Predictive Analytics for Diabetes Risk Assessment Using Machine Learning on BRFSS Health Indicators

Andres De Los Santos and Gilber Hernandez
Computer and Information Science Department
Fordham University
New York City, United States
{ald, gmh1}@fordham.edu

***Abstract*— The sedentary American lifestyle has significantly contributed to the increasing rate of diabetes cases. According to the CDC's National Diabetes Statistics Report of 2022, an estimated 37.3 million Americans have diabetes, representing 11.3% of the U.S. population. Of these, approximately 28.7 million have been diagnosed, while an estimated 8.6 million remain undiagnosed. Diabetes is characterized by elevated blood sugar levels due to the body's inability to use or store glucose properly. This study aims to perform a predictive analysis for diabetes risk assessment using machine learning techniques and to identify decisive risk factors contributing to diabetes prevalence.**

Utilizing data from the 2015 Behavioral Risk Factors Surveillance System (BRFSS), we developed multiple machine learning models, including logistic regression, decision trees, random forests, XGBoost, and K-Nearest Neighbors (KNN). The analysis revealed significant predictors of diabetes, such as Body Mass Index (BMI), physical activity levels, and hypertension. Identifying these key risk factors is crucial for guiding targeted interventions and policy recommendations to enhance diabetes prevention efforts. Through machine learning techniques, this study offers valuable insights into the early detection and prevention of diabetes, emphasizing the importance of addressing lifestyle factors to mitigate this growing health issue.

Keywords— type 2 diabetes, machine learning, predictive analysis, risk factors, Health Surveys Data.

I. INTRODUCTION.

Diabetes is a prevalent metabolic disorder characterized by elevated blood sugar levels resulting from insulin resistance and impaired glucose metabolism. According to the International Diabetes Federation (IDF), approximately 537 million adults aged 20-79 were affected by diabetes in 2021, with projections indicating a 46% increase to 784 million by 2045. In the United States alone, over 133 million individuals are affected by diabetes or prediabetes. As of 2019, 37.3 million people, representing 11.3% of the U.S. population, had diabetes, with over one-quarter of adults aged 65 and older being affected. Alarmingly, nearly 25% of adults with diabetes remain undiagnosed. Type 2 diabetes (T2DM) constitutes 90% to 95% of all diabetes cases.

Diabetes risk factors are categorized into modifiable and non-modifiable groups. Modifiable risk factors include overweight, obesity, physical inactivity, smoking, poor dietary habits, and certain medications. Non-modifiable risk factors encompass age, race/ethnicity, family history, and gestational diabetes mellitus (Naranjo et al., 2021). Preventive strategies recommended by the Pan American Health Organization (PAHO) emphasize maintaining a healthy body weight, engaging in regular physical activity, adhering to a balanced diet, limiting sugar and saturated fat intake, and avoiding tobacco use (PAHO, 2021).

Machine learning (ML) techniques have proven valuable in analyzing large datasets and uncovering complex patterns and relationships in healthcare research. Algorithms such as Decision Trees, Random Forests, Gaussian Naïve Bayes, XGBoost, and Logistic Regression have shown promise for predicting disease outcomes and identifying critical risk factors.

Despite significant advancements in diabetes research, there remains a need for robust predictive models to accurately assess the risk of developing type 2 diabetes. This study addresses this gap by utilizing the Behavioral Risk Factor Surveillance System (BRFSS) Health Indicators datasets and various machine learning models to develop predictive models for diabetes risk assessment. The research focuses on constructing and evaluating models—including K-Nearest Neighbors (KNN), Decision Trees, Naive Bayes, Random Forest, Linear Regression, Gaussian Naïve Bayes, XGBoost, and Logistic Regression—to detect the likelihood of diabetes onset and identify the most relevant risk factors. Additionally, K-modes clustering analysis will be employed to identify key characteristics distinguishing individuals with diabetes from those without, enhancing the understanding of diabetes risk and informing preventive strategies.

A. Background.

Diabetes, particularly type 2 diabetes (T2DM), has become a significant public health challenge due to its rising prevalence and severe complications. Type 2 diabetes is primarily driven by lifestyle factors such as a genetic predisposition, making it a complex condition to manage and prevent. The disease leads to severe complications, including cardiovascular disease, nerve damage, kidney failure, and vision problems, substantially impacting individuals' quality of life and posing a significant burden on healthcare systems worldwide.

According to the American Diabetes Association (ADA), diabetes is classified into several types based on its underlying causes and characteristics:

- **Prediabetes:** elevated blood sugar levels that are not yet high enough to be classified as diabetes.
- **Type 1 Diabetes (T1DM):** An autoimmune condition where the immune system destroys insulin-producing

cells in the pancreas, typically diagnosed in children and young adults.

- **Type 2 Diabetes (T2DM):** Characterized by insulin resistance and insufficient insulin production. It is the most common form of diabetes and can develop at any age, often linked to overweight, obesity, and a family history of diabetes. Lifestyle changes, such as weight management and increased physical activity, can help prevent T2DM.
- **Gestational Diabetes:** occurs during pregnancy and usually resolves after childbirth but increases the risk of developing T2DM later in life.
- **Other Types:** includes monogenic diabetes caused by genetic mutations and diabetes resulting from pancreatic damage due to conditions like cystic fibrosis or pancreatitis.

