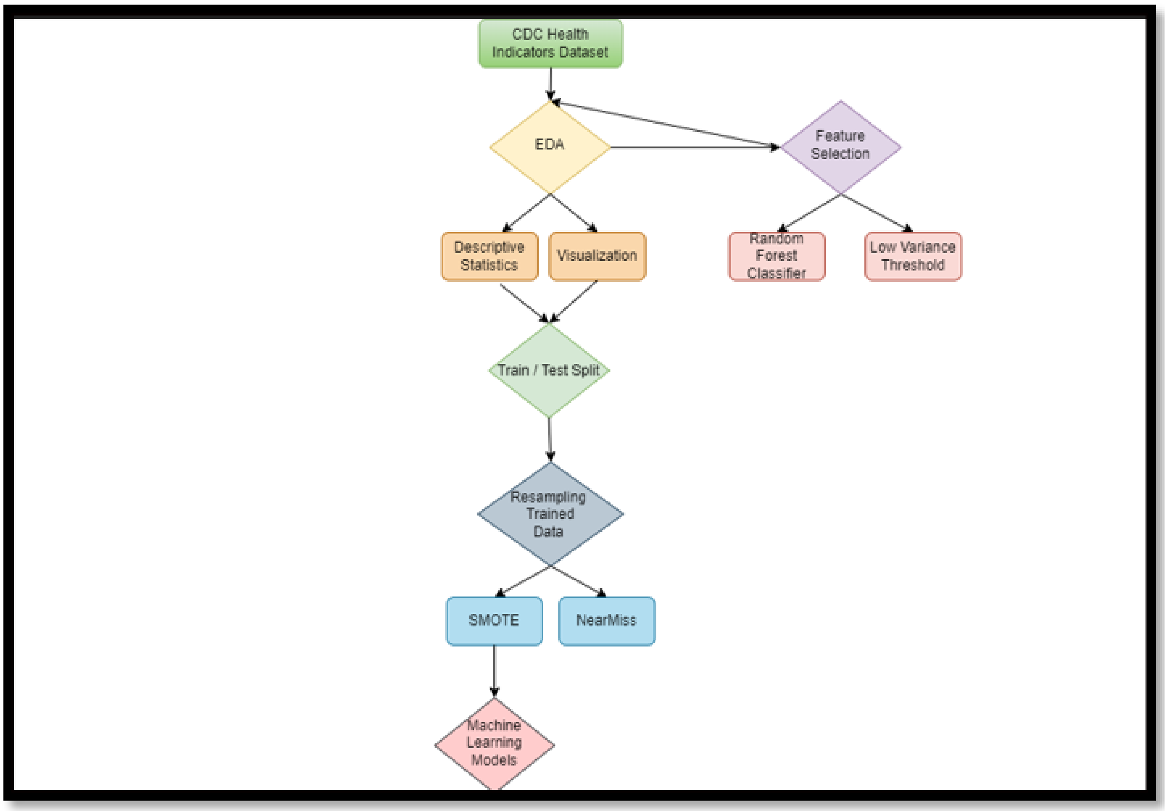


Figure1: Flow Chart of our Analysis Process

2.1.3 Modelling:

Machine learning models (Random Forest, Logistic Regression, Decision Tree, XGBoost) were performed on the training set on three types of datasets: raw data, features from RFC, and features from Low Threshold (LT)[Figure 2]. To reciprocate the results, the seed was set to 50/55. Perfect evaluation scores as results, emphasized on the chance of overfitting. Thus, class imbalance was checked and the data set was heavily imbalanced with about 86:14 ratio of Tagert 0 and 1 respectively. Class imbalance was addressed by applying oversampling (SMOTE) and undersampling (NearMiss) techniques to ensure balanced representation of classes and prevent bias towards the majority class (target 0).

To mitigate class imbalance, resampling techniques were employed, including Synthetic Minority Over-sampling Technique (SMOTE) and NearMiss on the training dataset and tested on the raw dataset (20% of the original data). After conducting multiple attempts, SMOTE emerged as the preferred method due to its higher accuracy in balancing the classes. However, the imbalance is more than the regular rule of 80:20 thus even the resampled data was shaky. Utilizing the resampled data, the models were trained and evaluated on three different feature sets: raw data with all features, features extracted using Random Forest Classifier (RFC), and features obtained from the Low Threshold (LT) method. Among the four models utilized, LR, RF, XGB and DT, Random Forest (RF) consistently outperformed the others, demonstrating superior predictive capability. Further comparison was conducted among the three types of features used (Raw, RFC, and LT). Surprisingly, all three feature sets yielded similar evaluation metrics. However, given the principle of Occam's razor, which favors simpler explanations when competing hypotheses are equally valid, the Random Forest model trained on RFC features with 8 variables[11] was deemed the most suitable choice.

After performing Random Forest classification using the features obtained from the RFC method, feature importance analysis was conducted. This analysis aimed to assess the relative importance of each feature in predicting diabetes risk within the selected subset of 8 features. By evaluating the contribution of each feature to the predictive accuracy of the model, we gained insights into which variables have the most significant impact on diabetes risk assessment within the reduced feature set. Following the feature importance analysis, visualizations was performed to discern trends and patterns specific to the 8 features with the target variable[12].

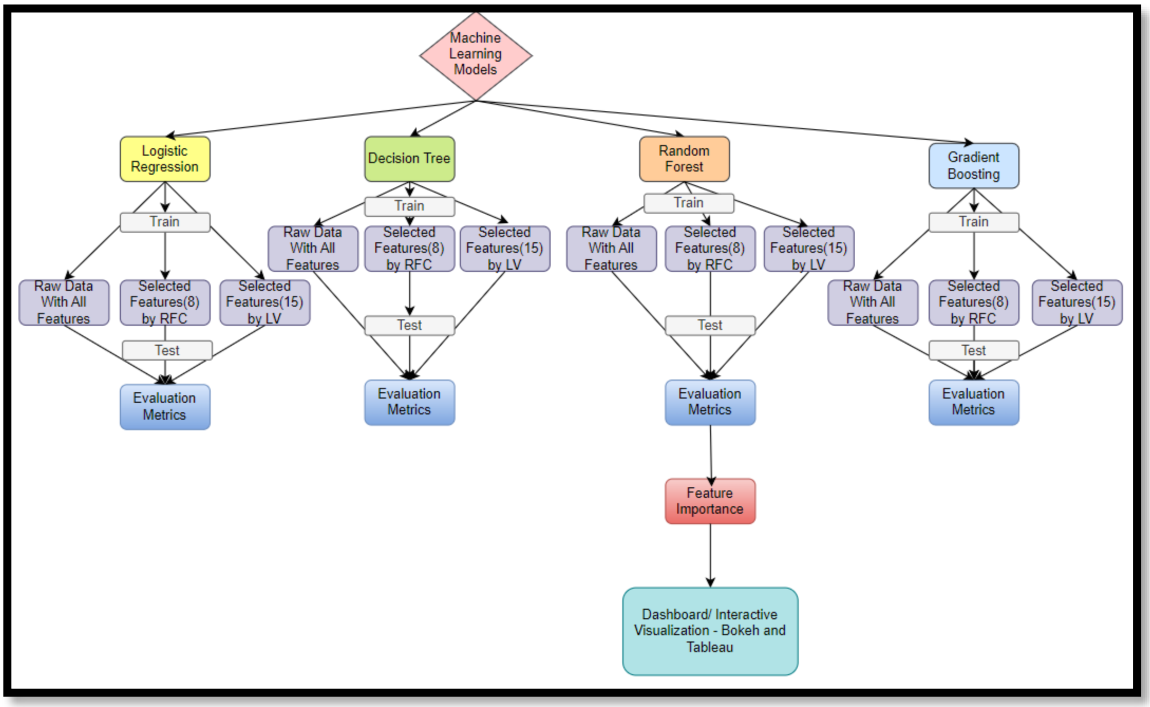

Machine Learning Models
Logistic Regression
Decision Tree
Random Forest
Gradient Boosting
Train
Raw Data With All Features
Selected Features(8) by RFC
Selected Features(15) by LV
Test
Evaluation Metrics
Feature Importance
Dashboard/ Interactive Visualization - Bokeh and Tableau

Figure2: Flow Chart of our Analysis Process (Contd.)

2.1.4 Evaluation:

The dataset was segregated to 80:20 train test split to evaluate the performance. The 4 models and on subsequent three sets of feature sets were performed on the resampled training data set and then tested on the original test dataset (from raw data). Resampling was done only on the training dataset to get the test results without any tweaking. Equal weights were assigned to standard evaluation metrics including accuracy, precision, recall, and F1-score for each model, and these metrics were compared across different models and datasets to assess performance.

Confusion Matrix was utilized to understand the False Positives and Negatives. K-fold cross-validation was subsequently performed exclusively for the Random Forest model to validate its performance and assess its generalization to unseen data. This rigorous validation process confirmed the reliability and robustness of the chosen model, enhancing confidence in its predictive power for identifying diabetes risk within the target population.

