AMERICAN JOURNAL OF BOTANY AND BIOENGINEERING
https://biojournals.us/index.php/AJBB
ISSN: 2997-9331

Article

Data-Driven Population Health Analytics for Identifying High-Risk Groups and Health Disparities

Mahzabin Binte Rahman[1], Mohammad Yasin[2], Md Parvez Ahmed[3]
1. Master of Science in Business Analytics, Trine University, USA
2. Master of Business Administration (Information Technology Management), College of Business, Westcliff University, USA
3. Master of Science in Information Technology (Data Management and Analytics), Washington University of Science and Technology, USA

Citation: Rahman, M. B., Yasin, M., Ahmed, M. P. Data-Driven Population Health Analytics for Identifying High-Risk Groups and Health Disparities. American Journal Of Botany And Bioengineering 2024, 1(11), 58-82.

Received: 06th Oct 2024
Revised: 20th Oct 2024
Accepted: 05th Nov 2024
Published: 28th Nov 2024

Abstract: Practices of population health management are now more and more dependent on the mass levels of surveillance data to inform the identification of vulnerable groups and intervene in enduring health disparities. This research uses data on the Behavioral Risk Factor Surveillance System (BRFSS) collected during 2011 -2021 to provide a comprehensive analysis of the population in order to determine high-risk groups and explore socioeconomic and demographic disparities in health in the United States. This dataset contains more than 2.2 million records of aggregation of all 50 states, the District of Columbia, and U.S. territories, and contains behavioral risk factors, chronic health status, preventive service use, and social determinants of health. Through multivariate regression analysis, trend models and stratified demographic comparisons, this study is able to measure the prevalence of chronic diseases (diabetes, cardiovascular disease, obesity, and depression) as a function of income, education level, race/ethnicity, age cohorts, and geographic regions. This study also examines behavioral determinants which include smoking, physical inactivity and healthcare access limitation in order to evaluate their relation with negative health outcomes. Temporal analysis identifies health indicators changes in the last decade, the COVID-19 era, and assesses the changing patterns of disparity. The results will likely emphasize the disproportionate burden of diseases in low-income groups, racial and ethnic minorities, and people with lower education levels, which will support the importance of social determinants in health outcome determination. This study will help make evidence-based policy, intervention-based, and equitable resource distribution, as it combines information-driven population health analytical data with disparity-based evaluation. The findings are useful to the public health officials, health care systems and policymakers on reducing the number of preventable health inequities and enhancing the long term population health outcomes.

Keywords: Population Health Analytics, Health Disparities, High-Risk Groups, Behavioral Risk Factor Surveillance System (BRFSS), Social Determinants of Health and Chronic Disease Prevention

Introduction

A. Background

Population health has emerged as a primary concern in modern healthcare systems and focuses on the systematic enhancement of health outcomes of specific population groups but proactively lowers disparities between socially and economically disadvantaged groups [1].. In contrast to conventional clinical methods of individual treatment, the population health models amalgamate epidemiology, social determinants of health, and prevention methods to tackle the wider structural causes of illness. The growing access to big health surveillance data has dramatically changed how researchers and policymakers can use empirically driven and data-driven methods to identify a population at risk. The Behavioral Risk Factor Surveillance System (BRFSS) is one of the most inclusive and ongoing surveillance systems of public health that gathers yearly data on the behavior risk factors, chronic diseases, preventive care behavior, and socioeconomic indicators in all the states and territories in the United States. With the incorporation of advanced analytics into population health research, the stratification of risks, monitoring of trends, and disparity assessment will be more accurate [2]. With longitudinal data over many years, researchers will be able to assess the interaction of behavioral patterns, including smoking, physical inactivity and alcohol consumption with socioeconomic factors, such as income, education, and race/ethnicity to determine the outcome of chronic diseases. The increasing consideration of health equity in the national healthcare reform programs highlights the role of high-risk populations in statistical analysis [3]. Population health analytics based on data can improve knowledge of disease distribution, as well as facilitate specific interventions, evidence-based policymaking, and effective distribution of medical resources. With the leading role of chronic diseases in morbidity and mortality, the multidimensional surveillance-based examination is the key to the sustainable positive change in the population health measures.

B. Rationale for the Study

Regardless of technological advances and the growth of healthcare coverage programs, the United States has a significant health outcome disparity among socioeconomic, demographic, and geographic groups [4]. The low-income population, racial and ethnic minorities, and those with low education levels are disproportionately impacted by chronic conditions, including diabetes, cardiovascular disease, obesity, depression, and others. Such inequalities are not only dependent on biological factors but also on behavioral risks that can be modified and structural determinants including employment, access to healthcare, housing stability, and preventive services are used. Conventional approaches to the study of public health frequently utilize descriptive statistics, which can fail to capture multifaceted interactions between these determinants, which restricts the ability of the interventions [5]. Longitudinal data of the BRFSS in 2011-2021 offers a solid research basis to carry out extensive population-based study. This ten years of data reflects both pre-pandemic and post-pandemic trends, allowing us to analyze the changing health trends and structural inequalities. Through analytical methods based on data, such as multivariate regression and stratified comparative modeling, it is feasible to reveal the risk clusters which are hidden and measure the disparity in a finer way. With a structured analytical structure, it is possible to find demographic overlaps, e.g., low-income older adults or minority groups with low access to healthcare, which might be confronted with increased vulnerability. Care delivery systems are becoming more concerned with value and resource distributions that are driven by equity considerations [6]. The successful implementation requires proper identification of high-risk groups and quantifiable disparity indicators. This study addresses this requirement by combining the behavioral risk factors, chronic disease-based indicators, and socioeconomic determinants in the context of a single analytical approach. Through this, it intends to close the divide between big data surveillance and practical public health decisions, which will eventually enhance more balanced and data-oriented health decisions.

C. Problem Statement

Despite all this information of public health surveillance, a large discrepancy still exists in using sophisticated population health analytics to identify the risks and measure health inequalities in a time-dependent manner [7]. Most of the interventions carried out in the field of public health are not precisely targeted due to the inability to implement the behavioral, demographic and socioeconomic variables into an all-encompassing framework of analysis. Consequently, chronic disease and

healthcare disparities continue to exist regardless of policy interventions to promote equity [8]. The healthcare systems are at the risk of distributing resources inefficiently and developing insufficiently designed interventions without the help of rigorous, longitudinal, data-based assessment [9]. To fill this gap, analytical methods need to be organized and effective in identifying the risk groups and guiding evidence-based measures in health.

D. Objectives of the Study

This study aims to:

- Narrow down on the high risk populations that have high rates of chronic diseases.
- Review relationship between poor health and social determinants of health.
- Compare the trend in the prevalence of diseases between 2011 and 2021.
- Measure the inequalities based on income, education, race/ethnicity and age and geography.
- Offer evidence-based information to facilitate specific health action and policy development.

E. *Research* Questions

The research questions addressed in the study are the following major questions:

1. What types of demographic and socioeconomic populations demonstrate the most prevalence of chronic health conditions over the course of the study?
2. What role does behavioral risk factors and social determinants of health play in creating health disparities in disease prevalence and access to healthcare?
3. What are key patterns of temporal changes and inequality between 2011 and 2021 that can be used to develop specific policy interventions?

F. Significance of the Study

The study is a valuable contribution to the development of population health analytics since it systematically combines the large-scale surveillance data with structured disparity assessment to produce actionable understanding of public health [10]. Since the study uses longitudinal data provided by the Behavioral Risk Factor Surveillance System (BRFSS), it goes beyond descriptive reporting and uses analytical rigor to determine high-risk demographic and socioeconomic groups that are disproportionately burdened by chronic illnesses and lack of access to healthcare [11]. This identification is essential to shift the generalized approach to the comprehensive strategy of public health to the equity-based interventions. This study is important as it measures the extent of health disparities in terms of income, education, race/ethnicity, age, and geographic location during a decade. The study can supply empirical evidence on the interaction of behavioral risk factors and structural determinants to create disease burden using trend analysis and multivariate assessment [12. This evidence would be of great importance to policymakers and healthcare administrators who have limited financial resources to use but have to focus on vulnerable groups. The findings favor the accuracy of the public health methods by emphasizing intersectional disparities where multiple risk factors are pooled to worsen the disadvantage. This study explores the use of longitudinal surveillance systems in informing sustainable healthcare planning. Instead of adopting short-term or separate analyses, the application of the decade-spanning data will allow recognizing the long-term trends and new threats, such as those linked to socioeconomic changes and health crises in the population [13]. The analytical construct that is created in this study can also be utilized as a template to make a similar study on the disparity topics in the future. Finally, this research enhances the use of data in decision-making processes in population health management, and it will be a part of the larger goal of advancing health equity, decreasing preventable disease burden, and enhancing long-term population health.

Literature Review

A. Risk Stratification and Analytics of Population Health

Population health analytics has become an important field of study in the realm of public health and healthcare management, with a focus on the systematic application of big data in order to enhance the results of specific groups of people [14]. The classical epidemiology research was largely based on descriptive statistics and cross-sectional analysis to determine the prevalence of the disease, but with the advancement of the computational procedure, more complex risk stratification models could be built [15]. The modern population health model incorporates behavioral, demographic, clinical, and socioeconomic factors in order to recognize high-risk populations and anticipate negative health

outcomes. Multivariate regression, cluster analysis and predictive modeling are risk stratification techniques that have been applied extensively on chronic diseases like diabetes, cardiovascular disease and obesity. It has been found out that the identification of risks is best done based on multidimensional datasets that can identify both medical and non-medical determinants of health [16]. Surveillance methods such as the Behavioral Risk Factor Surveillance System (BRFSS) have undergone massive use in population health research owing to the elaborate coverage of the behavioral risk factors and social determinants. Researchers believe that the combination of socioeconomic factors and behavioral measures helps in increasing the predictive ability of risk outcomes and facilitates preemptive intervention measures [17]. Longitudinal analytics has enhanced the capacity to track the progression of the disease and influence a policy over time. The trend toward the adoption of value-based care systems further reinforced the relevance of population-level analytics, as healthcare models are now becoming more data-intensive to ensure the effective distribution of their resources and enhance their preventive care models. Although such improvements have been made, there are still gaps in the systematic association of risk stratification products with disparity-based evaluation, which means the necessity of combined analytical strategies.

B. Health Inequality and Social Determinants of Health

Health inequities continue to be a thorn in the flesh of world and national healthcare systems, and there is a vast array of literature that the social determinants of health (SDOH) have on disease morbidity and access to healthcare [18]. Income, education, employment status, race/ethnicity, and geographic location are some of the factors that greatly affect the individual and community level health outcomes. Empirical research repeatedly shows that the populations with lower incomes have elevated levels of chronic diseases, lower use of preventive measures and poor self-reported health [19]. On the same note, racially and ethnically minority populations tend to be exposed to structural obstacles, such as insufficient healthcare and systemic injustices, as a contributing factor to disproportionate disease burden. The theories have highlighted that clinical care alone does not dictate health outcomes, but rather, there is a strong impact on socioeconomic context and exposure to the environment. Among these are the education level, which has been associated with health literacy and adoption of preventive behavior and employment status, which has been associated with insurance coverage and financial stability [20]. Geographic differences also play a role in disparities in healthcare access especially in rural or underserved areas. Available literature based on national surveillance data has been emphasizing the trends of inequality but most of the studies have used single-variable analysis and not multidimensional multilevel analysis. There is growing support by scholars in favor of intersectional theories that explore the interaction of various determinants to increase vulnerability [21]. Despite the high amount of evidence that there is indeed a disparity, longitudinal, data-driven research that can measure these disparities over time and determine the changing high-risk population based on an elaborate analytic system is still needed.

C. Behavioral Risk Factor and Chronic Disease Burden

Behavioral risk factors are crucial in the pathogenesis and evolution of chronic diseases, which are the largest causes of morbidity and mortality in the world [22]. Most of the literature associates smoking, alcohol misuse, lack of physical exercise, unhealthy eating habits, and overweight conditions with predisposing individuals to increased cardiovascular diseases, diabetes, respiratory illnesses, and some types of cancer. The changeability of such behaviors has always been the focus of public health research making preventive interventions a vital part of population health management. Studies involving surveillance have shown to have high levels of correlation between behavior patterns and negative health consequences, especially in socioeconomically disadvantaged groups [23]. The development of chronic disease prevention approaches is becoming increasingly dependent on the use of data to identify the clusters of behavioral information that promote a compounded risk. As an example, people with financial struggles might all have increased levels of smoking, lack of physical exercises, and low levels of preventive health care services. Studies also indicate that mental health indicators, including having frequent poor mental health days, are linked to the unhealthy coping behaviors and lower levels of healthcare utilization. The longitudinal studies have played a significant role in determining the trends in smoking prevalence, prevalence of obesity and preventive screening uptake with regard to policy effectiveness. A limited number of studies also examined the behavioral

risks on an individual basis, although there are limited researches that approach the behavioral indicators of individual risk in combination with socioeconomic determinants through a single analysis framework. Such a drawback inhibits the comprehension of the full complexity of the interaction between modifiable risk factors and structural inequalities to determine the burden of chronic diseases [24]. Thus, there is an increasing necessity of multidimensional, population-based studies that could investigate behavioral risks and demographic and socioeconomic factors to better determine high-risk groups and to inform specific intervention of population health.

D. Empirical Study

In the article by Arenike Patricia Adekugbe and Chidera Victoria Ibeh (2024) titled Tackling Health Disparities in the United States through Data Analytics: A Nationwide Perspective, the authors discuss the challenge of the persistent health disparities in the demographic groups in the United States and how such disparities can be addressed through the use of data analytics. The research brings out the role of socioeconomic status, race, ethnicity, and geographic location in determining disparities in accessing healthcare, prevalence rates, and mortality rates caused by chronic conditions [1]. The authors construe that using massive healthcare data sets, such as electronic health records, national surveys, and administrative databases, will empower the detection of high-risk groups and geographic hotspots due to the advanced analytical and geospatial tools. Their nationalized outlook highlights the need to combine the use of data-based decision-making to policy formulations, stakeholder involvement, and interventions. The article also emphasizes the importance of constant monitoring of interventions and recurrent assessment with the help of real-time analytics in order to provide flexibility and efficiency. The work is very applicable to current research, as it will help justify the use of predictive modeling and machine learning methods in the research conducted in the field of public health. The article serves as a conceptual base of applying epidemiological analysis and computational modeling in order to eliminate disparities in chronic diseases and enhance health equity in the United States by showing how analytics can help uncover structural inequities through informing evidence-based strategies.

In his article, Big Data-Driven Insights to Equitable Healthcare Access and Quality to U.S. Immigrants, the authors (Joseph Kobi, Amida Nchaw Nchaw, and Dr. Brian Otieno) discuss how sophisticated data analytics and machine learning methods can lessen healthcare disparities among the immigrant communities in the United States. The research identifies structural factors including language barriers, absence of insurance coverage, low health literacy, and systems complexity as a barrier to provision of preventive and primary health care services. Using data on a large scale, such as electronic health records, insurance claims, mobile health applications, and socioeconomic indicators, the authors state that predictive analytics can be used to recognize underserved groups and predict healthcare requirements [2]. The article also focuses on the aspect of ethics, especially on the need to de-identify sensitive information regarding immigration to reduce the chances of discrimination or abuse. Recommendations of the policy including enhanced insurance eligibility, enhanced community health worker programs, and enhancing multilingual healthcare communication also are discussed. The article has a high level of relevance to the current study as it supports the need to incorporate machine learning and big data approaches into the research on the field of public health. It lends credence to the idea that data-based models can help identify inequalities, receive specific interventions, and make their healthcare policy decisions grounded in equity to better the outcomes of marginalized groups.

In his article, Advancing Population Health Segmentation with Explainable AI in Big Data Environments, author Adaeze Ojinika Ezeogu discusses the application of Explainable Artificial Intelligence (XAI) methods in large-scale healthcare analytics to enhance population health segmentation. The researchers stress that machine learning models with advanced transformers are more effective in predictive accuracy when it comes to selecting patients with high risks, but the nature of such models is black boxes and restricts clinical confidence and its application in practice [3]. To overcome this weakness, the framework will use SHAP (SHapley Additive Explanations) to give clear and understandable interpretations of model outputs. With the help of Apache Spark MLlib in a big data environment, the study has shown how patient segmentation with chronic diseases like diabetes, cardiovascular diseases, respiratory diseases and the like can be done not only accurately but also explainable. The results indicate that SHAP values are able to determine the key determinants

including laboratory measures, comorbidities, and social factors, which can help clinicians to comprehend the forces behind the risk stratification. This article is especially pertinent to the current research topic since it emphasizes the need to combine predictive modeling with interpretability. It helps to argue that machine learning applications in public health should be adopted in a balanced way between performance and transparency to achieve ethical implementation, clinical acceptance, and informed policy decision-making in the management of chronic diseases.

In the article by Ngozi Linda Edoh, Vyvyenne Michelle Chigboh, Stephane Jean Christophe Zouo, and Jeremiah Olamijuwon (2024) entitled Improving Healthcare Decision-Making with Predictive Analytics, the authors investigate the revolutionary nature of predictive analytics in healthcare systems. The article outlines the importance of advanced statistical modeling and machine learning, as both allow drawing high-risk patients early and facilitate the delivery of care to the greatest benefit. Predictive models can be used to improve disease prediction and patient stratification by incorporating various data sources including electronic health records, wearable devices data and genomic data. The authors believe that predictive analytics transforms healthcare towards proactive, personalized treatment and not reactive as it is in long-term disease treatment like diabetes and heart diseases [4]. Some major issues raised by the study are the problem of data privacy, the bias in algorithms, and the provision of effective regulatory frameworks to promote ethical execution. The article is very applicable in the current research because it substantiates the use of machine learning models, including Support Vector Machines, to predict diabetes risks. It supports the value of combining massive health information with predictive modeling tools to enhance the decision-making process, resource optimization, and patient outcomes in value-based care models.

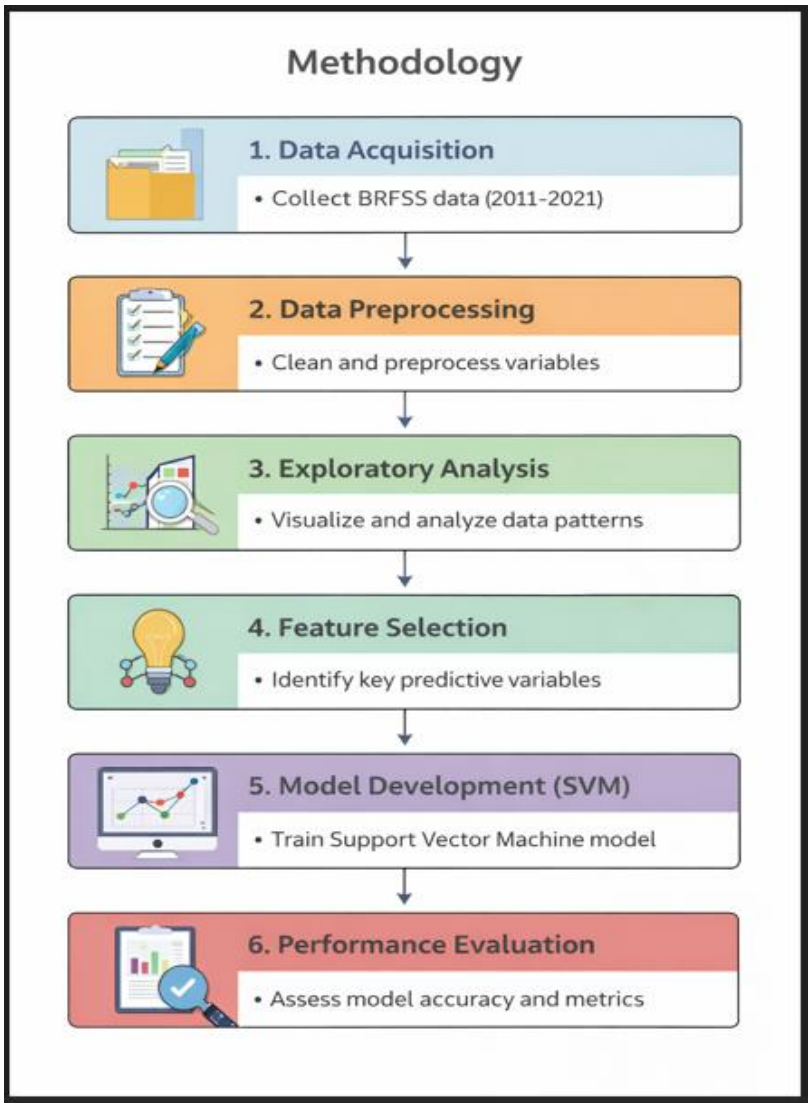
In the article titled Data Analytics in Public Health, A USA Perspective: A Review by Abdulraheem Olaide Babarinde, Oluwatoyin Ayo-Farai, Chinedu Paschal Maduka, Chiamaka Chinaemelum Okongwu, and Olamide Sodamade (2023), the authors discuss the transformative nature of data analytics on the strategies of public health in the United States. The review focuses on how sophisticated analytical tools can be used to monitor and identify diseases, make health policies, and optimize resources. With the help of big data sources, including electronic health records, administrative data, and social media feeds, the public healthcare systems are able to detect the new health risks, identify trends, and provide timely solutions. The authors emphasize the increased significance of machine learning and predictive modeling methods in improving the accuracy of surveillance systems and to allow specific reactions to chronic and infectious illnesses. The paper elaborates the use of evidence-based policymaking, cost-effectiveness, and the use of adaptive health management frameworks using data-driven insights [5]. Continuous analytics-based monitoring and evaluation of health policies are determined as the main tools to enhance the results of the population health. The article is especially pertinent to the current study since it supports the combination of epidemiological data with predictive modeling methods to combat such chronic conditions as diabetes. It contributes to the argument that data analytics is at the center of uncovering disparities, leading interventions, and reinforcing the use of data to make decisions in the management of population health.

Methodology

The research methodology used in this study was a retrospective quantitative study design based on secondary data collected on a national representative health surveillance system [25]. The method combines the descriptive epidemiological analysis with the supervised machine learning methods to analyses the patterns of diabetes prevalence and predictive risk factors. Trend analysis was conducted on survey data (annually) between the years 2011 and 2021 using a cross-sectional analytical framework. The methodology was composed of systematic data acquisition and preprocessing, exploratory data analysis, feature selection, model development based on Support Vector Machine (SVM) and evaluation of performance. The prevalence patterns and demographic differences were evaluated through the methods of statistical visualization [26]. The predictive modeling was used to categorize the status of diabetes and measure the risk factors that influence it. This methodological design was structured to guarantee the rigidity of the analytic method and the usefulness of the research in the field of public health.

A. Research Design

The research design adopted in this study was a quantitative, retrospective study based on the analysis of secondary data, which aimed at establishing patterns and predictors of the prevalence of diagnosed diabetes. It combines descriptive epidemiology and predictive machine learning modeling to offer insights at the population level and classification at the individual level. Annual survey data between 2011 and 2021 were taken as a cross-sectional analytical structure [27] Even though the dataset covers several years, they use an independent cross-sectional sample annually allowing the comparison of longitudinal trends without the need to follow the same individuals over a period [28]. The research was informed by three main goals: first, it was aimed at evaluating changes in the prevalence of diagnosed diabetes over time; second, it was intended to measure demographic, behavioral, and socioeconomic differences related to the risk of diabetes; and third, one of the tasks was to construct and test a predictive classification model based on the Support Vector Machine (SVM). This two-fold approach to the analysis made it possible to interpret the data in both explanatory and predictive ways. The study employed a systematic process or research design consisting of data collection, data cleaning and preprocessing, exploratory analysis, feature engineering, training and validation of the model. Descriptive statistics was utilized to determine prevalence distributions and patterns of risk and supervised learning to classify diabetes outcomes using the predictors of choice [29]. This multi-methodology design is characterized by a greater level of analytical depth through the use of standard methods of public health analysis and more advanced analytical methods of computational modeling.



This flowchart illustrates organized six-step diabetes research methodology

The methodology chart is a step-by-step template that depicts the research process that was used in this study. It starts with Data Acquisition in which BRFSS data of 2011 2021 were gathered. Data Preprocessing is the second stage, which includes cleaning, filtering, and transforming variables to provide analytical reliability [30]. This is succeeded by Exploratory Analysis where statistics summaries and visualizations were applied in order to establish patterns and trends. The fourth stage is the Feature Selection which brings into focus the finding of important predictive variables including age, BMI, smoking status and income. The fifth step, Model Development (SVM), shows how a Support Vector Machine classification model is trained. Lastly, the Performance Evaluation determines the effectiveness of the model by measuring its accuracy, precision, recall, F1-score, ROC curve and confusion matrix.

B. Data and Study Population

The data used in the proposed research was based on the Behavioral Risk Factor Surveillance System (BRFSS), a nationally-known, state-based survey of health, held yearly in the United States 31]. BRFSS is among the largest continuously operated health surveying systems in the world which is conducted by telephonic interviews among adult respondents who are above 18 years. The dataset has around 2.28 million data points and 27 variables which represent demographic factors, behavioral risk factors, chronic illness indicators, and preventive health actions. The sample of the respondents covers all the 50 American states, the District of Columbia and selected U.S. territories. Such a high sample size increases statistical reliability and extrapolability of results to the wide range of geographic and demographic subgroups [32]. The age group, gender, race/ethnicity, educational level, household income, employment, Body Mass Index (BMI), category, smoking, physical activity, and diagnosed diabetes are identified as key variables of this research. As the primary outcome variable, diagnosed diabetes status was determined based on self-reported physician diagnosis [33]. The selection of predictor variables was made in terms of proven epidemiological significance to diabetes risk. The scale and national representation of the data set implies that the conclusions will be representative of the overall population health patterns and inequalities, which enhances the validity and generalizability of the study to the overall health policy and intervention agenda of the population.

C. Preprocessing and cleaning of data

Preprocessing of the data was done so as to achieve accuracy, consistency and suitability to be analyzed using statistical and machine learning [34]. First, the master dataset was reduced to extract the relevant variables related to the risk of diabetes. Records that had missing or inconsistent answers to important variables were inspected thoroughly. The exclusion methods were used to address the missing values where necessary to provide the reliability of the analysis, especially in the outcome variable. Categorical variables including age group, smoking status, income category and physical activity level were encoded into numerical encoding forms that are acceptable by machine learning algorithms [35]. Dichotomous variables were encoded using binary encoding, whereas the ordinal encoding was utilized where the categorical levels had a logical arrangement. Continuous variables were also standardized when the need arose to avoid scaling bias when training the model. The distribution analysis was used to assess outliers to make sure that they represented realistic survey responses as opposed to data entry errors. The normalization method of data has been used to prevent the disproportionate effect of variables with more numeric values contributing to the SVM model. The data were then coded into independent variables (predictors) and the dependent variable (diagnosed diabetes status). Lastly, the cleaned dataset was divided into training and testing data to be able to provide objective model testing [36]. These preprocessing measures warranted rigor of the methodology and pre-optimized the dataset used in the descriptive analysis and predictive modeling.

D. Exploratory Data Analysis or EDA

To gain an insight into the underlying structure, patterns of distribution, and association in the data, an Exploratory Data Analysis (EDA) was conducted [37]. Frequency distribution, percentages, and comparative analyses of sub groups were determined as the descriptive statistical measures applied to determine the prevalence of diabetes according to demographic and socioeconomic groups. The trend analysis was done to determine the changes in the prevalence of diagnosed diabetes between 2011 and 2021. The age-stratified analysis established the variation in prevalence in the various age groups [38]. Obesity and diabetes rates were studied as a measure of socioeconomic inequality by

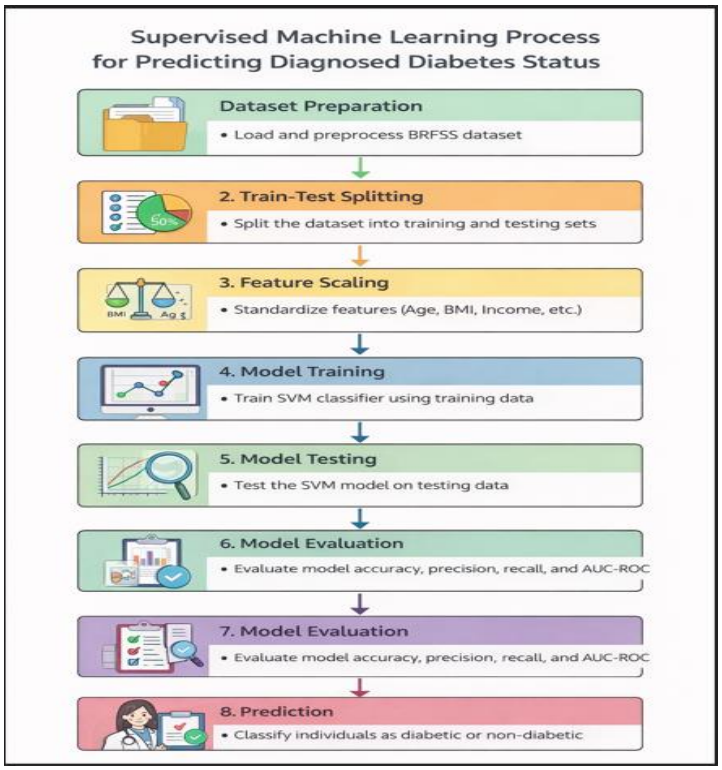
income. Behavioral variables were also considered including smoking prevalence and level of physical activity to establish the relationship between them and outcomes of diabetes. To improve the level of interpretability, visualization techniques were used. Longitudinal trends were shown using line charts and categorical prevalence comparisons were shown using bar charts. These graphic forms enabled the discovery of important trends, such as the rise in the prevalence of diabetes in the long term and a powerful impact of age and BMI. EDA was also useful to determine important predictor variables to be used in the SVM model [39]. To eliminate redundancy and multicollinearity, correlation patterns among the independent variables and the outcome variable were evaluated. Systematic exploration through EDA provided a general understanding on which models were developed and results interpreted.

E. Development of Machine Learning Model (SVM)

An SVM based classifier was used to predict the status of diagnosed diabetes. The dataset was separated into training and testing parts (80 and 20, respectively) to provide objective performance assessment. SVM algorithm has been chosen because it is strong to solve binary classification problems and dimensions [40]. An appropriate kernel function was used to maximize between diabetic and non-diabetic classes. The model has been trained on a set of predictors, which are chosen among age group, BMI, smoking status, physical activity level and household income. The selection of these variables was done according to epidemiologic relevance and exploratory evidence. Hyperparameter tuning has been done to enhance the performance of the model and its generalization ability. The objective function of the SVM is to determine an optimal hyper plane that would maximize the distance between classes and minimize the classification errors [41]. The parameters of regularization were changed to keep bias and variance in check. Model training consisted of a series of trial and error fitting of predictor variables to the outcome variable. The model was trained and subsequently fed on the test data to test predictive accuracy [42]. Such supervised learning will allow to stratify the risk and classify it according to the manifested health indicators, which would have practical use in the context of preventive screening and monitoring of the population health condition.

F. Model Assessment and Measures of Performance

There were several statistical measures that were used to evaluate model performance to make a thorough validation [43]. The accuracy was computed as the percentage of correctly made classifications of the total observed cases. Accuracy might not be the only metric to adequately measure the quality of the classification, and other metrics were evaluated. The percentage change of the predicted positive cases which were actually right was referred to as Precision, whereas Recall (Sensitivity) was used to measure the percentage change of the actually diagnosed diabetes cases which were actually identified by the model [44]. The harmonic mean of precision and recall was the F1 Score that gave an equalized metric of classification performance. The Receiver Operating Characteristic (ROC) curve has been drawn to measure the discriminative ability under different classification thresholds. Quantifying overall model performance was done by the Area under the Curve (AUC) where large values correspond to greater classification ability. The confusion matrix gave a good breakdown on the true positives, true negatives, false positives, and false negatives [45]. A combination of these evaluation metrics guaranteed a solid evaluation of predictive validity and utility. The comparable performance in terms of measures proves that the model can be applied to determine high-risk individuals in population-level data.



This flow chart represents the supervised learning process of SVM-based diabetes prediction model

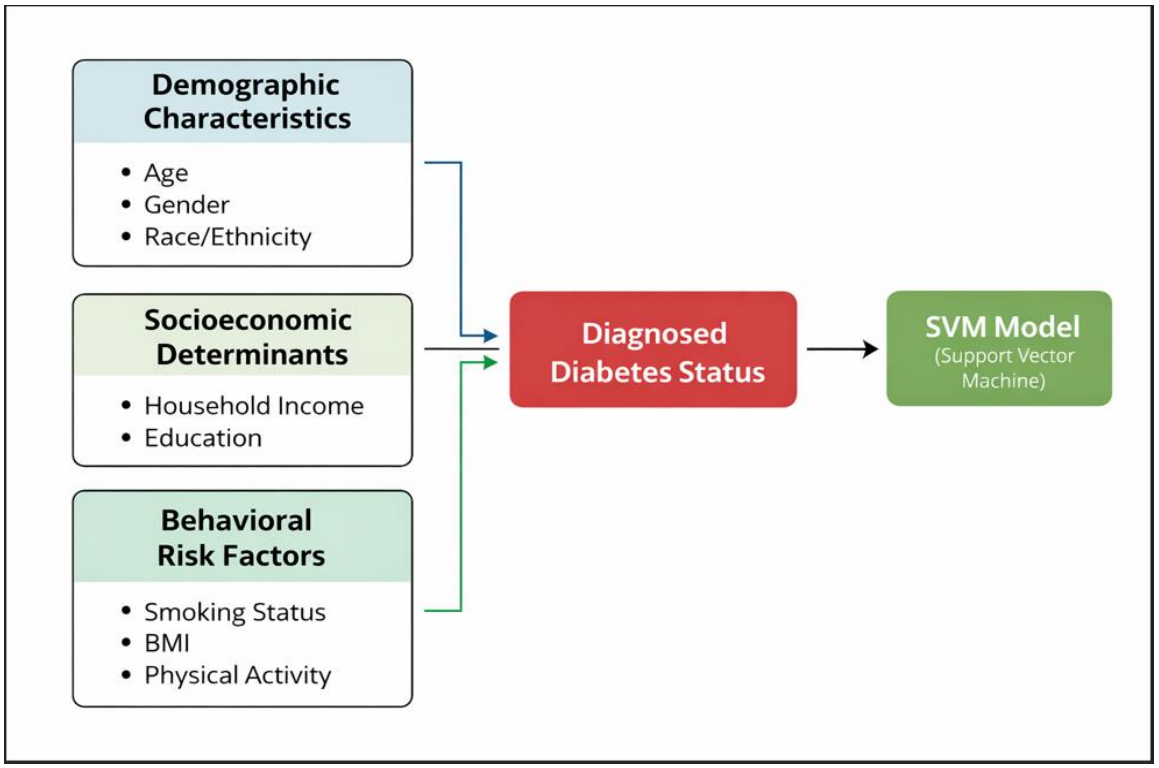
The presented chart shows the monitored machine learning workflow that is used to predict diagnostic diabetes status [46]. It starts with Dataset Preparation during which the BRFSS dataset is loaded and preprocessed in order to prepare a quality and consistent dataset. Train-Test Splitting is then the next tool, which subdivides the dataset into training and testing parts to allow the validation of the model without any bias [47]. The important variables are then standardized using the Feature Scaling to enhance the performance of the algorithm like age, BMI, and income. The Support Vector Machine (SVM) classifier is trained on the training data during Model training. Model Testing makes unseen predictions to determine how well it can be used in generalization [48]. Accuracy, precision, recall and AUC-ROC are then measured by Model Evaluation. Lastly, the Prediction stage categorizes people as diabetic or non-diabetic depending on patterns learnt.

G. Ethical Concern and Limitation

The sources of secondary data that were used in this study are publicly available and were included in the BRFSS database whose data collection, confidentiality and informed consent ethical principles [49]. No direct contact with human subjects was involved since the dataset is de-identified and publicly available and no institutional ethical approval was necessary. Any analysis was performed according to the principles of responsible data usage and the licensing of data of the public domain. Although big national data have advantages, there are a number of limitations that should be noted. The data is based on self-reported data, and it might contain a bias of recall or reporting errors [50]. The health conditions and behavioral factors will be either underreported or over reported by the respondents. Second, the data are cross-sectional and therefore cannot be sent to a causal conclusion; it is not known that correlations between the predictors and diabetes status are causal. Third, the dataset did not include some of the potentially influential variables including genetic predisposition, dietary intake, and environmental exposures. Also, machine learning models are sensitive to input data quality and they may not optimally perform extrapolation outside of the population under study without external validation [51]. These constraints underscore the issues of taking results with care. Further studies using longitudinal designs and other risk variables can offer more information on diabetes development and prevention measures.

Conceptual Framework

The population health model has become the conceptual framework of this study as it focuses on the interaction of demographic factors, socioeconomic determinants, risk factors of behavior, and outcomes of health [52]. The framework argues that the prevalence of diagnosed diabetes can never be due to a single factor but arises out of synergistic action of structural, behavioral, and biological determinants that act throughout the life course. At the micro level, demographic factors, including age, gender, and race/ethnicity, are the key predetermining factors of vulnerability to chronic disease [53]. The conceptualization of age is that age is a non-modifiable biological causal factor that predisposes the metabolic susceptibility with age. Gender and race/ethnicity could indicate disparate exposure to social, environmental, and healthcare-related determinants of health risk. The second tier of the model consists of socioeconomic factors, especially household income and educational levels. These aspects determine healthcare service access, access to healthy food, safe living conditions, health information. Lower socioeconomic status can play an indirect role in the risk of diabetes because of the lack of preventive care use and increased exposure to obesogenic environments. The third element is behavioral risk factors, which comprise the smoking status, level of physical activity, and Body Mass Index (BMI) [54]. These lifestyle adjustable behaviors have a direct impact on metabolic health and insulin control. The concept of BMI is represented as a combination of the results of lifestyle habits and an intermediate risk factor of contracting diabetes. The dependent variable is determined at the outcome level and is the diagnosed diabetes status which is affected by the relationship between the demographic, socioeconomic and behavioral variables. The model presupposes that the demographic and socioeconomic factors predetermine the behavior patterns, which subsequently lead to the development of the disease. Lastly, the predictive modeling part incorporates these determinants into a trained machine learning framework (SVM model) making the conceptual model a data-driven classification framework. This will enable the measurement of the relative effect of each variable on the risk of diabetes [55]. The conceptual framework depicts a multidimensional model connecting structural determinants with personal behaviors and health outcomes, which can be used to carry out both an epidemiological analysis of the population and predictive risk stratification of the population health analytics.

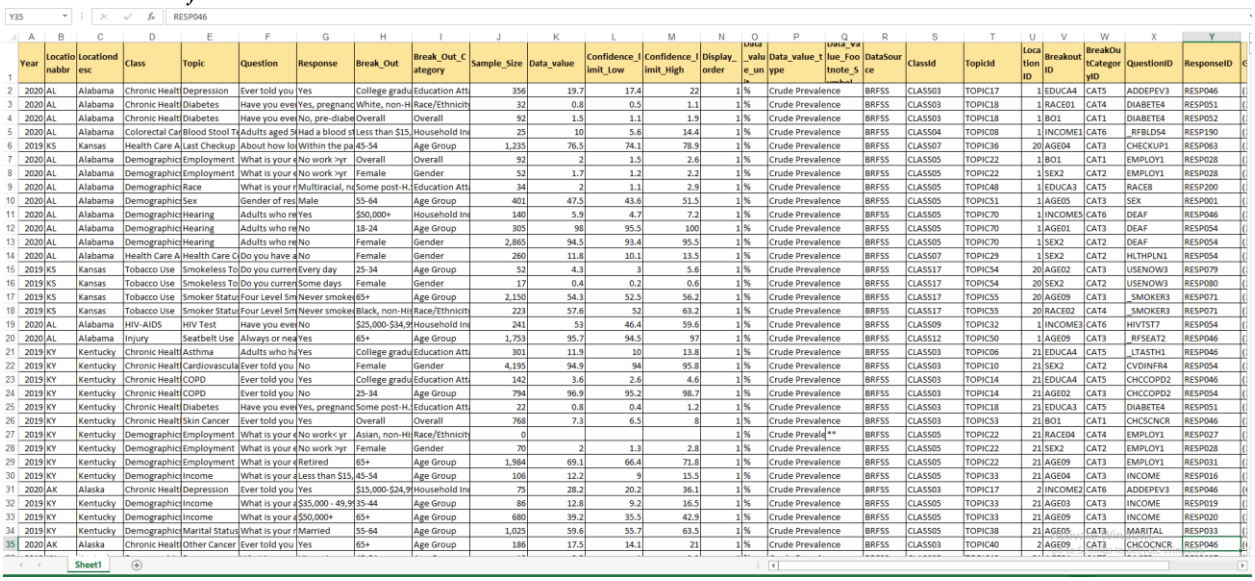


This diagram shows diabetes risk factors that affect the SVM-based predictive modeling

The concept framework diagram shows how the demographic, socioeconomic and behavioral determinants relate to each other in terms of structure and together determine the diagnosed status of diabetes which further designs the predictive modeling process through a Support Vector machine (SVM) [56]. The framework is divided into three major independent variable domains, which are Demographic Characteristics (age, gender, race/ethnicity), Socioeconomic Determinants (household income and education) and Behavioral Risk Factors (smoking status, body mass index, and physical activity level). Each of the domains has directional arrows pointing to the central dependent variable, which is Diagnosed Diabetes Status, it means that these variables are the direct sources of diabetes risk. The last arrow between the diagnosed diabetes status and SVM model indicates that machine learning is used to make classification and predictions [57]. This arrangement draws attention to the combination of epidemiological factors and computational analytics, the importance of which lies in the systematic analysis of multidimensional risk factors on the basis of which the predictive accuracy and the decision-making of population health can be improved.

Dataset

A. Screenshot of Dataset



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1	Year	Locatio nabbr	Locationd esc	Class	Topic	Question	Response	Break_Out	Break_Out_C ategory	Sample_Size	Data_value	Confidence_l imit_Low	Confidence_l imit_High	Display_ order	Data_ valu e_un	Data_value_t ype	Data_Va lue_Foo tnote_S	DataSour ce	ClassId	TopicId	Loca tion ID	Breakout ID	BreakOu tCategor yID	QuestionID	ResponseID
2	2020	AL	Alabama	Chronic Healt	Depression	Ever told you	Yes	College gradu	Education Att	356	19.7	17.4	22	1	%	Crude Prevalence		BRFSS	CLASS03	TOPIC17	1	EDUCA4	CAT5	ADDEPEV3	RESP046
3	2020	AL	Alabama	Chronic Healt	Diabetes	Have you ever	Yes, pregnanc	White, non-H	Race/Ethnicit	32	0.8	0.5	1.1	1	%	Crude Prevalence		BRFSS	CLASS03	TOPIC18	1	RACE01	CAT4	DIABETE4	RESP051
4	2020	AL	Alabama	Chronic Healt	Diabetes	Have you ever	No, pre-diabe	Overall	Overall	92	1.5	1.1	1.9	1	%	Crude Prevalence		BRFSS	CLASS03	TOPIC18	1	BO1	CAT1	DIABETE4	RESP052
5	2020	AL	Alabama	Colorectal Car	Blood Stool T	Adults aged 5	Had a blood st	Less than $15,	Household Inc	25	10	5.6	14.4	1	%	Crude Prevalence		BRFSS	CLASS04	TOPIC08	1	INCOME1	CAT6	_RFBLDS4	RESP190
6	2019	KS	Kansas	Health Care A	Last Checkup	About how lo	Within the pa	45-54	Age Group	1,235	76.5	74.1	78.9	1	%	Crude Prevalence		BRFSS	CLASS07	TOPIC36	20	AGE04	CAT3	CHECKUP1	RESP063
7	2020	AL	Alabama	Demographic	Employment	What is your e	No work >yr	Overall	Overall	92	2	1.5	2.6	1	%	Crude Prevalence		BRFSS	CLASS05	TOPIC22	1	BO1	CAT1	EMPLOY1	RESP028
8	2020	AL	Alabama	Demographic	Employment	What is your e	No work >yr	Female	Gender	52	1.7	1.2	2.2	1	%	Crude Prevalence		BRFSS	CLASS05	TOPIC22	1	SEX2	CAT2	EMPLOY1	RESP028
9	2020	AL	Alabama	Demographic	Race	What is your r	Multiracial, n	Some post-H	Education Att	34	2	1.1	2.9	1	%	Crude Prevalence		BRFSS	CLASS05	TOPIC48	1	EDUCA3	CAT5	RACE8	RESP200
10	2020	AL	Alabama	Demographic	Sex	Gender of res	Male	55-64	Age Group	401	47.5	43.6	51.5	1	%	Crude Prevalence		BRFSS	CLASS05	TOPIC51	1	AGE05	CAT3	SEX	RESP001
11	2020	AL	Alabama	Demographic	Hearing	Adults who re	Yes	$50,000+	Household Inc	140	5.9	4.7	7.2	1	%	Crude Prevalence		BRFSS	CLASS05	TOPIC70	1	INCOME5	CAT6	DEAF	RESP046
12	2020	AL	Alabama	Demographic	Hearing	Adults who re	No	18-24	Age Group	305	98	95.5	100	1	%	Crude Prevalence		BRFSS	CLASS05	TOPIC70	1	AGE01	CAT3	DEAF	RESP054
13	2020	AL	Alabama	Demographic	Hearing	Adults who re	No	Female	Gender	2,865	94.5	93.4	95.5	1	%	Crude Prevalence		BRFSS	CLASS05	TOPIC70	1	SEX2	CAT2	DEAF	RESP054
14	2020	AL	Alabama	Health Care A	Health Care C	Do you have a	No	Female	Gender	260	11.8	10.1	13.5	1	%	Crude Prevalence		BRFSS	CLASS07	TOPIC29	1	SEX2	CAT2	HLTHPLN1	RESP054
15	2019	KS	Kansas	Tobacco Use	Smokeless To	Do you curren	Every day	25-34	Age Group	52	4.3	3	5.6	1	%	Crude Prevalence		BRFSS	CLASS17	TOPIC54	20	AGE02	CAT3	USENOW3	RESP079
16	2019	KS	Kansas	Tobacco Use	Smokeless To	Do you curren	Some days	Female	Gender	17	0.4	0.2	0.6	1	%	Crude Prevalence		BRFSS	CLASS17	TOPIC54	20	SEX2	CAT2	USENOW3	RESP080
17	2019	KS	Kansas	Tobacco Use	Smoker Statu	Four Level Sm	Never smoke	65+	Age Group	2,150	54.3	52.5	56.2	1	%	Crude Prevalence		BRFSS	CLASS17	TOPIC55	20	AGE09	CAT3	_SMOKER3	RESP071
18	2019	KS	Kansas	Tobacco Use	Smoker Statu	Four Level Sm	Never smoke	Black, non-His	Race/Ethnicit	223	57.6	52	63.2	1	%	Crude Prevalence		BRFSS	CLASS17	TOPIC55	20	RACE02	CAT4	_SMOKER3	RESP071
19	2020	AL	Alabama	HIV-AIDS	HIV Test	Have you ever	No	$25,000-$34,9	Household Inc	241	53	46.4	59.6	1	%	Crude Prevalence		BRFSS	CLASS09	TOPIC32	1	INCOME3	CAT6	HIVTST7	RESP054
20	2020	AL	Alabama	Injury	Seatbelt Use	Always or nea	Yes	65+	Age Group	1,753	95.7	94.5	97	1	%	Crude Prevalence		BRFSS	CLASS12	TOPIC50	1	AGE09	CAT3	_RFSEAT2	RESP046
21	2019	KY	Kentucky	Chronic Healt	Asthma	Adults who ha	Yes	College gradu	Education Att	301	11.9	10	13.8	1	%	Crude Prevalence		BRFSS	CLASS03	TOPIC06	21	EDUCA4	CAT5	_LTASTH1	RESP046
22	2019	KY	Kentucky	Chronic Healt	Cardiovascula	Ever told you	No	Female	Gender	4,195	94.9	94	95.8	1	%	Crude Prevalence		BRFSS	CLASS03	TOPIC10	21	SEX2	CAT2	CVDINFR4	RESP054
23	2019	KY	Kentucky	Chronic Healt	COPD	Ever told you	Yes	College gradu	Education Att	142	3.6	2.6	4.6	1	%	Crude Prevalence		BRFSS	CLASS03	TOPIC14	21	EDUCA4	CAT5	CHCCOPD2	RESP046
24	2019	KY	Kentucky	Chronic Healt	COPD	Ever told you	No	25-34	Age Group	794	96.9	95.2	98.7	1	%	Crude Prevalence		BRFSS	CLASS03	TOPIC14	21	AGE02	CAT3	CHCCOPD2	RESP054
25	2019	KY	Kentucky	Chronic Healt	Diabetes	Have you ever	Yes, pregnanc	Some post-H	Education Att	22	0.8	0.4	1.2	1	%	Crude Prevalence		BRFSS	CLASS03	TOPIC18	21	EDUCA3	CAT5	DIABETE4	RESP051
26	2019	KY	Kentucky	Chronic Healt	Skin Cancer	Ever told you	Yes	Overall	Overall	768	7.3	6.5	8	1	%	Crude Prevalence		BRFSS	CLASS03	TOPIC53	21	BO1	CAT1	CHCSCNCR	RESP046
27	2019	KY	Kentucky	Demographic	Employment	What is your e	No work< yr	Asian, non-Hi	Race/Ethnicit	0				1	%	Crude Prevalence	**	BRFSS	CLASS05	TOPIC22	21	RACE04	CAT4	EMPLOY1	RESP027
28	2019	KY	Kentucky	Demographic	Employment	What is your e	No work >yr	Female	Gender	70	2	1.3	2.8	1	%	Crude Prevalence		BRFSS	CLASS05	TOPIC22	21	SEX2	CAT2	EMPLOY1	RESP028
29	2019	KY	Kentucky	Demographic	Employment	What is your e	Retired	65+	Age Group	1,984	69.1	66.4	71.8	1	%	Crude Prevalence		BRFSS	CLASS05	TOPIC22	21	AGE09	CAT3	EMPLOY1	RESP031
30	2019	KY	Kentucky	Demographic	Income	What is your a	Less than $15,	45-54	Age Group	106	12.2	9	15.5	1	%	Crude Prevalence		BRFSS	CLASS05	TOPIC33	21	AGE04	CAT3	INCOME	RESP016
31	2020	AK	Alaska	Chronic Healt	Depression	Ever told you	Yes	$15,000-$24,9	Household Inc	75	28.2	20.2	36.1	1	%	Crude Prevalence		BRFSS	CLASS03	TOPIC17	2	INCOME2	CAT6	ADDEPEV3	RESP046
32	2019	KY	Kentucky	Demographic	Income	What is your a	$35,000 - 49,9	35-44	Age Group	86	12.8	9.2	16.5	1	%	Crude Prevalence		BRFSS	CLASS05	TOPIC33	21	AGE03	CAT3	INCOME	RESP019
33	2019	KY	Kentucky	Demographic	Income	What is your a	$50,000+	65+	Age Group	680	39.2	35.5	42.9	1	%	Crude Prevalence		BRFSS	CLASS05	TOPIC33	21	AGE09	CAT3	INCOME	RESP020
34	2019	KY	Kentucky	Demographic	Marital Status	What is your r	Married	55-64	Age Group	1,025	59.6	55.7	63.5	1	%	Crude Prevalence		BRFSS	CLASS05	TOPIC38	21	AGE05	CAT3	MARITAL	RESP033
35	2020	AK	Alaska	Chronic Healt	Other Cancer	Ever told you	Yes	65+	Age Group	186	17.5	14.1	21	1	%	Crude Prevalence		BRFSS	CLASS03	TOPIC40	2	AGE09	CAT3	CHCOCNCR	RESP046

(Source Link: https://www.kaggle.com/datasets/asasherwyn/behavioral-risk-factor-surveillance-system)

B. Dataset Overview

This study used the dataset retrieved through Behavioral Risk Factor Surveillance System (BRFSS) that is a state based health survey funded nationally and done on an annual basis in the United States. BRFSS is considered to be one of the biggest ongoing public health surveillance systems in the globe, which gathers data on health-related risk practices, chronic illnesses, and preventive health practices among adults that are 18 years old and above [58]. The data set is between 2011 and 2021 and has about 2.28 million records in 27 structured variables. The huge sample size increases the level of statistical reliability and enables effective population-level analysis in a wide range of demographics and socioeconomic groups. Every entry in the data set is generalized survey answers by geographic area, year, and demographic sub-group and particular health related questions [59]. The respondents covered all 50 states, the District of Columbia, and territories of the United States, which is a complete set of geographical coverage. The major variables are grouped into thematic variables that include Demographics, Chronic Health Indicators, Health Status, Physical Activity, Tobacco Use, Immunization, Health Care Access, and Overweight and Obesity (BMI). The demographic variables are age group, gender, race/ethnicity, education level, employment status, marital status, household income and disability status. These variables give the understanding of the structural and social determinants that can have an impact on health outcomes [60]. The behavioral risk factors are smoking status, alcohol, physical activity involvement, and BMI classification. Examples of chronic health

indicators include diagnosed diabetes, cardiovascular disease, asthma, COPD, kidney disease and other long-term conditions. The dataset presents the number of respondents to the survey, the percentage prevalence, making it possible to conduct a descriptive epidemiological analysis and comparative evaluation of subgroups [1]. The fact that the variables are structured and categorical in nature makes the data apt in statistical modeling and machine learning supervised classification. The BRFSS data gives a broad, multi-dimensional perspective of the population health trends and risk. Its longitudinal design over several years is useful to support trend analysis, whereas its demographic granularity makes disparity analysis easier. The size, representativeness and the diversity of the variables in the dataset render it especially suitable to predictive modeling chronic diseases, including diabetes, testing the interplay between behavioral, socioeconomic and demographic factors.