The chances of developing type 2 diabetes depend on a combination of risk factors. Although, you can't change risk factors related to family history, age, race, or ethnicity, you may be able to avoid some risk factors by maintaining a healthy weight and being physically active. Type 2 diabetes can be developed at any age, even during childhood. The disease is likely to develop if you become overweight or obese, have 35 years or more, family history with diabetes, ethnicity, not physically active, have prediabetes, history of gestational diabetes, among others.

The Behavioral Risk Factor Surveillance System (BRFSS) is one of the largest ongoing health surveys globally, conducted by the Centers for Disease Control and Prevention (CDC). Established in 1984, the BRFSS collects data on health-related risk behaviors, chronic health conditions, and the use of preventive services from the U.S. population. The comprehensive nature of the BRFSS dataset, which includes various health indicators, makes it an invaluable resource for public health research and policymaking.

Previous studies have utilized the BRFSS dataset to examine the prevalence of diabetes and associated risk factors. Research has consistently shown that factors such as high body mass index (BMI), physical inactivity, poor diet, and smoking are strongly correlated with an increased risk of developing type 2 diabetes. Additionally, demographic factors like age, race/ethnicity, and socioeconomic status have been identified as significant contributors to diabetes risk.

II. METHODOLOGY.

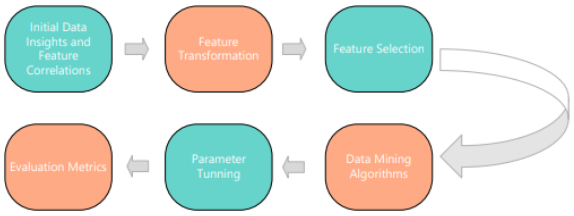

Figure 1. methodology flowchart

In this section, we outline the methodology employed to develop and evaluate Machine Learning (ML) models for predicting prediabetes or type 2 diabetes, using the 2015 BRFSS dataset. Our approach encompasses several key stages, starting with a detailed description of the dataset, followed by obtaining initial insights and feature correlation, to understand the underlying patterns and characteristics of the data. We then proceed with feature transformation and selection, crucial steps to enhance the model's performance and interpretability.

Additionally, we discuss the data mining algorithms applied in our study, highlighting their relevance and application in the context of diabetes prediction. Finally, we detail the assessment metrics and procedures used to evaluate the models, ensuring a comprehensive analysis of their accuracy, robustness, and clinical relevance. Through this methodology, we aim to establish a reliable framework for predicting type 2 diabetes.

A. Dataset description.

We retrieved the publicly available diabetes data from 2015, collected by the Behavioral Risk Factor Surveillance System (BRFSS) of the Centers for Disease Control and Prevention (CDC) of the United States. This dataset, sourced from Kaggle, has been pre-cleaned and transformed, including 21 features and a target variable related to diabetes. The original dataset comprises 330 features and a target variable. The data used in this research consists of 253,680 records, with 39,977 (15.7%) representing cases of individuals diagnosed with prediabetes or type 2 diabetes. The selection of this dataset is pivotal for our analysis due to its comprehensive nature and alignment with our research objectives on diabetes risk factors.

TABLE 1. DETAILED DESCRIPTION OF FEATURES.

Feature name	Description
HighBP	High Blood Pressure,
HighChol	High Cholesterol, categories: 0 = No high chol; 1 = High chol
CholCheck	Cholesterol Check in the last 5 years.
BMI*	Body Mass Index.
Smoker	Have you smoked at least 100 cigarettes in your entire life?
Stroke	(Ever told) you had a stroke.
HeartDiseaseorAttack	coronary heart disease (CHD) or myocardial infarction (MI).
PhysActivity	physical activity in past 30 days - not including job.
Fruits	Consume Fruit 1 or more times per day.
Veggies	Consume Vegetables 1 or more times per day.
HvyAlcoholConsump	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week).
AnyHealthcare	Have any kind of health care coverage.
NoDocbcCost	Needed to see a doctor but could not because of the cost, in the last 12 months?
GenHtlh*	General Health.
MentalHealth*	How many days during the past 30 days was your mental health not good?
PhysHlth*	how many days during the past 30 days was your physical health not good?
DiffWalk	Do you have serious difficulty walking or climbing stairs?
Sex	Sex
Age	Age
Education*	Education level
Income*	Income level.

** Non binary features. Please, visit BRFSS codebook 2015 Codebook Report.*

B. *Initial Data Insights and Feature Correlation*.

To thoroughly comprehend the dataset, we embarked on an exploratory data analysis (EDA) to examine the characteristics of the features, their data types, and their distributions. We initiated this process by importing the dataset and conducting initial inspections to understand its structure and data quality. Our preliminary steps involved addressing missing values and duplicates, which were managed through appropriate imputation techniques to maintain the integrity of our analysis. Subsequently, we employed histograms and boxplots to visualize the distribution of numerical features, which allowed us to identify the range, central tendencies, and the presence of outliers in the data.