2.1.5 Interactive Visualizations Dashboard Creation:

Interactive visualizations using tools like Bokeh and Tableau were created to explore relationships between features and the target variable. These dashboards also explain the proportion of people affected by diabetes and not affected by diabetes within each category of the features. These visualizations enhanced interpretation by allowing users to interactively explore the data and uncover insights not readily apparent from static plots.

4 Results and Discussions

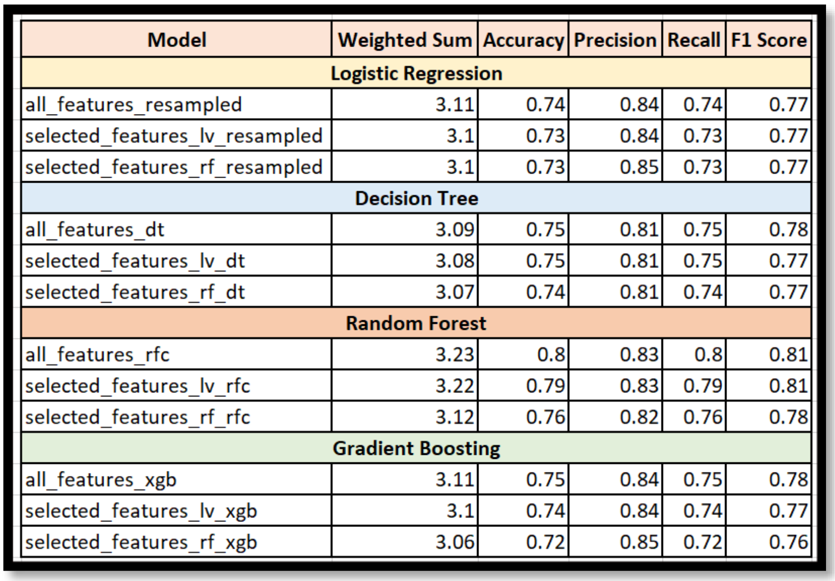
4.1 Insights from Data Model:

The data set was preprocessed for missing values and outliers, there were no missing values in the dataset and there were no significant outliers. Feature Selection was performed using Random Forest classifier which resulted in 8 features out of 21 - HighBP, BMI, GenHlth, MentHlth, PhysHlth, Age, Education, Income. And the Low Variance Threshold method with 0.1 threshold resulted in the following 15 features - HighBP, HighChol, BMI, Smoker, PhysActivity, Fruits, Veggies, GenHlth, MentHlth, PhysHlth, DiffWalk, Sex, Age, Education, Income. We have used 0.1 threshold purely due to small correlation values between the target and the features and we do not want to lose information. The variable types were converted from int to category to abide the data type rules, they were already label and one hot encoded, there are a total of 6 categorical variables out of 21. Only BMI was on the continuous scale and the rest 14 variables are binary and one binary target variable.

EDA was performed and the count and proportion of each feature with respect to target were calculated. The proportions were off, resulting in more healthy people and less affected people by diabetes in each of the features. The distribution of BMI was analyzed and the bar plots were plotted for other features to get a better understanding of the data. These plots indicated that more people were in certain categories, such as people with General Health 4 were high, people with 0 Mental and Physical Health were high in this dataset, people with education level 6 and income level 8 were more, no blood pressure were high and people between the age groups of 55 to 69 were high. Lastly, for the BMI, there are more individuals better the range of 20 to 40.

The data is split on a 80:20 ration for train and split. The models were performed on the raw data, the scores were around 90% but since we did the EDA, it looked as if the model was overfitted. We check for the class proportion and as expected the model was heavily imbalanced following a 86:14 ratio of not having diabetes: having diabetes. To balance the data, we employed SMOTE and NearTres. The models were employed on both the resampled datasets with different seeds, the SMOTE technique worked the best comparatively. The resampling techniques were only performed on the trained data set and a new trained resampled dataset was used while the testing was done on the original 20 split to avoid manipulation of data. The result suggested different evaluation metrics and confusion matrix a total of 12 different evaluation metric sets, detailed in our Python Codebook with output. Of the 4 different ML models performed, the rankings were based on assigning equal weights and summing up the 4 different metrics, Accuracy, Precision, Recall and F1 Score[9].

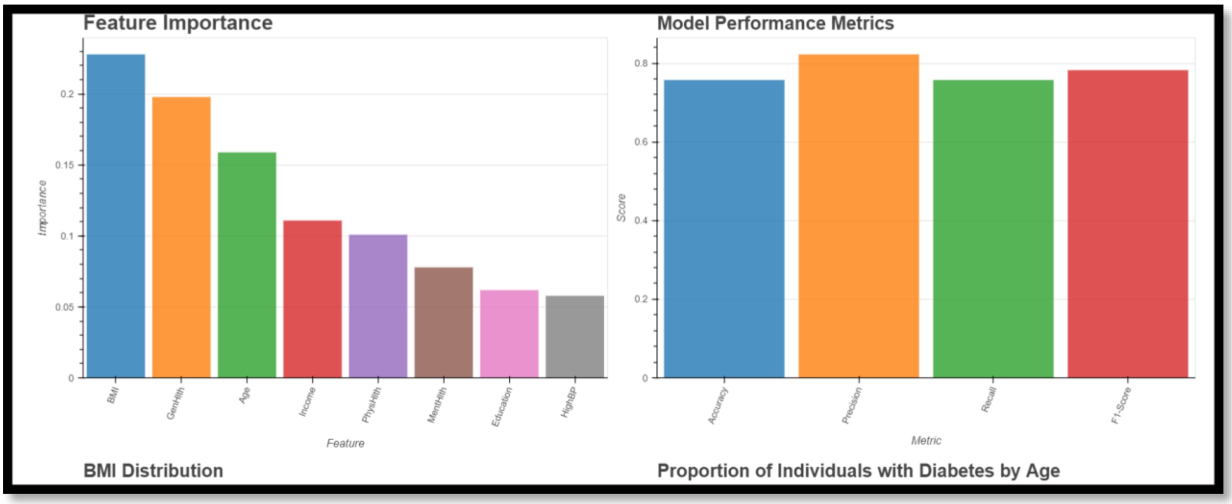
We initially anticipated that the Gradient Boosting method would rank first, given its renowned ability to perform well with imbalanced data. However, to our surprise, the Random Forest model outperformed the other four models, including Gradient Boosting, across our evaluation metrics.



Model	Weighted Sum	Accuracy	Precision	Recall	F1 Score
Logistic Regression					
all_features_resampled	3.11	0.74	0.84	0.74	0.77
selected_features_lv_resampled	3.1	0.73	0.84	0.73	0.77
selected_features_rf_resampled	3.1	0.73	0.85	0.73	0.77
Decision Tree					
all_features_dt	3.09	0.75	0.81	0.75	0.78
selected_features_lv_dt	3.08	0.75	0.81	0.75	0.77
selected_features_rf_dt	3.07	0.74	0.81	0.74	0.77
Random Forest					
all_features_rfc	3.23	0.8	0.83	0.8	0.81
selected_features_lv_rfc	3.22	0.79	0.83	0.79	0.81
selected_features_rf_rfc	3.12	0.76	0.82	0.76	0.78
Gradient Boosting					
all_features_xgb	3.11	0.75	0.84	0.75	0.78
selected_features_lv_xgb	3.1	0.74	0.84	0.74	0.77
selected_features_rf_xgb	3.06	0.72	0.85	0.72	0.76

Table 3: Evaluation Metrics

In selecting the optimal model for our application, we chose the Random Forest model utilizing the 8 features identified through the Random Forest classifier's feature selection (selected_features_rf_rfc). This decision was based on a balance between model performance and simplicity. While the Random Forest model with all features (all_features_rfc) achieved slightly higher metrics—Accuracy: 0.80, Precision: 0.83, Recall: 0.80, F1 Score: 0.81—compared to the selected_features_rf_rfc model with Accuracy: 0.76, Precision: 0.82, Recall: 0.76, and F1 Score: 0.78, the difference is minimal. The selected model's weighted sum of 3.12 is also only marginally lower than the 3.23 of the all-features model [Table 3]. Additionally, the feature importance analysis revealed that the 8 chosen features were significantly more influential in predicting outcomes compared to the remaining features, which were relatively less important. By focusing on these key features, the model is simplified, making it easier to interpret and faster to train and predict, and avoids the risk of including less informative features that might add noise rather than valuable information. This streamlined model thus provides a practical and efficient solution while maintaining robust performance, justifying its selection over the other 11 evaluated models. In the analysis of factors affecting diabetes prevalence, the full output file can be accessed on GitHub [9].