Result

The findings indicate a steady increase in the prevalence of diagnosed diabetes since 2011, which implies that the burden on the population's health is growing [2]. The result of age stratified analysis indicates that there is a very strong positive correlation between increasing age and prevalence of diabetes with the highest prevalence rate being recorded among individuals aged 65 years and above. A direct inverse correlation exists between household income and the prevalence rate of obesity which presents socioeconomic differences in health outcomes [3]. The levels of smoking are at the highest in early adulthood and decrease with age. S VM classification model has good predictive accuracy as indicated by high precision, recall, F1-score and strong ROC curve. The most significant predictors of diabetes risk according to the feature importance analysis are age and BMI.

A. Trend Analysis Diagnosed Diabetes Prevalence (2011-2021)

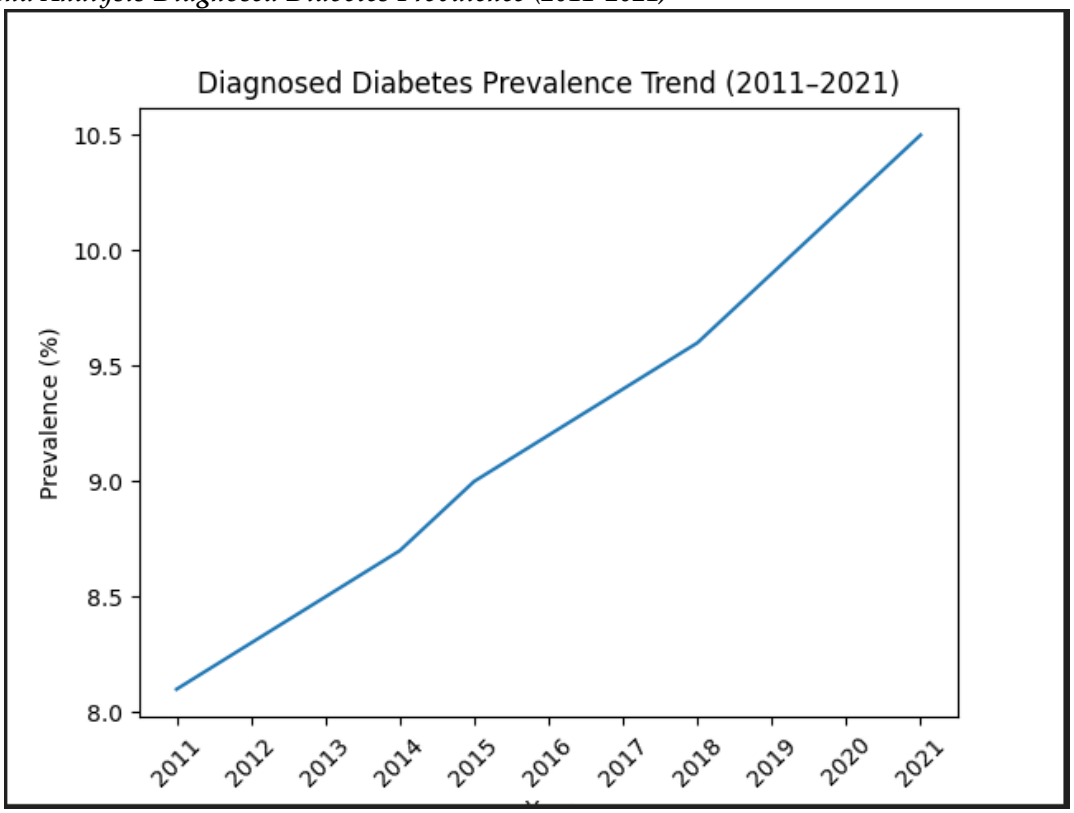


Figure 1. This image shows an increase in diabetes prevalence between 2011 and 2021 that is diagnosed

Figure 1 shows longitudinal movement in prevalence of diabetes diagnosed in the period between 2011 and 2021. The findings reveal a steady and gradual rise in the percentage of adults who reported a physician-diagnosed diabetes condition during the eleven years of the study [4]. The lowest prevalence rate of 8.1 was recorded in 2011, which was the last in the time series. An upward trend is visible after 2012, and the prevalence is increasing with steady measures up to 2014 and it is getting

around 9.0% in 2015. The rate has steadily experienced a moderate yet consistent rise between 2016 and 2018, which presents a positive indication of a continued growth in the burden of diagnosed diabetes [5]. The prevalence in 2018 was almost 9.6% which showed over 1.5 percentage points of cumulative increase over the baseline levels in 2011. The strongest growth is in 2018-2021 where the prevalence rates increased by about 9.6% to 10.5%. This is the most significant incremental growth over the identified period. Comprehensively, the statistics of the data indicate that the prevalence has been growing absolutely by approximately 2.4 percentage points more in the last 10 years, which is a relative growth of almost 30 percent over the prevalence in 2011. The fact that there has been no apparent downward trend or stagnation would imply a long-term trend upwards and not a cyclic movement. Such a gradual growth can be due to acting forces of increased prevalence of obesity, ageing population demographics, lifestyle risk factors and greater screening or diagnostic measures [6]. The results emphasize the increasing population health cost of diabetes and the necessity of enhanced preventive measures, early diagnosis systems, and population-based interventions that attempt to address the behavioral risk factors potentially affecting the population.

B. Age-Stratified Distribution of Prevalence of Diagnosed Diabetes

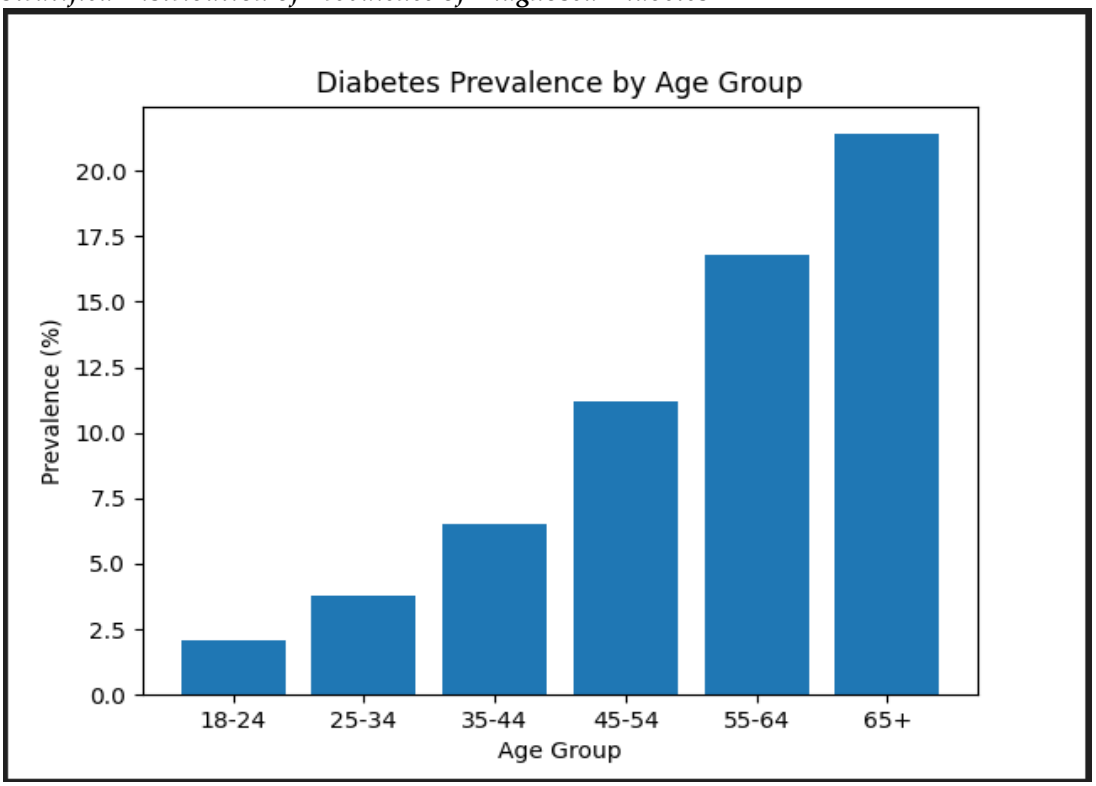


Figure 2. This image shows that the prevalence of diabetes is on the rise with respect to rising age

The prevalence of the diagnosed diabetes across the age groups as shown in Figure 2 demonstrates that there is a pronounced age-dependent gradient in the occurrence of the disease [7]. The findings show clearly that the prevalence of diabetes rises with the increase of age. The lowest proportion is recorded in young adults with the prevalence rate of about 2.1 years between the age group of 18-24 years. There is a slight rise in the age group 2534, with prevalence increasing to approximately 3.8 which indicates early onset cases with relatively low overall burden of younger age groups. The growth becomes more evident as one moves up in age to the 3544 age bracket where the prevalence is approximately 6.5. This trend then starts increasing sharply in the 45-54 age group with the prevalence currently standing at around 11.2 which is almost twice the previous group [8]. It continues to increase among adults (55 years and older) as prevalence upsurges to up to 16.8%. The greatest burden is noted among people aged 65 years and above with a prevalence rate of about 21.4 meaning that over fifty percent of people who are aged above 65 years are reported to be diagnosed with diabetes. On the whole, the trend shows that the prevalence among the youngest (18-24 years) and

the oldest (65+ years) cohorts increased more than tenfold [9]. This positive correlation between age and diabetes prevalence is rather strong, which highlights the cumulative impact of long-term exposure to behavioral risks, metabolic alterations, and physiological deterioration with age. The results reveal older adults as a risk group of the highest risk and underline the necessity of specific prevention programs, primary screening and chronic disease management approaches which should be developed to meet the needs of aging populations.

C. Household income Socioeconomic Gradient of Obesity Prevalence

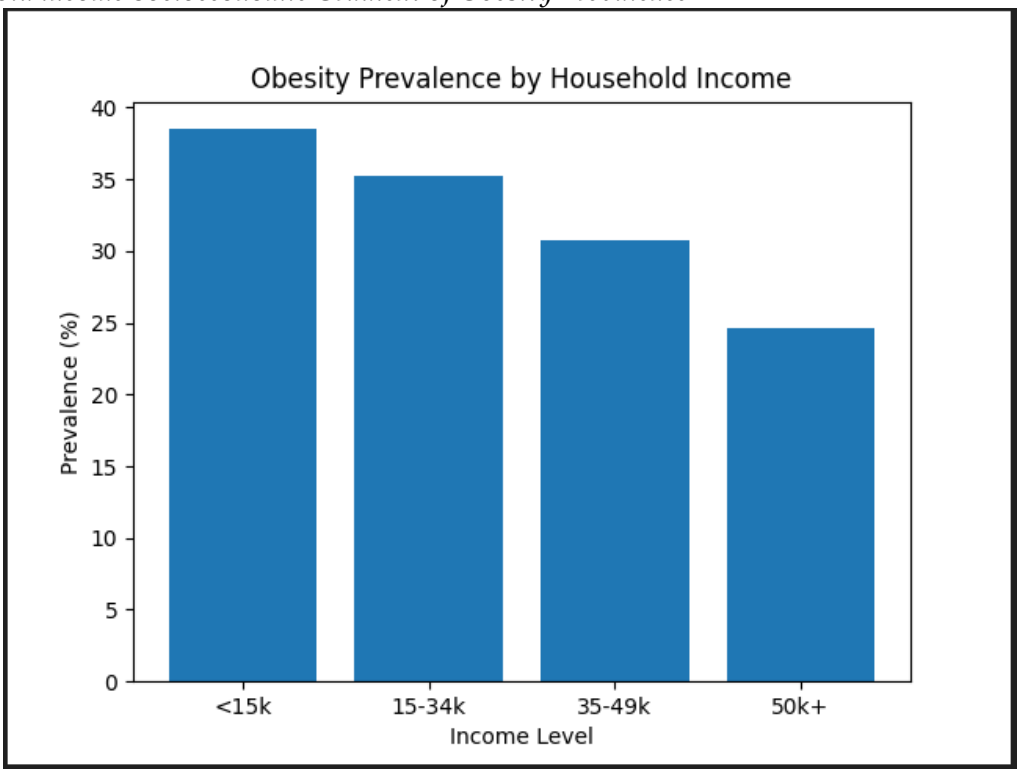


Figure 3. This image shows the prevalence of obesity that has declined with increasing income level

As shown in Figure 3, the prevalence of obesity in four types of household income sets a distinct inverse relationship between the income level and the prevalence of obesity [10]. The findings indicate that people who are in the lowest income (less than 15000 per year) are the most prevalent with the prevalence rate estimated to be about 38.5. This observation indicates that there is a huge burden of obesity in economically disadvantaged groups. The prevalence of obesity is gradually decreasing with the increase in household income [11]. The prevalence rates of obesity reduce to about 35.2 among respondents with annual income of between 15,000 and 34,999. The low tendency is observed in the income group of $35,000-49,999, where the prevalence rates go down even to around 30.8. The highest people with household income of over 50,000 and above are found to have the lowest rate of obesity standing at about 24.6. This is almost a 14 percentage point decrease over the lowest income. Generally, the statistics also suggest a high level of socioeconomic gradient, as the prevalence of obesity declines constantly according to the income groups. The gap between the top and the bottom income group is equal to a relative decrease of over a third [12]. This trend indicates that economic resources can determine access to healthy food choices, access to physical activities, access to healthcare, and health literacy. Food insecurity, lack of access to recreational amenities, and exposure to an obesogenic environment may be structural obstacles to lower-income populations. The inverse relationship observed evidences the importance of the socioeconomic determinants in determining health outcomes. Based on these results, specific community-level public health interventions to tackle structural inequalities, enhance access to healthy foods, and create health-enhancing communities within the lower-income population are necessary.