As we delved deeper into the analysis, we focused on the relationships between features and the target variable, using different methods based on the nature of the variables. For continuous features, we computed Pearson correlations, while for binary features, we utilized the phi coefficient. This correlation analysis revealed that features such as BMI, General Health (GenHlth), Physical Health (Physhlth), and Age exhibited the highest correlations with the target variable, suggesting their significant role in predicting diabetes status. In contrast, binary variables showed very low correlations with the target variable, indicating they may have limited predictive power in their current form.

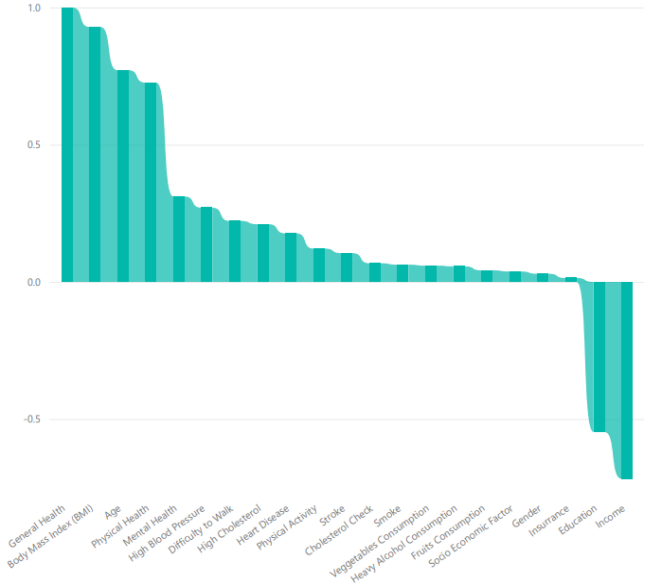

Figure 2. Correlation of features with the target variable.

A critical insight from our EDA was the distribution of the target variable, categorized into prediabetes (2%), diabetes (14%), and no diabetes (84%). This distribution highlighted a significant class imbalance, with a pronounced skew towards the "no diabetes" category. Addressing this imbalance is crucial for our modeling efforts, as it may impact the performance of predictive algorithms. Overall, our comprehensive EDA provided valuable insights into the dataset's structure, feature relationships, and distribution characteristics, laying a solid foundation for subsequent data modeling and analysis.

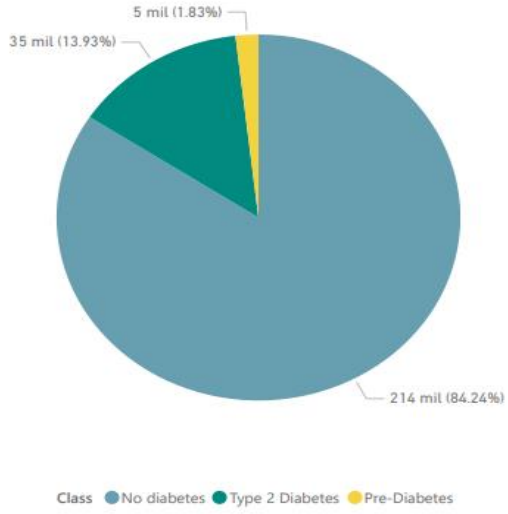

Figure 3. Distribution of the target variable.

C. *Feature transformation.*

After conducting the initial data insights and feature correlation section, we transformed several risks factors whose classifications were either non-binary or grouped into multiple categories. These variables included BMI, Age, Mental Health, Physical Health, Income, General Health, and Education. The goal of these transformations was to reduce these variables to a maximum of 4 or 5 categories to simplify the prediction models.

For Body Mass Index (BMI), which ranged from 13 to 98 in the dataset, we adopted the CDC classification system: underweight, healthy weight, overweight, and obesity. For income, the dataset originally contained clusters ranging from 1 to 13, with 1 representing low income (less than $10k per year) and 13 representing high income (more than $75k per year). This was transformed into the commonly accepted scale of low class, middle class, and high class, based on the 2018 article "The American middle class is stable in size, but losing ground financially to upper-income families" by Rakesh Kochhar from the Pew Research Center, which relies on U.S. Census Bureau studies.

Similarly, the dataset recorded the number of days of poor health per month for Mental Health and Physical Health, ranging from 0 to 30 days. We transformed these into categories: excellent, very good, good, fair, and poor. Additionally, we created a new column called "unhealthy days," representing the sum of days of poor physical or mental health over a month. This metric is widely used in economic studies related to health conditions, as implemented by the CDC.

For Age, we performed value mapping to assign different weights based on age ranges, recognizing that individuals aged 50 to 80 and older are more likely to develop diabetes than younger individuals. Furthermore, we transformed the Age feature by creating an interaction term with BMI, multiplying the two values. This aimed to strengthen the correlation with the target variable, as age alone does not provide sufficient predictive information. By incorporating BMI, we gain a more comprehensive understanding of an individual's health condition and their risk of developing diabetes, regardless of age.