Feature Importance

Model Performance Metrics

BMI Distribution

Proportion of Individuals with Diabetes by Age

Figure 3: Feature Importance and Evaluation Metrics Chart

We then performed feature importance only for these 8 features. The feature importance analysis reveals that certain variables play a more significant role than others in predicting outcomes. Among the selected features, BMI (Body Mass Index) holds the highest importance with a score of 0.23, followed closely by GenHlth (General Health) at 0.20 and Age at 0.16. Other notable contributors include Income (0.11) [Figure

3], PhysHlth (Physical Health, 0.10), and MentHlth (Mental Health, 0.08). Education and the presence of High Blood Pressure (HighBP) are also considered, albeit with relatively lower importance scores of 0.06 each. These insights highlight the critical variables influencing the model's predictions, providing valuable guidance for further analysis and decision-making.

These insights highlight the critical variables influencing the model's predictions, providing valuable guidance for further analysis and decision-making. Although features such as Mental Health, Education, and High Blood Pressure have relatively lower importance scores (less than 0.1), it is reasonable to include all 8 features for our intervention. Given that our dataset is relatively small, with only 21 features, discarding too many features could lead to a loss of valuable information. Even if we had included all 21 features, the feature importance plot clearly shows that these 8 features are significantly more important compared to the remaining 13.

Analyzing feature importance helps us understand which features have the strongest influence on the model's predictions, which is crucial for designing targeted interventions. Therefore, we have ensured that our targeted interventions are focused on both the highly important features and the relatively lower important ones. This balanced approach allows us to maximize the effectiveness of our interventions based on the insights gained from the feature importance analysis.

4.2 Model Performance Comparison:

While the model using all features in the dataset ("all_features_rfc") outperforms others with a weighted sum of 3.23, the "selected_features_rf_rfc" model, utilizing only a subset of 8 features, demonstrates commendable performance, achieving a slightly lower weighted sum of 3.12. Despite the reduced feature set, it maintains strong metrics, including an accuracy of 0.76, precision of 0.82, recall of 0.76, and F1 score of 0.78. Opting for the "selected_features_rf_rfc" model aligns with the principle of simplicity and efficiency. By leveraging only 8 selected features, the model reduces complexity, making it easier to interpret and implement. This streamlined approach enhances efficiency in both training and inference phases, facilitating quicker decision-making processes.

The decision to focus on a selected set of features underscores a deliberate effort to prioritize the most influential variables for the task. These features, identified by the Random Forest classifier, are deemed crucial for predicting the target outcome. By honing in on these key factors, the model facilitates more targeted interventions and decision-making, potentially leading to more impactful outcomes. Utilizing a smaller set of features in the "selected_features_rf_rfc" model may mitigate the risk of overfitting and improve generalization to unseen data. This enhances the model's robustness and reliability in real-world applications, ensuring that it can effectively generalize to new scenarios beyond the training data. Despite the superior performance of the "all_features_rfc" model, the decision to prioritize the "selected_features_rf_rfc" model is rooted in its simplicity, efficiency, and focus on key factors. This approach closely aligns with the objectives of the analysis, emphasizing practicality and ease of implementation while maintaining strong predictive performance.

4.3 Further Insights:

Understanding the confusion matrix is essential, especially given the pronounced data imbalance, with only 14% of the dataset representing the minority class. Despite applying resampling techniques, their efficacy was limited due to the severe class imbalance, as these methods typically yield optimal results when the ratio is above the general rule of 80:20. For the chosen Random Forest model with selected features, the confusion matrix portrays a nuanced picture. It identifies 3,621 true positives (correctly predicted instances of class 1) and 34,820 true negatives (correctly predicted instances of class 0). However, the model also generates 8,826 false positives (instances of class 0 incorrectly identified as class 1) and 3,469 false negatives (instances of class 1 incorrectly identified as class 0). Despite these challenges, the model demonstrates a commendable sensitivity to the minority class, albeit with some instances of over-classification. Overall, the model's reasonable ability to navigate the intricacies of imbalance while maintaining practical utility underscores its effectiveness in this context.

Utilizing interactive plots generated through Bokeh [9] and a highlight table created using Tableau [13], we sought to explore the impact of various categorical and binary variables on diabetes prevalence. The tableau table, specifically, focuses on individuals unaffected by diabetes, highlighting the percentage distribution across different classes. With Bokeh, we visualized features against the target variable, uncovering notable trends in their relationships. Despite the vast dataset size and computational power, generating partial residual plots became time-consuming. However, through visual analysis, clear trends emerged. Notably, BMI showcased a significant association with diabetes, with a higher prevalence observed between the ages of 20 to 40, peaking around 30 and gradually declining thereafter. Similarly, individuals reporting Fair or Poor General Health exhibited a higher likelihood of diabetes, whereas better General Health correlated with lower diabetes risk. Age emerged as a significant risk factor, with the 55-69 age group showing elevated diabetes prevalence. Income and education followed similar patterns, with higher income and education associated with reduced diabetes risk. While the trends for Physical and Mental Health were less clear, both suggested that better health correlated with lower diabetes risk. Moreover, individuals with high blood pressure showed an increased likelihood of diabetes [Figure 4,5,6].

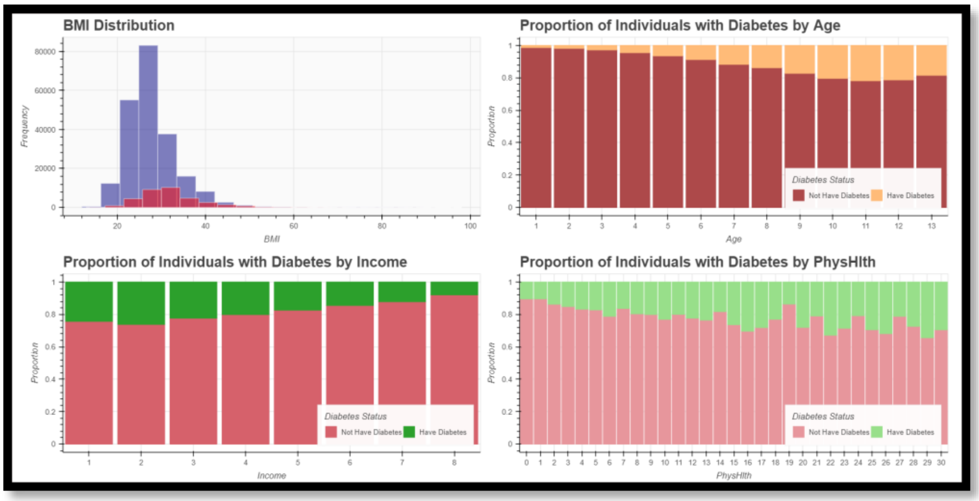


Figure 4: Bokeh Plots for Features vs Target

Importantly, this dataset primarily comprises a healthy population from the United States. Comparing the prevalence rate to the American Diabetes Association's statistics from 2021 (38.4 million Americans, or 11.6% of the population, had diabetes), we find reasonable alignment. However, effective intervention requires balanced data aligned with our goals. Given that the dataset was sourced from Kaggle and not constructed for specific research objectives, limitations must be acknowledged.

A fundamental principle guiding our intervention proposals is the Central Limit Theorem, which underscores the importance of larger sample sizes for stable conclusions. With more data available within specific categories, we can more confidently recommend interventions, as our models are naturally drawn towards these bins. This inclination towards categories with larger datasets aligns with statistical principles, ensuring that our proposed interventions are grounded in robust empirical evidence. By leveraging the abundance of data within these categories, we can formulate targeted strategies aimed at addressing prevalent health concerns effectively.