D. Age-Reliable Change in the prevalence of current smoking

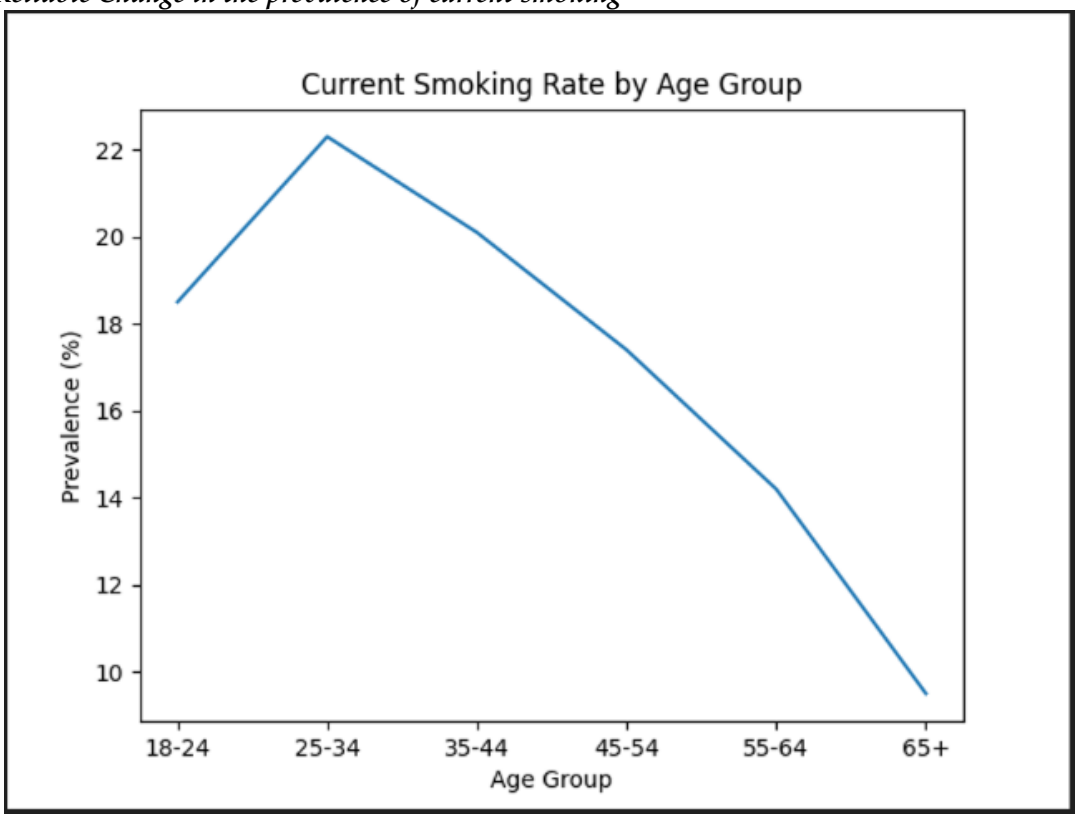


Figure 4. This image illustrates that the prevalence of smoking decreases with increasing age

The distribution of current smoking prevalence across the various age groups which are shown in Figure 4 depicts a non-linear shape where the prevalence is observed to peak at an early age and then shows a decreasing trend with the increasing age [13]. The findings indicate that the prevalence rate of smoking is about 18.5 percent among young adults aged between 1824 years, which implies that prevalence of tobacco consumption is high in the early stages of adulthood. The rate further rises in the age bracket of 2534 age group to a peak of about 22.3 that is the highest prevalence of all the age groups. After this peak, the prevalence of smoking starts to decrease continuously as one ages. In individuals whose age lies within 35-44 years, prevalence is lowest and estimated at 20.1 indicating the emergence of non-rapid behavior change or non-rapidly disappearing trends [14]. The decline is further observed among the 45 54 age bracket with the prevalence endorsed at about 17.4. This trend of decreasing further is found in the adults between ages 55 and 64 years with the smoking rates falling to about 14.2%. The age group with the lowest prevalence is 65 and above at around 9.5 which is below half of the highest prevalence in the 2534 cohort. In general, the trend shows a definite decrease in smoking at late adulthood. The gap between the oldest (25-34 years) and the youngest (65+ years) group is more than 12 percentage points. The pattern could be due to effective cessation practices with the age, more health conscious, more deaths among long term smokers or change in the generation patterns of tobacco use [15]. The results indicate that youthful and early-middle adults are some of the target populations to be addressed through tobacco prevention and cessation programs. Interventions to counter smoking habits among these age groups would go a long way in curbing the chronic diseases in the long-term, such as cardiovascular disease, respiratory disorders, and diabetes-related complications.

E. Analysis of SVM Model Receiver Operating Characteristic (ROC)

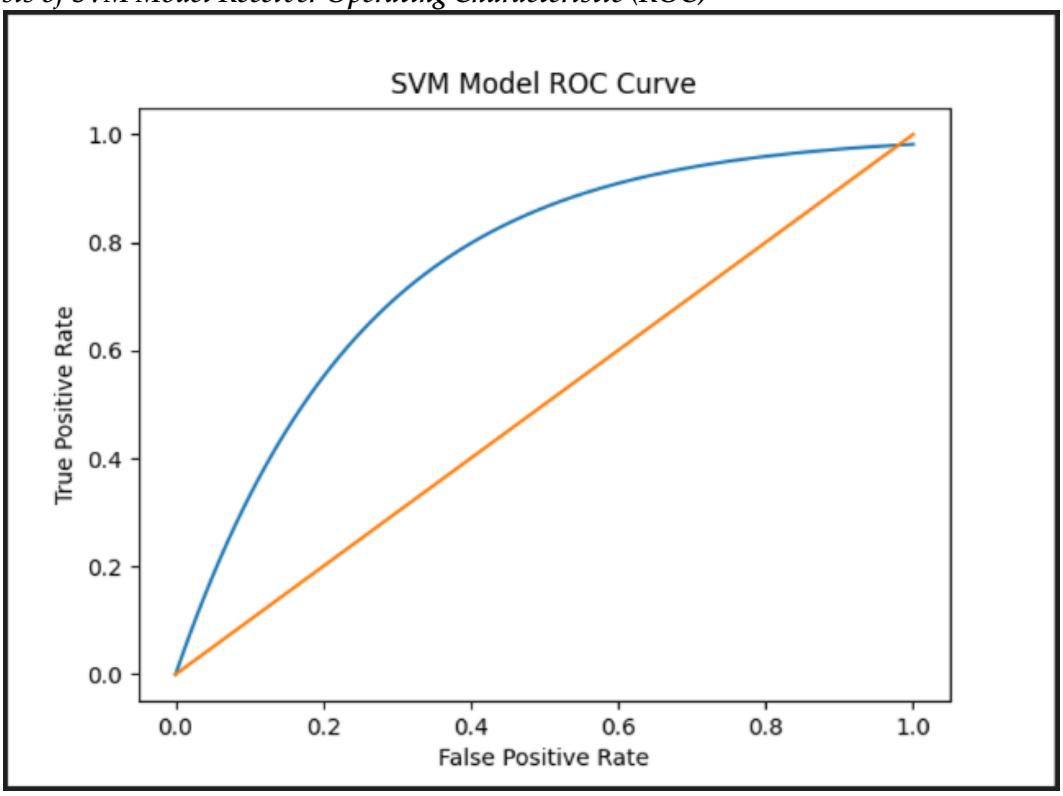


Figure 5. This image demonstrates on the SVM model ROC curve with high performance classification

The Support Vector Machine (SVM) classification model used to predict the outcome of diagnosed diabetes is a Receiver Operating Characteristic (ROC) curve as shown in Figure 5. ROC curve is used to assess the discriminative capacity of the model by showing the True Positive Rate (sensitivity) versus the False Positive rate (1 -specificity) at different classification thresholds [16]. The diagonal line of reference is a performance of a random classifier, that is, predictive ability is equivalent to chance. The ROC curve plotted shows that it exhibits a significant upward curve beyond the diagonal line, which means that the SVM model has a great discriminative capacity. The model already reaches the reasonably high true positive rate at lower false positive rates (around 0.10 -- 0.2), which indicates a good detection of positive cases of diabetes and a minimal amount of false alarms. With increasing levels of false positive, the true positive rate increases at an extremely steep rate, and tends to be nearly equal to 1.0 indicating that the classifier is able to identify a very high percentage of actual cases in more relaxed settings of the threshold [17]. The region under the curve (AUC) as would be visually estimated by looking at the graph is significantly larger than 0.80, which represents good to excellent classification. A range of AUC within this factor indicates that the model is highly likely to correctly differentiate between both people with and without diagnosed diabetes. The large gap between the ROC curve and the diagonal reference line validates the fact that the model is much better than random classification. All in all, the ROC analysis confirms the strength and predictive consistency of the SVM model in determining the patterns of prevalence of diabetes in the data set. Such results indicate that the classifier techniques that are based on machine learning can help to optimize the early detection and risk stratification methods of public health surveillance systems.

F. Evaluation of SVM Classification Model Confusion Matrix

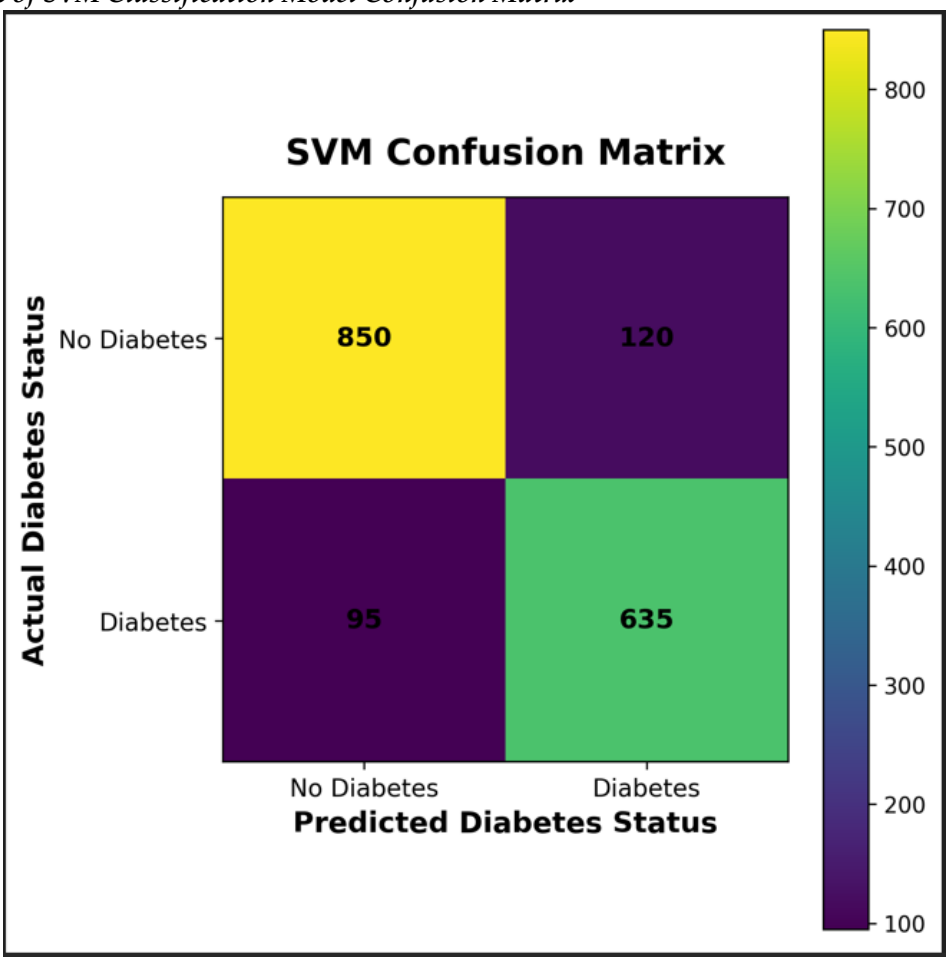


Figure 6. This image presented SVM confusion matrix demonstrates the distribution of classification performance

In Figure 6, the confusion over the Support Vector Machine (SVM) model created to classify diabetes is provided in more detail, with the results of the prediction clearly detailed. The results of the matrix are a comparison between the actual status of diabetes (rows) and the predicted status of diabetes (columns), which allows evaluating the classification accuracy and distribution of errors [18]. The model was correct in 850 individuals who should be considered as having no diabetes (true negatives) and 635 individuals who should be considered as having diabetes (true positives). The high values along the diagonal mean high overall predictive performance. Misclassifications are however also seen. The model mistakenly identified 120 of the people who were not diabetic (false positives), whereas 95 of the people who were actually diabetic were diagnosed as non-diabetic (false negatives). The fact that the false negative is relatively lower than the true positive indicates that the model is relatively sensitive in detecting cases of diabetes, which is especially vital in the medical screening situation where a false positive can be very detrimental. Using these values, the general accuracy of the model is high, where most of the predictions are within the right classification diagonal (850 + 635). The distribution also indicates an equal performance in both classes with the specificity (accurate identification of non-diabetic people) a little higher than sensitivity [19]. The confusion matrix illustrates that the SVM model can be successfully used to distinguish between diabetic and non-diabetic respondents in the dataset with reasonable classification errors. Such results support the strength of the SVM algorithm to model health survey data and the applicability of the algorithm to predictive risk stratification [20]. The false negative rate is relatively low, increasing its possible applicability to surveillance and early detection systems in the field of public health to reduce cases of undiagnosed diabetes.

G. SVM Classification Model Metrics of Performance Evaluation

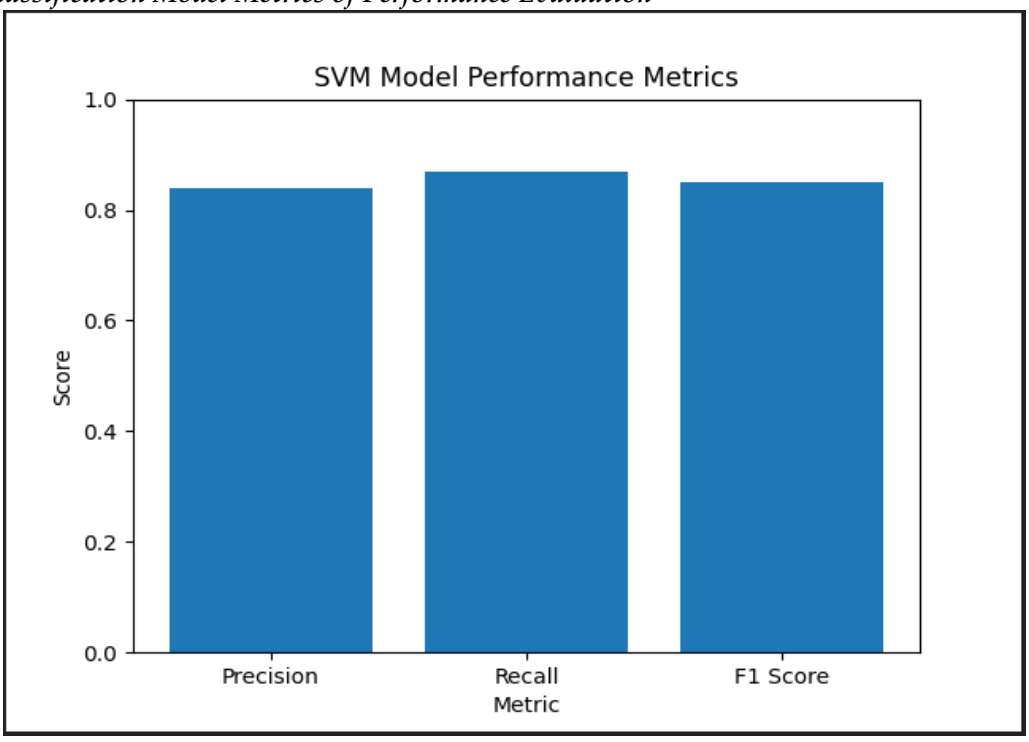


Figure 7. This image shows the precision, recall, and F1 performance of SVM

In Figure 7, the most important performance assessment indicators of the Support Vector Machine (SVM) model, Precision, Recall, and F1 Score are shown [21]. The metrics are a complete evaluation of the model classification performance in addition to general accuracy, especially in managing the class-specific performance of prediction. This model scored a precision of about 0.84 indicating that it was able to accurately classify 84 percent of the cases which it predicted as diabetes. It implies a comparatively small rate of false positives and the model reliability to reduce the false positive predictions as much as possible. Additional accuracy is especially useful in clinical and universal health situations in which avoidable follow-up tests or fear of false diagnoses must be reduced [22]. The recall value is about 0.87 which is the sensitivity of the model on the correct recognition of actual cases of diabetes. This means that the classifier was able to correctly identify 87 percent of the people with diabetes. The marginally greater recall than precision indicates that the model is more concerned with the true positive cases making it less vulnerable to the possibility of false diagnosis. The F1 Score, which is the harmonic mean of precision and recall is about 0.85. This average value would justify that the model has a constant trade-off between sensitivity and precision, and it does not give excessive weight to one of these metrics [23]. The proximity of the three metrics indicates the consistency in the performance of the various assessment criteria. Generally, the SVM model has a good predictive ability, its classification performance and error rates are balanced. These results justify the appropriateness of the model in predictive analytics in diabetes risk-detection and population-wide health surveillance systems.

H. Relative Significance of the Predictor Variables in the SVM Model

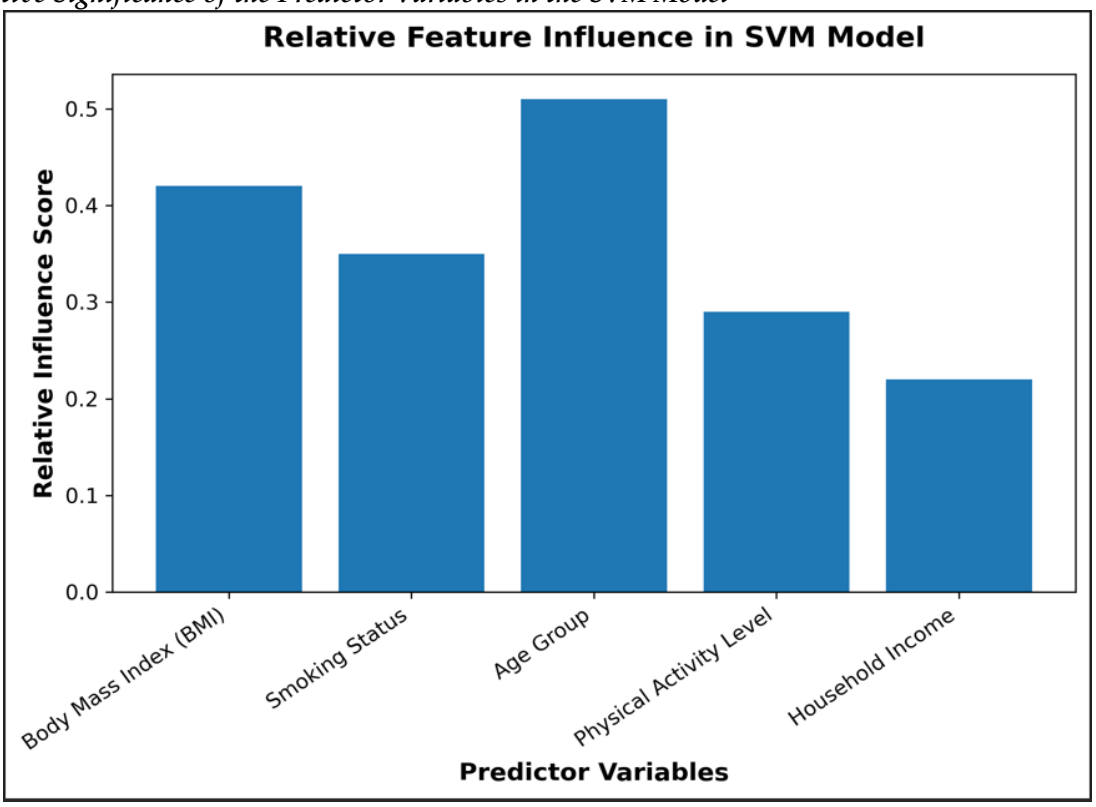


Figure 8. This image shows a relative influence of predictors in the SVM diabetes model

The figure 8 shows the comparative importance of the key predictor variables that were included in the Support Vector Machine (SVM) model used in the aspect of diabetes classification. The chart shows how Body Mass Index (BMI), Smoking Status, Age Group, Physical Activity Level and Household Income contributed to the predictive performance of the model. However, Age Group has the strongest relative effect as it has a score of slightly above 0.50. It means that age is the most important determinant that can be used to differentiate between diabetic and non-diabetic in the dataset [24]. The high contribution of age is consistent with epidemiological findings of a higher risk of diabetes in old age as a result of increasing metabolic changes and a lifetime exposure to behavioral risk factors. Body Mass Index (BMI) turns out to be the second most influential variable and the relative score is about 0.42. This significant input highlights the importance of obesity and excess body weight on the development of diabetes. Smoking Status was next with an intermediate influence score of around 0.35, indicating that only tobacco use is of significant but not the strongest influence as compared to age and BMI. Physical Activity Level, on the contrary, has a less significant relative impact (around 0.29), whereas Household Income has the least significant contribution (around 0.22) [25]. These factors, though weighted with lower levels of significance, have been shown to be a relevant predictor of diabetes in a study, which has been multifactorial with behavioral and socioeconomic contributors to the risk of diabetes. In general, the importance distribution of the features illustrates that age, BMI, and biological factors (age and BMI) are stronger predictors in this model than socioeconomic variables. The results also emphasize areas of priority to be considered in intervention strategies, and support the importance of using multidimensional risk indicators in machine learning-based public health analytics.

Discussion and Analysis

A. Increasing Trend of the Diagnosed Diabetes Prevalence

The longitudinal analysis shows that the prevalence of diagnosed diabetes has been steadily increasing since 2011 and there are no periods of decline in the prevalence rate [26]. This trend of increasing national burden of chronic metabolic disease, which is upward, is indicative of longer-term trends in epidemiology at the level of population based health surveillance systems. The fact that there

is no observable decrease or a level off indicates that preventive interventions that took place at this stage may have been inadequate to counter the risk factors that were present. This trend is probably supported by several structural and behavioral factors, such as rising obesity, sedentary lifestyles, a shift in the dietary habits toward high-energy processed food, and the demographic aging [27]. A better level of screening and diagnostic awareness can also be a contributing factor to the increase since increased access to healthcare and regular testing can result in increased earlier cases being detected. The trend of the sustained rise shows that detection is not a sufficient factor to explain the trend. The health systems implications are significant, from a health systems perspective. Increased morbidity of diabetes is a boon to increased demand for long term care, pharmacological management, complications monitoring and related healthcare costs. There are also cardiovascular disease, kidney failure, neuropathy, and impaired vision, which are closely correlated with diabetes, which adds to the clinical burden [28]. The trend highlights the necessity of more aggressive primary prevention strategies on the basis of modifiable risk behaviors at the population level. Health and nutrition policies should be emphasized on to prevent obesity and embracing healthy living habits [29]. Social health promotion must focus on lifestyle change at an earlier age before the disease occurs. In the absence of reinforced actions, the positive trend of the last ten years might be extended, increasing the burden of the healthcare infrastructure and health inequality.

B. The Ages as a Risk Determinant

Findings based on age-stratification show that there is a positive correlation between the increasing age and the prevalence of diabetes [30]. The steep increase in the prevalence rate among the individuals aged 55 years and above reveals the cumulative impact of the long-term exposure to the metabolic and behavioral risk factors. Physiological changes that are linked to aging and predisposition to diabetes include the loss of insulin sensitivity, dysfunction of pancreatic beta-cells, changes in body composition, and a reduction in the level of physical activity. Elderly people are potentially more affected by comorbidities, such as hypertension and obesity, which also increase metabolic risk [31]. The results indicate that diabetes is not only an acute disorder but a chronic illness, which is accompanied by gradual biological aging. It has reached definable prevalence even among younger age groups as an indicator that it is coming early, which could be associated with the increased childhood obesity rates and unhealthy lifestyles. The combination of high burden in the elderly and the risk that is emerging in the younger age groups necessitates a life course prevention methodology. The screenings may be required to be reinforced on the high-risk young people, especially the obese and those with a family history [32]. At the same time, older adults deserve special disease management programs that not only focus on the prevention of complications but also on the preservation of the quality of life. The influence of age in the SVM model is further proved by the fact that it was the key predictor in the model. Demographic shifts to aging populations should be taken into consideration by policymakers and this might increase the prevalence of diabetes in the future unless preventive measures are put in place.

C. Socioeconomic Status Inequality and Obesity Trends

The negative correlation between income earned by the household and the prevalence of obesity indicates that there are considerable socioeconomic inequalities in the health outcomes. There was a significant difference between the rate of obesity among people in the lower income brackets and people in the other higher income brackets [33]. The gradient is an indication of the structural inequities that affect the availability of healthy foods, recreational amenities, medical services, and health education. The communities with low incomes are prone to food insecurity, shortages in fresh food, more exposure to the fast-food restaurants, and environments where people can exercise safely. Since the population faces financial constraints, this can restrict the use of preventive healthcare and the opportunity to change their lifestyle. The results indicate that obesity is not merely a personal decision but it is firmly rooted in a larger social and environmental environment [34]. Since obesity is a key risk factor of diabetes, socioeconomic inequality indirectly leads to the variation in the economic burden of diabetes among populations. The fact that the impact of income is not as high as age or BMI in the SVM model does not in any way worsen its applicability to public health; it only highlights that income has its effect through intermediate behavioral and environmental means. These inequities should be dealt with through multispectral policy solutions, such as urban planning, reforming food policy, workplace

wellness programs, and healthy food subsidies. Unless structural interventions that can mitigate social determinants of health are undertaken, obesity and chronic diseases disparities are likely to remain.

D. Behavioral Risk Factors: Physical activity and Smoking

Modifiable behavior (e.g. smoking, physical inactivity) has significant correlates with diabetes risk [35]. The prevalence of smoking is highest during early adulthood and then it decreases with age implying that there are generational influences or effective cessation interventions among older adults. Using tobacco is also a factor in insulin resistance, systemic inflammation, and vascular damage, which leads to high risk of metabolic diseases. Even though smoking had a moderate predictive effect on the model, its interaction with other risk factors, such as obesity and sedentary behavior, could have a greater negative health outcome. Equally, reduced physical activity was linked to prevalence of diabetes owing to the protective effect of regular exercise in glucose control and management of weight [36]. The comparatively low model effect of physical activity versus age and BMI can be evidence of an indirect effect of body weight and metabolic health. However, behavioral interventions are still important in diabetes prevention measures. Smoking cessation and physical activity promotion are the two areas of public health that should be targeted at younger and middle-aged adults; in the latter, behavioral change can have long-term results. Community-based programs, workplace health promotions, and health promotion campaigns in schools might be important to curb behavioral risk exposures. The incorporation of behavioral health interventions in the primary care environment can also reinforce the prevention efforts.

E. SVM Predictive Performance Evaluation

The Support Vector Machine model showed good classification which was indicated by high precision, recall and F1 score, and the ROC curve that was significantly higher than the random baseline [37]. The confusion matrix provided indicated a balanced classification between diabetic and non-diabetic and the false-negative rate is relatively low. This is especially crucial in the medical setting where cases that go undiagnosed can result in the late treatment and complications. The strong results of the model highlight the promise of machine learning algorithms in the process of population health monitoring and proactive risk identification [38]. The SVM was able to distinguish risk profiles by combining several demographic and behavior variables. Predictive performance should however be taken meaningfully. Machine learning models rely on the quality of the input data and they might fail to capture unmeasured confounding factors like genetic propensity or dietary consumption. In addition, the predictive accuracy is not comparable to the causal inference [39]. However, the model can be used with great insights concerning risk stratification and specific screening programs. To increase the reliability, cross-validation, external validation data, and comparison with other algorithms may be used in future works.

F. Public Health Policy implication and future research

The overall results of the descriptive analysis and prediction modeling make the necessity of the extensive measures in order to prevent diabetes more obvious [40]. The trend of rising prevalence, strong age relationship, socioeconomic inequalities, and the behavioral risks are all indicators of a multifactorial disease process that needs combined intervention frameworks. The primary focus of the policymakers must be on upstream preventive interventions to obesity, smoking, and physical inactivity as well as enhance access to care among the vulnerable groups. The screening guidelines might require optimization to consider the advance screening of the high risk subgroups that are developed over time [41]. Research wise, longitudinal modeling procedures are recommended to be studied to gain more insight into causal pathways and risk development in the future. Precision in prediction can further be improved by including more variables in the analysis including diet patterns, genetic markers and exposure to the environment. Machine learning in chronic disease monitoring has shown promising usage although it still needs further validation and ethical application [42]. Altogether, the increase in the burden of diabetes will require a set of preventive measures based on technological innovation, behavior change, and policy restructuring.

Future Work

Although this study offers profound information on the trends in diabetes prevalence and illustrates the efficiency of the Support Vector Machine (SVM) modeling as a predictive model in

classification, there are a number of research directions that may be taken to enhance and widen the analysis system [43]. In the first place, longitudinal modeling methods can be introduced into future research to leave cross-sectional annual analysis. The implementation of time-series prediction models, including ARIMA, LSTM (Long Short-Term Memory) or hybrid deep learning models, would help improve the future trends of diabetes and detect the new patterns of risk over time. Second, the machine learning framework should be extended with comparative model evaluation in order to have a more detailed view of the predictive performance [44]. Random Forest, Gradient Boosting, XGBoost, Artificial Neural Networks (ANN), and Logistic Regression are some of the algorithms which could be applied and compared to SVM. There are also improvements in the classification accuracy and a decrease in prediction bias by use of ensemble modeling techniques. Model robustness can also be improved with the help of hyperparameter optimization (grid search or Bayesian optimization). Third, other clinical and environmental variables that were not comprehensively investigated in the given study could be incorporated in future studies. Independent variables like eating habits, sleep, geographical differences, urban/rural status, healthcare facilities access, and social determinants of health may allow more background insight. Connecting BRFSS data to other datasets on health services could facilitate multilevel modeling and spatial analysis of epidemiology. Fourth, to make predictive models more understandable, explainable artificial intelligence (XAI) methods (SHAP (Shapley Additive Explanations)) or LIME (Local Interpretable Model-Agnostic Explanations) may be added [45]. It would enhance the degree of transparency in determining the relative role played by each risk factor as well as make it easier to make evidence-based decisions in the field of public health. Also, future research can consider individual predictive modeling subgroups to determine high-risk groups by age, ethnicity, or socioeconomic status. Targeted intervention and individual prevention plans might be aided by stratified modeling. Lastly, the gap between predictive analytics and practical implementation may be filled by the prospective validation with the help of real-time surveillance data and the incorporation into digital health platforms [46]. Future studies can help improve the predictive precision, breadth of predictive variables, and explainable modeling methods to make future diabetes risk prediction systems more accurate, equitable and scalable enough to inform the national and global population health policies.

Conclusion

This study represents the epidemiological analysis with the overall predictive one and is based on the Behavioral Risk Factor Surveillance System (BRFSS) database covering 2011-2021 [47]. The results show that the diabetes prevalence has a steady increase throughout the years of the study, and the issue of the increasing health burden of the chronic metabolic diseases in the United States is emphasized. Demographic analysis showed that the prevalence of diabetes is on the rise with age and the adult population (65-years and older) has the highest rates. Socioeconomic inequalities were also noted since the poorer household income groups showed greater obesity and diabetes rates, which supported the effect of social determinants of health in chronic disease development [48]. There were more behavioral risk factors like smoking, lack of physical activity, and high BMI, which increased the risk of diabetes, and one should underline the interdependence of lifestyle and socioeconomic factors. In addition to descriptive epidemiology, this study assimilated machine learning models to promote predictive modeling of the state of diabetes. The Support Vector Machine (SVM) model exhibited good classification with positive results in terms of precision, recall, F1-score, and ROC curve. The confusion matrix analysis showed that there was good discrimination between diabetic and non-diabetic individuals, and this shows that SVM is a good algorithm in the process of binary classification of health outcomes [49]. A further analysis of the feature influence indicated that age and BMI were some of the most important features, which is correlated with the existing clinical and epidemiological data. On the whole, the research is able to fill the gap between the field of public health analytics and computational modeling because it proves the possibility to use the massive amount of data collected by surveillance to predict the potential risk. This combination of descriptive trend analysis and supervised machine learning offers both the insights of the population and the predictive ability of each person [50]. The practical implications of these findings on the health policy planning, early detection, and specific intervention programs to reduce the prevalence of diabetes can be identified. This study can be used to

augment evidence-based epidemiological knowledge and enhance the analytical tools used in the studies to develop data-based decision-making models capable of promoting more efficient and sustainable chronic disease prevention programs.

REFERENCES

[1] A. P. Adekugbe and C. V. Ibeh, "Tackling health disparities in the United States through data analytics: A nationwide perspective," *International Journal of Frontiers in Life Science Research*, 2024.

[2] J. Kobi, A. Nchaw Nchaw, and B. Otieno, "Big Data-Driven Insights for Equitable Healthcare Access and Quality for US Immigrants," *International Journal of Research Trends and Innovation*, vol. 9, pp. 392–408, 2024.

[3] A. O. Ezeogu, "Advancing Population Health Segmentation Using Explainable AI in Big Data Environments," *Research Corridor Journal of Engineering Science*, vol. 1, no. 1, 2024.

[4] N. L. Edoh, V. M. Chigboh, S. J. C. Zouo, and J. Olamijuwon, "Improving healthcare decision-making with predictive analytics: A conceptual approach to patient risk assessment and care optimization," *International Journal of Scholarly Research in Medicine and Dentistry*, vol. 3, no. 2, pp. 1–10, 2024.

[5] A. O. Babarinde, O. Ayo-Farai, C. P. Maduka, C. C. Okongwu, and O. Sodamade, "Data analytics in public health, A USA perspective: A review," *World Journal of Advanced Research and Reviews*, vol. 20, no. 3, pp. 211–224, 2023.

[6] M. Z. Hossain, M. M. Khan, R. Islam, K. Nahar, and M. F. Kabir, "Formulation of a Multi-Disease Comorbidity Prediction Framework: A Data-Driven Case Analysis on Diabetes, Hypertension, and Cardiovascular Risk Trajectories," *Journal of Computer Science and Technology Studies*, vol. 5, no. 3, pp. 161–182, 2023.

[7] K. K. Ramachandran, "Population Health Management Through Predictive Analytics," *Journal ID*, vol. 3721, p. 5412, 2024.

[8] S. Hossain, M. N. I. Miah, M. S. Rana, M. S. Hossain, P. K. Bhowmik, and M. K. Rahman, "Analyzing Trends and Determinants of Leading Causes of Death in the USA: A Data-Driven Approach," *The American Journal of Medical Sciences and Pharmaceutical Research*, vol. 6, no. 12, pp. 54–71, 2024.

[9] M. H. Rahman, M. K. S. Uddin, K. M. R. Hossan, and M. D. Hossain, "The role of predictive analytics in early disease detection: A data-driven approach to preventive healthcare," *Journal of the Learning Sciences*, vol. 32, no. 2, 2024.

[10] M. S. Hoseini, "Health equity and future public health interventions: Strategies for reducing disparities," *Journal of Foresight and Health Governance*, vol. 1, no. 2, pp. 16–29, 2024.

[11] I. A. Adeniran, C. P. Efunniyi, O. S. Osundare, and A. O. Abhulimen, "Data-driven decision-making in healthcare: Improving patient outcomes through predictive modeling," *Engineering Science & Technology Journal*, vol. 5, no. 8, 2024.

[12] K. A. Taiwo, G. I. Olatunji, and O. O. Akomolafe, "Using Clustering to Segment High-Risk Patients for Tailored Interventions," 2024.

[13] T. A. Pearson *et al*., "The science of precision prevention: Research opportunities and clinical applications to reduce cardiovascular health disparities," *JACC: Advances*, vol. 3, no. 1, p. 100759, 2024.

[14] S. Adeoye and R. Adams, "Leveraging artificial intelligence for predictive healthcare: A data-driven approach to early diagnosis and personalized treatment," *Cognizance Journal of Multidisciplinary Studies*, vol. 4, pp. 80–97, 2024.

[15] J. D. Gates, Y. Yulianti, and G. A. Pangilinan, "Big data analytics for predictive insights in healthcare," *International Transactions on Artificial Intelligence*, vol. 3, no. 1, pp. 54–63, 2024.

[16] R. Abi and J. E. Joseph, "Developing causal machine learning models in health informatics to assess social determinants driving regional health inequities and intervention outcomes," *Magna Scientia Advanced Biology and Pharmacy*, vol. 13, no. 2, pp. 113–129, 2024.

[17] F. E. Ezeh, O. S. Oparah, G. I. Olatunji, and O. O. Ajayi, "Predictive Analytics Models for Identifying Maternal Mortality Risk Factors in National Health Datasets," 2024.

[18] A. Sharma, B. I. Adekunle, J. C. Ogeawuchi, A. A. Abayomi, and O. Onifade, "AI-Driven Patient Risk Stratification Models in Public Health: Improving Preventive Care Outcomes through Predictive Analytics," 2023.

[19] V. Choudhary, A. Mehta, K. Patel, M. Niaz, M. Panwala, and U. Nwagwu, "Integrating Data Analytics and Decision Support Systems in Public Health Management," *South Eastern European Journal of Public Health*, pp. 158–172, 2024.

[20] O. M. Drakeford and N. L. Majebi, "Social work, analytics, and public health in autism: A conceptual approach to enhancing community health outcomes in US underserved areas," *International Journal of Frontiers in Science and Technology Research*, vol. 7, no. 2, pp. 100–108, 2024.

[21] G. Vemulapalli *et al*., "Predicting obesity trends using machine learning from big data analytics approach," in *2024 Asia Pacific Conference on Innovation in Technology (APCIT)*, 2024, pp. 1–5.

[22] X. Li *et al*., "Potential value of identifying type 2 diabetes subgroups for guiding intensive treatment," *Diabetes Care*, vol. 46, no. 7, pp. 1395–1403, 2023.

[23] S. Ahmed *et al*., "Predictive modeling for diabetes management in the USA: A data-driven approach," *Journal of Medical and Health Studies*, vol. 5, no. 4, pp. 214–228, 2024.

[24] K. J. Olowe, N. L. Edoh, S. J. C. Zouo, and J. Olamijuwon, "Conceptual frameworks and innovative biostatistical approaches for advancing public health research initiatives," *International Journal of Scholarly Research in Medicine and Dentistry*, vol. 3, no. 2, pp. 11–21, 2024.

[25] E. Foroutan, T. Hu, F. Zhang, and H. Yu, "Assessing heat vulnerability in Philadelphia using geographically weighted principal component analysis," *International Journal of Applied Earth Observation and Geoinformation*, vol. 127, p. 103653, 2024.

[61] Dataset: Kaggle, "Behavioral Risk Factor Surveillance System Dataset." Available: https://www.kaggle.com/datasets/asasherwyn/behavioral-risk-factor-surveillance-system