To conclude, we combined class 1 (prediabetes) with class 2 (diabetes) into one class (class 1), to simplify the model. This

decision was made because the prediabetes class constituted only 1.8% of the total records, while the diabetes class accounted for 13.9%. Additionally, according to the American Diabetes Association (ADA), prediabetes serves as a warning sign for future diabetes, sharing similar risk factors with the condition. Moreover, given that this research is centered on predicting and identifying significant risk factors for type 2 diabetes, incorporating the prediabetes class label would introduce additional noise to our models. This is especially problematic considering the high imbalance rate already present in the target variable. This is particularly pertinent for a project seeking to prevent or analyze the stage's characteristics preceding the onset of diabetes.

D. Feature selection.

For feature selection, after transforming the data, we employed the XGBoost and Logistic Regression algorithms, using both Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS) methodologies. The aim was to identify features that contributed valuable information for predicting diabetes, with accuracy as the primary evaluation metric. We then selected the common features that demonstrated the highest accuracy across both methods.

Next, we evaluated these features against well-established risk factors for type 2 diabetes as identified in the literature. This step was essential to ensure that the selected features not only enhanced model performance but also aligned with recognized medical knowledge, thereby simplifying the prediction models. We identified and selected 14 key features: High Blood Pressure, High Cholesterol, Cholesterol Check, Smoker, Stroke, Heart Disease or Attack, Physical Activity, Heavy Alcohol Consumption, No Doctor Because of Cost, General Health, Difficulty to Walk, Sex, Age_Mapped, and BMI Status Number.

E. Data mining algorithms.

To predict type 2 diabetes using the 2015 BRFSS dataset, we evaluated several supervised machine learning classifiers, including Logistic Regression, Decision Tree, Random Forests, XGBoost, Gaussian Naïve Bayes, and K-Nearest Neighbors (KNN). The transformed dataset, which included the selected features, was divided into three distinct subsets: 70% for training, 15% for testing, and 15% for validation. The training set was used to develop and train the model, while the test set provided an initial evaluation of model performance. The validation set was employed to verify if our model was overfitted, and adjustments to hyperparameters were made based on these results to mitigate overfitting.

Given that only 16% of the dataset consisted of individuals with type 2 diabetes or prediabetes, we applied the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance. SMOTE increased the number of minority class instances (diabetes/prediabetes) to match the majority class (no diabetes), reducing potential model bias. We utilized Python and the Scikit-Learn and IMBlearn libraries to develop prediction models with the SMOTE-balanced training data.

Additionally, k-modes clustering analysis was employed to identify key characteristics distinguishing individuals with diabetes from those without. K-modes clustering is particularly suited for categorical data and helped uncover distinct clusters, enhancing our understanding of diabetes risk and informing preventive strategies.

To further optimize model performance, we employed 5-fold cross-validation. This technique involved dividing the training data into five subsets, training the model on four subsets, and validating it on the remaining subset. This process was repeated five times, providing a robust estimate of the model's generalizability and reducing the risk of overfitting. Hyperparameter tuning was conducted using GridSearch, which systematically explored a range of hyperparameter values for each classifier to identify the optimal configuration, thus enhancing model accuracy and performance.

F. Evaluation metrics.

For model evaluation, we calculated a range of metrics, including accuracy, precision, recall (sensitivity), F1-score, support, and Area Under the Curve (AUC). Each metric provides unique insights into the performance of our models:

- **Accuracy:** measures the proportion of correctly classified instances (both positive and negative) out of the total number of instances. It is calculated as:

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Instances}$$

- **Precision:** measures the proportion of true positive predictions among all positive predictions made by the model. It is calculated as:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

- **Recall (Sensitivity):** measures the proportion of true positive predictions among all actual positive cases. It is calculated as:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

- **F1-Score:** is the harmonic mean of precision and recall, providing a single metric that balances both aspects. It is calculated as:

$$F1 - Score = 2\ x\ \frac{Precision\ x\ Recall}{Precision + Recall}$$

- **Support:** refers to the number of actual occurrences of each class in the dataset. It is a measure of how many instances belong to each class and is used to weight the metrics accordingly. Support helps understand the distribution of classes and their impact on metric calculations.
- **Area Under the Curve (AUC):** measures the ability of the model to distinguish between classes across all threshold levels. It is derived from the Receiver Operating Characteristics (ROC) curve, which plots the true positive rate (recall) against the false positive rate at various thresholds. AUC values range from 0 to 1, where a higher AUC indicates better model performance. AUC provides a summary measure of the model's discriminatory power.

We placed particular emphasis on **recall**, **F1-score**, and **AUC** for several reasons:

- **Recall** is critical for evaluating how well the model detects all relevant cases of type 2 diabetes and prediabetes, as missing these cases could have serious health implications.
- **F1-score** balances precision and recall, offering a comprehensive view of the model's performance, particularly in scenarios where both false positives and false negatives are of concern.
- **AUC** assesses the model's overall ability to discriminate between positive and negative cases, providing insight into its performance across all possible classification thresholds.

By focusing on these metrics, we aimed to develop models that are highly effective in identifying individuals at risk for type 2 diabetes, ensuring both comprehensive detection and reliable classification.

III. Results

The performance for the 6 machine learning algorithms implemented for the prediction of type 2 diabetes, was evaluated using a range of metrics, including accuracy, Area Under the Curve (AUC-ROC), precision, recall, F1-Score, and support. The results for each algorithm are summarized in the following table:

TABLE 2. PERFORMANCE METRICS FOR EACH MODEL.