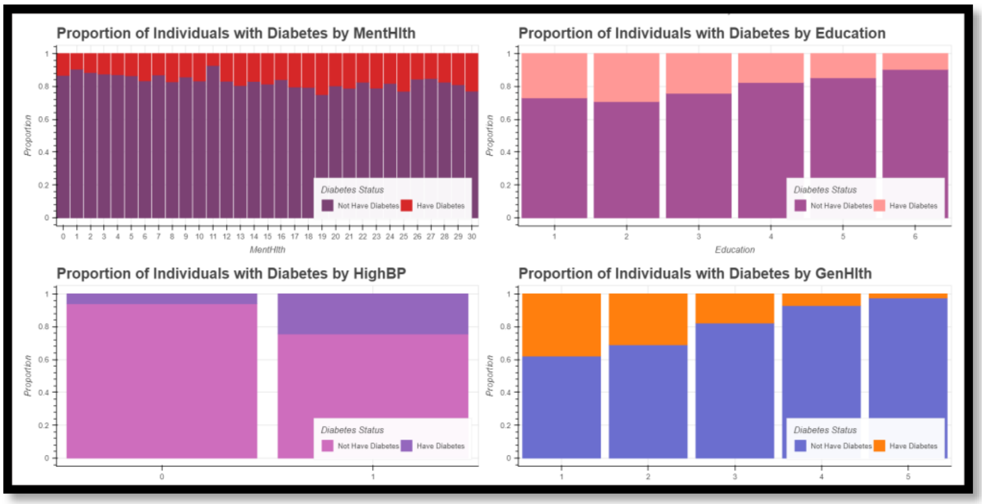
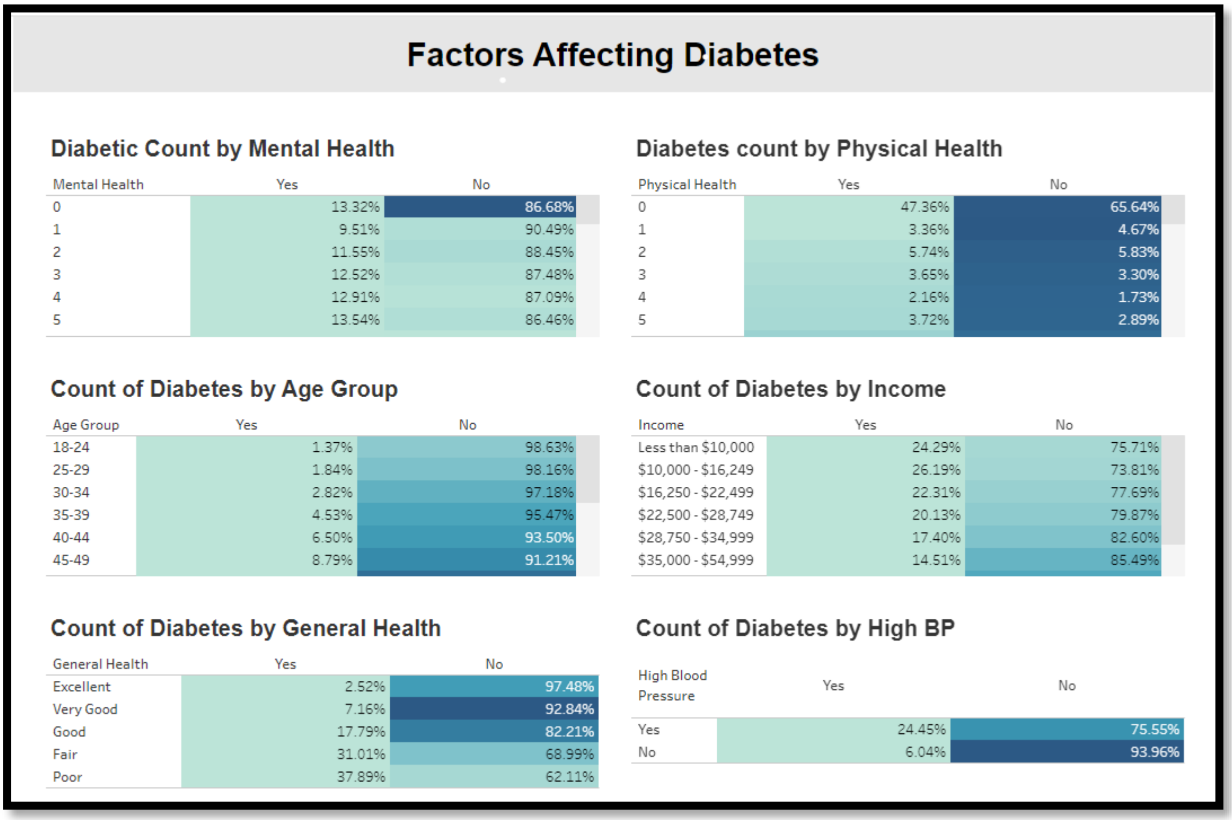


Figure 5: Bokeh Plots for Features vs Target

However, while leveraging larger datasets within specific categories aligns with statistical principles, it is crucial to acknowledge that this approach may inadvertently bias our results. The abundance of data in certain categories can influence model predictions, potentially skewing intervention recommendations towards these overrepresented groups. This phenomenon highlights the importance of ensuring dataset balance and considering potential biases when interpreting results and formulating interventions. By maintaining awareness of these nuances, we can strive to mitigate the impact of skewed data on our analyses and ensure more equitable and effective intervention strategies.

Furthermore, it is essential to recognize that our intervention efforts will likely yield additional data, potentially balancing the dataset and leading to more robust results. By actively collecting data through our interventions, we can enhance the representativeness of our dataset and refine our analyses, ultimately contributing to more accurate and effective intervention strategies. This iterative process underscores our commitment to continuous improvement and the pursuit of impactful outcomes in addressing health disparities.

We acknowledge the sensitive nature of working with diabetes and health data, which has the potential to profoundly impact individuals' well-being. Given this responsibility, we are committed to meticulously collecting and handling data with the utmost care. Recognizing the significance of the confusion matrix and evaluation metrics, we prioritize accuracy in our analyses. These metrics serve as critical tools in guiding intervention strategies and decision-making processes. While we recognize that our current dataset may have some inherent biases, it reflects the information available to us at present. We are aware of the consequences and accept this reality as we strive to make the best use of the data we have. Moving forward, we are dedicated to improving our approach through continuous data collection, monitoring, and iterative refinement of our models. By embracing this process, we aim to enhance the reliability and effectiveness of our interventions, ultimately making meaningful strides in addressing diabetes and related health challenges.



Factors Affecting Diabetes

Diabetic Count by Mental Health

Mental Health	Yes	No
0	13.32%	86.68%
1	9.51%	90.49%
2	11.55%	88.45%
3	12.52%	87.48%
4	12.91%	87.09%
5	13.54%	86.46%

Diabetes count by Physical Health

Physical Health	Yes	No
0	47.36%	65.64%
1	3.36%	4.67%
2	5.74%	5.83%
3	3.65%	3.30%
4	2.16%	1.73%
5	3.72%	2.89%

Count of Diabetes by Age Group

Age Group	Yes	No
18-24	1.37%	98.63%
25-29	1.84%	98.16%
30-34	2.82%	97.18%
35-39	4.53%	95.47%
40-44	6.50%	93.50%
45-49	8.79%	91.21%

Count of Diabetes by Income

Income	Yes	No
Less than $10,000	24.29%	75.71%
$10,000 - $16,249	26.19%	73.81%
$16,250 - $22,499	22.31%	77.69%
$22,500 - $28,749	20.13%	79.87%
$28,750 - $34,999	17.40%	82.60%
$35,000 - $54,999	14.51%	85.49%

Count of Diabetes by General Health

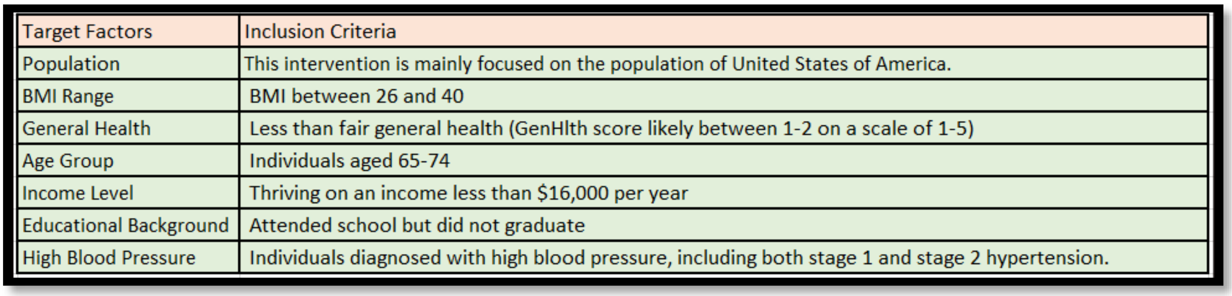
General Health	Yes	No
Excellent	2.52%	97.48%
Very Good	7.16%	92.84%
Good	17.79%	82.21%
Fair	31.01%	68.99%
Poor	37.89%	62.11%

Count of Diabetes by High BP

High Blood Pressure	Yes	No
Yes	24.45%	75.55%
No	6.04%	93.96%

Figure 6: Tableau Highlight Table of Features Vs Target

5 Target Audience Description



Target Factors	Inclusion Criteria
Population	This intervention is mainly focused on the population of United States of America.
BMI Range	BMI between 26 and 40
General Health	Less than fair general health (GenHlth score likely between 1-2 on a scale of 1-5)
Age Group	Individuals aged 65-74
Income Level	Thriving on an income less than $16,000 per year
Educational Background	Attended school but did not graduate
High Blood Pressure	Individuals diagnosed with high blood pressure, including both stage 1 and stage 2 hypertension.

Table 4: Target Intervention Audience

Based on an empirical analysis of feature importance, our intervention targets individuals with a higher Body Mass Index (BMI) between 26 and 40, as this is a significant risk factor for diabetes (importance score: 0.23). Higher BMI indicates overweight or obesity, which markedly increases the likelihood of developing diabetes. Furthermore, individuals reporting less than fair general health (GenHlth score: 1-2) are prioritized, given their potential underlying health challenges contributing to diabetes risk (importance score: 0.20). The age group of 65-74 years is also a key focus, as this demographic is more susceptible to diabetes due to age-related factors and potential co-morbidities (importance score: 0.16). Low income (less than $16,000 per year) is another critical factor, as financial constraints can limit access to healthcare services, nutritious food, and opportunities for physical activity, all essential for diabetes prevention and management (importance score: 0.11)[Table 4].

