MACHINE AND DEEP LEARNING APPROACH FOR TYPE 2 DIABETES PREDICTION USING THE CDC'S BRFSS DATASET:

A RETROSPECTIVE ANALYSIS

A THESIS IN
Bioinformatics

Presented the faculty of the University
of Missouri-Kansas City in partial fulfillment of
the requirement for the degree

MASTER OF SCIENCE

BY
JUSTIN NGOYI MPANGA

B.S., University of Missouri-Kansas City, 2019

Kansas City, Missouri
2022

MACHINE AND DEEP LEARNING APPROACH FOR TYPE 2 DIABETES PREDICTION USING THE CDC'S BRFSS DATASET: A RETROSPECTIVE ANALYSIS

Justin Ngoyi Mpanga, Candidate for the Master of Science Degree
University of Missouri-Kansas City, 2022

ABSTRACT

Type 2 diabetes mellitus (T2DM) is a complex metabolic disease which is characterized by persistent hyperglycemia caused by insulin resistance. It is the most prevalent type of diabetes mellitus (DM). T2DM presents a heterogenous etiology with social, environmental, behavioral, and genetic risk factors. It is associated with serious microvascular and macrovascular complications which are also associated with increased morbidity, mortality, and health expenditure. However, early detection, lifestyle changes and treatment may prevent or delay the onset of associated long-term complications.

This study used the 2020 Behavioral Risk Factor Surveillance System (BRFSS) dataset to train different machine learning (ML) and neural network or multilayer perceptron classifier (NN) model(s) and test their performance on predicting the risk for T2DM. A copy of the dataset was transformed to have balanced classes in the outcome variable to allow further comparison of performance for each predictive model when trained with either the original or transformed dataset. A cross-sectional data analysis using chi-square was employed to investigate the association of selected predictors or risk factors with T2DM.

Metrics used to assess model performance included accuracy, area under the curve-receiver operating characteristics (ROC-AUC), precision, recall, and F1-score.

When models were trained on the original train dataset (data with significant outcome variable class imbalance), accuracy ranged from 71.6% to 81%, ROC-AUC from 0.57 to 0.75, precision from 0% to 55.7%, recall from 0% to 38.3%, and F1-score from 0% to 38%. ROC-AUC for Decision Tree Classifier (DT) was 0.57, K-Nearest Neighbors Classifier (KNN) was 0.65, and Support Vector Classifier (SVC) was 0.68 which interpreted to a failed or poor predictive models. But these models had satisfactory or good accuracy. Training models on the original train dataset caused models to overfit the majority class. Thus, they had poor recall or sensitivity, precision and F1-score values which are crucial in detecting positive, false positives and false negative classes for T2DM. Also, time it took a model to train on training data and score on test data was evaluated and SVC had the longest times for both training and scoring while NN model took long to train but was faster to score. When models were trained on transformed data (data with balanced outcome variable classes), accuracy ranged from 66.7% to 82.5%, ROC-AUC from 0.73 to 0.91, precision from 66.9% to 79.7%, recall from 66.4% to 92.1%, and F1-score from 66.5% to 83.2%. This comparison clearly showed Random Forest Classifier (RF) to be the best performing model with consistently good and excellent fit across all metrics (accuracy: 82.5%, ROC-AUC: 0.91, precision: 79.7%, recall: 87.0%, and F1-score: 83.2%). Gaussian Naïve Bayes classifier (GNB) had the poorest fit across all metrics. Again, SVC was the worst model time wise. All models showed significant increase in recall, precision and F1-score values suggesting that significant outcome class imbalance has a negative effect on all models. RF, KNN, and DT had F1-score values of 83.2%, 80.9%, and 78.7%, recall values of 87.0%, 92.1% and, 83.0% and precision values of 79.7%, 72.2%, and 74.7%, respectively.

Of all models, RF, KNN, and DT showed high performance across all metrics. KNN had the fastest training but longest testing time, RF and DT slightly slower train and fast testing time. These models are good candidates for initial T2DM screening, but RF is the model of choice.

APPROVAL PAGE

The faculty listed below, appointed by the Dean of the School of Medicine have examined a thesis titled "Machine and Deep Learning Approach for Type 2 Diabetes Prediction Using the CDC's BRFSS Dataset: A Retrospective Analysis," presented by Justin Ngoyi Mpanga, candidate for the Master of Science degree, and certify that in their opinion it is worthy of acceptance

Supervisory Committee

Monica Gaddis, Ph.D., Committee Chair
Department of Biomedical and Health Informatics

Betty Drees, MD, Research Advisor
Department(s) of Internal Medicine,
Department of Biomedical and Health Informatics

Stephen Simon, Ph.D., Research Advisor
Department of Biomedical and Health Informatics

CONTENTS

ILLUSTRATIONS

TABLES

ACKNOWLEDGMENTS

Foremost, I would like to express my sincere gratitude to my advisor, mentor and chair of my thesis committee, Dr. Monica Gaddis PhD, for the continuous support throughout my graduate journey, for her patience, enthusiasm, motivation, care, immense knowledge, and her time even when it was family time. Her guidance helped me in all the time of authoring this thesis.

Besides my advisor, I would like to express my deep and sincere gratitude to my mentors and members of my thesis committee, Dr. Stephen Simon PhD and Dr. Betty Drees MD, whose separate works in clinical research and related fields are the inspiration behind this thesis. They found time to help and to answer my inquiries regardless of time and day of the week and they did not hesitate to ask challenging questions or give constructive critics.

Lastly, I would like thank God for his unconditional love. I would also like to thank my family and friends for their support, motivation, and patience. I could not have imagined having a better support system. Thank you!

CHAPTER 1

INTRODUCTION

T2DM is the most prevalent type of diabetes mellitus (DM) among the three broad classifications of the disease.[1] The other two are Type 1 Diabetes Mellitus (T1DM) and Gestational Diabetes Mellitus (GDM).[2] DM is a heterogeneous metabolic disorder involving carbohydrates, lipids, and protein metabolism. It is characterized by destruction of insulin producing cells and/or cells resistance to insulin causing persistent hyperglycemia, a condition portrayed with increased glucose levels.[3,4] Despite the longevity of its existence and abundant knowledge from various research studies, there is no specific cure for T2DM. Management includes interventions for lifestyle change and treatments for its symptoms.

Studies have shown factors such as body mass index (BMI), exercise, family history, race, age, gestational diabetes, and high blood pressure to be good predictors for occurrence of T2DM.[5–9] For instance, a person with a BMI of greater than or equal to 25, no exercise, and has history of gestational diabetes has an increased risk for the disease compared to someone with a BMI of less than 25 who exercises at least 30 minutes on most days and had no personal or family history of diabetes. Again, the disease cannot be cured but it can be prevented and controlled with lifestyle changes.

T2DM affects millions of people worldwide. In the United States, more than 30 million people (about the population of Texas) are affected by and exposed to high medical costs. These costs prevent a sizable number of affected individuals from accessing the care and treatment needed to maintain blood glucose levels.[10,11] Untreated hyperglycemia leads to serious aberrations in prominent parts of the body causing various microvascular and macrovascular complications which are associated with increased morbidity and mortality.[12]

Knowledge of risk factors or predictors for T2DM has resulted in effective diabetes prevention programs. Availability of programs that promote exercise and healthy lifestyles suggests that T2DM can be prevented, or associated complications can be prevented or delayed as well, with important lifestyle changes. This study aimed to compare performance of different T2DM predictive models, show how accuracy alone can be a misleading measure of predictive models' performance, and discuss how the best performing models can be used in T2DM prevention or detection of high-risk persons who need care and treatment.

CHAPTER 2

REVIEW OF LITERATURE

Ancient History of Diabetes Mellitus

Accounts of DM dates to ancient Egyptians, Indians, Chinese and Arab physicians who described its symptoms from 1500 BC to the 11th century AD.[13] They observed distinctive features of DM and suggested various treatment options. An Egyptian patient with excessive thirst and copious urination was treated with plant extracts. An Indian surgeon observed honey-like urine which was sticky and sweet, attracting ants. The surgeon also made an association between the prevalence of DM and high socioeconomic classes (rich castes). They consumed excess foods such as sweets and rice. Chinese physicians observed similar symptoms with an addition of significant weight loss and proposed refraining from sex, wine, and salt to treat the disease. Arab physicians described complications associated with DM such as gangrene and sexual dysfunction.[14]

Aristaeus of Cappadocia, a Greco-Roman physician, coined and accurately described diabetes. Thomas Willis, an English anatomist, and physician, coined the term mellitus.[14] Claude Bernard, a French physiologist, discovered glycogenic actions of the liver illuminating the pathway of gluconeogenesis and promoting the study of diabetes.[14] Oskar Minkowski and Joseph Mering's experiment on dogs demonstrated the key role of the pancreas in the maintenance of glucose homeostasis. Minkowski and Mering's experiment paved the way for the discovery of insulin.[15] Initially, Frederick Banting, a Canadian surgeon, and a Charles Best, a medical student, were collaborating on an experiment involving ligation of dogs' pancreatic ducts which caused degeneration of the pancreas. The degenerated pancreases were later removed for insulin extraction. From there on, insulin became an effective treatment for DM’s undisputed symptom, hyperglycemia.[16] For the past century, we have not crossed from treating symptoms for T2DM to treating T2DM itself. Modern studies have shown a different approach to fighting the epidemic by studying the association between environmental/lifestyle risk factors and T2DM.

Modern Discovery

More recently, researchers have recognized social, environmental, and behavioral risk factors for T2DM that need to be considered when focusing on improving the health and wellbeing of affected individuals. Social and environmental risk factors are conditions in which a person is born, lives, grows and work. These include socioeconomic status, race and/or ethnicity, air pollution, poor water quality, climate change and disease-causing microbes. On the other hand, behavioral risk factors are unhealthy behaviors that can be changed such as sedentary lifestyle, tobacco, and excessive alcohol use. Understanding the association of these risk factors and T2DM is crucial when creating predictive models and dealing with the disease.

Studies and Reviews on Risk Factors for T2DM

A study in India showed decreased level of exercise and high BMI to be associated with increased risk for T2DM.[17] A systemic review of 60 study articles retrieved from Scopus, Science Direct, Pub Med and Web of Science showed walkability, air pollution, food, roadway proximity and physical activity environment to be the most studied environmental risk factors for the T2DM. Noise and pollution were associated with increased risk, while increased walkability and green spaces with reduced risk for the disease.[18] Evidence-based reviews discussed the positive association between T2DM and environmental or lifestyle risk factors such as poor diet quality and quantity, reduced physical activity, increased screen viewing time, exposure to noise and fine dust, poor sleep or sleep deprivation, smoking, stress or depression, and low socioeconomic status. These factors were then associated with increased BMI.[6,11,19]

Studies on food intake and T2DM

There are extensive studies that evaluate and show the link between individual food items and T2DM. Intake of cereal fibers, whole grains, dairy products, higher green leafy vegetables, anthocyanin-rich foods, and coffee have been associated with reduced risk for T2DM.[20,21] But intake of white rice (processed grains), red and processed meat, sugar sweetened beverages and heavy alcohol consumption have been associated with increased risk for the disease.[21–23] Certain foods had different association with T2DM in different parts of the world or simply different populations. This may have been due to different methods of preparation. For instance, high fish and/or seafood consumption was associated with increased risk in North America and Europe but decreased risk in Asia. Also, heavy alcohol consumption was associated with reduced risk among overweight individuals and increased risk among normal weight individuals. [21] These differences emphasize the importance of a generalizable sample in T2DM prediction studies.

Studies on Physical Activities, Socioeconomic Status and T2DM

Physical inactivity has been associated with increased risk for T2DM in meta-analysis studies. Sedentary behaviors such as increased screen viewing time are an increasing issue due to increased adoption of online shopping and binge-watching television series. Meta-analysis studies showed the association of moderate to high intensity exercise with reduced T2DM.[21,24–26] Physical activity was associated with socioeconomic status (SES) in terms of lack of green or exercising space in lower SES neighborhoods.[26] Determinants for SES included levels of education, occupation, and income. Lower SES was associated with increased risk for T2DM in a meta-analysis of studies from countries in Europe, Africa, Asia, and the Americas as well as studies within the US.[21,26] The effects of having a lower SES include but are not limited to lack of access to healthy foods and places to exercise as well as living unhealth lifestyles[11]. Thus, lower SES is a contributing factor to an increased risk of T2DM.

Studies on Sleep, Depression, Smoking, Heavy Alcohol Consumption and T2DM

Sleep quality and quantity has been associated with T2DM in various meta-analysis cohort studies. Sleep times shorter than six hours, longer than eight hours, and trouble initiating or staying asleep have been associated with increased risk for the disease.[21,27,28] Also, extended night shift positions increased the risk for T2DM. Sleep apnea was associated with overweight individuals then with T2DM. These sleep habits may increase fatigue and promote a sedentary lifestyle.[21] Poor sleep was also associated with depression, heavy alcohol consumption, and smoking.[29] All these unhealthy lifestyles were among major behavioral risk predictors for T2DM.[21]

Studies Predicting T2DM

A study by Korean researchers used longitudinal data to create T2DM models to predict subsequent disease occurrence in the following year. The study used deidentified electronic health record (EHR) data from Hanaro Medical foundation in Seoul, South Korea with 535169 instances, 253395 subjects and 1444 features.[30] Feature selection was used to select the most noteworthy predictors for the final machine learning (ML) model. Features of interest were fasting plasma glucose (FPG), Hemoglobin A1c (HbA1c), triglycerides, body mass index (BMI), gamma glutamyl transpeptidase (gamma-GTP), sex, age, uric acid, smoking, drinking, physical activity, and family history. The prediction models were created using random forest (RF), XGBoost (XGB), and support vector machine (SVM) methods. RF and SVM models had the highest accuracy (73%) followed by GBT (72%). The models were satisfactory at predicting the incidence of T2DM in the Korean population prior to developing the disease.[30] This study used accuracy, precision, recall, and F1-score to measure model performance. The study was specific to the Korean population.

A 2021 systemic review of 90 peer reviewed articles on T2DM and ML and NN from both Pub-Med and Web of Science showed that there was no agreement regarding specific features for predictive models. Some studies had up to 70 features in their models. However, predictors like lifestyle, SES and diagnostic data produced better models. Most studies used SVM, RF, GBT and deep neural network (DNN) methods to create predictive models. However, there was considerable heterogeneity for optimal validation metrics among studies, making it harder to compare models. RF had the highest average AUC (ROC) of 0.98. RF together with DNN had the advantage of dealing with big and “dirty” data. ML models presented better performances with fewer observations. On the other hand, DL (Deep Learning) models performed better with more than 70000 observations according to the review.[31] This review only assessed model accuracy and AUC(ROC) to compare model performance.

A 2019 study created and compared a total of eight predictive models for T2DM using the 2014 Behavioral Risk Factor Surveillance System data. The models included SVM, (DT), logistic regression (LR), RF, Multilayer Perceptron Classifier (NN), GNB, univariate and multivariate LR models. The results showed NN to have the highest AUC (0.90). However, DT had the highest sensitivity and was preferred for initial screening for T2DM.[32]

There are no blueprints when it comes to T2DM prediction models. There are no predefined features or limitations to the number of features a study can use. However, features or predictors such as BMI, sex, age, smoking, drinking, physical activity, and family history appeared in almost all studies. Some models are sensitive to the number of features while others are not. Some models are preferred for smaller datasets while others produce better results with big data.

Research Objectives

This study aimed to compare performances of different T2DM predictive models and show how model accuracy alone could be a misleading measure of T2DM models' performance. The hypothesis was that T2DM predictive models trained on the dataset with significant outcome class imbalance would have poor sensitivity, precision, and F1-score values.

CHAPTER 3

METHODOLOGY

Data Source and Study Population

The study used the 2020 Behavioral Risk Factor Surveillance System (BRFSS) dataset. BRFSS was established in 1984 and it is the Unites States' premier system of health-related telephone surveys. It is a collaboration of 53 U.S. states and territories and the Centers for Disease Control (CDC). The intent of the data set is to monitor health related risk behaviors, chronic medical conditions, and use of preventive services. The target population is US adults aged 18 years and older with working cellular phones who resided in a private residence or college housing.[33]

Sample Description

A randomly selected telephone number was a sample. Participating states had to show that a sample record was a representation or probability of all telephone owning households in the state. The requirement was met in 2020. Fifty-one states used a disproportionate stratified sample (DSS) design while Guam and Puerto Rico used a simple random-sample design. Telcordia database and 1,000 banks were the basis of the sampling frame. The BRFSS drew a sample from one of the created intervals. An interval (K) equals total telephone numbers in frame (N) divided by sample size (n), or $\mathrm{K} = \frac{\mathrm{N}}{\mathrm{n}}$. Most of the states sampled from strata corresponding to sub-state regions.[33]

Integrity of Data Collection

Data was collected via telephone surveys. Repeated BRFSS questionnaire and procedures training were required and offered to every interviewer before approval to conduct interviews. All BRFSS surveillance sites monitored interviewers' performance per

BRFSS guidelines. Monitoring techniques included listening to the interviewer on-site or listening to both the interviewer and interviewee remotely. The questionnaire consisted of three parts arranged in the following order: core components (questions that all states used), optional BRFSS modules and State-added questions. This arrangement was key for comparability across states. The CDC provided states with core components and optional modules. States chose a module(s) specific to their programs then, compiled a questionnaire with all components which was then submitted to the CDC. States also had the option to translate the questionnaire into other languages.[33] The BRFSS data is reliable because of its robust data collection guidelines and implementation techniques.

Study Variable and Data Preprocessing

Variables used in previous years were absent in the 2020 BRFSS dataset. Using exact variables as previous studies for comparison was not feasible. Instead, the study selected variables based on prior literature and availability. Selected variables were age, race, sex, BMI, general health, physical health, mental health, smoking status, income, education, location, flu shot, employment, relationship status, sleep time and exercise. Variables names were changed from BRFSS codes to normal language (Table 6). They were also assessed for completeness and skewness.

Python and R were the only software tools used in this study. The BRFSS dataset from the CDC's website was an XPT file, a data format from SAS. R was only used to convert the XPT file to a comma-separated file (CSV) while Python was used for data preprocessing, visualization, and analysis. The original dataset had 401958 rows and 280 columns. A new dataset was created with seventeen selected features and the target column. The dataset was assessed and cleared of missing values and duplicate rows. Independent and dependent variable(s) were each assessed. Age younger than 40 was excluded to reduce chances for Type 1 diabetes Mellitus (T1DM) diagnosis in the dependent variable. Records in any column showing “refused” or “don’t know” were removed. Data was normalized using ‘Yeo-Johnson,’ an improved version of Box-Cox introduced by Yeo and Johnson.[34] Finally, variables categories were coded as shown in Table 5. After data preprocessing and feature selection, there were 69467 rows and 17 columns.

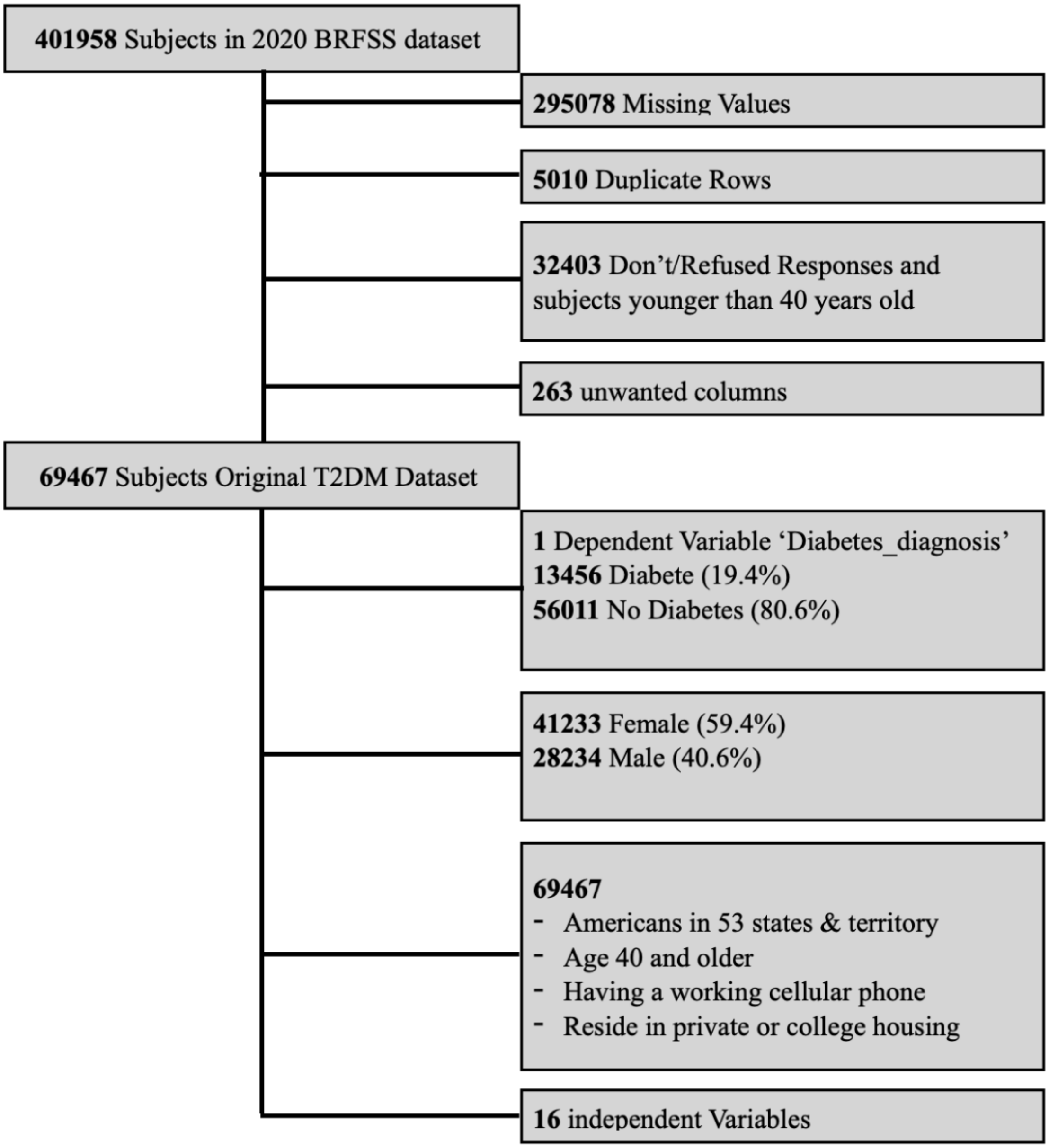

401958 Subjects in 2020 BRFSS dataset
295078 Missing Values
5010 Duplicate Rows
32403 Don't/Refused Responses and subjects younger than 40 years old
263 unwanted columns
69467 Subjects Original T2DM Dataset
1 Dependent Variable 'Diabetes_diagnosis'
13456 Diabete (19.4%)
56011 No Diabetes (80.6%)
41233 Female (59.4%)
28234 Male (40.6%)
69467
- Americans in 53 states & territory
- Age 40 and older
- Having a working cellular phone
- Reside in private or college housing
16 independent Variables

Figure 1: Diagram of data preprocessing and analytical population

Dependent Variable

Type 2 Diabetes diagnosis was the dependent or outcome variable with two classes ("diabetes" and "no diabetes"). There was a significant class imbalance.

Statistical Analysis

Data was categorical. Therefore, chi-square tests were used to evaluate the association between individual predictors and the outcome variable to determine feature importance.

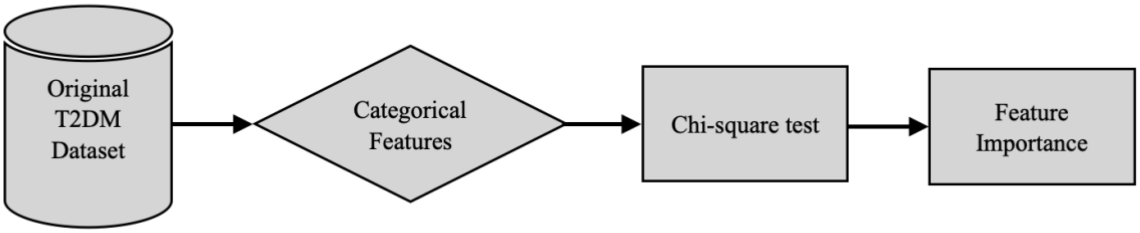


Figure 2: Process for feature importance
BRFSS: Behavioral Risk Factor Surveillance System

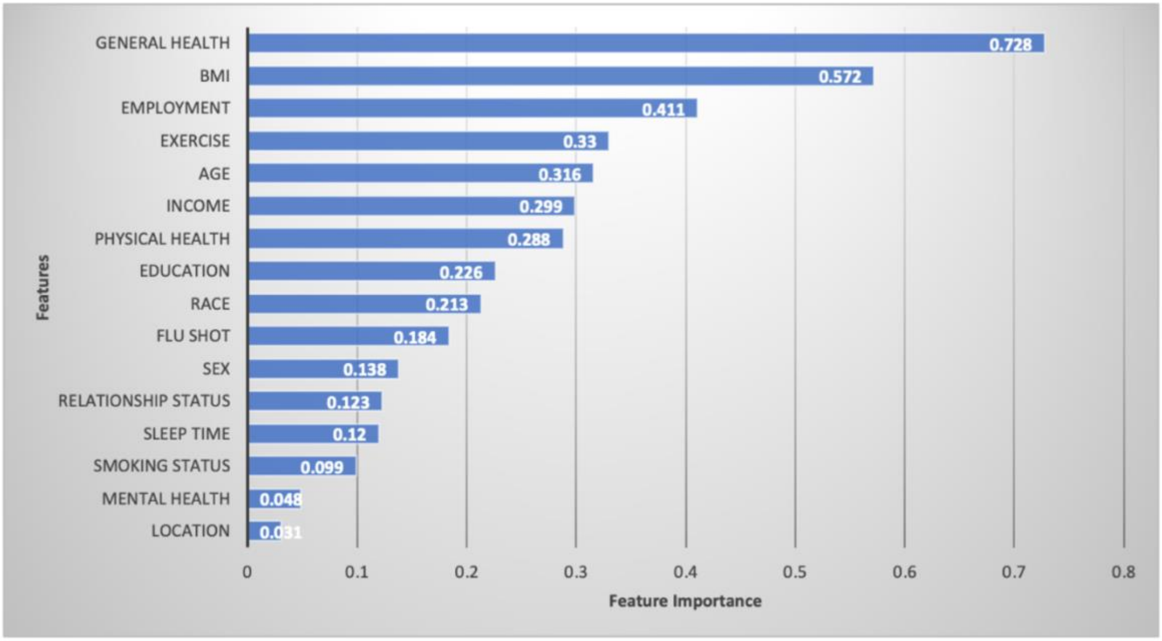


Figure 3: Feature importance ranking

Prediction Models, Training and Testing

The study compared eight ML and NN prediction models prediction of DM. The models included KNN, RF, GNB, LR, DT, XGB, SVC and NN (Figure 4). The new dataset was split into a training dataset (80%) and a validation dataset (20%). All data transformations such as normalizing, and oversampling were performed on the training set. After fitting or training models, each model was fed new, unseen, and untransformed data (validation set) to measure model accuracy (Figure 5). All models were also used in their default form. No parameter tuning was employed. For comparison, F1-scores, ROC-AUC, recall, accuracy, and precision of models on the validation set were used.

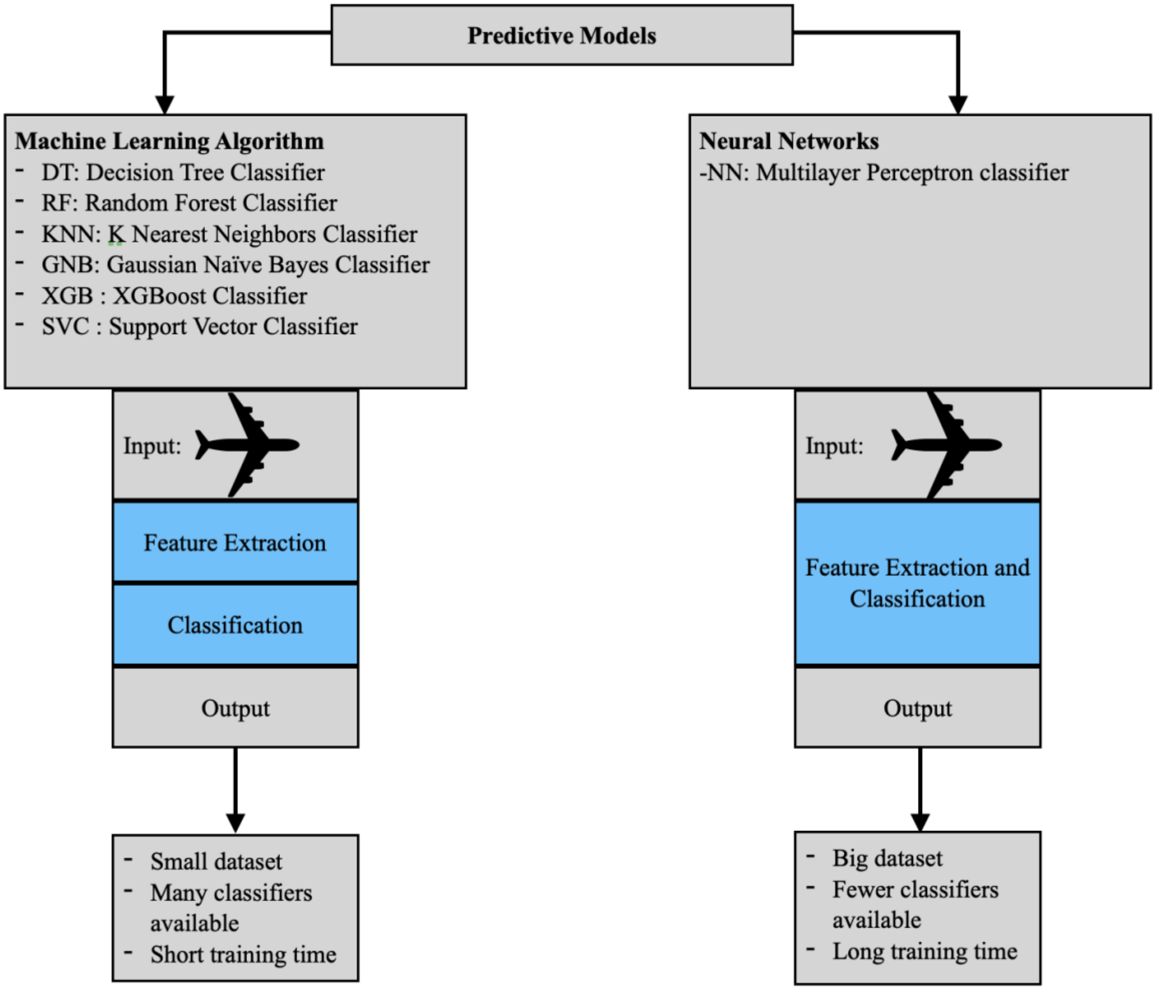

Predictive Models
Machine Learning Algorithm
- DT: Decision Tree Classifier
- RF: Random Forest Classifier
- KNN: K Nearest Neighbors Classifier
- GNB: Gaussian Naïve Bayes Classifier
- XGB : XGBoost Classifier
- SVC : Support Vector Classifier
Neural Networks
-NN: Multilayer Perceptron classifier
Input:
Feature Extraction
Classification
Output
Input:
Feature Extraction and Classification
Output
- Small dataset
- Many classifiers available
- Short training time
- Big dataset
- Fewer classifiers available
- Long training time

Figure 4: Predictive models overview

Class imbalance in the outcome variable prompted another experiment to further compare predictive models. The models were trained on two different datasets, the new dataset (dataset with imbalance target classes or original dataset), and a copy of the new dataset with balanced target classes (transformed dataset) to observe model performance across all metrics. For transformed data, Synthetic Minority Over-sampling Technique (SMOTE) was used to oversample the minority class until it matched the majority class (Figure 5). Having balanced classes of the target variable is important for both ML and NN models. Balanced classes prevent models from overfitting the majority class. However, SMOTE benefits models at the cost of introducing noise to the data. Another method used by other researchers to balance classes is to under-sample the majority class. This method also risks losing important data even if it is believed to benefit the models' performance. Comparing models' performance on both the original and transformed dataset gave the study an in depth understanding of how models' performance is affected by the distribution of the outcome variable.

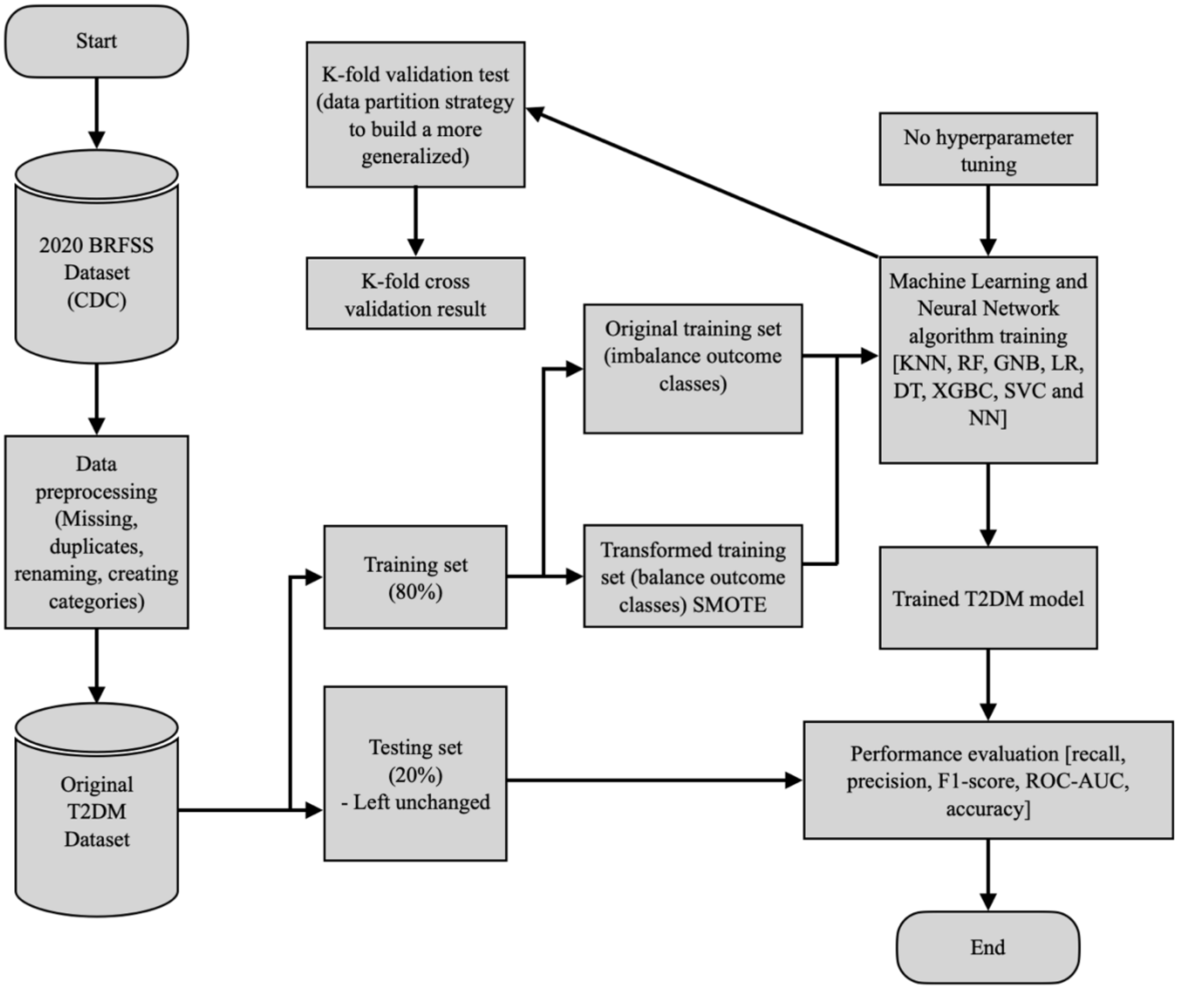

Start
2020 BRFSS Dataset (CDC)
Data preprocessing (Missing, duplicates, renaming, creating categories)
Original T2DM Dataset
K-fold validation test (data partition strategy to build a more generalized)
K-fold cross validation result
Training set (80%)
Testing set (20%) - Left unchanged
Original training set (imbalance outcome classes)
Transformed training set (balance outcome classes) SMOTE
No hyperparameter tuning
Machine Learning and Neural Network algorithm training [KNN, RF, GNB, LR, DT, XGBC, SVC and NN]
Trained T2DM model
Performance evaluation [recall, precision, F1-score, ROC-AUC, accuracy]
End

Figure 5: The architecture of the prediction model

There were three main model comparisons: comparison of models’ performance on original train dataset, comparison of models’ performance on transformed train dataset, and comparison of models’ performance in general. There is more to model performance than only accuracy can tell. Accuracy can tell how good of a predictor a model is, but it does not tell how good a model predicts the class of interest like predicting the risk for diabetes. The results from the study showed models with the highest accuracy to have the lowest sensitivity when models were trained on the original train data.

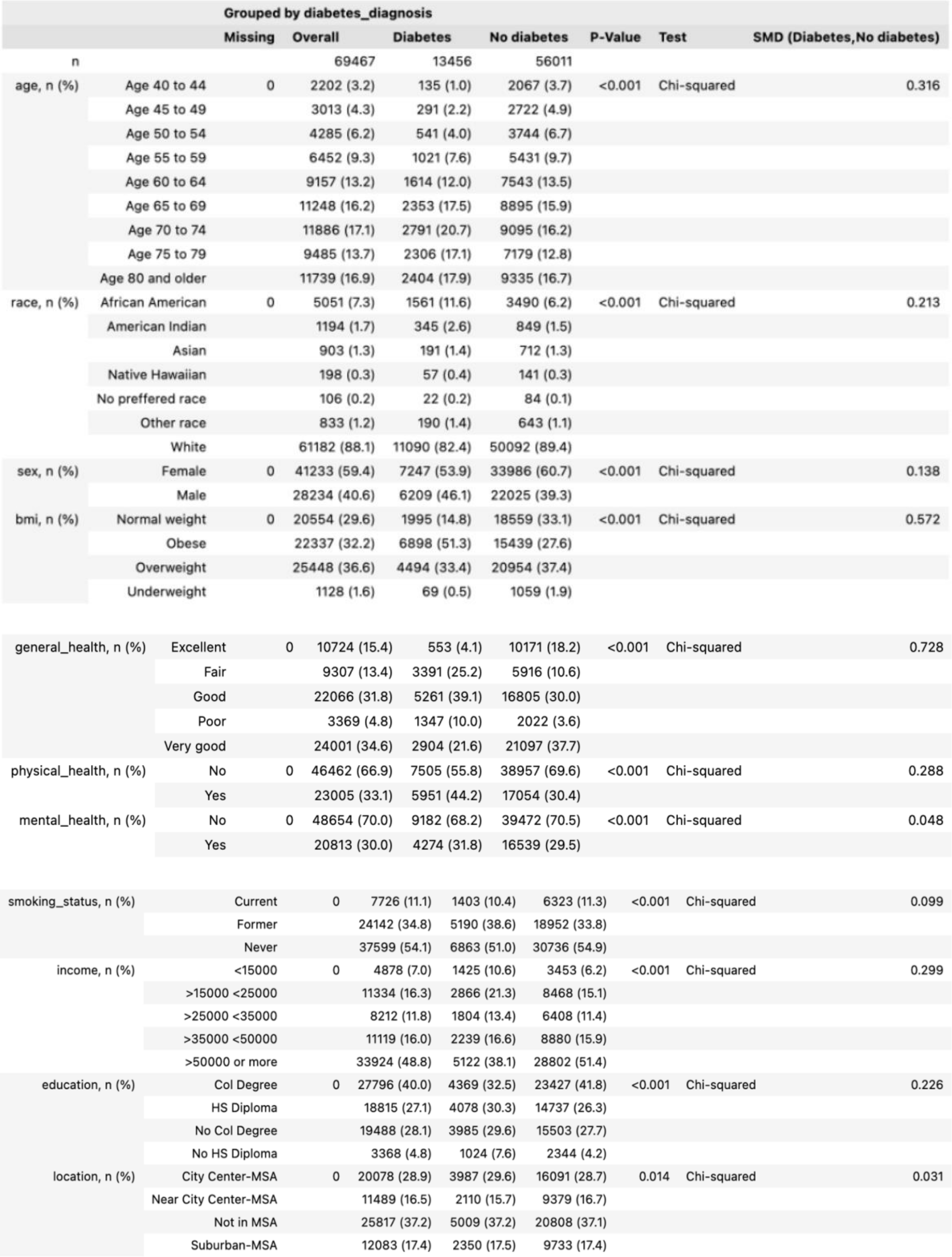
CHAPTER 4

RESULTS

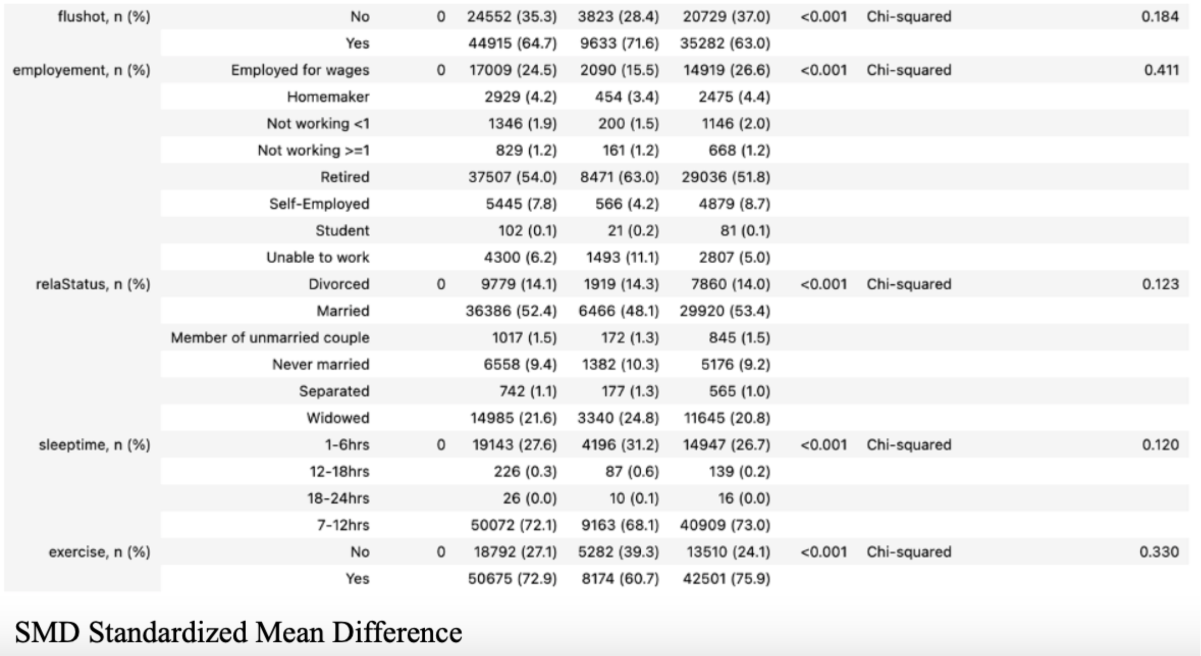
Baseline Characteristic

Each row or record in the dataset represented a person (subject). In total, there were 69467 subjects evaluated for DM. Only 13556 of subjects reported have DM. The subjects were 40 years and older. There were more females (59.4%) than males (40.6%) and about 20% of all subjects reported having DM. Chi-square test was used to evaluate the association between a predictor and the outcome variable (Table 1). The p-value for all predictors was less than 0.001 except for location ($p = 0.014$) indicating that all predictors except location influenced the outcome. Standard mean difference (SMD) ranged from 0.031 to 0.728 with general health (0.728) having the highest followed by BMI (0.572), employment (0.411), exercise (0.330), age (0.316), income (0.299), physical health (0.288), education (0.226), race (0.213), flu shot (0.184), relationship status (0.123), sleep time (0.120), smoking status (0.099), mental health (0.048), and location (0.031).

Table 1: Variable Overview



		Grouped by diabetes_diagnosis						
		Missing	Overall	Diabetes	No diabetes	P-Value	Test	SMD (Diabetes,No diabetes)
n			69467	13456	56011			
age, n (%)	Age 40 to 44	0	2202 (3.2)	135 (1.0)	2067 (3.7)	<0.001	Chi-squared	0.316
	Age 45 to 49		3013 (4.3)	291 (2.2)	2722 (4.9)			
	Age 50 to 54		4285 (6.2)	541 (4.0)	3744 (6.7)			
	Age 55 to 59		6452 (9.3)	1021 (7.6)	5431 (9.7)			
	Age 60 to 64		9157 (13.2)	1614 (12.0)	7543 (13.5)			
	Age 65 to 69		11248 (16.2)	2353 (17.5)	8895 (15.9)			
	Age 70 to 74		11886 (17.1)	2791 (20.7)	9095 (16.2)			
	Age 75 to 79		9485 (13.7)	2306 (17.1)	7179 (12.8)			
	Age 80 and older		11739 (16.9)	2404 (17.9)	9335 (16.7)			
race, n (%)	African American	0	5051 (7.3)	1561 (11.6)	3490 (6.2)	<0.001	Chi-squared	0.213
	American Indian		1194 (1.7)	345 (2.6)	849 (1.5)			
	Asian		903 (1.3)	191 (1.4)	712 (1.3)			
	Native Hawaiian		198 (0.3)	57 (0.4)	141 (0.3)			
	No preffered race		106 (0.2)	22 (0.2)	84 (0.1)			
	Other race		833 (1.2)	190 (1.4)	643 (1.1)			
	White		61182 (88.1)	11090 (82.4)	50092 (89.4)			
sex, n (%)	Female	0	41233 (59.4)	7247 (53.9)	33986 (60.7)	<0.001	Chi-squared	0.138
	Male		28234 (40.6)	6209 (46.1)	22025 (39.3)			
bmi, n (%)	Normal weight	0	20554 (29.6)	1995 (14.8)	18559 (33.1)	<0.001	Chi-squared	0.572
	Obese		22337 (32.2)	6898 (51.3)	15439 (27.6)			
	Overweight		25448 (36.6)	4494 (33.4)	20954 (37.4)			
	Underweight		1128 (1.6)	69 (0.5)	1059 (1.9)			
general_health, n (%)	Excellent	0	10724 (15.4)	553 (4.1)	10171 (18.2)	<0.001	Chi-squared	0.728
	Fair		9307 (13.4)	3391 (25.2)	5916 (10.6)			
	Good		22066 (31.8)	5261 (39.1)	16805 (30.0)			
	Poor		3369 (4.8)	1347 (10.0)	2022 (3.6)			
	Very good		24001 (34.6)	2904 (21.6)	21097 (37.7)			
physical_health, n (%)	No	0	46462 (66.9)	7505 (55.8)	38957 (69.6)	<0.001	Chi-squared	0.288
	Yes		23005 (33.1)	5951 (44.2)	17054 (30.4)			
mental_health, n (%)	No	0	48654 (70.0)	9182 (68.2)	39472 (70.5)	<0.001	Chi-squared	0.048
	Yes		20813 (30.0)	4274 (31.8)	16539 (29.5)			
smoking_status, n (%)	Current	0	7726 (11.1)	1403 (10.4)	6323 (11.3)	<0.001	Chi-squared	0.099
	Former		24142 (34.8)	5190 (38.6)	18952 (33.8)			
	Never		37599 (54.1)	6863 (51.0)	30736 (54.9)			
income, n (%)	<15000	0	4878 (7.0)	1425 (10.6)	3453 (6.2)	<0.001	Chi-squared	0.299
	>15000 <25000		11334 (16.3)	2866 (21.3)	8468 (15.1)			
	>25000 <35000		8212 (11.8)	1804 (13.4)	6408 (11.4)			
	>35000 <50000		11119 (16.0)	2239 (16.6)	8880 (15.9)			
	>50000 or more		33924 (48.8)	5122 (38.1)	28802 (51.4)			
education, n (%)	Col Degree	0	27796 (40.0)	4369 (32.5)	23427 (41.8)	<0.001	Chi-squared	0.226
	HS Diploma		18815 (27.1)	4078 (30.3)	14737 (26.3)			
	No Col Degree		19488 (28.1)	3985 (29.6)	15503 (27.7)			
	No HS Diploma		3368 (4.8)	1024 (7.6)	2344 (4.2)			
location, n (%)	City Center-MSA	0	20078 (28.9)	3987 (29.6)	16091 (28.7)	0.014	Chi-squared	0.031
	Near City Center-MSA		11489 (16.5)	2110 (15.7)	9379 (16.7)			
	Not in MSA		25817 (37.2)	5009 (37.2)	20808 (37.1)			
	Suburban-MSA		12083 (17.4)	2350 (17.5)	9733 (17.4)			

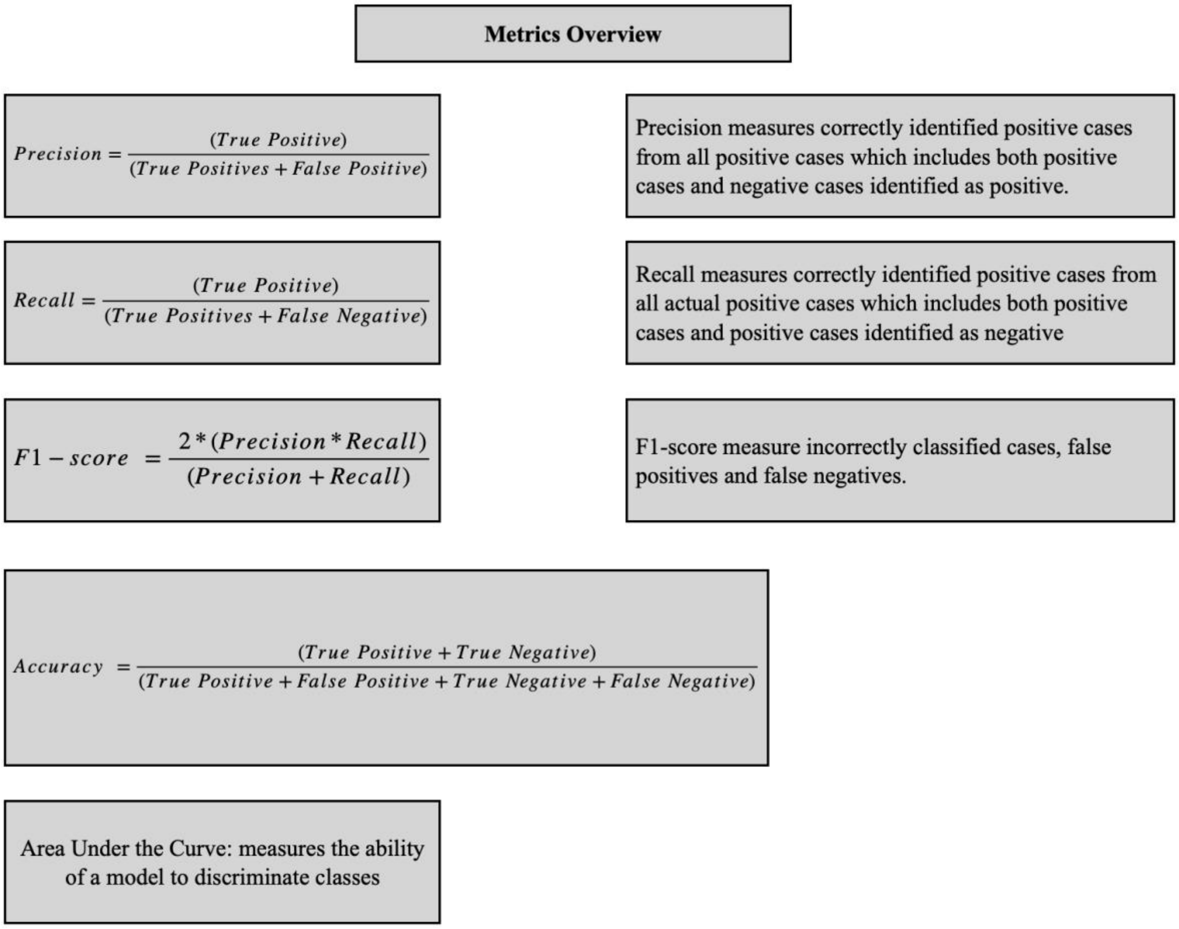


flushot, n (%)	No	0	24552 (35.3)	3823 (28.4)	20729 (37.0)	<0.001	Chi-squared	0.184
	Yes		44915 (64.7)	9633 (71.6)	35282 (63.0)			
employement, n (%)	Employed for wages	0	17009 (24.5)	2090 (15.5)	14919 (26.6)	<0.001	Chi-squared	0.411
	Homemaker		2929 (4.2)	454 (3.4)	2475 (4.4)			
	Not working <1		1346 (1.9)	200 (1.5)	1146 (2.0)			
	Not working >=1		829 (1.2)	161 (1.2)	668 (1.2)			
	Retired		37507 (54.0)	8471 (63.0)	29036 (51.8)			
	Self-Employed		5445 (7.8)	566 (4.2)	4879 (8.7)			
	Student		102 (0.1)	21 (0.2)	81 (0.1)			
	Unable to work		4300 (6.2)	1493 (11.1)	2807 (5.0)			
relaStatus, n (%)	Divorced	0	9779 (14.1)	1919 (14.3)	7860 (14.0)	<0.001	Chi-squared	0.123
	Married		36386 (52.4)	6466 (48.1)	29920 (53.4)			
	Member of unmarried couple		1017 (1.5)	172 (1.3)	845 (1.5)			
	Never married		6558 (9.4)	1382 (10.3)	5176 (9.2)			
	Separated		742 (1.1)	177 (1.3)	565 (1.0)			
	Widowed		14985 (21.6)	3340 (24.8)	11645 (20.8)			
sleeptime, n (%)	1-6hrs	0	19143 (27.6)	4196 (31.2)	14947 (26.7)	<0.001	Chi-squared	0.120
	12-18hrs		226 (0.3)	87 (0.6)	139 (0.2)			
	18-24hrs		26 (0.0)	10 (0.1)	16 (0.0)			
	7-12hrs		50072 (72.1)	9163 (68.1)	40909 (73.0)			
exercise, n (%)	No	0	18792 (27.1)	5282 (39.3)	13510 (24.1)	<0.001	Chi-squared	0.330
	Yes		50675 (72.9)	8174 (60.7)	42501 (75.9)			

SMD Standardized Mean Difference

Model Comparison

Models were trained using either a training set from the original dataset (significant imbalance in the outcome variable classes: 20 to 80%) or a transformed dataset (balanced outcome classes). Both training sets were scaled and normalized. The performance of each model was evaluated on the test set of each dataset that did not undergo outcome class imbalance correction, scaling, or normalization. This approach showed how model accuracy alone could be misleading as a measure of performance because accuracy only shows correct total predictions and does not consider how well a model discriminates between outcome classes. To validate the comparison of different models, multiple metrics were used with much focus on ROC-AUC, recall or sensitivity, precision, and F1-score of each model. Also, K-fold validation was employed.



Metrics Overview

$$Precision = \frac{(True\ Positive)}{(True\ Positives + False\ Positive)}$$

Precision measures correctly identified positive cases from all positive cases which includes both positive cases and negative cases identified as positive.

$$Recall = \frac{(True\ Positive)}{(True\ Positives + False\ Negative)}$$

Recall measures correctly identified positive cases from all actual positive cases which includes both positive cases and positive cases identified as negative

$$F1 - score = \frac{2 * (Precision * Recall)}{(Precision + Recall)}$$

F1-score measure incorrectly classified cases, false positives and false negatives.

$$Accuracy = \frac{(True\ Positive + True\ Negative)}{(True\ Positive + False\ Positive + True\ Negative + False\ Negative)}$$

Area Under the Curve: measures the ability of a model to discriminate classes

Figure 6: Metrics

Area Under the Curve measures the ability of model to discriminate classes.

Model accuracy is a good metric for comparison in the absence of classes imbalance. However, datasets for real life problems like T2DM classification do not guarantee balanced classes. Therefore, evaluating metrics other than accuracy is particularly important to determine model performance.

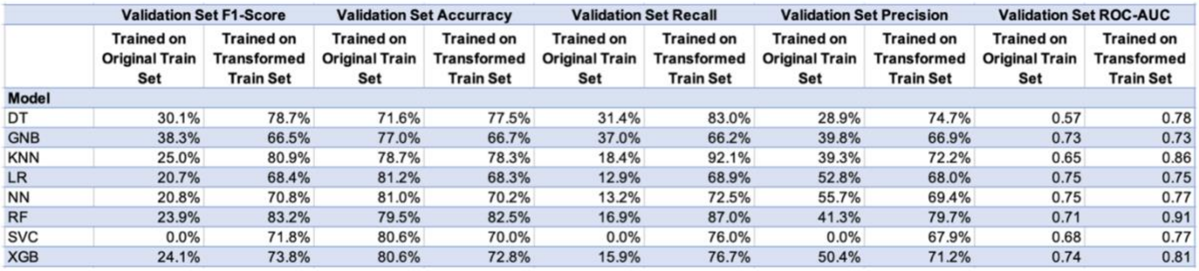
Model Comparison on Original Dataset

For the first type of model comparison, models' accuracy ranged from 71.6% to 81%, ROC-AUC from 0.57 to 0.75, precision from 0% to 55.7%, recall from 0% to 38.3%, and F1-score from 0% to 38%. ROC-AUC for DT (0.57), KNN (0.65), and SVC (0.68) and interpreted as failed or poor predictive models (Table 2 and Figure 7). GNB, LR, RF, NN and XGB had satisfactory to good ROC-AUC and accuracy values. Time it took for the models to train on training data and score on test data was another important evaluation. SVC was the worst model timewise. It took longer to fit and score these models as indicated in Table 3. The NN model took longer to train but it was faster to score. Again, SVC had the third highest accuracy (80.6) but overfitted on the majority class and failed to discriminate between classes. SVC had an F1-score, recall and precision value of 0. Training models on the original dataset caused models to overfit on the majority class. Thus, they had poor sensitivity, precision and F1-Score which is crucial in detecting actual positive cases for T2DM. These results confirmed the hypothesis and failed to produce a good model.

Model Comparison on Transformed Dataset

For the second type of model comparison using transformed data, the models' accuracy ranged from 66.7% to 82.5%, ROC-AUC from 0.73 to 0.91, precision from 66.9% to 79.7%, recall from 66.4% to 92.1%, and F1-score from 66.5% to 83.2% (Table 2 and Figure 7). This comparison clearly showed RF to be the best performing model with consistently good and excellent fit across all metrics (accuracy: 82.5%, ROC-AUC: 0.91, precision: 79.7%, recall: 87.0%, and F1-score: 83.2%). On the other hand, GNB had the poorest fit across all models (Figure 7). SVC was again the worst model timewise. It took longer to fit and score as indicated in Table 3. The NN model had satisfactory scores across all metrics, and it took longer to train but faster to score. With this comparison, all models showed increased values of recall, precision and F1-score suggesting that significant outcome class imbalance had a negative effect on all models. RF, KNN, and DT had F1-score values of 83.2%, 80.9%, and 78.7%, recall values of 87.0%, 92.1% and, 83.0% and precision values of 79.7%, 72.2%, and 74.7%, respectively. KNN had the fastest training, but longest testing time and RF and DT had slightly slower training time and faster testing time (Table 3). KNN, RF and DT models are good candidates for T2DM screening because they discriminated between classes of the outcome variable (high recall and precision) and detected false negative and positive cases (high F1-Score).

Table 2 : Predictive models' performance across metrics



	Validation Set F1-Score		Validation Set Accurracy		Validation Set Recall		Validation Set Precision		Validation Set ROC-AUC	
	Trained on Original Train Set	Trained on Transformed Train Set	Trained on Original Train Set	Trained on Transformed Train Set	Trained on Original Train Set	Trained on Transformed Train Set	Trained on Original Train Set	Trained on Transformed Train Set	Trained on Original Train Set	Trained on Transformed Train Set
Model										
DT	30.1%	78.7%	71.6%	77.5%	31.4%	83.0%	28.9%	74.7%	0.57	0.78
GNB	38.3%	66.5%	77.0%	66.7%	37.0%	66.2%	39.8%	66.9%	0.73	0.73
KNN	25.0%	80.9%	78.7%	78.3%	18.4%	92.1%	39.3%	72.2%	0.65	0.86
LR	20.7%	68.4%	81.2%	68.3%	12.9%	68.9%	52.8%	68.0%	0.75	0.75
NN	20.8%	70.8%	81.0%	70.2%	13.2%	72.5%	55.7%	69.4%	0.75	0.77
RF	23.9%	83.2%	79.5%	82.5%	16.9%	87.0%	41.3%	79.7%	0.71	0.91
SVC	0.0%	71.8%	80.6%	70.0%	0.0%	76.0%	0.0%	67.9%	0.68	0.77
XGB	24.1%	73.8%	80.6%	72.8%	15.9%	76.7%	50.4%	71.2%	0.74	0.81

DT: Decision Tree Classifier, GNB: Gaussian Naïve Bayes Classifier, KNN: K-Nearest Neighbors Classifier, LR: Logistic Regression, NN: Multi-layer perceptron Classifier, RF: Random Forest Classifier, SVC: Support Vector Classifier, XGB : XGBoost Classifier

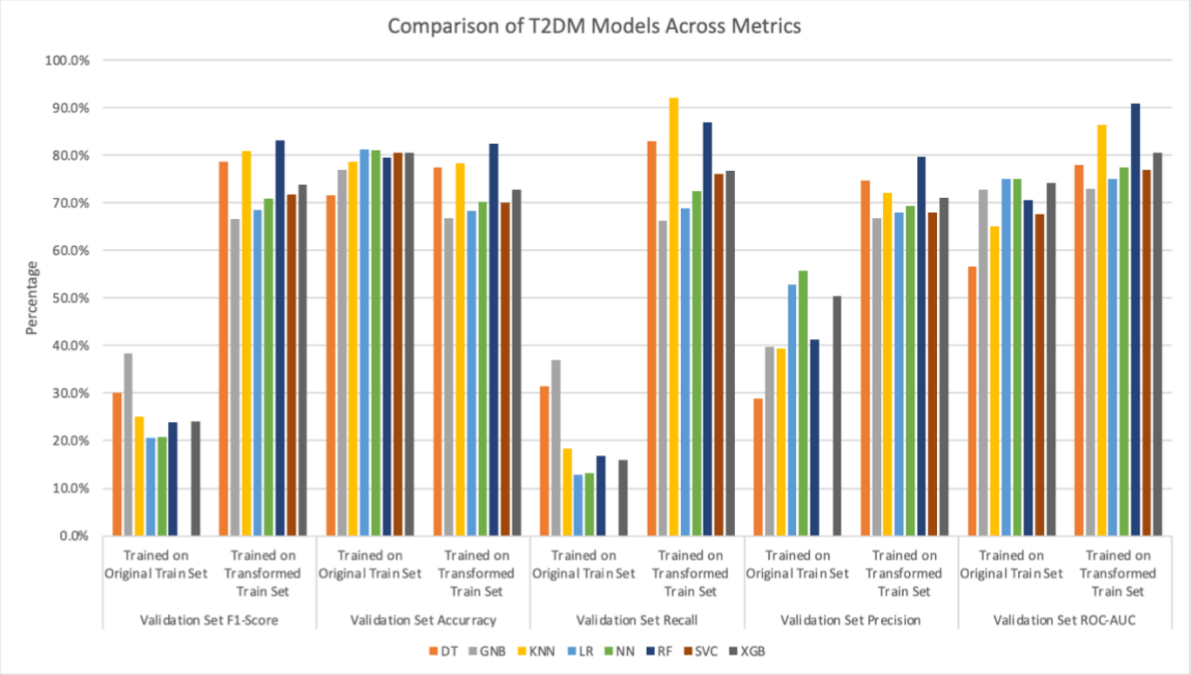
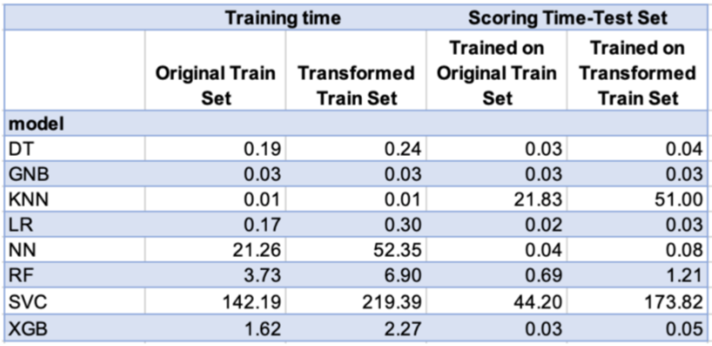


Figure 7: Plot of predictive models' performance across metrics
DT: Decision Tree Classifier, GNB: Gaussian Naïve Bayes Classifier, KNN: K-Nearest Neighbors Classifier, LR: Logistic Regression, NN: Multi-layer perceptron Classifier, RF: Random Forest Classifier, SVC: Support Vector Classifier, XGB : XGBoost Classifier

Table 3: Predictive models' training and testing times



	Training time		Scoring Time-Test Set	
	Original Train Set	Transformed Train Set	Trained on Original Train Set	Trained on Transformed Train Set
model				
DT	0.19	0.24	0.03	0.04
GNB	0.03	0.03	0.03	0.03
KNN	0.01	0.01	21.83	51.00
LR	0.17	0.30	0.02	0.03
NN	21.26	52.35	0.04	0.08
RF	3.73	6.90	0.69	1.21
SVC	142.19	219.39	44.20	173.82
XGB	1.62	2.27	0.03	0.05

DT: Decision Tree Classifier, GNB: Gaussian Naïve Bayes Classifier, KNN: K-Nearest Neighbors Classifier, LR: Logistic Regression, NN: Multi-layer perceptron Classifier, RF: Random Forest Classifier, XGB : XGBoost Classifier, SVC: Support Vector Classifier

Comparison With Previous Studies.

The results of this study were compared with a similar study conducted in 2019 that used the 2014 BRFSS dataset. They also transformed their training data using SMOTE. They found DT to have the highest sensitivity (51.6%) for T2DM.[32] For this study, DT was also among models with the best sensitivity (82.5%). The significant difference in sensitivity values may have been caused by differences in the outcome variable's class distribution, number and choice of predictors, and data cleaning and/or feature engineering techniques.

CHAPTER 5

CONCLUSIONS/DISCUSSION

There are many published studies about T2DM that have produced predictive models. Some studies used longitudinal data and had full control of the type of data to collect. As a result, they produced models with ROC-AUC greater than 0.95. In most cases, these models were created specifically for clinical use. This study evaluated performance of eight models being KNN, RF, GNB, LR, DT, XGB, SVC and NN using readily available predictors to find the best performing predictive model. Training models on the original training dataset confirmed the association between model performance and distribution of classes in the outcome variable. This also shows why model accuracy alone should not be used as a measure of performance. Overall, ML and NN models performed better when trained on the transformed training data. KNN, RF, and DT were good candidates for initial T2DM screening because of their high recall, precision, and F1-score.

Clinical Implication

T2DM affects millions of people worldwide. It is associated with serious microvascular and macrovascular complications such as ischemic heart disease, nephropathy, peripheral vascular disease, cerebrovascular disease, retinopathy, and neuropathy. These complications are associated with increased morbidity, mortality, and health cost. It is important to have a reliable predictive model available for everyone to help detect or screen for T2DM. Literature shows that early detection of risk and early treatment may prevent or delay the onset of the disease for those at risk or may reduce associated complications for those with the disease. Thus, preventing or reducing the effects of T2DM through early detection will reduce the burden of this epidemic on people and entities that are directly or indirectly affected.

Limitations

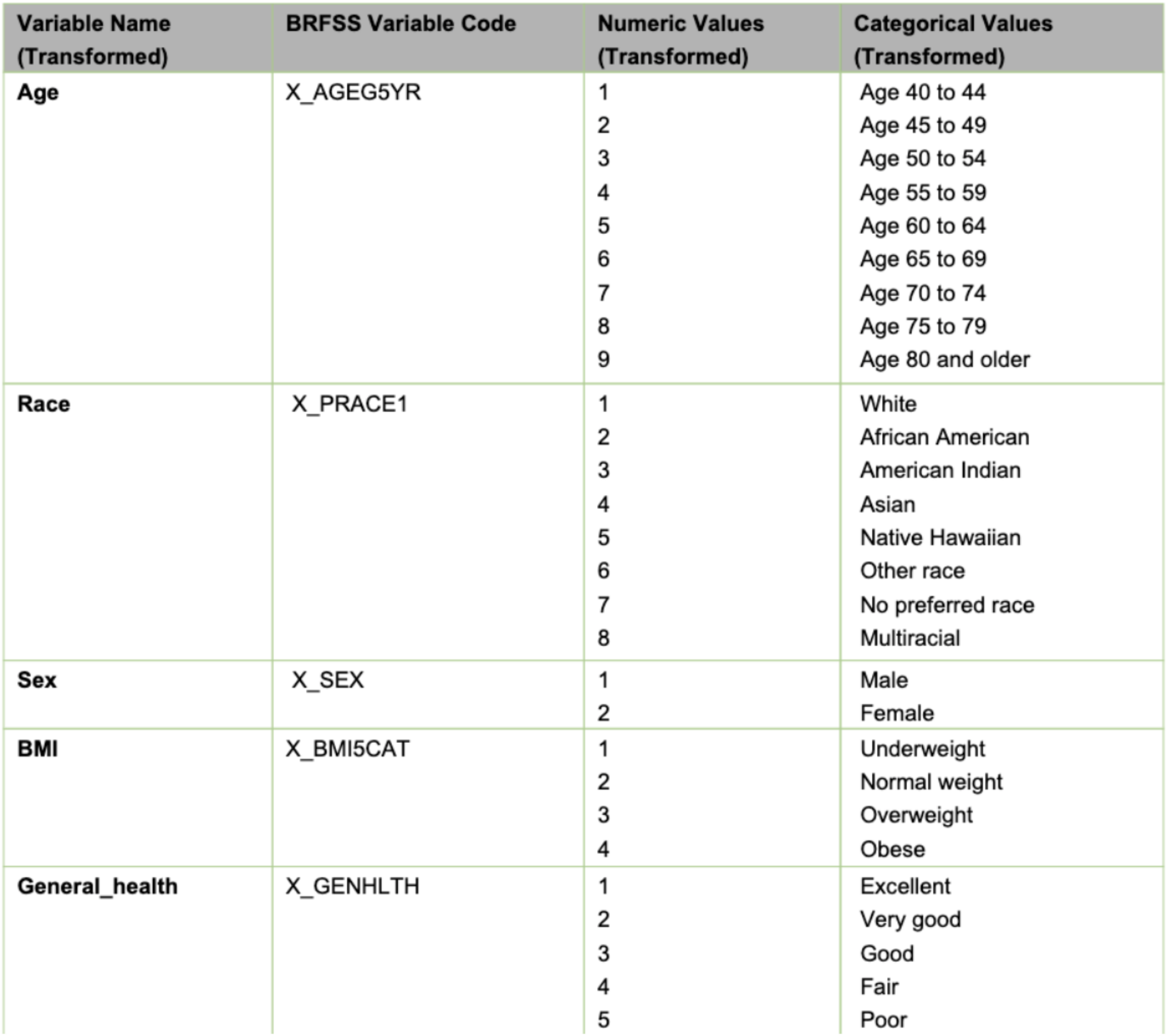
The findings of this study should be interpreted in the context of limitations. The 2020 BRFSS dataset did not clarify whether all DM cases were T1DM or T2DM. The study was carried out under the assumption that subjects who were 40 years and older represented T2DM cases. Anyone younger than 40 years of age was excluded. The best performing models were trained on transformed data. This data corrected the significant training data outcome class imbalance issue. SMOTE was used to correct the imbalance by randomly oversampling the minority class to match the majority class. This method might have added noise to the data. However, testing or scoring of model performance was done on untransformed data. Finally, all models were compared in their default form. Some models may have had an advantage because heterogeneity or homogeneity of parameters was not assessed.

Conclusions

The creation of predictive models used readily available variables for public use unlike longitudinal studies using variables specifically for clinical use. Future work will include feature selection, tuning of model parameters and using additional metrics like confidence interval to compare models.

APPENDIX

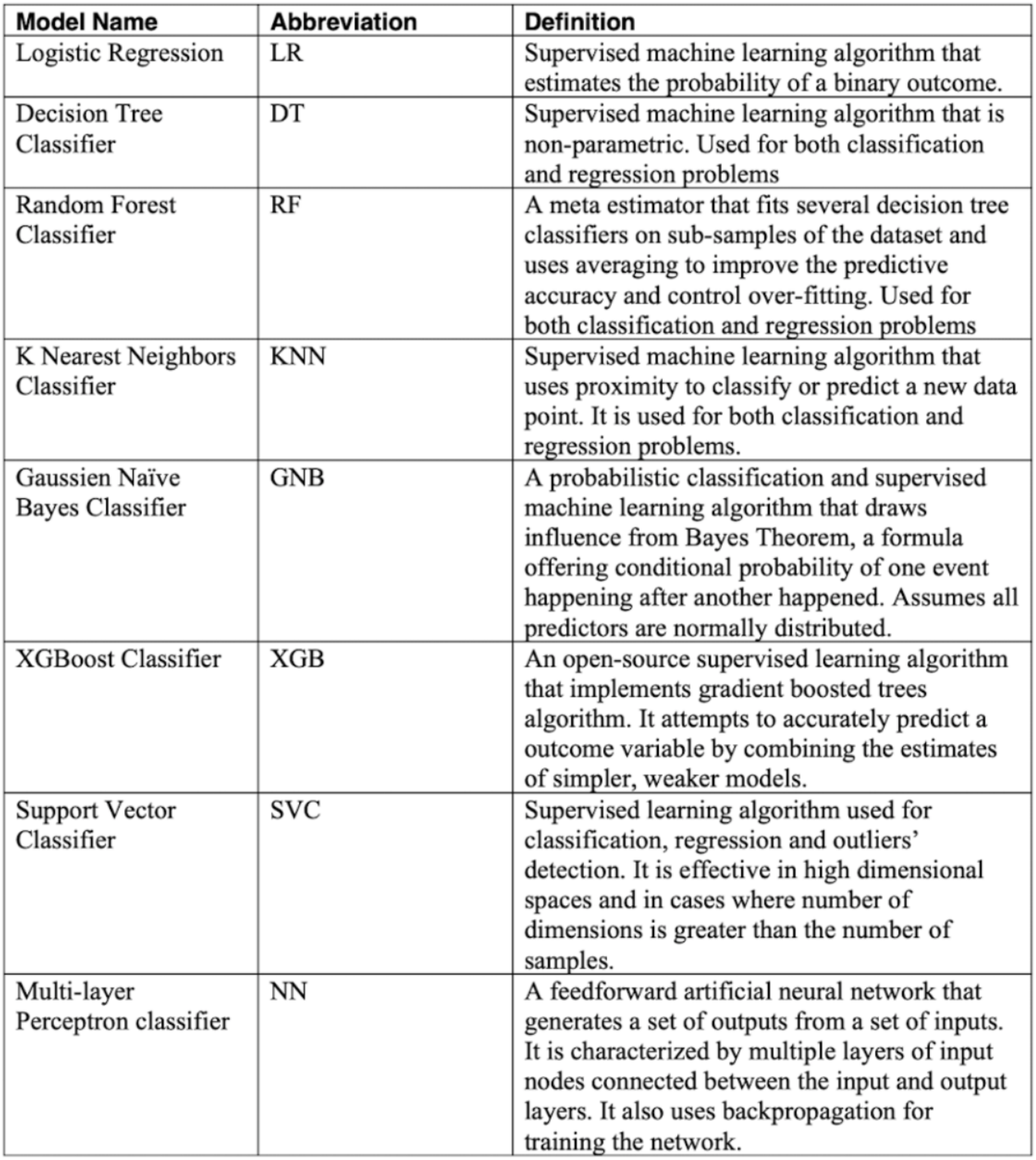
Table 4: Variables overview



Variable Name (Transformed)	**BRFSS Variable Code**	**Numeric Values (Transformed)**	**Categorical Values (Transformed)**
Age	X_AGEG5YR	1	Age 40 to 44
		2	Age 45 to 49
		3	Age 50 to 54
		4	Age 55 to 59
		5	Age 60 to 64
		6	Age 65 to 69
		7	Age 70 to 74
		8	Age 75 to 79
		9	Age 80 and older
Race	X_PRACE1	1	White
		2	African American
		3	American Indian
		4	Asian
		5	Native Hawaiian
		6	Other race
		7	No preferred race
		8	Multiracial
Sex	X_SEX	1	Male
		2	Female
BMI	X_BMI5CAT	1	Underweight
		2	Normal weight
		3	Overweight
		4	Obese
General_health	X_GENHLTH	1	Excellent
		2	Very good
		3	Good
		4	Fair
		5	Poor

Physical_health	X_PHYS14D	0	No
		1	Yes
Mental_health	X_MENT14D	0	No
		1	Yes
Smoking_status	X_SMOKER3	1	Current
		2	Former
		3	Never
Income	X_INCOMG	1	<15000
		2	>15000 <25000
		3	>25000 <35000
		4	>35000 <50000
		5	>50000 or more
Education	X_EDUCAG	1	No HS Diploma
		2	HS Diploma
		3	No Col Degree
		4	Col Degree
location	MSCODE	1	City Center-MSA
		2	Near City Center-MSA
		3	Suburban-MSA
		4	Not in MSA
Flushot	flushot7	0	No
		1	Yes
Employment	EMPLOY1	1	Employed for wages
		2	Self-Employed
		3	Not working >=1
		4	Not working <1
		5	Homemaker
		6	Student
		7	Retired
		8	Unable to work
RelStatus	MARITAL	1	Married
		2	Divorced
		3	Widowed
		4	Separated
		5	Never married
		6	Member of unmarried couple
		7	
Sleeptime	SLEPTIM1	1	1-6hrs
		2	7-12hrs
		3	12-18hrs
		4	18-24hrs
Exercise	EXERANY2	0	No
		1	Yes
Diabetes_diagnosis	DIABETE4	0	No diabetes
		1	Diabetes

Table 5: Model description



Model Name	Abbreviation	Definition
Logistic Regression	LR	Supervised machine learning algorithm that estimates the probability of a binary outcome.
Decision Tree Classifier	DT	Supervised machine learning algorithm that is non-parametric. Used for both classification and regression problems
Random Forest Classifier	RF	A meta estimator that fits several decision tree classifiers on sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. Used for both classification and regression problems
K Nearest Neighbors Classifier	KNN	Supervised machine learning algorithm that uses proximity to classify or predict a new data point. It is used for both classification and regression problems.
Gaussien Naïve Bayes Classifier	GNB	A probabilistic classification and supervised machine learning algorithm that draws influence from Bayes Theorem, a formula offering conditional probability of one event happening after another happened. Assumes all predictors are normally distributed.
XGBoost Classifier	XGB	An open-source supervised learning algorithm that implements gradient boosted trees algorithm. It attempts to accurately predict a outcome variable by combining the estimates of simpler, weaker models.
Support Vector Classifier	SVC	Supervised learning algorithm used for classification, regression and outliers' detection. It is effective in high dimensional spaces and in cases where number of dimensions is greater than the number of samples.
Multi-layer Perceptron classifier	NN	A feedforward artificial neural network that generates a set of outputs from a set of inputs. It is characterized by multiple layers of input nodes connected between the input and output layers. It also uses backpropagation for training the network.

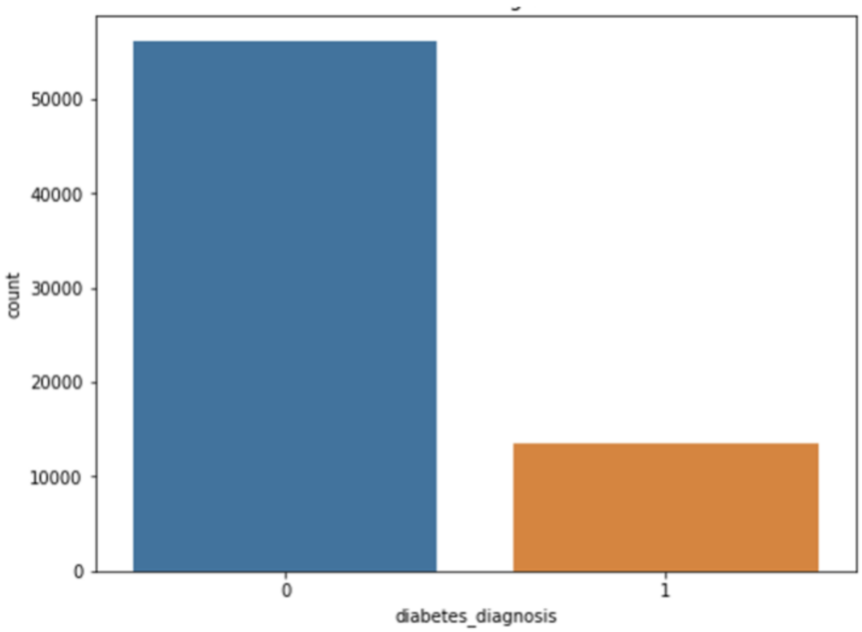
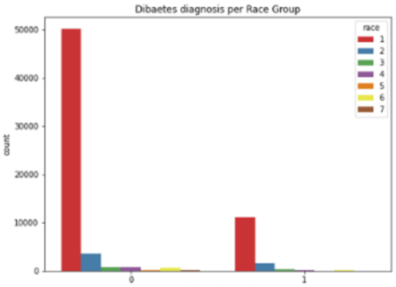
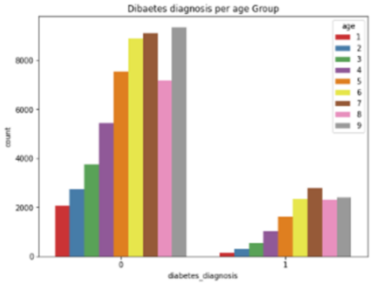
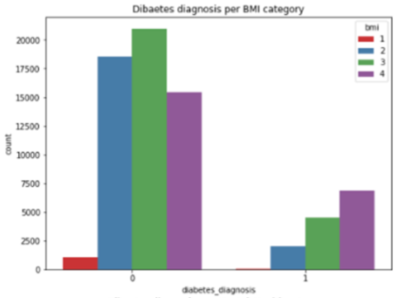
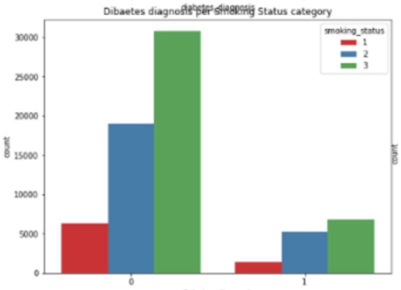


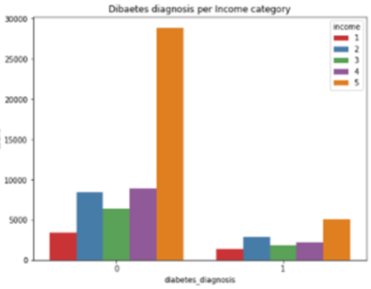
Figure 8: Distributes of the dependent variable
0: No Diabetes, 1: Diabetes

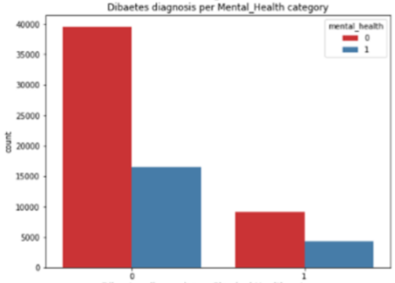

Dibaetes diagnosis per Race Group
race
1
2
3
4
5
6
7
count
50000
40000
30000
20000
10000
0
0
1
diabetes_diagnosis

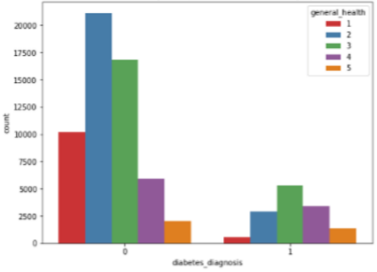

Dibaetes diagnosis per age Group
age
1
2
3
4
5
6
7
8
9
count
8000
6000
4000
2000
0
0
1
diabetes_diagnosis

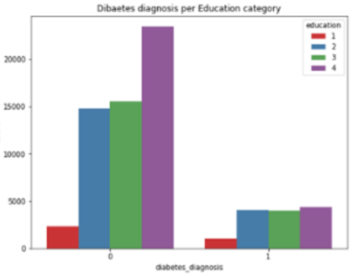

Dibaetes diagnosis per BMI category
bmi
1
2
3
4
count
20000
17500
15000
12500
10000
7500
5000
2500
0
0
1
diabetes_diagnosis

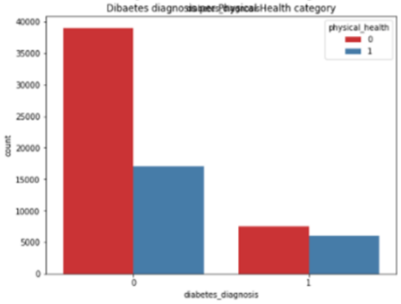

Dibaetes diagnosis per Smoking Status category
smoking_status
1
2
3
count
30000
25000
20000
15000
10000
5000
0
0
1
diabetes_diagnosis


Dibaetes diagnosis per Income category
income
1
2
3
4
5
count
30000
25000
20000
15000
10000
5000
0
0
1
diabetes_diagnosis


Dibaetes diagnosis per Mental_Health category
mental_health
0
1
count
40000
35000
30000
25000
20000
15000
10000
5000
0
0
1


Dibaetes diagnosis per General Health category
general_health
1
2
3
4
5
count
20000
17500
15000
12500
10000
7500
5000
2500
0
0
1
diabetes_diagnosis


Dibaetes diagnosis per Education category
education
1
2
3
4
count
20000
15000
10000
5000
0
0
1
diabetes_diagnosis


Dibaetes diagnosis per Physical_Health category
physical_health
0
1
count
40000
35000
30000
25000
20000
15000
10000
5000
0
0
1
diabetes_diagnosis

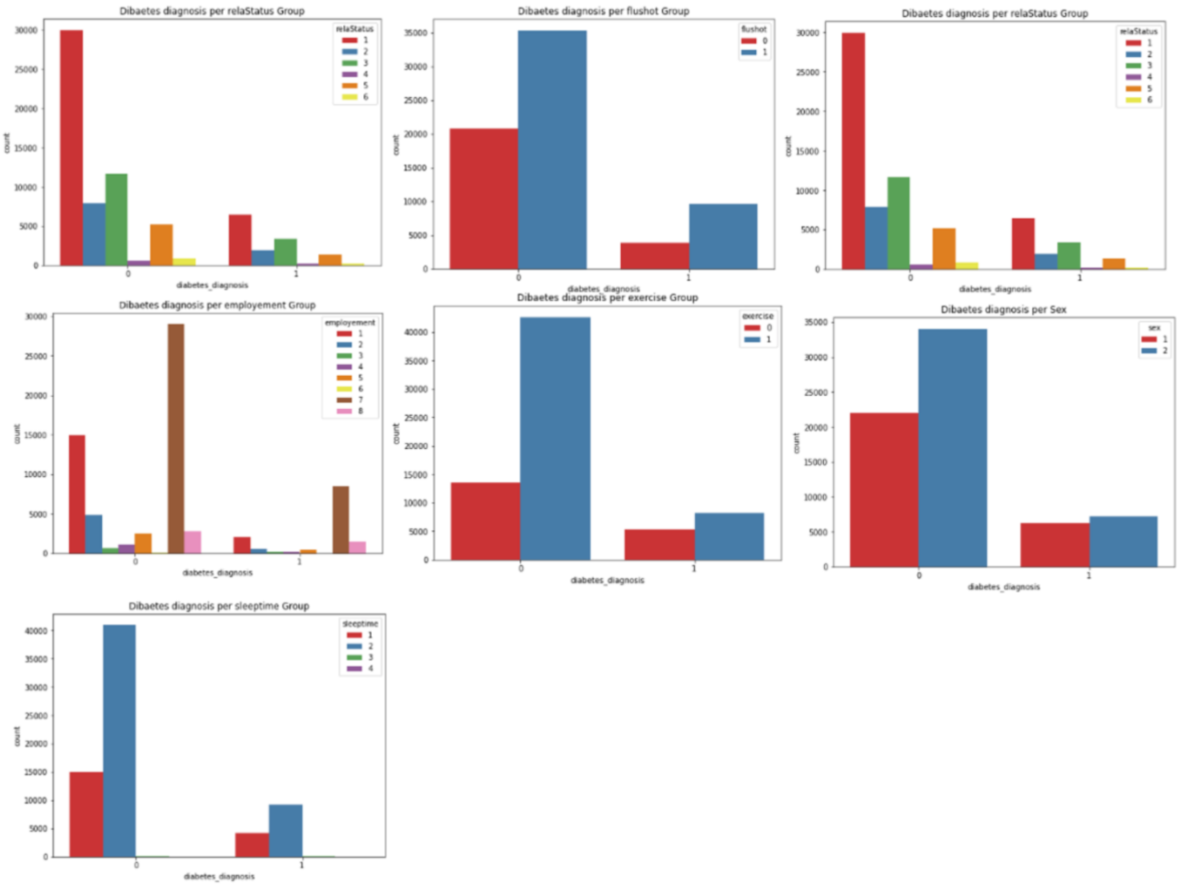


Figure 9: Predictors grouped by the outcome

REFERENCES

1. Khan MAB, Hashim MJ, King JK, Govender RD, Mustafa H, Al Kaabi J. Epidemiology of Type 2 Diabetes – Global Burden of Disease and Forecasted Trends: *J Epidemiol Glob Health*. 2019;10(1):107. doi:10.2991/jegh.k.191028.001
2. Sarría-Santamera A, Orazumbekova B, Maulenkul T, Gaipov A, Atageldiyeva K. The Identification of Diabetes Mellitus Subtypes Applying Cluster Analysis Techniques: A Systematic Review. *Int J Environ Res Public Health*. 2020;17(24):9523. doi:10.3390/ijerph17249523
3. Wang Q, Zhang X, Fang L, Guan Q, Guan L, Li Q. Prevalence, awareness, treatment and control of diabetes mellitus among middle-aged and elderly people in a rural Chinese population: A cross-sectional study. *PloS One*. 2018;13(6):e0198343. doi:10.1371/journal.pone.0198343
4. Keller R, Hartmann S, Teepe GW, et al. Digital Behavior Change Interventions for the Prevention and Management of Type 2 Diabetes: Systematic Market Analysis. *J Med Internet Res*. 2022;24(1):e33348. doi:10.2196/33348
5. Ismail L, Materwala H, Al Kaabi J. Association of risk factors with type 2 diabetes: A systematic review. *Comput Struct Biotechnol J*. 2021;19:1759-1785. doi:10.1016/j.csbj.2021.03.003
6. Kolb H, Martin S. Environmental/lifestyle factors in the pathogenesis and prevention of type 2 diabetes. *BMC Med*. 2017;15(1):131. doi:10.1186/s12916-017-0901-x
7. Deshpande AD, Harris-Hayes M, Schootman M. Epidemiology of diabetes and diabetes-related complications. *Phys Ther*. 2008;88(11):1254-1264. doi:10.2522/ptj.20080020

8. Wu Y, Ding Y, Tanaka Y, Zhang W. Risk factors contributing to type 2 diabetes and recent advances in the treatment and prevention. *Int J Med Sci*. 2014;11(11):1185-1200. doi:10.7150/ijms.10001
9. Bellou V, Belbasis L, Tzoulaki I, Evangelou E. Risk factors for type 2 diabetes mellitus: An exposure-wide umbrella review of meta-analyses. *PloS One*. 2018;13(3):e0194127. doi:10.1371/journal.pone.0194127
10. Willner S, Whittemore R, Keene D. “Life or death”: Experiences of insulin insecurity among adults with type 1 diabetes in the United States. *SSM - Popul Health*. 2020;11:100624. doi:10.1016/j.ssmph.2020.100624
11. Hill J, Nielsen M, Fox MH. Understanding the social factors that contribute to diabetes: a means to informing health care and social policies for the chronically ill. *Perm J*. 2013;17(2):67-72. doi:10.7812/TPP/12-099
12. Cade WT. Diabetes-related microvascular and macrovascular diseases in the physical therapy setting. *Phys Ther*. 2008;88(11):1322-1335. doi:10.2522/ptj.20080008
13. Lakhtakia R. The history of diabetes mellitus. *Sultan Qaboos Univ Med J*. 2013;13(3):368-370. doi:10.12816/0003257
14. Karamanou M, Protogerou A, Tsoucalas G, Androutsos G, Poulakou-Rebelakou E. Milestones in the history of diabetes mellitus: The main contributors. *World J Diabetes*. 2016;7(1):1-7. doi:10.4239/wjd.v7.i1.1
15. Vecchio I, Tornali C, Bragazzi NL, Martini M. The Discovery of Insulin: An Important Milestone in the History of Medicine. *Front Endocrinol*. 2018;9:613. doi:10.3389/fendo.2018.00613

16. Lewis GF, Brubaker PL. The discovery of insulin revisited: lessons for the modern era. *J Clin Invest*. 2021;131(1):142239. doi:10.1172/JCI142239

17. Barik A, Mazumdar S, Chowdhury A, Rai RK. Physiological and behavioral risk factors of type 2 diabetes mellitus in rural India. *BMJ Open Diabetes Res Care*. 2016;4(1):e000255. doi:10.1136/bmjdrc-2016-000255

18. Dendup T, Feng X, Clingan S, Astell-Burt T. Environmental Risk Factors for Developing Type 2 Diabetes Mellitus: A Systematic Review. *Int J Environ Res Public Health*. 2018;15(1):78. doi:10.3390/ijerph15010078

19. Spruijt-Metz D, O'Reilly GA, Cook L, Page KA, Quinn C. Behavioral Contributions to the Pathogenesis of Type 2 Diabetes. *Curr Diab Rep*. 2014;14(4):475. doi:10.1007/s11892-014-0475-3

20. Ardisson Korat AV, Willett WC, Hu FB. Diet, Lifestyle, and Genetic Risk Factors for Type 2 Diabetes: A Review from the Nurses' Health Study, Nurses' Health Study 2, and Health Professionals' Follow-Up Study. *Curr Nutr Rep*. 2014;3(4):345-354. doi:10.1007/s13668-014-0103-5

21. Ley SH, Schulze MB, Hivert MF, Meigs JB, Hu FB. Risk Factors for Type 2 Diabetes. In: Cowie CC, Casagrande SS, Menke A, et al., eds. *Diabetes in America*. 3rd ed. National Institute of Diabetes and Digestive and Kidney Diseases (US); 2018. Accessed July 30, 2022. http://www.ncbi.nlm.nih.gov/books/NBK567966/

22. Schulze MB, Manson JE, Willett WC, Hu FB. Processed meat intake and incidence of Type 2 diabetes in younger and middle-aged women. *Diabetologia*. 2003;46(11):1465-1473. doi:10.1007/s00125-003-1220-7

23. van Dam RM, Willett WC, Rimm EB, Stampfer MJ, Hu FB. Dietary Fat and Meat Intake in Relation to Risk of Type 2 Diabetes in Men. *Diabetes Care*. 2002;25(3):417-424. doi:10.2337/diacare.25.3.417
24. Colberg SR, Sigal RJ, Fernhall B, et al. Exercise and Type 2 Diabetes. *Diabetes Care*. 2010;33(12):e147-e167. doi:10.2337/dc10-9990
25. Hamilton MT, Hamilton DG, Zderic TW. Sedentary Behavior as a Mediator of Type 2 Diabetes. In: Goedecke JH, Ojuka EO, eds. *Medicine and Sport Science*. Vol 60. S. Karger AG; 2014:11-26. doi:10.1159/000357332
26. Talaei M, Rabiei K, Talaei Z, et al. Physical activity, sex, and socioeconomic status: A population based study. *ARYA Atheroscler*. 2013;9(1):51-60.
27. Shan Z, Ma H, Xie M, et al. Sleep Duration and Risk of Type 2 Diabetes: A Meta-analysis of Prospective Studies. *Diabetes Care*. 2015;38(3):529-537. doi:10.2337/dc14-2073
28. Lu H, Yang Q, Tian F, et al. A Meta-Analysis of a Cohort Study on the Association between Sleep Duration and Type 2 Diabetes Mellitus. *J Diabetes Res*. 2021;2021:8861038. doi:10.1155/2021/8861038
29. Metse AP, Clinton-McHarg T, Skinner E, Yogaraj Y, Colyvas K, Bowman J. Associations between Suboptimal Sleep and Smoking, Poor Nutrition, Harmful Alcohol Consumption and Inadequate Physical Activity ('SNAP Risks'): A Comparison of People with and without a Mental Health Condition in an Australian Community Survey. *Int J Environ Res Public Health*. 2021;18(11):5946. doi:10.3390/ijerph18115946

30. Deberneh HM, Kim I. Prediction of Type 2 Diabetes Based on Machine Learning Algorithm. *Int J Environ Res Public Health*. 2021;18(6):3317. doi:10.3390/ijerph18063317
31. Fregoso-Aparicio L, Noguez J, Montesinos L, García-García JA. Machine learning and deep learning predictive models for type 2 diabetes: a systematic review. *Diabetol Metab Syndr*. 2021;13(1):148. doi:10.1186/s13098-021-00767-9
32. Xie Z, Nikolayeva O, Luo J, Li D. Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques. *Prev Chronic Dis*. 2019;16:E130. doi:10.5888/pcd16.190109
33. CDC, BRFSS. Behavioral Risk Factor Surveillance System Overview. :11.
34. Yeo IK. A new family of power transformations to improve normality or symmetry. *Biometrika*. 2000;87(4):954-959. doi:10.1093/biomet/87.4.954

VITA

Justin Ngoyi Mpanga was born on May 28th, 1996, in Lubumbashi, Democratic Republic of Congo. In 2013, he relocated to Kansas City, MO. During his two years of high school, he was enrolled in the Early College Academy, a partnership between Kansas City public Schools and Metropolitan Community College and graduated from East High Schoo1 in 2015. That same year, Mpanga started his college education at the University of Missouri-Kansas City (UMKC). Four years later, Mpanga earned his Bachelor of Science in Biology with minors in Chemistry and Psychology.

He worked as a lab technician and certified phlebotomist at Saint Luke's Hospital in Kansas City, MO during his undergraduate years. Upon graduation, he worked as a life enrichment coordinator in Blue Springs, MO, and then as a research assistant at the University of Missouri-Kansas City. His work experience led him to find a strong passion for evidence based medical decisions which are achieved with medical data analysis.

In 2020, Mpanga started his graduate education toward a Master of Science in Biomedical and Health Informatics – Computational emphasis at the University of Missouri-Kansas City School of Medicine. Upon completion of his degree requirements, Mpanga intends to pursue a medical degree.