Model	Recall	F1-Score	Precision	Accuracy	AUC
KNN	0.42	0.40	0.39	0.81	0.75
Decision Tree	0.79	0.45	0.32	0.70	0.80
Logistic Regression	0.76	0.47	0.34	0.72	0.81
XGBoost	0.78	0.46	0.33	0.71	0.82
Gaussian Naïve Bayes	0.82	0.43	0.29	0.66	0.79
Random Forest	0.88	0.42	0.28	0.62	0.82

A. *Best parameters using Gridsearch.*

To optimize the performance of the implemented machine learning models, a comprehensive hyperparameter tuning process was conducted using GridSearch. The optimal hyperparameters for each model were determined based on cross-validation results. Table 3 summarizes the best parameters identified for all the models implemented in the study.

TABLE 3. OPTIMAL HYPERPARAMETERS IDENTIFIED VIA GRIDSEARCH.

Model	Parameters	Values
KNN	Cross Validation	3
	K	50
	Distance	Euclidean
	Smote	Yes
	Weight	Uniform
Decision Tree	Cross Validation	5
	Criteria	Entropy
	Max_depth	5
	Min_samples_leaf	1
	Min_samples_split	2
	Smote	Yes
Logistic Regression	CV	5
	Regularization	L1
	Inverse of Regularization	0.08
	Solver	Liblinear
	Class_weight	None
	Smote	Yes
XGBoost	CV	5
	Alpha	0
	Colsample_bytree	0.7
	Eta	0.1
	Gamma	0.1
	Lambda	1
	Learning_rate	0.1
	Max_depth	5
	Min_child_weight	3
	N_estimators	100
	Scale_pos_weight	5.345705086
	Subsample	0.7
Gaussian Naïve Bayes	CV	5
	Nb_var_smoothing	1.00E-09
	Preprocessor_scaler	StandardScaler
	Smote	Yes
Random Forest	CV	5
	Max_depth	10
	Min_samples_leaf	1
	Max_features	Sqrt
	Min_samples_split	2
	N_estimators	50
	Smote	Yes

B. *Accuracy and AUC-ROC.*

The performance of the machine learning models was assessed based on accuracy and AUC. The K-Nearest Neighbors (KNN) algorithm achieved the highest accuracy of 80%, indicating its strong capability in correctly classifying instances of diabetes and non-diabetes. Additionally, KNN recorded an AUC of 0.75, demonstrating a solid ability to distinguish between the two classes. Gaussian Naïve Bayes exhibited a robust AUC of 0.79, indicating effective class separation, although its accuracy was relatively lower at 66%. Logistic Regression performed notably well, with an accuracy of 76% and an AUC of 0.81, underscoring its superior ability to differentiate between diabetic and non-diabetic cases. XGBoost, comparable to the Decision Tree model, achieved an accuracy of 78% and an AUC of 0.82, reflecting its efficient performance in classifying diabetes status.

C. *Precision, Recall, and f1-Score:*

In terms of precision, Recall, and F1-Score, there were distinct performance variations across models. The Random Forest model achieved the highest recall at 0.88, highlighting its effectiveness in identifying true positive cases of diabetes. However, it had the lowest precision at 0.28, indicating a higher rate of false positives. Logistic Regression, while having a lower recall of 0.76, demonstrated the highest F1-Score of 0.47, reflecting a balanced performance between precision and recall. The XGBoost model, with a recall of 0.78 and an F1-Score of 0.46, also showed a strong balance in performance. Gaussian Naïve Bayes had a recall of 0.81 but a lower precision of 0.29, which impacted its F1-Score. The Decision Tree model, despite its good recall of 0.79 and an F1-Score of 0.45, exhibited an overall stable performance.

D. Cluster Analysis.

The k-mode cluster analysis revealed distinct characteristics among individuals with and without diabetes, highlighting significant differences in health indicators and behaviors. Cluster 0, representing non-diabetic individuals, showed lower rates of high blood pressure (HighBP) and high cholesterol (HighChol), higher physical activity, and lower heavy alcohol consumption. These individuals exhibited moderate general health, less difficulty walking, and intermediate values for Age_Mapped and BMI_Status_Number. Diabetes prevalence in this cluster was 11.6%. Cluster 1, also non-diabetic, featured the lowest rates of HighBP and HighChol, higher physical activity, but slightly worse general health and more difficulty walking compared to Cluster 0. This cluster was younger with similar BMI_Status_Number, having a diabetes prevalence of 13.4%. Cluster 2, with the highest diabetes prevalence of 27.1%, exhibited the highest rates of HighBP and HighChol, lowest physical activity, slightly higher heavy alcohol consumption, worst general health, highest difficulty walking, oldest Age_Mapped, and highest BMI_Status_Number.