Educational background plays a role, with individuals who attended school but did not graduate being at higher risk, likely due to limited access to health information and resources that can impact health outcomes, including diabetes risk (importance score: 0.06). Additionally, high blood pressure is included, as hypertension is a common comorbidity of diabetes and can exacerbate its complications. Including individuals diagnosed with high blood pressure, including both stage 1 and stage 2 hypertension, helps address this critical risk factor (importance score: 0.06). While mental and physical health are significant factors in the model's diabetes risk assessment, their direct relationship with diabetes may not follow a clear trend. Nonetheless, addressing these aspects holistically remains essential for comprehensive health promotion and disease prevention efforts (importance scores: Mental Health 0.08, Physical Health 0.10).

Finally, our intervention initially focuses on major cities in the United States, with the potential for program expansion to other areas. This structured and evidence-based approach ensures we target the most influential factors in diabetes prevention effectively. This group has a higher risk of developing diabetes compared to the rest of the population, as indicated by the data model. Any individual meeting one or more of these criteria is included, as each factor contributes significantly to the increased risk. Additionally, our intervention program will be open to all with main focus on this particular populations due to susceptibility.

5.1 Proposed Intervention (HI-5 Aligned):

This project explores potential interventions to reduce diabetes risk in a high-needs population. Utilizing a random forest model, we analyzed relevant health indicators to identify modifiable risk factors. The empirical evidence gained from this analysis provides valuable insights for developing targeted strategies that maximize impact and improve health outcomes.

5.1.1 HI-5 Bucket 1: Positive Health Impacts:

The following are our proposed Community-based programs promoting healthy lifestyle changes (Interventions):

1) Health Education Workshops: Tailored workshops conducted weekly to educate participants on nutrition, physical fitness, and general health maintenance. Curriculum includes interactive sessions on meal planning, portion control, and understanding nutritional labels. Guest speakers, such as dietitians and fitness trainers, provide expert advice and practical tips.

2) Support Groups: Weekly support groups focus on emotional well-being, stress management, and social support. Facilitated discussions address coping strategies, mindfulness techniques, and resilience-building exercises. Specialized sessions on diabetes disease management cover topics like medication adherence, blood glucose monitoring, and symptom recognition.

3) Affordable Fitness Classes: Weekly fitness classes offer affordable options like dancing and Zumba to engage participants in enjoyable physical activities. Classes scheduled three times per week to accommodate diverse schedules and preferences. Registration and payment options available on the program website, with monthly membership fees. Weekly surveys collect feedback on class enjoyment, perceived health benefits, and suggestions for additional fitness activities. Expansion of class offerings based on participant feedback and demand.

4) Financial Assistance for Accessing Health Services and Healthy Foods: Job training and education programs provided to improve socioeconomic status and access to healthcare resources. Eligibility criteria include income verification and participation in community health programs. Data on income levels and program participation used to assess eligibility and track outcomes.

5) Health Screening Camps: Monthly health screening camps held at various locations to assess diabetes risk factors and provide immediate medical assistance. Program website features camp schedules, location maps, and registration details. Comprehensive health assessments conducted by medical professionals, including BMI measurement, blood pressure checks, and blood glucose testing. Urgent cases referred to local hospitals with program coverage for treatment expenses.

We are planning to propose on a rotational workshop method. Structured rotation of workshops ensures comprehensive coverage of health topics throughout the month. Each week focuses on a different aspect of health, ranging from nutrition and fitness to mental wellness and community support. Schedule posted on the program website [14] with detailed information and registration instructions [Figure 7].

5.1.2 HI-5 Bucket 2: Achieving Results Within Five Years:

Our program's primary objective is to proactively prevent diabetes onset and effectively manage the condition among diagnosed individuals. Success is defined as a reduction in the incidence of diabetes diagnosis over a five-year period. Additionally, successful outcomes include controlling diabetes levels among diagnosed individuals through lifestyle modifications and appropriate medical management.

The interventions outlined target modifiable lifestyle behaviors, focusing on improving participants BMI, general health, income status, physical fitness, and mental well-being. These interventions, including education workshops, support groups, and fitness classes, aim to instill sustainable habits conducive to positive health outcomes, such as improved dietary choices, increased physical activity, and enhanced stress management.

Collaboration with local community centers and healthcare providers is paramount to the success of our program. These partnerships facilitate program delivery, participant recruitment, and access to suitable venues for organizing support groups, fitness classes, and health screenings. By leveraging existing community resources, we can maximize program reach and effectiveness.

A crucial aspect of our further plan involves continuous monitoring and evaluation of program outcomes. Tracking program participation, behavior changes, and pre-diabetes/diabetes rates over time allows us to assess effectiveness and iteratively update our intervention strategies. By integrating data-driven insights into our approach, we can optimize program delivery and enhance community health outcomes. A pivotal role of our ongoing strategy entails the continual acquisition of comprehensive health data and the iterative refinement of our predictive model. Through systematic monitoring and evaluation, we aim to collect a rich dataset encompassing various health indicators, including but not limited to BMI, blood pressure, physical health, mental health, and general health perceptions.

Our approach emphasizes interventions that are practical and easily integrated into the community, taking into consideration the tribal mentality prevalent among our target population. By understanding and respecting cultural norms and community values, we aim to develop initiatives such as tailored education workshops, support groups, and fitness classes that resonate deeply within the community. This approach ensures that our interventions are both accessible but effective in promoting sustainable health behaviors and reducing diabetes risk.

This expansive dataset serves as the cornerstone for enhancing the accuracy and predictive power of our model. By tracking program participation, behavior changes, and the incidence of pre-diabetes/diabetes cases over time, we can assess the effectiveness of our interventions as-well as identify new insights and trends within the data. Leveraging advanced analytics techniques, such as machine learning algorithms and predictive modeling, we extract actionable insights from this wealth of information. Our data-centric approach enables us to adapt our interventions dynamically, tailoring them to the evolving needs and characteristics of our target population. Through iterative model updates and refinements, we continuously enhance the predictive capabilities of our model, ensuring its alignment with real-world outcomes. By harnessing the power of data-driven decision-making, we aim optimize program delivery, maximize intervention impact, such as in the case of merging technology and human interaction avoiding technocentrism and ultimately foster positive health outcomes within our community.

5.1.3 HI-5 Bucket 3: Cost-Effectiveness and Cost Savings:

We will conduct a comprehensive evaluation of the cost-effectiveness of our community programs compared to healthcare costs associated with diabetes complications. This assessment will involve analyzing permitted medical records, counseling data, and insights derived from our dataset along with our intervention program cost. Collaboration with insurance firms will enable us to obtain accurate healthcare cost data related to diabetes diagnosis and management.

Program outcomes will inform advocacy efforts aimed at promoting policies that increase access to affordable healthy foods and physical activity opportunities. Based on identified factors significantly impacting diabetes risk, such as BMI, income level, and education, policy recommendations will be tailored to address local community needs. Implementation of these policies will be phased, starting after the fourth year of the intervention, once a substantial sample group has demonstrated improvement.

Cultivating alliances with health insurers presents an opportunity to incentivize program engagement and proactive health management. Leveraging our updated data model, we can identify high-risk individuals, such as those aged 55-69, and propose tailored incentives, such as premium discounts or wellness rewards. By aligning with insurers, we aim to foster cost-effective prevention strategies and reduce long-term healthcare expenses associated with diabetes complications.

Overall, our research paper outlines a comprehensive approach to achieving positive health impacts, cost-effectiveness, and sustainable outcomes within a five-year timeframe. By integrating multidimensional interventions, leveraging community partnerships, and advocating for policy change, we aspire to create lasting improvements in diabetes prevention and management within our target population.

5.1.4 HI-5 Based Intervention Chosen:

Our project focusing on diabetes prevention and management, the most appropriate HI-5 intervention category would be:

Multi-Component Worksite Obesity Prevention

Our interventions align well with the multi-component approach to obesity prevention. According to the intervention, strategies at the workplace include information and education, behavioral and social strategies, environmental components, and financial incentives, which is much closer to our proposed intervention, wherein it is community based, includes education workshop, support group, affordable fitness, health screening camps, and financial assistance to prevent diabetes. Obesity and diabetes share many common risk factors, and interventions targeting obesity often overlap with those for diabetes prevention and management.

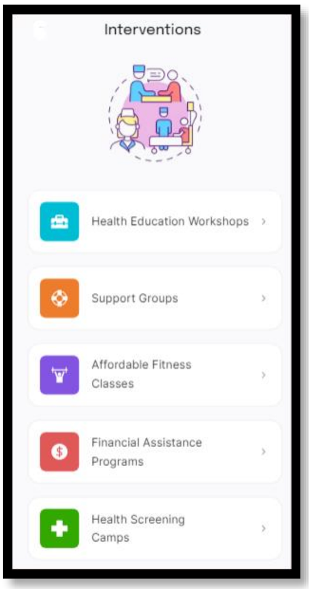

Interventions
Health Education Workshops
Support Groups
Affordable Fitness Classes
Financial Assistance Programs
Health Screening Camps

Figure 7: Intervention Web/ App Prototype

6 Application of CDC Population Health Competencies to Our Intervention Program

6.1 Assessing Health Status of Populations:

Our intervention program begins with a thorough assessment of the health status of our target population using CDC's data source on diabetes health indicator data, which comprises of people information, lab results, surveys. By analyzing these data sources, we identify high-risk individuals within major cities across the United States who are aged 55-69, have a lower educational background, report less than fair general health, earn less than $16,000 per year, have a BMI between 26 and 40, and are diagnosed with high blood pressure. This comprehensive assessment allows us to tailor our interventions to meet the specific needs of our population.

6. 2 Socioeconomic, Environmental, and Cultural Determinants:

Our program recognizes the significant impact of socioeconomic, environmental, cultural, and other population-level determinants on health. For instance, individuals with lower educational attainment may have limited access to health information and resources, impacting their ability to manage their health effectively. Similarly, low-income individuals may face barriers to accessing nutritious food and healthcare services, which are critical for diabetes prevention and management. By addressing these determinants through our community-based interventions, we aim to reduce health disparities and improve health outcomes.

6.3 Integrating Biologic and Genetic Risk with Population-Level Factors:

While our primary focus is on population-level factors, we also consider emerging information on individuals' biologic and genetic risks when developing prevention and treatment strategies. For example, we acknowledge that age-related factors and High Blood Pressure Condition increase the susceptibility to diabetes in our target population. Our health education workshops and support groups incorporate this knowledge to provide personalized advice and strategies for managing diabetes risk.

6.4 Appraising Quality of Evidence:

We base our interventions on a thorough appraisal of the quality of evidence from peer-reviewed medical and public health literature. This evidence informs our program design and ensures that our strategies are grounded in proven practices. For instance, our health education workshops and support groups are modeled on evidence-based approaches that have demonstrated effectiveness in promoting healthy lifestyle changes and improving disease management.

6.5 Primary and Secondary Prevention Strategies:

Our program emphasizes both primary and secondary prevention strategies to improve the health of individuals and populations. Primary prevention efforts include health education workshops, affordable fitness classes, and financial assistance for accessing healthy foods and health services. Secondary prevention strategies focus on early detection and management of diabetes through regular health screening camps and support groups for individuals already diagnosed with the condition. By integrating these strategies, we aim to reduce the incidence of diabetes and enhance disease management among diagnosed individuals.

6.6 Identifying Community Assets and Resources:

We actively identify and leverage community assets and resources to enhance the effectiveness of our interventions. Collaborations with local community centers, healthcare providers, and other organizations facilitate program delivery and participant recruitment. These partnerships provide access to venues for workshops, support groups, and fitness classes, as well as medical professionals for health screening camps. By utilizing existing community resources, we maximize the reach and impact of our program.

6.7 Community-Engagement Strategies:

Community engagement is a cornerstone of our intervention program. We employ strategies such as participatory planning, stakeholder consultations, and continuous feedback mechanisms to ensure that our program aligns with community needs and priorities. Engaging community members in the design and implementation of interventions helps to build trust, increase participation, and enhance the sustainability of our efforts. Our inclusive approach welcomes all individuals, regardless of whether they meet the high-risk criteria, ensuring that everyone benefits from our initiatives aimed at preventing diabetes and managing the condition.

Overall, our intervention program is a comprehensive, community-based effort designed to address the diabetes epidemic in major cities across the United States. By focusing on high-risk populations and employing evidence-based strategies, we aim to prevent the onset of diabetes and improve disease management. Our approach aligns with the CDC's Population Health Competencies, ensuring a robust and effective intervention that addresses the diverse needs of our target population. Through continuous assessment, community engagement, and leveraging local resources, we strive to create lasting improvements in health outcomes and reduce health disparities.

7 Data Engineering Plan for Diabetes Prevention and Management Initiative

Diabetes being a chronic condition with significant public health implications. This data engineering plan outlines the process of collecting, processing, analyzing, and utilizing data to drive effective diabetes prevention and management interventions. The plan focuses on ensuring deployability, scalability, reliability, and maintainability while integrating newly collected data with existing datasets.

7. 1 Data Collection:

Survey Data: Utilize Google Forms to gather demographic information, health-related data, and intervention feedback [Figure 8]].

Medical Checkup Data: Collaborate with healthcare providers who are performing regular screening for the participants to collect BMI, blood pressure, and blood glucose data using Wi-Fi-enabled medical devices, securely stored in an EHR system like Epic.

Patients' Health Records: Integrate with Epic EHR systems to access historical health data relevant to diabetes diagnoses and treatments.

Workshop Organizer Data: Collect workshop data using Google Forms or a custom-built platform to track attendance and outcomes.

Financial Assistance Data: Gather program data using Salesforce, focusing on eligibility, participation, and outcomes.

Participant Login Data: Implement Auth0 for secure participant login and activity tracking on the program website.

7.2 Data Transmission and Storage:

Cloud-Based Storage: Utilize Amazon S3 for scalable and secure storage of collected data.

Secure Data Transmission: Implement SSL/TLS encryption for secure data transmission over Ethernet and Wi-Fi networks, with AWS Direct Connect for dedicated connections.

Access Control: Manage access control using AWS IAM to restrict access to sensitive data stored in S3.

7.3 Data Preprocessing and Integration:

Data Cleaning: Use Python with Pandas and NumPy to clean and preprocess data, handling missing values and outliers.

Data Integration: Integrate newly collected data with existing datasets using Apache NiFi for ETL processes.

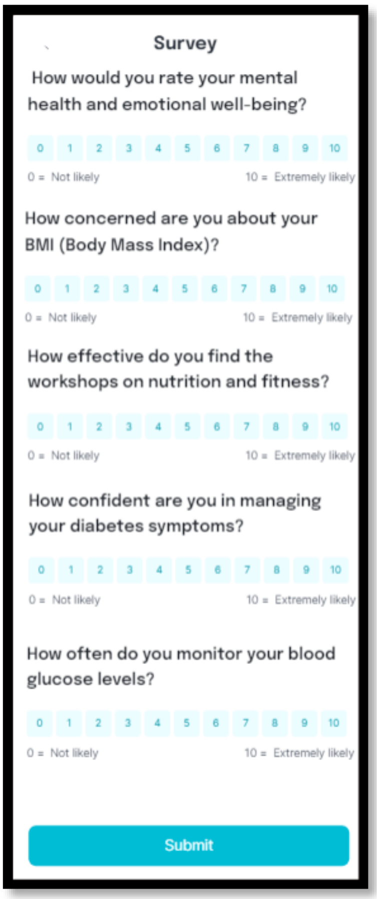
Data Enrichment: Enhance data with additional contextual information using external APIs like Google Maps and Census Bureau.

7.4 Data Analysis and Modeling:

Exploratory Data Analysis (EDA): Conduct EDA using Jupyter Notebooks or Advanced Python platforms such as Google Colab with Matplotlib and Seaborn to identify patterns and correlations specific to diabetes risk factors and intervention outcomes.

Predictive Modeling: Build machine learning models using scikit-learn to predict diabetes risk factors and assess intervention effectiveness. Apply in-database machine learning (ML), for predictive modeling and analysis directly within the database environment, if required.

Model Deployment: Deploy models using Docker containers orchestrated with Kubernetes for scalable and reliable deployment.



Survey

How would you rate your mental health and emotional well-being?

0 1 2 3 4 5 6 7 8 9 10

0 = Not likely 10 = Extremely likely

How concerned are you about your BMI (Body Mass Index)?

0 1 2 3 4 5 6 7 8 9 10

0 = Not likely 10 = Extremely likely

How effective do you find the workshops on nutrition and fitness?

0 1 2 3 4 5 6 7 8 9 10

0 = Not likely 10 = Extremely likely

How confident are you in managing your diabetes symptoms?

0 1 2 3 4 5 6 7 8 9 10

0 = Not likely 10 = Extremely likely

How often do you monitor your blood glucose levels?

0 1 2 3 4 5 6 7 8 9 10

0 = Not likely 10 = Extremely likely

Figure 8: Survey Data Collection - Sample Web/ App Prototype

7.5 Testing:

Unit Testing for Data Pipelines: Implement unit tests for each component of the data pipelines (data cleaning scripts, ETL processes in NiFi) to ensure they function correctly and handle edge cases gracefully.

Integration Testing: Test the integration of different data sources and systems (Epic EHR integration, Salesforce data integration) to verify data consistency and integrity throughout the pipeline.

Performance Testing: Conduct performance tests using simulated data loads to ensure our infrastructure (AWS S3 storage) can handle expected volumes without degradation.

Scalability Testing: Specifically test scalability of data processing pipelines (using Apache Spark) to ensure they can scale horizontally with increasing data volumes and user traffic.

7.6 Validation:

Data Quality Validation: Use tools like Apache Airflow or custom scripts to implement data quality checks (completeness, correctness, consistency) at different stages of data processing.

Model Validation: Validate machine learning models using techniques such as cross-validation, train/ test split, holdout validation, or A/B testing where applicable. Ensure models generalize well to new data and produce reliable predictions.

Ethical and Bias Evaluation: Evaluate models for biases and ethical considerations, especially in healthcare contexts, to ensure fairness and avoid unintended consequences.

User Acceptance Testing (UAT): Involve end-users (healthcare providers, program administrators) in UAT to validate that the implemented data-driven interventions meet their functional and usability requirements.

7.7 Monitoring and Maintenance:

Performance Monitoring: Monitor data pipelines and storage using Amazon CloudWatch for performance metrics and anomaly detection.

Data Quality Assurance: Implement data quality checks using Apache Airflow to ensure accuracy and completeness.

System Maintenance: Perform regular maintenance using Ansible for automated configuration management and updates.

7.8 Compliance and Security:

Regulatory Compliance: Ensure compliance with GDPR, HIPAA, and CCPA using AWS HIPAA Eligible Services.

Security Controls: Apply security best practices using AWS KMS and Inspector to safeguard sensitive data.

7.9 Network Infrastructure:

Cloud Infrastructure: Utilize AWS VPC for scalable and isolated network environments tailored to data transmission needs.

Connectivity Options: Use Ethernet for reliable, high-speed transmission and secure Wi-Fi networks for remote data collection locations.

7.10 Collaboration and Communication:

Project Management Tools: Utilize Jira for project management and Slack for real-time communication and collaboration.

7.11 Scalability and Flexibility:

Scalable Architecture: Design scalable data engineering pipelines using Apache Spark and Kafka to handle growing data volumes and user traffic.

Flexible Workflows: Implement flexible workflows to accommodate changes in data sources and processing requirements.

7.12 Utilization of Analyzed Data for Intervention Enhancement:

After completing the data engineering cycle and analyzing the collected data, it is crucial to leverage the insights gained to enhance diabetes prevention and management interventions effectively. Here are potential steps for utilizing the analyzed data:

1) Identify Key Modifiable Risk Factors: Utilize the findings from exploratory data analysis and predictive modeling to identify key modifiable risk factors contributing to diabetes prevalence and progression.
2) Tailor Intervention Strategies: Customize intervention strategies based on identified risk factors and target demographics. This could include developing targeted educational materials, lifestyle modification programs, or community outreach initiatives.
3) Personalized Patient Interventions: Implement personalized patient interventions by leveraging predictive models to identify individuals at high risk of developing diabetes or experiencing complications. Provide targeted support, counseling, and resources to these individuals to improve health outcomes.
4) Continuous Monitoring and Evaluation: Establish mechanisms for continuous monitoring and evaluation of intervention effectiveness using real-time data feedback loops. Track key

performance indicators (KPIs) such as changes in health behaviors, clinical outcomes, and program adherence to assess intervention impact and refine strategies over time.

5) Collaborative Decision-Making: Foster collaboration among healthcare providers, researchers, policymakers, and community stakeholders to inform decision-making processes regarding intervention design, implementation, and adaptation. Encourage multidisciplinary discussions and knowledge sharing to drive innovation and best practices in diabetes prevention and management.

By utilizing the analyzed data to inform intervention enhancement strategies, we can optimize the effectiveness of diabetes prevention and management initiatives, ultimately improving patient outcomes and reducing the burden of diabetes on individuals and communities. Initially, the data volume for this initiative is expected to be relatively low, given that it is an initiation program. However, if the program expands and the data volume increases significantly, we may consider leveraging Hadoop for distributed storage and processing capabilities to handle large-scale data efficiently.

1) Assessment of Data Volume: Regularly assess the volume of collected data as said in testing part as well to determine if it is approaching the capacity of existing storage and processing systems.
2) Evaluation of Expansion: If the data volume exceeds the capacity of current systems or is projected to increase significantly due to program expansion, consider the implementation of Hadoop.
3) Hadoop Integration: Integrate Hadoop into the existing data engineering pipeline to leverage its distributed storage and processing capabilities.
4) Data Migration: Transfer existing data to the Hadoop Distributed File System (HDFS) and modify data processing workflows to utilize Hadoop's MapReduce or Spark for parallel processing.
5) Scalability Testing: Conduct scalability testing to ensure that the Hadoop cluster can handle the increased data volume effectively.
6) Monitoring and Maintenance: Implement monitoring tools to track the performance of the Hadoop cluster and perform regular maintenance to optimize its efficiency.
7) Training and Documentation: Provide training for team members on Hadoop usage and update documentation to reflect changes in the data engineering workflow.
8) Continuous Assessment: Continuously assess the performance and scalability of the Hadoop cluster as data volume and processing requirements evolve over time.

Following the initial modeling and proposal of intervention strategies using our existing dataset, we recognize the imbalance within our data, where a larger proportion of individuals are considered healthy compared to those affected by diabetes. This initial dataset serves as the foundation for our proposed interventions. Subsequent to the implementation of our interventions, we aim to gather additional data to supplement our dataset. By leveraging the proposed data engineering plan, we can integrate the newly collected data with our existing dataset, thus mitigating the limitations posed by the initial imbalance. Through this iterative process, we can refine our intervention strategies based on a more comprehensive understanding of the diverse health profiles within the target population. This approach enables us to continuously improve our interventions, ensuring they are tailored to the needs of individuals at risk or affected by diabetes, ultimately leading to more effective health outcomes and enhanced population health management.

Thorough documentation and regular reporting will be necessitated to ensure clarity, facilitate effective communication, and support ongoing optimization of the diabetes prevention and management initiative

8 Limitations

While our intervention represents a comprehensive effort to address diabetes prevention and management, several limitations warrant consideration. Firstly, our reliance on self-reported data for certain demographic and health-related variables may introduce inherent biases and inaccuracies. Additionally, the generalizability of our findings may be limited by the specific characteristics of our target population and the geographic scope of our program. Furthermore, despite our best efforts to mitigate imbalance and improve model accuracy, the predictive power of our analytical models may be constrained by the complexity and multifactorial nature of diabetes risk. Notably, the dataset is predominantly composed of healthy individuals, which could skew our results and highlight the need for further data collection strategies to address this imbalance. Moreover, the effectiveness of our interventions may be influenced by external factors such as socioeconomic trends, policy changes, and cultural norms, which are beyond our control. Despite meticulous feature selection techniques and model choice, it is essential to acknowledge that feature importance scores, even for the most significant variables, may not be sufficiently high.

This project explores potential interventions to reduce diabetes risk in a high-needs population. We utilized a random forest model to analyze relevant health indicators and identify modifiable risk factors. The model, while a valuable starting point, has limitations. The dataset exhibits imbalance (falling outside the 80:20 rule), and the model achieves an accuracy of approximately 75% with a precision of 82%. Additionally, the model struggles with false negatives and false positives, particularly for individuals diagnosed with diabetes. We acknowledge that the model's accuracy is not optimal and requires further development. This will involve acquiring additional data to address the imbalance and improve the model's performance. Resampling techniques can only mitigate the issue to a certain extent. However, despite these limitations, the data analysis and empirical evidence gained from the model can still inform valuable intervention strategies. By prioritizing evidence-based decision-making, we aim to develop targeted strategies that maximize impact and improve health outcomes.

Additionally, the confusion matrix prediction issue, particularly prevalent in health-related datasets, underscores the need for ongoing refinement and evaluation of our models. With health data, achieving utmost accuracy is paramount, and we acknowledge the critical importance of selecting the most appropriate model for our intervention strategies. Importantly, the choice of our analytical model and feature selection technique is contingent upon the specific characteristics of our dataset and research objectives. As such, it is imperative to recognize that our findings and intervention strategies are based on the chosen model, which may evolve with the incorporation of more advanced or alternative modeling approaches. Lastly, while we have conducted extensive literature reviews and evidence-based research to inform our approach, ongoing evaluation and refinement are essential to address emerging challenges and opportunities in the field of diabetes prevention and management.

While we identified associations between predictors and the target variable, correlation does not imply causation. The relationships observed do not necessarily indicate causal effects. Although the random forest model is powerful and flexible, it has several limitations. One key limitation is that it can be difficult to interpret. While the model provides variable importance scores, it does not provide clear insights into the nature of the relationships between predictors and the target variable. Additionally, random forests can be computationally intensive, especially with large datasets and many trees, and may require significant resources for training and prediction. To address these limitations, future work should consider complementing random forest models with more interpretable models, to better understand feature contributions. Nonetheless, the model's ability to handle complex interactions and non-linear relationships provided valuable insights that informed our interventions. Despite these limitations, our commitment to data-driven decision-making, community engagement, and continuous improvement positions us to overcome obstacles and achieve meaningful impact in the fight against diabetes.

9 Future Project Extensions

In this analysis, we primarily focused on the relationship between the target variable (diabetes status) and individual predictor variables. While this approach provides valuable insights into the direct associations between predictors and the target variable, it has several limitations and opportunities for further improvement. We have to extensively explore the relationships between the predictor variables themselves. Understanding how features interact with each other can provide deeper insights and help identify confounding variables or interactions that might influence the target variable.

Future work would include a detailed analysis of the relationships between features. This can be achieved through visualizations such as pair plots, correlation matrices, and network diagrams to identify potential interactions and dependencies. Comprehensive Statistical Analysis, while we employed some descriptive statistics, there is room for a more comprehensive set of statistical analyses. An extended will include applying appropriate statistical tests such as Chi-square tests for independence, t-tests for comparing means between groups, and ANOVA for comparing means across multiple groups to uncover more specific associations and insights.

Future work would also incorporate advanced visualizations that highlight the relationships between features and the target variable, as well as among the features themselves. To better understand the structure of the data and identify potential subgroups within the dataset, future work should include dimensionality reduction techniques such as Principal Component Analysis (PCA). PCA can help reduce the complexity of the data by identifying the principal components that explain the most variance. This, in turn, can aid in identifying clusters of similar data points, which could be beneficial for targeted interventions or personalized treatments. Also, these clusters can provide valuable information for segmenting the population and understanding different subgroups, which can inform more tailored public health strategies and interventions.

10 Ethical Considerations in Diabetes Prevention and Management Initiative

Implementing a robust data engineering plan for diabetes prevention and management demands rigorous adherence to ethical principles to safeguard participant rights, ensure data integrity, and promote fairness throughout the initiative. Obtaining explicit consent from participants is crucial to respecting their autonomy and ensuring transparency in data collection practices. It involves clearly communicating the purpose, scope, and potential uses of collected data to build trust and foster informed decision-making among participants. To prevent control creep—where data usage extends beyond initial consent or unauthorized expansion of data collection—our initiative emphasizes strict governance frameworks. Regular audits will ensure that data usage aligns with agreed-upon boundaries, protecting participant privacy and maintaining ethical standards. Emphasizing data quality through rigorous validation and cleaning processes is paramount. High-quality data minimizes the risk of misleading conclusions and enhances the effectiveness of interventions based on accurate insights into diabetes risk factors and outcomes.

Safeguarding against data leaks and breaches is imperative. We deploy strong encryption, access controls, and conduct regular security assessments to protect participant information throughout its lifecycle—from collection through storage and analysis. Ensuring user-friendly interfaces and accessible tools for data collection enhances participant engagement and data accuracy. HCI principles guide the design of our digital platforms to facilitate seamless interaction while respecting participant privacy and usability. While leveraging technology for personalized interventions and predictive analytics, we mitigate risks such as discriminatory pricing, algorithmic bias, and unintended consequences. Ethical AI design and diverse data representation are integral to promoting fairness and equity in healthcare delivery. Proactively identifying and managing threats associated with data handling—such as digital footprints and surveillance implications—demonstrates our commitment to responsible data stewardship. Clear policies on threat ownership and mitigation strategies ensure participant trust and data security. Guarding against bias in data collection, analysis, and decision-making processes is critical. We implement bias detection algorithms and ensure diverse representation in our datasets to mitigate the risk of unfair treatment based on demographic or health characteristics. Recognizing the impact of digital footprints on participant privacy, we prioritize transparency in data collection practices. Mitigating the panopticon effect—where continuous surveillance erodes trust—requires ethical oversight and participant empowerment throughout our data-driven interventions. Overall, integrating these ethical considerations into our diabetes prevention and management initiative is essential. By upholding principles of informed consent, data integrity, security, fairness, and participant privacy, we ensure that our initiative maximizes positive health outcomes while minimizing potential harms. This ethical foundation not only strengthens participant trust but also sets a precedent for responsible and impactful healthcare interventions.

11 Conclusion

In conclusion, our targeted intervention for diabetes prevention and management embodies a holistic approach to addressing the complex challenges faced by individuals across diverse demographics. By targeting specific risk factors and leveraging community-based strategies, we aim to empower individuals with the knowledge, resources, and support needed to combat diabetes effectively. This project is driven by empirical insights, emphasizing rigorous evaluation and optimization of intervention strategies. Moving forward, our commitment remains steadfast in conducting future endeavors without bias or pre-assumptions. By adhering to these principles, we aim to continually enhance our approach to diabetes prevention and management, ensuring that interventions are evidence-based and effective in improving patient outcomes. Through rigorous data analysis, evidence-based practices, and collaboration with stakeholders, our program seeks to achieve sustainable improvements in health outcomes and reduce disparities in diabetes prevalence. Furthermore, our commitment to continuous refinement and adaptation ensures that our interventions remain responsive to evolving community needs and emerging evidence. As we embark on this journey to promote health equity and enhance population health, we recognize the importance of ongoing collaboration, innovation, and engagement. Together, we can make significant strides in preventing diabetes, improving disease management, and ultimately fostering a healthier future for all.

12 References

[1] Centers for Disease Control and Prevention. (2023, September 25). CDC diabetes health indicators. https://www.archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators

[2] Rajesh,S. (20024, June 13). Detailed Code Book of 8 Selected Features. GitHub. https://github.com/shrinirajesh05/HDS_Final_Project/blob/main/Detailed%20Code%20Book%20of%208%20Selected%20Fe

[3] American Diabetes Association Professional Practice Committee. (2024). Prevention or delay of diabetes and associated comorbidities: Standards of care in diabetes—2024. Diabetes Care, 47(Supplement_1), S43–S51. https://diabetesjournals.org/care/article/47/Supplement_1/S43/153945

[4] Tomic, D., Shaw, J. E., & Magliano, D. J. (2022). The burden and risks of emerging complications of diabetes mellitus. Nature Reviews Endocrinology, 18(6), 525–539. https://www.nature.com/articles/s41574-022-00690-7

[5] Habehh, H., & Gohel, S. (2021). Machine learning in healthcare. Current Genomics, 22(4), 291–300. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8822225/

[6] DeSalvo, K. B., Wang, Y. C., Harris, A., Auerbach, J., Koo, D., & O'Carroll, P. (2017). Public Health 3.0: A call to action for public health to meet the challenges of the 21st century. Preventing Chronic Disease, 14. https://www.cdc.gov/pcd/issues/2017/17_0017.htm

[7] Khalifa, M., & Albadawy, M. (2024). Artificial intelligence for diabetes: Enhancing prevention, diagnosis, and effective management. Artificial Intelligence in Medicine, 5, Article 100141. https://www.sciencedirect.com/science/article/pii/S2666990024000089#:~:text=In%20this%20domain%2C%20Artificial%20Intelligence,care%20and%20outcomes%20%5B8%5D

[8] Carney, T. J., Wiltz, J. L., Davis, K., Briss, P. A., & Hacker, K. (2023). Advancing chronic disease practice through the CDC Data Modernization Initiative. Preventing Chronic Disease, 20. https://www.cdc.gov/pcd/issues/2023/23_0120.htm

[9] Rajesh, S. (2024). HDS_Output_File. GitHub. https://github.com/shrinirajesh05/HDS_Final_Project/blob/main/HDS_Output_File.ipynb

[10] Rajesh, S. (2024). Detailed Literature Reviews. GitHub. https://github.com/shrinirajesh05/HDS_Final_Project/blob/main/Detailed%20Literature%20Reviews.docx

[11] Rajesh, S. (2024). Detailed Code Book of 8 Selected Features and Target Variables.docx. GitHub. https://github.com/shrinirajesh05/HDS_Final_Project/blob/main/Detailed%20Code%20Book%20of%208%20Selected%20Features%20and%20Target%20Variables.docx

[12] Rajesh, S. (2024). HDS_Code.ipynb. GitHub. https://github.com/shrinirajesh05/HDS_Final_Project/blob/main/HDS_Code.ipynb

[13] Rajesh, S. (2024). Factors Affecting Diabetes. https://public.tableau.com/app/profile/shrinidhi.rajesh/viz/FactorsAffectingDiabetes/FactorsAffectingDiabetes

[14] Patel, K. (2024). Intervention WebPage Prototype. Visily. https://app.visily.ai/projects/b4cda24b-76ea-4531-8430-31116169d7bf/boards/1003769/presenter