When focusing on diabetic clusters, Cluster 0 (Diabetic) showed higher rates of HighBP and HighChol compared to non-diabetics, with high physical activity but significant heavy alcohol consumption. This cluster also had worse general health, significant difficulty walking, a higher proportion of females, older Age_Mapped, and higher BMI_Status_Number, with a diabetes prevalence of 100%. Cluster 1 (Diabetic) had lower rates of HighBP and HighChol compared to Cluster 0 diabetics, lower physical activity, less heavy alcohol consumption, worse general health, more difficulty walking, higher proportion of females, slightly younger Age_Mapped but higher BMI_Status_Number, and a diabetes prevalence of 100%. Cluster 2 (Diabetic) featured the highest rates of HighBP and HighChol, higher physical activity than Cluster 1 but still low, similar general health and difficulty walking, balanced gender distribution, oldest Age_Mapped, and highest BMI_Status_Number, with a diabetes prevalence of 100%. The analysis indicates that diabetic clusters generally have higher rates of HighBP, HighChol, worse general health, lower physical activity, older age, higher BMI, and a higher proportion of females, underscoring critical risk factors for diabetes.

E. Discussion:

Despite the prevalent use of logistic regression and other traditional methods for predicting type 2 diabetes, this study offers a comprehensive analysis by employing a variety of machine learning algorithms. The models assessed include K-Nearest Neighbors (KNN), Decision Trees, Logistic Regression, XGBoost, Gaussian Naïve Bayes, and Random Forests. Their performance was evaluated based on metrics such as Recall, F1-Score, Precision, Accuracy, and AUC.

The Random Forest model demonstrated the highest Recall at 0.88, indicating its superior ability to identify individuals with type 2 diabetes. It also achieved a high AUC of 0.82, reflecting strong discriminatory power. XGBoost followed with a Recall of 0.78 and an AUC of 0.82, showing its effectiveness in predicting diabetes cases. Logistic Regression achieved a notable AUC of 0.81, reflecting its strong overall performance, but had a lower Recall of 0.76 compared to Random Forest and XGBoost. This suggests that while Logistic Regression excels at distinguishing between diabetic and non-diabetic individuals, it is somewhat less effective at identifying all individuals with diabetes.

On the other hand, the Gaussian Naïve Bayes model performed well with a high Recall of 0.82 and an AUC of 0.79. However, it exhibited lower Precision (0.29) and Accuracy (0.66) relative to other models, indicating that while it effectively identifies diabetes cases, it may also include more false positives. The Decision Tree model achieved a Recall of 0.79 and an Accuracy of 0.70, striking a balance between identifying diabetes cases and overall accuracy. KNN, while showing the highest Accuracy of 0.81, had a lower Recall of 0.42 and an F1-Score of 0.39, suggesting that although KNN is good at classifying the overall data, it is less effective at identifying all individuals with diabetes.

The results of this capstone project underscore the significant impact of various risk factors on predicting diabetes. By employing Decision Tree, Random Forest, XGBoost, and Logistic Regression, we gained valuable insights into feature importance. General Health, BMI, and Age were consistently highlighted as significant predictors across all models. Specifically, General Health was crucial in Logistic Regression and XGBoost, BMI in Random Forest and XGBoost, and Age in all models, reflecting its established link to diabetes risk. Regular cholesterol checks also emerged as important, indicating their role in monitoring health status.

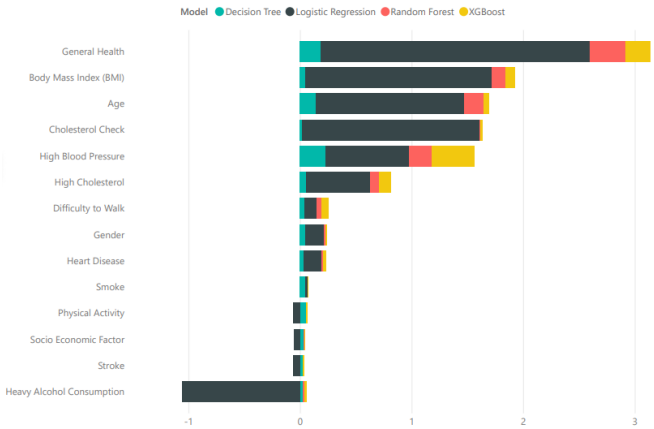

Figure 3. Comparative analysis of feature importance across different algorithms.

High Blood Pressure and High Cholesterol were influential across all models, reinforcing their critical role in diabetes prediction. Features like Difficulty in Walking and Gender showed varying importance, with Difficulty in Walking being notable in Random Forest and Gender in XGBoost and Logistic Regression. Overall, the consistency in feature importance across models strengthens the reliability of these predictors and underscores their value in developing targeted diabetes screening and prevention strategies.

For future work and model enhancement, it is essential to incorporate additional data sources beyond national surveys. Integrating laboratory results, clinical measurements, and other health indicators could significantly improve the accuracy and robustness of the predictive models. Expanding the dataset to include diverse and comprehensive health information would enable more precise risk assessment and personalized intervention strategies. This approach could enhance our ability to identify diabetes risk factors more accurately and provide more effective early detection and prevention measures.

IV. Related work

The application of machine learning techniques to predict type 2 diabetes risk has garnered significant attention in recent years. This section reviews key studies that have contributed to this field, highlighting their methodologies, findings, and implications.

Xie et al. (2019) developed risk prediction models for type 2 diabetes using various machine learning techniques in their study published in *Preventing Chronic Disease*. They utilized demographic data, lifestyle factors, and medical history to train models, including logistic regression and decision trees. Their research demonstrates the effectiveness of data-driven approaches in diabetes risk assessment, highlighting how machine learning can identify high-risk individuals with improved precision.

Mujumdar and Vaidehi (2020) investigated multiple machine learning algorithms, such as Support Vector Machines (SVM), Random Forests, and K-Nearest Neighbors (KNN), for diabetes prediction. Their study, conducted at the Vellore Institute of Technology, underscores the importance of algorithm selection and feature engineering in enhancing predictive accuracy. The comparative analysis provided insights into the strengths and limitations of different models, emphasizing the need for appropriate algorithmic choices to achieve reliable predictions.

Dinh et al. (2019) employed a data-driven approach to predict both diabetes and cardiovascular disease in their study published in *BMC Medical Informatics and Decision Making*. They combined electronic health records with advanced machine learning algorithms, such as gradient boosting machines and neural networks. This research highlights the interconnected nature of chronic diseases and the benefits of integrated predictive models in addressing multiple health conditions simultaneously.

Tasin et al. (2023) explored diabetes prediction using machine learning and explainable AI techniques in their study published in *Healthcare Technology Letters*. Supported by the National Institutes of Health (NIH), their research emphasizes not only the accuracy of predictive models but also their interpretability. The use of explainable AI techniques provides transparency in the prediction process, facilitating better clinical decision-making by clarifying the "black-box" nature of machine learning models.

Okolo (2022), in a study conducted at the University of Louisiana at Lafayette, examined various machine learning algorithms for diabetes prediction. The research highlights the critical role of feature engineering and model tuning in optimizing predictive performance. Documented in a comprehensive thesis, Okolo's work contributes to the understanding of advanced analytics in healthcare, emphasizing methods to refine predictive models for diabetes.

Islam et al. (2022) focused on identifying risk factors for type 2 diabetes and developing predictive models using machine learning techniques in their study published in *Health Systems*. They employed algorithms like logistic regression, SVM, and ensemble methods to analyze a broad range of risk factors, including genetic, behavioral, and environmental variables. This study provides a thorough overview of diabetes predictors and evaluates the effectiveness of different models in risk prediction, offering valuable insights into diabetes risk factors and prediction methodologies.

V. Conclusion.

This study employed various machine learning techniques to predict type 2 diabetes risk using the 2015 BRFSS dataset, focusing on models such as Logistic Regression, Decision Trees, Random Forests, XGBoost, and K-Nearest Neighbors (KNN). Key predictors identified include Body Mass Index (BMI), physical activity, and hypertension, which align with established diabetes risk factors. The application of Synthetic Minority Over-sampling Technique (SMOTE) effectively balanced the dataset, enhancing model performance metrics like recall, F1-score, and AUC. Additionally, feature selection and transformation were crucial for improving model accuracy and interpretability.

K-modes clustering provided additional insights by differentiating between diabetes and non-diabetes individuals based on distinct characteristics. This study highlights the effectiveness of machine learning in diabetes risk prediction and underscores the importance of addressing lifestyle factors for diabetes prevention. The findings offer valuable implications for public health strategies and suggest that further research could refine these models and evaluate targeted interventions to mitigate diabetes risk.

VI. References.

[1] Xie, Z., Nikolayeva, O., Luo, J., & Li, D. (2019). Building risk prediction models for type 2 diabetes using machine learning techniques. *Preventing Chronic Disease, 16*, 190109. http://dx.doi.org/10.5888/pcd16.190109

[2] Mujumdar, A., & Vaidehi, V. (2020). Diabetes prediction using machine learning algorithms. *Vellore Institute of Technology, Chennai, India*. Available online 27 February 2020.

[3] Dinh, A., Miertschin, S., Young, A., & Mohanty, S. D. (2019). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Medical Informatics and Decision Making, 19*, 211. https://doi.org/10.1186/s12911-019-0918-5

[4] Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2023). Diabetes prediction using machine learning and

explainable AI techniques. *Healthcare Technology Letters, 10*(1-2), 1-10. https://doi.org/10.1049/htl2.12039

[5] Okolo, C. (2022). Diabetes prediction using machine learning algorithms. *University of Louisiana at Lafayette*. https://doi.org/10.13140/RG.2.2.25215.18084/2

[6] Islam, M. M., Rahman, M. J., Abedin, M. M., Ahammed, B., Ali, M., Ahmed, N. A. M. F., & Maniruzzaman, M. (2022). Identification of the risk factors of type 2 diabetes and its prediction using machine learning techniques. *Health Systems, 12*(2), 243-254. https://doi.org/10.1080/20476965.2022.2141141

[7] Ismail, L., Materwala, H., & Al Kaabi, J. (2021). Association of risk factors with type 2 diabetes: A systematic review. *Computational and Structural Biotechnology Journal, 19*, 1759-1785. https://doi.org/10.1016/j.csbj.2021.03.003

[8] Diabetes and High Blood Pressure. (n.d.). *Johns Hopkins Medicine*. Retrieved from https://www.hopkinsmedicine.org/health/conditions-and-diseases/diabetes/diabetes-and-high-blood-pressure

[9] Diabetes and hypertension: Connection, complications, risks. (n.d.). *Medical News Today*. Retrieved from https://www.medicalnewstoday.com/articles/317631

[10] Diabetes, Hypertension, and cardiovascular disease: Clinical Insights and Vascular Mechanisms. (n.d.). *National Institutes of Health*. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5648369/

[11] Cholesterol and Diabetes. (n.d.). *American Heart Association*. Retrieved from https://www.heart.org/en/health-topics/diabetes/prevention-treatment-of-diabetes/cholesterol-and-diabetes

[12] Cholesterol and Glucose Metabolism: What You Should Know. (n.d.). *Healthline*. Retrieved from https://www.healthline.com/health/diabetes/cholesterol-and-diabetes

[13] Cholesterol testing and results information. (n.d.). *Mount Sinai - New York*. Retrieved from https://www.mountsinai.org/health-library/tests/cholesterol-testing-and-results

[14] Health checks for people with diabetes. (n.d.). *American Diabetes Association*. Retrieved from https://www.diabetes.org/healthy-living/medication-treatments/checkup-exams

[15] Excess weight and type 2 diabetes. (n.d.). *HonorHealth*. Retrieved from https://www.honorhealth.com/healthy-living/excess-weight-and-type-2-diabetes

[16] Smoking and Diabetes. (n.d.). *Centers for Disease Control and Prevention*. Retrieved from https://www.cdc.gov/tobacco/campaign/tips/diseases/diabetes.html

[17] How smoking can increase the risk for and affect diabetes. (n.d.). *Food and Drug Administration*. Retrieved from https://www.fda.gov/tobacco-products/health-effects-tobacco-use/how-smoking-can-increase-risk-and-affect-diabetes

[18] Diabetes and Stroke: What Are the Connections? (n.d.). *National Institutes of Health*. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5676740/

[19] Pathophysiology of Physical Inactivity-Dependent Insulin Resistance: A Theoretical Mechanistic Review Emphasizing Clinical Evidence. (n.d.). *National Institutes of Health*. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4676078/

[20] National diabetes statistics report, 2024. *Centers for Disease Control and Prevention*. Retrieved from https://www.cdc.gov/diabetes/data/statistics-report/index.html

[21] Risk factors for type 2 diabetes. (n.d.). *National Institute of Diabetes and Digestive and Kidney Diseases*. Retrieved from https://www.niddk.nih.gov/health-information/diabetes/overview/risk-factors-type-2-diabetes

[22] National diabetes statistics report, 2022. *Centers for Disease Control and Prevention*. Retrieved from https://www.cdc.gov/diabetes/data/statistics-report/index.html

[23] Prevalence of both diagnosed and undiagnosed diabetes. *National diabetes statistics report, 2022*. Retrieved from https://www.cdc.gov/diabetes/data/statistics-report/index.html

[24] Methods. *National diabetes statistics report, 2022*. Retrieved from https://www.cdc.gov/diabetes/data/statistics-report/methods.html

[25] Prevalence of prediabetes among adults. *National diabetes statistics report, 2022*. Retrieved from https://www.cdc.gov/diabetes/data/statistics-report/index.html

[26] About Adult BMI. (n.d.). *Centers for Disease Control and Prevention*. Retrieved from https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html

[27] Middle class keeps its size, loses financial ground to upper-income tier. (2018). *Pew Research Center*. Retrieved from https://www.pewresearch.org/fact-tank/2018/09/06/middle-class-keeps-its-size-loses-financial-ground-to-upper-income-tier/

[28] The income you need to fall in America's lower, middle and upper classes. (2020). *Yahoo Finance*. Retrieved from https://www.yahoo.com/now/the-income-you-need-to-fall-in-americas-lower-middle-and-upper-

classes-find-out-where-you-rank-and-how-these-social-levels-are-defined-214759066.html
[29] The American middle class: Key facts, data, and trends since 1970. (2018). *Pew Research Center*. Retrieved from https://www.pewresearch.org/social-trends/2018/09/06/the-american-middle-class-is-stable-in-size-but-losing-ground-financially-to-upper-income-families/
[30] NCBI. (2021). A new risk prediction model for type 2 diabetes. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8642577/.
[31] NCBI. (2021). Machine learning approaches for diabetes prediction. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8702133/.
[32] Diabetology & Metabolic Syndrome. (2021). Systematic review of diabetes prediction models. Retrieved from https://dmsjournal.biomedcentral.com/articles/10.1186/s13098-021-00767-9.
[33] BMC Public Health. (2021). Impact of socioeconomic factors on diabetes. Retrieved from https://doi.org/10.1186/s42492-021-00097-7.
[34] MDPI Healthcare. (2021). Healthcare utilization and diabetes management. Retrieved from https://doi.org/10.3390/healthcare9121712.
[35] Diabetology & Metabolic Syndrome. (2021). Predictive analytics in diabetes care. Retrieved from https://doi.org/10.1186/s13098-021-00767-9.
[36] A. Teboul, *Diabetes Health Indicators Dataset*, Kaggle, 2023. Retrieved from https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset.