PREDICTIVE MODELING FOR DETECTION OF DIABETES MELLITUS: A STUDY ON RISK FACTORS AND MACHINE LEARNING APPROACHES

MAH SEAU SHER 22115483
SUPERVISOR: DR NAZEAN BINTI JOMHARI

List of Abbreviations

Abbreviations	Full Name
BRFSS	Behavioral Risk Factor Surveillance System
KNN	k-Nearest Neighbors
SVM	Support Vector Machine
LR	Logistic Regression
RF	Random Forest
GB	Gradient Boosting
AdaBoost	Adaptive Boosting
EHR	Electronic Health Records
SHAP	SHapley Additive exPlanations
DT	Decision Tree
NN	Neural Network
XGBoost	Extreme Gradient Boosting
LightGBM	Light Gradient Boosting
RFECV	Recursive Feature Elimination with Cross Validation
ML	Machine Learning
EDA	Exploratory Data Analysis
BMI	Body Mass Index
CatBoost	Categorical Boosting
RC	Ridge Classifier
LassoR	Lasso Regression
Extra Tree	Extremely Randomized Trees
Bagging	Bagging Ensemble Method
GLM	General Linear Model

Overview

- Introduction
- Problem Statement
- Research Question & Research Objective
- Scope
- Literature Review
- Application of BRFSS Dataset in Research
- Methodology – SEMMA
- Tool
- Results and Discussions
- Conclusion
- Limitations and Future Works
- Reference

Introduction

Diabetes is a global public health concern, contributing significantly to morbidity and mortality across all demographics (Hossain, Al-Mamun, & Islam, 2024).

In the U.S., 29.3 million people are diagnosed with diabetes, with rising prevalence worldwide (Martin et al., 2024).

Classified into Type 1 (autoimmune β-cell destruction) and Type 2 (progressive β-cell dysfunction with insulin resistance) (American Diabetes Association, 2024).

Early detection of diabetes is crucial to prevent severe complications such as heart disease, stroke, eye complications, nerve damage, and kidney disease (Sah, Kulkarni, Sehgal, & Victor).

Problem Statement

Imbalanced datasets in diabetes prediction led to biased ML models and inaccurate results, highlighting the need for strategies to address class imbalance and improve model reliability (Talebi Moghaddam et al., 2024).

Many existing studies rely on limited ML models, which may not fully leverage the diverse range of features available for predicting diabetes (Lakshmi, Reddy, & Naidu, 2023; Prasetyo, Izdihar, & Nabiilah, 2024).

Crucial to develop a robust predictive model that is able to circumvent this challenge of imbalance and to determine the model's suitability; a model capable of analyzing various factors ranging from medical data to individual lifestyle variables to accurately detect and anticipate individuals with risk of diabetes mellitus.

Research Question

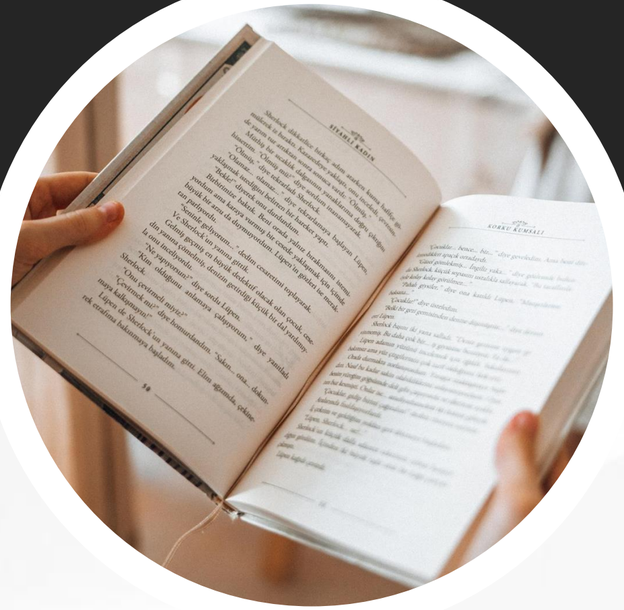
1. What are the factors that contribute to the likelihood of diabetes mellitus?
2. What are the methods that can circumvent the class imbalance in real-world datasets?
3. How effective are the ML models in diabetes mellitus prediction?

Research Objective

1. To identify the factors contributing to the likelihood of diabetes mellitus.
2. To develop predictive models capable of circumventing the class imbalance in the real-world dataset.
3. To evaluate the performance of the ML models.

Scope

The research project scope will focus on 50 states in the United States and the adult population more than 18 years old. This project may not address all aspects of diabetes mellitus prediction, such as genetic factors or specific medical tests.



Literature Review

Risk factors contributing to diabetes

Author	Risk factors
(Collins, Mallett, Omar, & Yu, 2011)	Age, Family history of diabetes, BMI, Hypertension, Waist circumference, Sex, Ethnicity, Fasting glucose level, Smoking status, Physical activity
(Fleitman, 2024)	Age, BMI, General Health Perception
(Lee et al., 2024)	Age, Ethnicity, Socioeconomic Status, Geographic Clustering, Insurance Type, Physical Inactivity, Obesity
(Olayeye, Bodunwa, & Adewole, 2024)	Age, blood pressure, cholesterol, fruits consumption, veggies consumption, difficulty in walking, general health, stroke, heart disease, health care coverage, physical activity, body mass Index
(Xie, Nikolayeva, Luo, & Li, 2019)	Sleeping time and frequency of checkup

Literature Review

Related works

Author	Dataset	Technique	Limitation
(Arslan & Özdemir, 2023)	BRFSS 2015	· GBM · XGBoost · LightGB · CatBoost · KNN · RF · RC · LR · GNB · DT	· No hyperparameter tuning · Limited features are selected · Oversampling has been performed on the whole dataset
(Burch et al., 2024)	BRFSS 2017	· LR	· Model underperform in unbalanced data set · No over/undersampling has been performed · Only 1 model has been performed · No feature selection
(Chang, Ganatra, Hall, Golightly, & Xu, 2022)	BRFSS 2015	· DT · RF · KNN · LR · NB	· No hyperparameter tuning · Cross-validation techniques should be used to improve the performance of the model · Oversampling has been performed on the whole dataset
(Chowdhury et al., 2024)	BRFSS 2021	· LR · RF · GB · AdaBoost · VC	· No feature selection · Research focus mainly on data augmentation techniques

Literature Review

Continuation … Related works

Author	Dataset	Technique	Limitation
(Dong, 2023)	BRFSS 2015	· RF	· Only 1 model has been performed · No over/undersampling has been performed · Other feature selection methods could be explored to improve the reliability and generalization ability of the model
(Hama Saeed, 2023)	PIMA dataset, BRFSS 2015	· Extra Tree · DT · AdaBoost · GB	· No hyperparameter tuning · Limited models are performed · DL models can be utilized
(Horestani, 2024)	BRFSS 2015	· LR	· Only 1 model has been performed
(Lakshmi et al., 2023)	BRFSS 2015	· KNN · SVM · LR	· Model underperform in unbalanced data set · No over/undersampling has been performed · Limited models have been performed · No feature selection
(Nguyen & Zhang, 2025)	BRFSS 2015	· DT · KNN · LR	· Model underperform in unbalanced data sets · No over/under sampling been performed · Evaluation purely based on accuracy

Literature Review

Continuation … Related works

Author	Dataset	Technique	Limitation
(Omoora et al., 2023)	BRFSS 2015	· SVM · NB · DT · RF · XGBoost	· Model underperform in unbalanced data sets · No over/under sampling been performed
(Pechprasarn et al., 2025)	BRFSS 2015	· 34 ML models including SVM, KNN, NN and etc	· Model underperform in unbalanced data set · No over/undersampling been performed
(Prasetyo et al., 2024)	BRFSS 2015	· GNB · DT · ANN	· Model underperform in unbalanced data set · No over/undersampling been performed · Limited models has been performed · No feature selection
(Shinde & Singh, 2023)	BRFSS 2015	· LR · NB · DT · RF · XGBoost · NN	· Oversampling has been performed on the whole dataset · Limited features have been used

Literature Review

Continuation … Related works

Author	Dataset	Technique	Limitation
(Su, 2023)	BRFSS 2015	· LR · GNB · DT · RF	· Model underperform in unbalanced data set · No over/undersampling been performed · No hyperparameter tuning
(Ullah et al., 2022)	BRFSS 2015	· KNN · RF · XGBoost · Bagging · AdaBoost	· No feature selection · Oversampling has been performed but no clarity whether it was applied only to the training set or to the entire dataset
(Wu, 2024)	BRFSS 2021	· GLM · LassoR · NN · DT · RF	· Model underperform in unbalanced data set · No over/undersampling been performed · No hyperparameter tuning

Applications of BRFSS Dataset in Research

Diabetes Prediction

The BRFSS 2015 dataset has been used in various studies applying machine learning techniques to predict diabetes and identify risk factors (Olayeye, Bodunwa, & Adewole, 2024).

Asthma Prediction

The 2019 BRFSS dataset was applied to predict asthma and assess risk factors using machine learning techniques (Budhathoki, Bhandari, Bashyal, & Lee, 2023).

Heart Disease Prediction

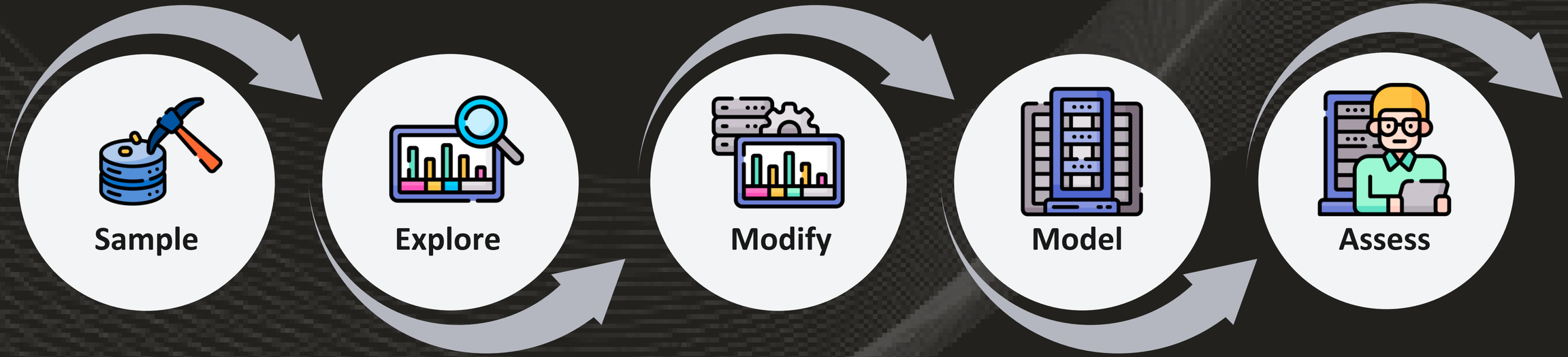
The 2020 BRFSS dataset has been utilized to compare the performance of machine learning models, including XGBoost, AdaBoost, and RF, for heart disease prediction (Mamun, Uddin, Tiwari, Islam, & Ferdous, 2022).

Stroke Risk Prediction

The 2022 BRFSS dataset has been used to evaluate multiple ML models for stroke risk prediction and assess the impact of various risk factors (Akter, Akter, & Pias, 2023)

Methodology - SEMMA

A data mining methodology developed by SAS that guides the process of transforming raw data into valuable insights through a systematic sequence: Sample, Explore, Modify, Model, and Assess (Hotz, 2024)



Retrieve BRFSS Dataset

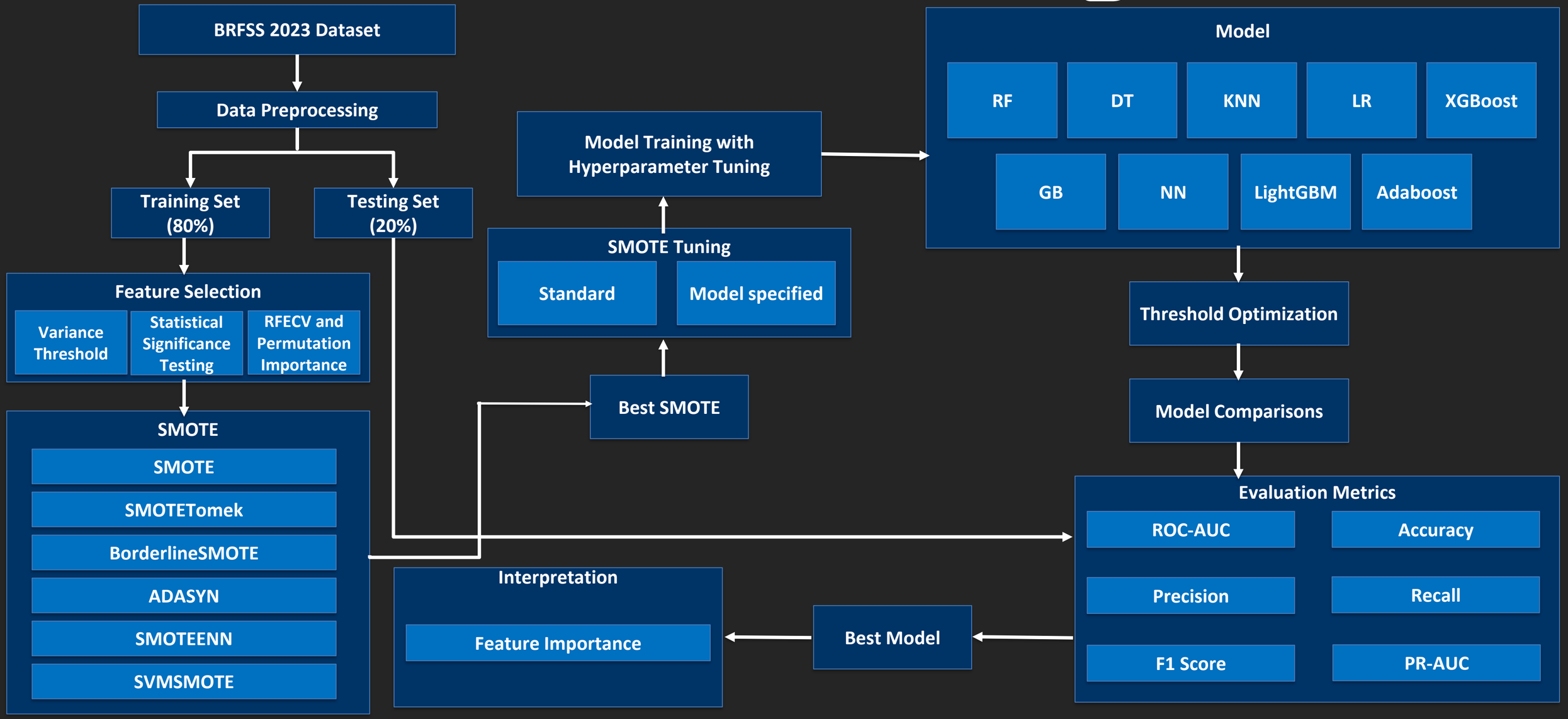
EDA: Use visualizations and statistical analysis to identify patterns, relationships, and anomalies such as missing values or outliers.

Clean and preprocess the data by handling missing values, transforming variables and creating new features if necessary.

Apply machine learning algorithms (e.g., RF, LR) to predict outcomes.

Evaluate the performance of the models using metrics such as accuracy, precision, recall, or AUC to ensure they meet the analysis objectives

Research Flow Diagram



Sample

Source : BRFSS, a large-scale, ongoing health survey carried out by CDC (CDC,2023)

Data Collection: Telephone interviews (both landline and cell phone)

Features: Demographics; Health Risk Behaviors; Chronic Health Conditions; Preventive Health Practices

Size: 433,323 individual records (2023 year data)

Explore

1

Gain an initial understanding of the dataset's features (distributions, central tendencies, potential outliers).

2

Identify potential relationships between variables visually.

3

Inform data cleaning and preparation choices.

Data Preprocessing

Variable Definitions and Transformations

Original Variable Name	New Variable Name	Definition	Mapping	Type
SEXVAR	Gender	Sex of respondent	1 = Male, 0 = Female	Categorical
GENHLTH	GeneralHealth	General health	1 = Poor, 2 = Fair, 3 = Good, 4 = Very good, 5 = Excellent	Ordinal
PHYSHLTH	PhysicalHealthDays	Number of days physical health not good	0–30	Continuous
MENTHLTH	MentalHealthDays	Number of days mental health not good	0–30	Continuous
CHECKUP1	LastCheckupTime	Time since last routine checkup	1 = Past 5 years or less, 0 = Never	Categorical
EXERANY2	PhysicalActivities	Exercise in the past 30 days	1 = Yes, 0 = No	Categorical
CVDINFR4	HadHeartAttack	Ever diagnosed with heart attack	1 = Yes, 0 = No	Categorical
CVDCRHD4	HadHeartDisease	Diagnosed with angina or coronary heart disease	1 = Yes, 0 = No	Categorical
CVDSTRK3	HadStroke	Diagnosed with stroke	1 = Yes, 0 = No	Categorical
ASTHMA3	HadAsthma	Diagnosed with asthma	1 = Yes, 0 = No	Categorical
CHCSCNC1	HadSkinCancer	Diagnosed with skin cancer (non-melanoma)	1 = Yes, 0 = No	Categorical
CHCCOPD3	HadCOPD	Diagnosed with C.O.P.D., emphysema or chronic bronchitis	1 = Yes, 0 = No	Categorical
ADDEPEV3	HadDepressiveDisorder	Diagnosed with depressive disorder	1 = Yes, 0 = No	Categorical

Data Preprocessing

Continuation ... Variable Definitions and Transformations

Original Variable Name	New Variable Name	Definition	Mapping	Type
CHCKDNY2	HadKidneyDisease	Diagnosed with kidney disease	1 = Yes, 0 = No	Categorical
HAVARTH4	HadArthritis	Diagnosed with arthritis	1 = Yes, 0 = No	Categorical
DIABETE4	HadDiabetes	Diagnosed with diabetes	1 = Yes (including borderline, pregnancy-related), 0 = No	Categorical
DIFFWALK	DifficultyWalking	Difficulty walking or climbing stairs	1 = Yes, 0 = No	Categorical
_SMOKER33	SmokerStatus	Smoking status	1 = Current smoker, 0 = Former/Never smoked	Categorical
ECIGNOW2	ECigaretteUsage	Use of e-cigarettes or vaping products	1 = Yes (daily/some days), 0 = No/never	Categorical
_RACEGR3	RaceEthnicityCategory	Race/ethnicity category	1 = Hispanic, 0 = Non-Hispanic (White, Black, Other)	Categorical
_AGEG5YR	AgeCategory	Age category	Age 18 to 24 = 21, Age 25 to 29 = 27, Age 30 to 34 = 32, Age 35 to 39 = 37, Age 40 to 44 = 42, Age 45 to 49 = 47, Age 50 to 54 = 52, Age 55 to 59 = 57, Age 60 to 64 = 62, Age 65 to 69 = 67, Age 70 to 74 = 72, Age 75 to 79 = 77, Age 80 or older = 85	Continuous

Data Preprocessing

Continuation … Variable Definitions and Transformations

Original Variable Name	New Variable Name	Definition	Mapping	Type
_BMI5	BMI	Body Mass Index	BMI < 25 = Normal, < 30 = Overweight, ≥ 30 = Obese	Ordinal
DRNKANY6	AlcoholDrinkers	Alcohol use in past 30 days	1 = Yes, 0 = No	Categorical
INCOME3	HouseholdIncome	Household income level	Low Income= < $25,000, Middle Income = $25k–<$100k, High Income = ≥ $100,000	Categorical
CHCOCNC1	HadOtherCancer	Diagnosed with other cancer types (melanoma/others)	1 = Yes, 0 = No	Categorical
EDUCA	EducationLevel	Level of education	High Education = College graduate, Low/Medium = Others	Categorical
_HLTHPL1	HadInsurance	Has any health insurance	1 = Yes, 0 = No	Categorical
_RFCHOL3	HighCholesterol	Diagnosed with high cholesterol	1 = Yes, 0 = No	Categorical
_RFHYPE6	HighBloodPressure	Diagnosed with high blood pressure	1 = Yes, 0 = No	Categorical

Data Cleaning

Standardize missing values

Apply 20% Threshold Rule

Duplicate Removal

Data Size = from 433,323 to 252,383

Variables	NaN Counts	NaN Percentages	Pass/Fail
Gender	0	0	Pass
GeneralHealth	1262	0.291238	Pass
PhysicalHealthDays	10785	2.488906	Pass
MentalHealthDays	8108	1.871122	Pass
LastCheckupTime	5781	1.334109	Pass
PhysicalActivities	1251	0.288699	Pass
HadHeartAttack	2568	0.59263	Pass
HadHeartDisease	4231	0.976408	Pass
HadStroke	1474	0.340162	Pass
HadAsthma	1701	0.392548	Pass
HadSkinCancer	2930	0.67617	Pass
HadCOPD	2066	0.476781	Pass
HadDepressiveDisorder	2587	0.597014	Pass
HadKidneyDisease	1892	0.436626	Pass
HadArthritis	2560	0.590783	Pass
HadDiabetes	984	0.227082	Pass
DifficultyWalking	17850	4.119329	Pass
SmokerStatus	23062	5.322127	Pass
ECigaretteUsage	23251	5.365743	Pass
RaceEthnicityCategory	9570	2.208514	Pass
AgeCategory	7779	1.795197	Pass
BMI	40535	9.354454	Pass
AlcoholDrinkers	29940	6.909396	Pass
HouseholdIncome	86623	19.9904	Pass
HadOtherCancer	2337	0.539321	Pass
EducationLevel	2325	0.536551	Pass
HadInsurance	18674	4.309487	Pass
Highcholesterol	55084	12.711995	Pass
HighBloodPressure	1919	0.442857	Pass

Tool

Python is the primary programming language to implement the machine learning models and data analysis processes.

scikit-learn, xgboost, lightgbm:
machine learning libraries

pandas:
data manipulation and analysis

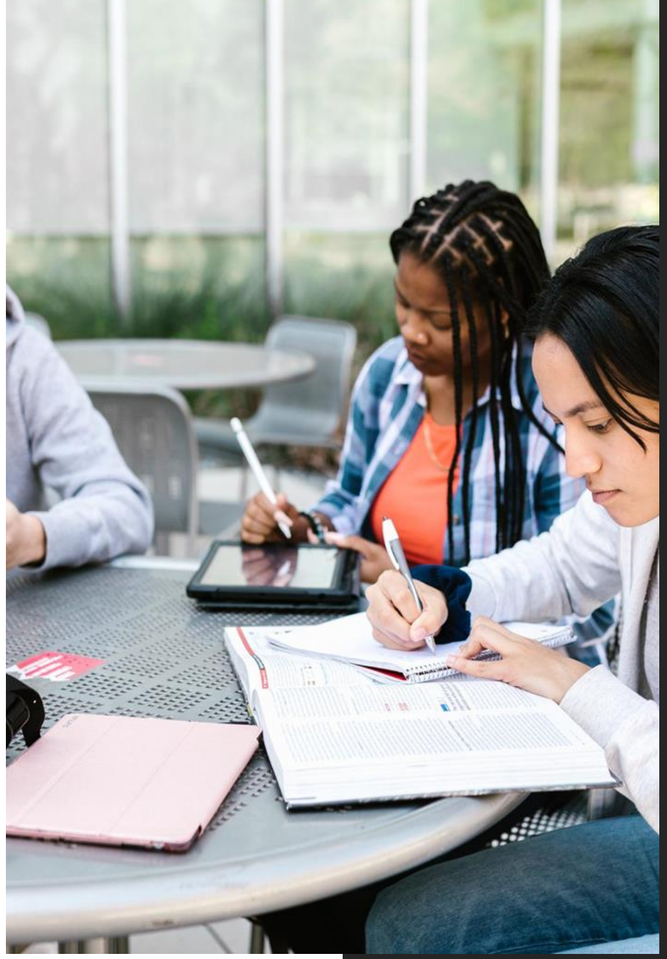
imbalanced-learn:
library specifically designed to handle imbalanced datasets

matplotlib/seaborn:
creating data visualizations

optuna:
Hyperparameter tuning

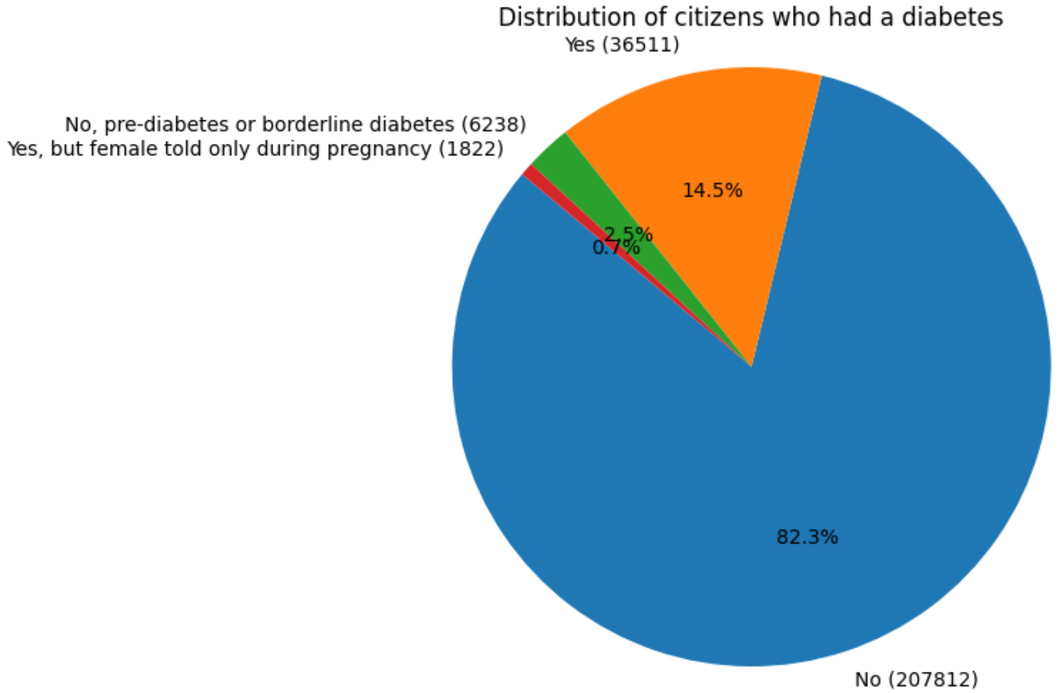
Streamlit:
Web app deployment

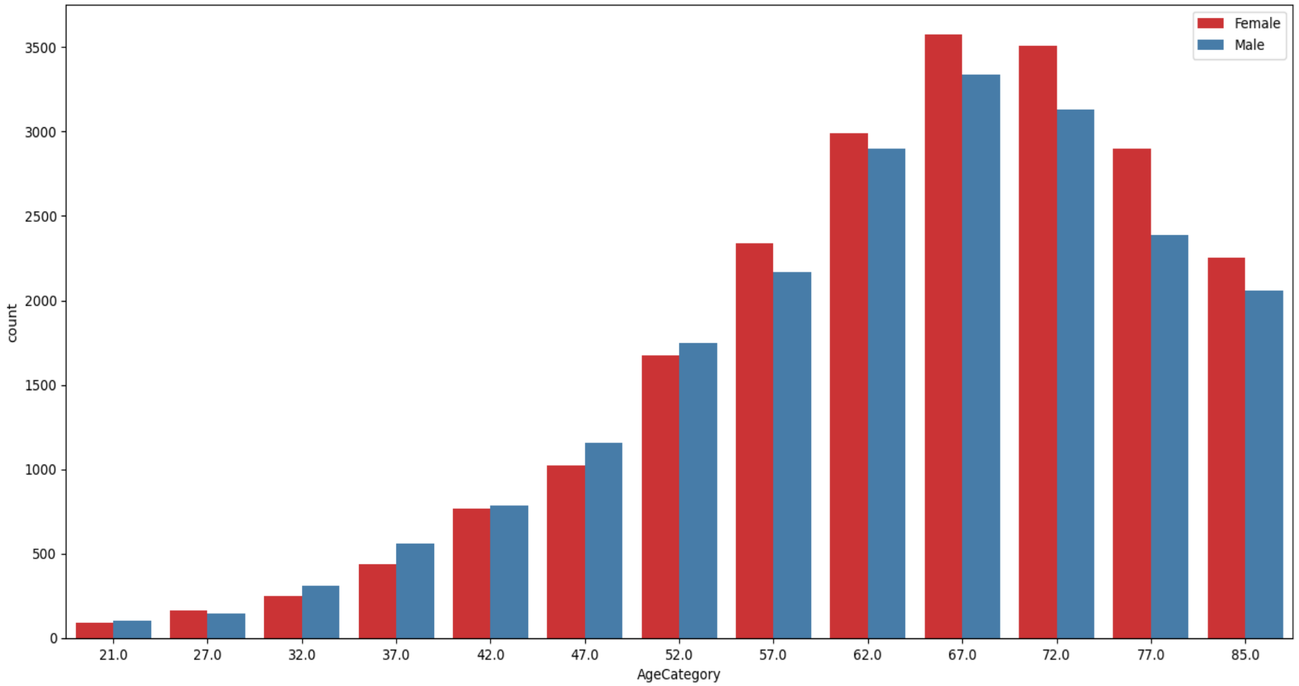
Github:
Version control and collaboration



Results and Discussions

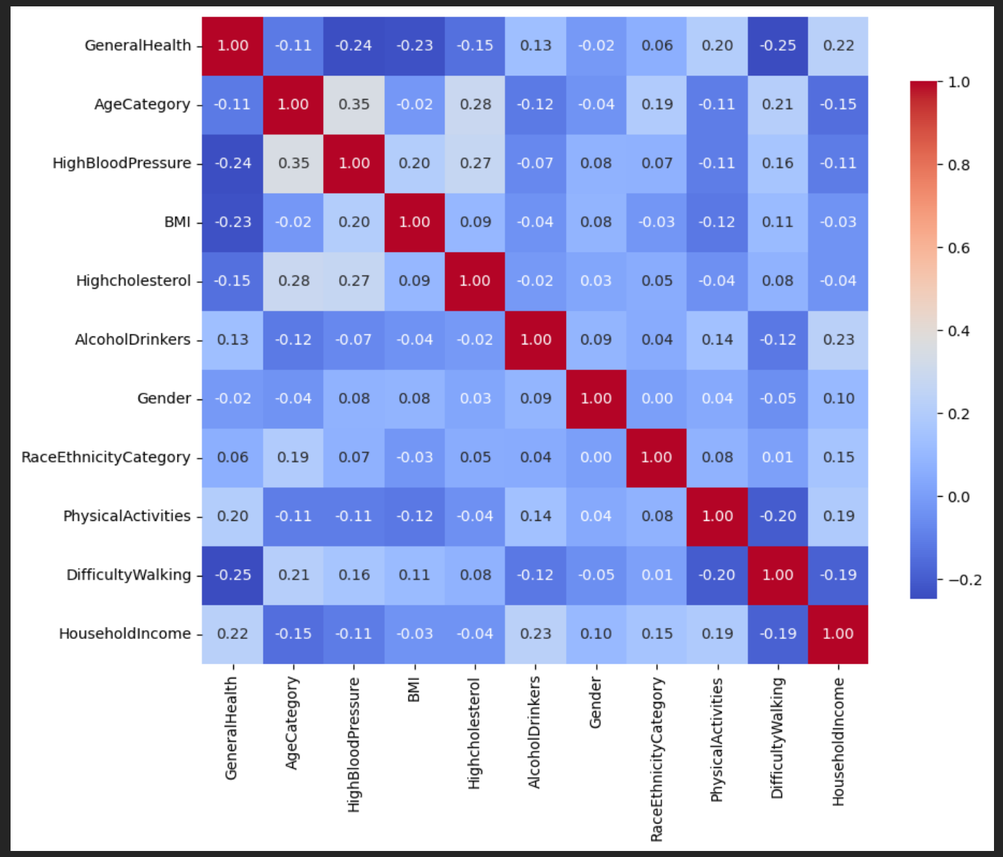
EDA


Distribution of citizens who had a diabetes
Yes (36511)
No, pre-diabetes or borderline diabetes (6238)
Yes, but female told only during pregnancy (1822)
14.5%
2.5%
0.7%
82.3%
No (207812)


Female
Male
count
AgeCategory
0
500
1000
1500
2000
2500
3000
3500
21.0
27.0
32.0
37.0
42.0
47.0
52.0
57.0
62.0
67.0
72.0
77.0
85.0

EDA

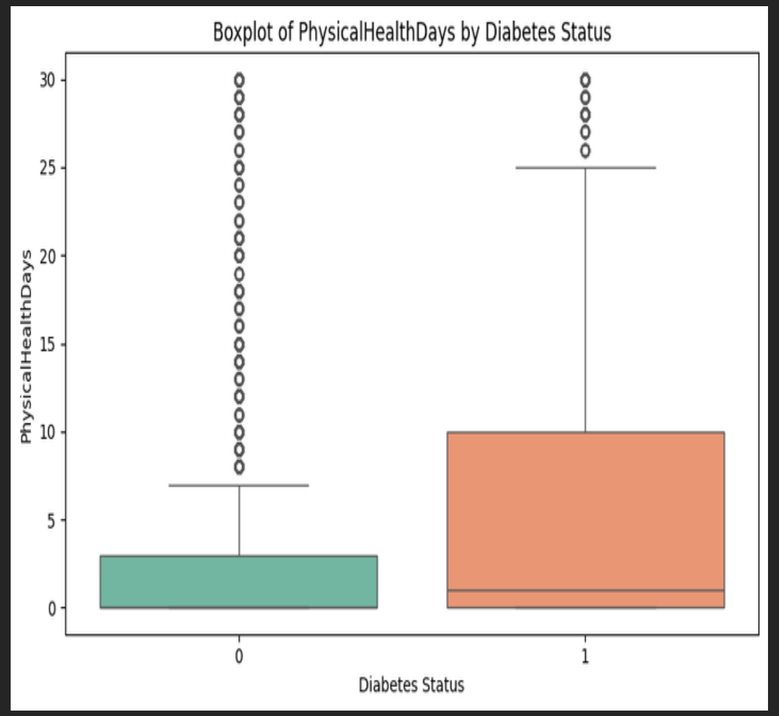
Continuation ...

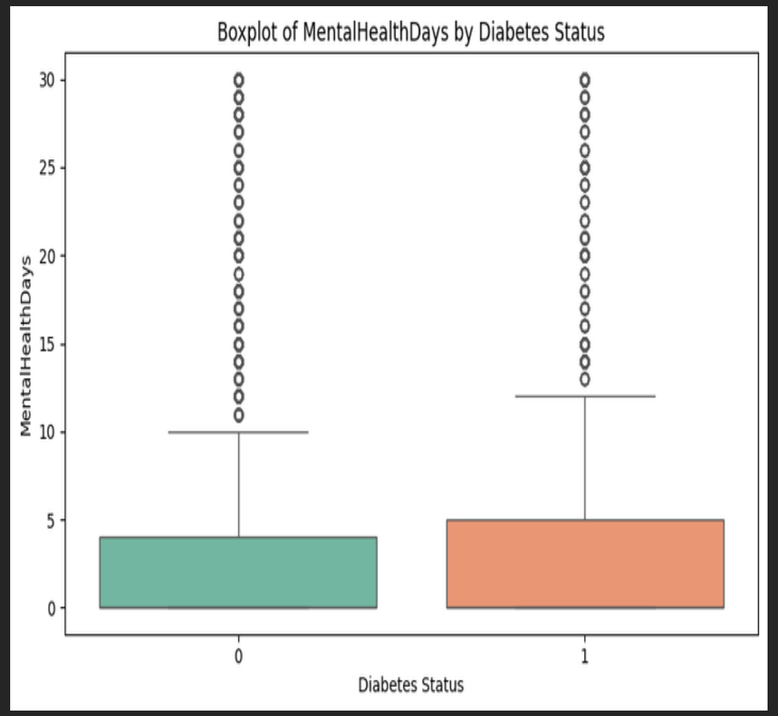


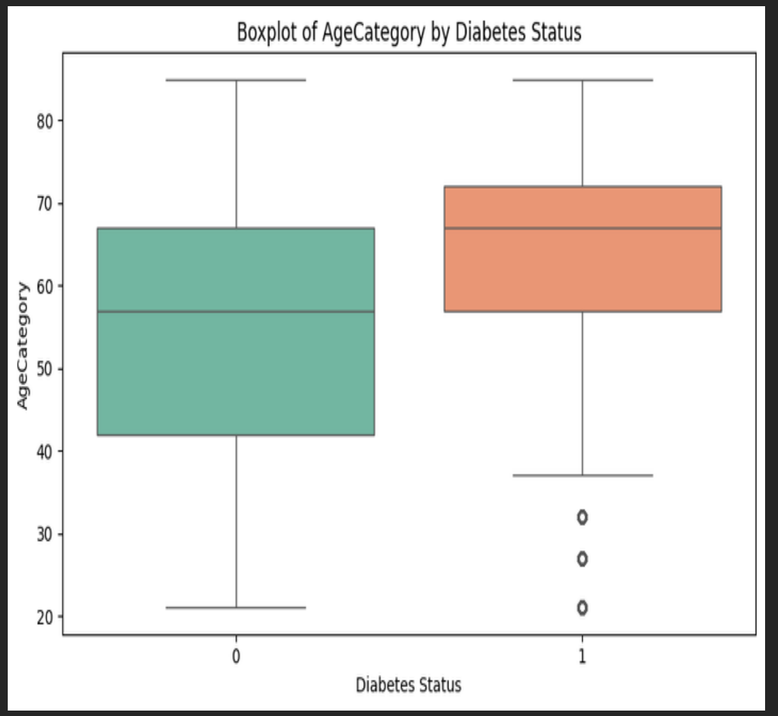
Feature	Phase	Count	Mean	Std	Min	25%	50%	75%	Max
PhysicalHealthDays	Before	201,906	4.31	8.60	0	0	0	3	30
	After	153,545	0.81	1.65	0	0	0	1	7
MentalHealthDays	Before	201,906	4.14	8.01	0	0	0	4	30
	After	153,545	1.22	2.43	0	0	0	2	10
AgeCategory	Before	201,906	57.01	17.07	21	42	57	72	85
	After	153,545	57.30	16.99	21	42	57	72	85

EDA

Continuation ...



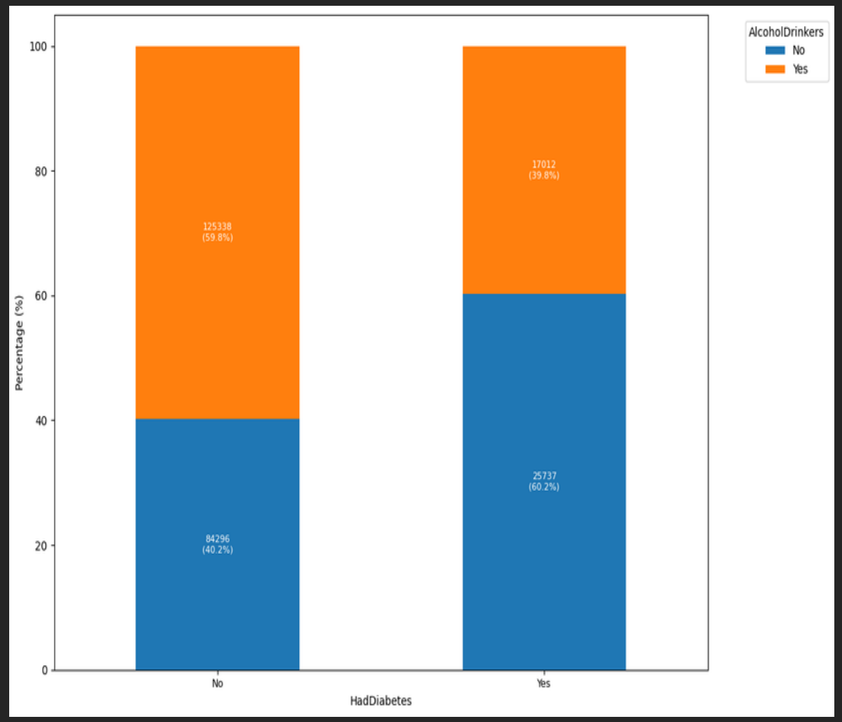




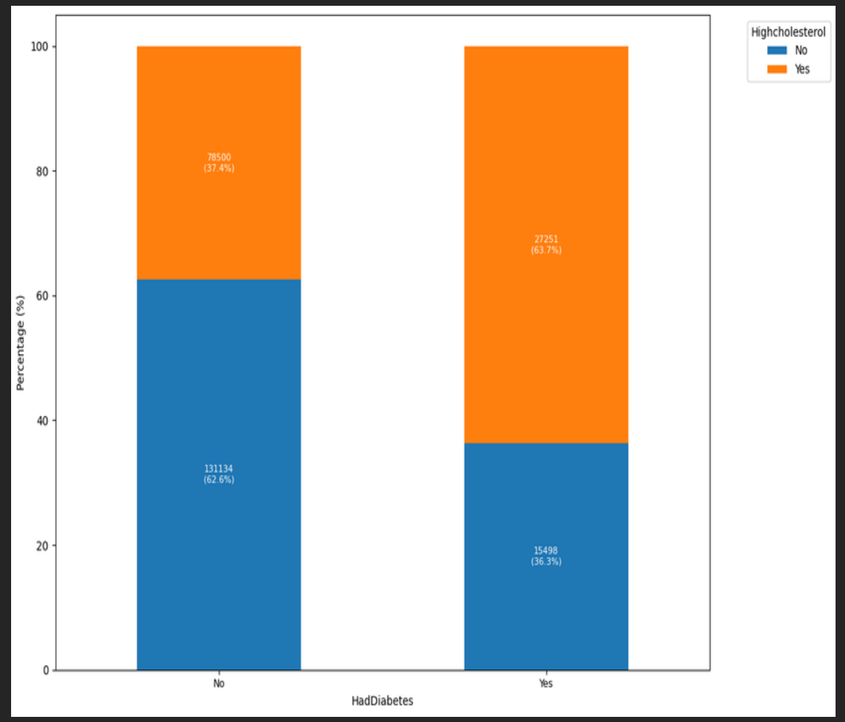
EDA

Continuation …

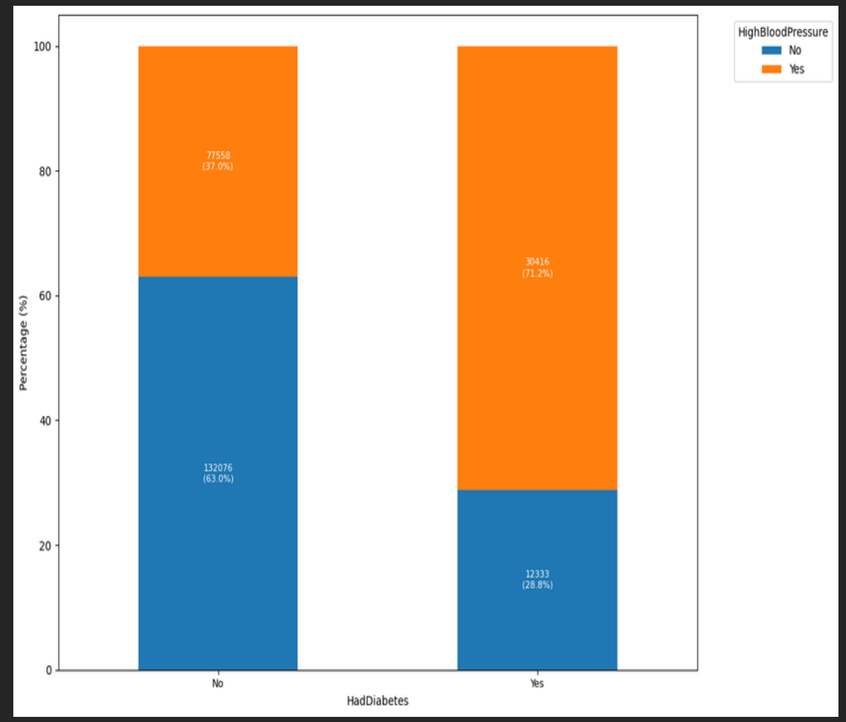
AlcoholDrinkers



HighCholesterol



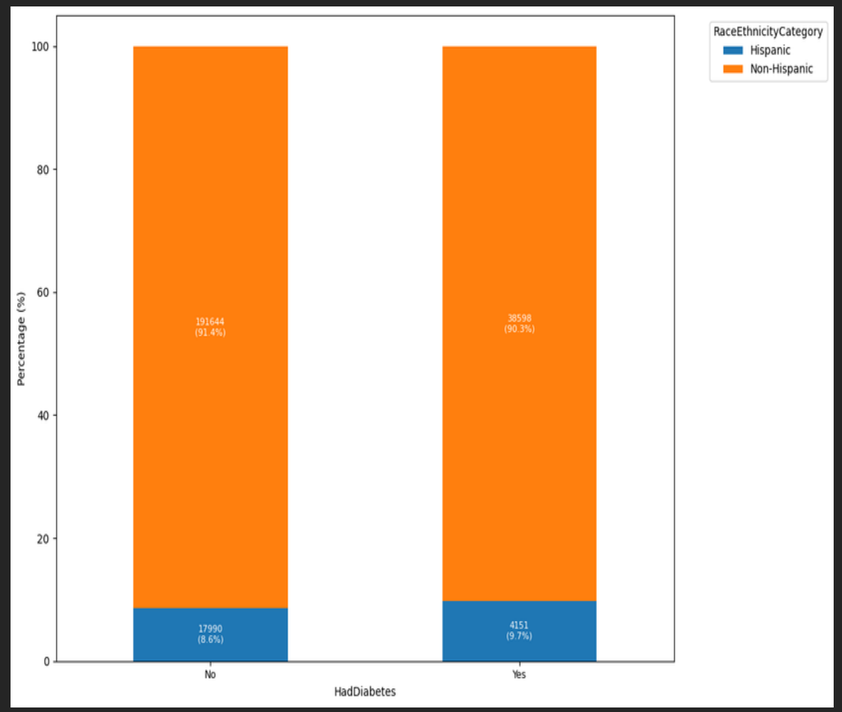
HighBloodPressure



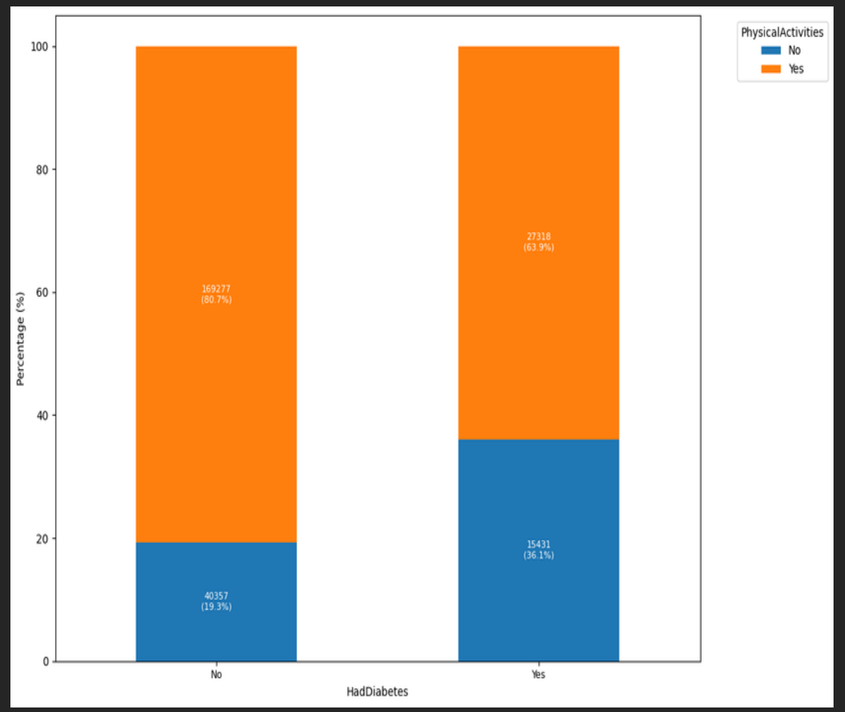
EDA

Continuation …

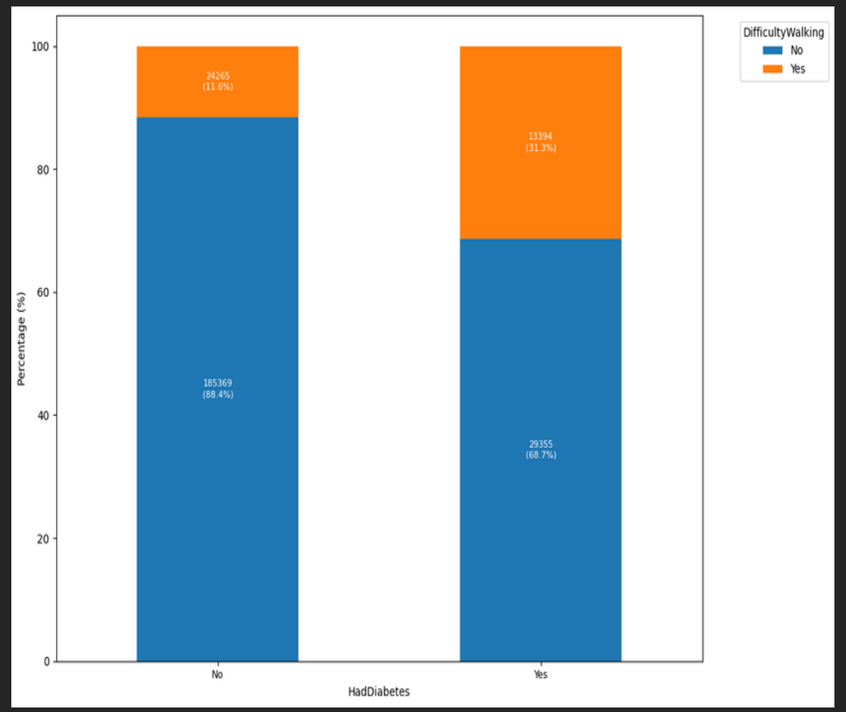
RaceEthnicityCategory



PhysicalActivities



DifficultyWalking



Handling Class Imbalance

Resampling	Classifier	Mean ROC-AUC	Mean PR-AUC
SVMSMOTE	XGBoost	0.790787	0.386115
SMOTE	XGBoost	0.786424	0.380721
SMOTETomek	XGBoost	0.786205	0.380082
BorderlineSMOTE	XGBoost	0.784348	0.358529
ADASYN	XGBoost	0.78291	0.369321
SMOTEENN	XGBoost	0.761859	0.352832
SVMSMOTE	RF	0.734566	0.313346
SMOTETomek	RF	0.724207	0.301932
SMOTE	RF	0.724061	0.301683
BorderlineSMOTE	RF	0.722848	0.28724
ADASYN	RF	0.71623	0.28768
SMOTEENN	RF	0.690679	0.257139
SVMSMOTE	NN	0.797031	0.393162
SMOTE	NN	0.795626	0.389482
SMOTETomek	NN	0.795185	0.389344
BorderlineSMOTE	NN	0.79409	0.380568
ADASYN	NN	0.792688	0.383159
SMOTEENN	NN	0.778213	0.365406
SVMSMOTE	LR	0.793963	0.380982
SMOTETomek	LR	0.792303	0.37538
SMOTE	LR	0.792302	0.375392
BorderlineSMOTE	LR	0.791678	0.372953
ADASYN	LR	0.790363	0.370632
SMOTEENN	LR	0.783047	0.366058

Resampling	Classifier	Mean ROC-AUC	Mean PR-AUC
SVMSMOTE	LightGBM	0.797527	0.394388
SMOTE	LightGBM	0.795062	0.390901
SMOTETomek	LightGBM	0.795052	0.391279
BorderlineSMOTE	LightGBM	0.794059	0.382657
ADASYN	LightGBM	0.792325	0.384601
SMOTEENN	LightGBM	0.774432	0.367736
SVMSMOTE	KNN	0.700129	0.266645
BorderlineSMOTE	KNN	0.699917	0.263657
ADASYN	KNN	0.699648	0.261636
SMOTETomek	KNN	0.698972	0.263142
SMOTE	KNN	0.698823	0.2632
SMOTEENN	KNN	0.62777	0.216593
SVMSMOTE	GB	0.79877	0.395147
SMOTETomek	GB	0.796633	0.39025
SMOTE	GB	0.796587	0.390084
BorderlineSMOTE	GB	0.7959	0.382742
ADASYN	GB	0.794252	0.384441
SMOTEENN	GB	0.779989	0.370849
SVMSMOTE	DT	0.694332	0.288041
SMOTETomek	DT	0.690933	0.283092
SMOTE	DT	0.690855	0.283053
ADASYN	DT	0.685936	0.273992
BorderlineSMOTE	DT	0.685411	0.271491
SMOTEENN	DT	0.647253	0.227013
SMOTE	AdaBoost	0.794341	0.385874
SMOTETomek	AdaBoost	0.794295	0.385827
BorderlineSMOTE	AdaBoost	0.7934	0.381152
SVMSMOTE	AdaBoost	0.793299	0.385212
ADASYN	AdaBoost	0.792049	0.380889
SMOTEENN	AdaBoost	0.782941	0.365454

Among the resampling methods, SVMSMOTE consistently achieved the highest scores across most ML models.

Standard vs. Model-Specific SVMSMOTE

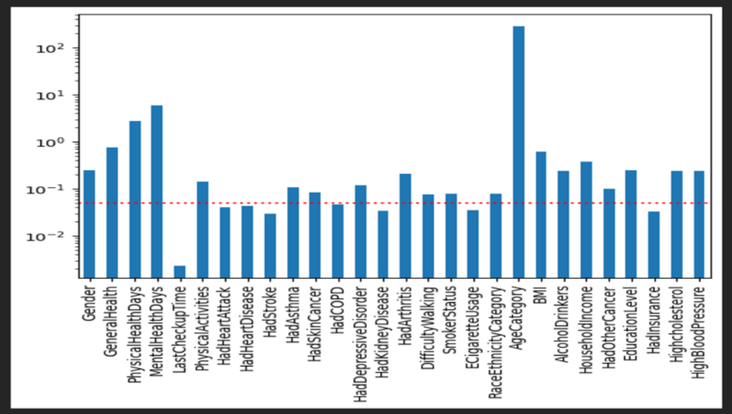
Model	Method	ROC-AUC	PR-AUC	Accuracy	F1	Precision	Recall
DecisionTreeClassifier	StandardSVMSMOTE	0.690287	0.321057	0.730451	0.418547	0.329764	0.572749
LogisticRegression	StandardSVMSMOTE	0.799965	0.434615	0.715474	0.470428	0.343511	0.746082
XGBoost	StandardSVMSMOTE	0.796519	0.436774	0.728193	0.472348	0.351879	0.718246
RandomForestClassifier	StandardSVMSMOTE	0.744288	0.355361	0.729025	0.427123	0.332703	0.596374
KNN	StandardSVMSMOTE	0.714131	0.312419	0.803633	0.357282	0.400902	0.322222
NeuralNetwork	StandardSVMSMOTE	0.803704	0.447457	0.737643	0.478273	0.360601	0.709942
GradientBoosting	StandardSVMSMOTE	0.804473	0.447931	0.716524	0.472363	0.34493	0.749123
AdaBoost	StandardSVMSMOTE	0.799452	0.438596	0.719179	0.470667	0.345713	0.737076
LightGBM	StandardSVMSMOTE	0.803766	0.446489	0.724904	0.477026	0.351794	0.740702
DecisionTreeClassifier	SVMSMOTE	**0.698359**	0.323282	0.756325	0.423834	0.353493	0.529123
LogisticRegression	SVMSMOTE	**0.800251**	0.436328	0.815916	0.444458	0.454623	0.434737
XGBoost	SVMSMOTE	**0.798641**	0.436555	0.807041	0.458829	0.437024	0.482924
RandomForestClassifier	SVMSMOTE	**0.753156**	0.370432	0.79664	0.406956	0.402101	0.41193
KNN	SVMSMOTE	**0.715355**	0.31696	0.812568	0.343396	0.422256	0.289357
NeuralNetwork	SVMSMOTE	**0.804694**	0.447888	0.801949	0.474478	0.430918	0.527836
GradientBoosting	SVMSMOTE	**0.805174**	0.449503	0.81966	0.448737	0.465277	0.433333
AdaBoost	SVMSMOTE	**0.800439**	0.438277	0.818016	0.44656	0.460487	0.43345
LightGBM	SVMSMOTE	**0.804599**	0.44637	0.821107	0.445672	0.468992	0.424561

Model-Specific SVMSMOTE

Model	smote_k	m_neighbors	smote_sampling	out_step
DT	13	13	0.7	0.9
LR	15	8	0.3	0.3
XGBoost	6	12	0.4	0.7
RF	13	12	0.4	0.2
KNN	9	5	0.3	0.2
NN	9	6	0.4	1.0
GB	13	9	0.3	0.6
AdaBoost	6	19	0.3	0.2
LightGBM	15	6	0.3	0.4

Feature Selection

Variance Thresholding



Statistical Significance Testing

Variable	Test Type	p-Value	Significance
PhysicalHealthDays	t-test	6.33E-62	Significant
MentalHealthDays	t-test	2.02E-73	Significant
AgeCategory	t-test	0	Significant
Gender	Chi-square	1.13E-45	Significant
PhysicalActivities	Chi-square	0	Significant
HadAsthma	Chi-square	2.51E-24	Significant
HadSkinCancer	Chi-square	2.53E-06	Significant
HadDepressiveDisorder	Chi-square	2.70E-11	Significant
HadArthritis	Chi-square	0	Significant
DifficultyWalking	Chi-square	0	Significant
SmokerStatus	Chi-square	0.1774274	Not significant
RaceEthnicityCategory	Chi-square	7.70E-06	Significant
AlcoholDrinkers	Chi-square	0	Significant
HadOtherCancer	Chi-square	2.25E-69	Significant
EducationLevel	Chi-square	3.76E-268	Significant
Highcholesterol	Chi-square	0	Significant
HighBloodPressure	Chi-square	0	Significant
GeneralHealth	Mann-Whitney U	0	Significant
BMI	Mann-Whitney U	0	Significant
HouseholdIncome	Mann-Whitney U	0	Significant

Model-based Methods

Model	Method	Feature Count	ROC-AUC	Selected Features
LR	Permutation	18	0.8001	GeneralHealth, AgeCategory, BMI, HighBloodPressure, Highcholesterol, AlcoholDrinkers, Gender, RaceEthnicityCategory, PhysicalActivities, HadSkinCancer, DifficultyWalking, MentalHealthDays, HouseholdIncome, EducationLevel, HadDepressiveDisorder, HadAsthma, HadOtherCancer, PhysicalHealthDays
KNN	Permutation	19	0.6971	AgeCategory, GeneralHealth, MentalHealthDays, BMI, PhysicalHealthDays, HouseholdIncome, HighBloodPressure, Highcholesterol, AlcoholDrinkers, Gender, EducationLevel, HadArthritis, PhysicalActivities, HadOtherCancer, HadSkinCancer, RaceEthnicityCategory, HadAsthma, HadDepressiveDisorder, DifficultyWalking
NN	Permutation	19	0.6768	GeneralHealth, AgeCategory, HighBloodPressure, BMI, Highcholesterol, AlcoholDrinkers, Gender, DifficultyWalking, RaceEthnicityCategory, PhysicalActivities, MentalHealthDays, HouseholdIncome, HadArthritis, HadSkinCancer, EducationLevel, PhysicalHealthDays, HadOtherCancer, HadDepressiveDisorder, HadAsthma
DT	RFECV	3	0.7743	GeneralHealth, AgeCategory, HighBloodPressure
RF	RFECV	3	0.7744	GeneralHealth, AgeCategory, HighBloodPressure
XGBoost	RFECV	11	0.7994	Gender, GeneralHealth, PhysicalActivities, DifficultyWalking, RaceEthnicityCategory, AgeCategory, BMI, AlcoholDrinkers, HouseholdIncome, Highcholesterol, HighBloodPressure
GB	RFECV	19	0.8065	Gender, GeneralHealth, PhysicalHealthDays, MentalHealthDays, PhysicalActivities, HadAsthma, HadSkinCancer, HadDepressiveDisorder, HadArthritis, DifficultyWalking, RaceEthnicityCategory, AgeCategory, BMI, AlcoholDrinkers, HouseholdIncome, HadOtherCancer, EducationLevel, Highcholesterol, HighBloodPressure
AdaBoost	RFECV	16	0.8030	Gender, GeneralHealth, MentalHealthDays, PhysicalActivities, HadAsthma, HadSkinCancer, HadDepressiveDisorder, DifficultyWalking, RaceEthnicityCategory, AgeCategory, BMI, AlcoholDrinkers, HouseholdIncome, EducationLevel, Highcholesterol, HighBloodPressure
LightGBM	RFECV	19	0.8059	Gender, GeneralHealth, PhysicalHealthDays, MentalHealthDays, PhysicalActivities, HadAsthma, HadSkinCancer, HadDepressiveDisorder, HadArthritis, DifficultyWalking, RaceEthnicityCategory, AgeCategory, BMI, AlcoholDrinkers, HouseholdIncome, HadOtherCancer, EducationLevel, Highcholesterol, HighBloodPressure

Model Performance

Hyperparameter Tuning

Model	Optimized Parameters
RF	n_estimators=150, criterion='gini', max_features='log2', max_depth=80, min_samples_split=5, min_samples_leaf=4, class_weight='balanced', bootstrap=True
DT	criterion='entropy', splitter='random', max_depth=20, min_samples_split=7, min_samples_leaf=2, max_leaf_nodes=50
KNN	n_neighbors=10, weights='uniform', algorithm='ball_tree', leaf_size=30, metric='manhattan'
LR	penalty='l2', C=2.3742, max_iter=1500, solver='saga', class_weight=None
XGBoost	learning_rate=0.05, n_estimators=250, subsample=0.6, scale_pos_weight=0.9, gamma=0.25, max_depth=5, min_child_weight=2, colsample_bytree=0.9, lambda=1.16, alpha=8.51
GB	n_estimators=200, learning_rate=0.0743, max_depth=4, min_samples_split=7, min_samples_leaf=7, subsample=0.825, max_features='log2'
AdaBoost	n_estimators=200, learning_rate=0.0515, base_estimator: {max_depth=7, min_samples_split=3, min_samples_leaf=5}
LightGBM	learning_rate=0.07, n_estimators=100, num_leaves=35, min_child_samples=15, subsample=0.9, colsample_bytree=0.9, boosting_type='gbdt'
NN	hidden_layer_sizes=(50,50), activation='logistic', solver='adam', alpha=0.00927, learning_rate='invscaling', max_iter=900

Model Performance

Before vs. After Hyperparameter Tuning

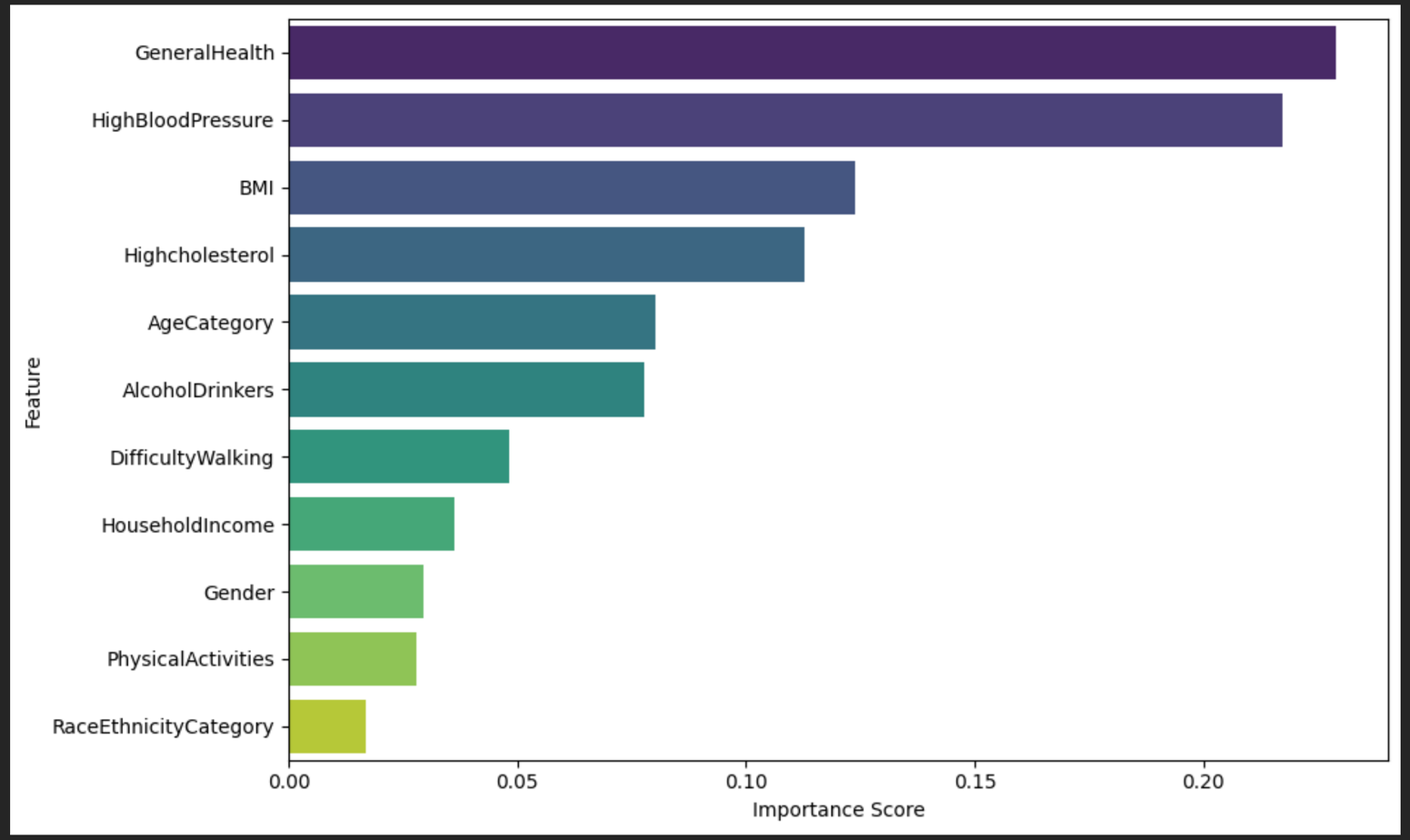
Model	Accuracy (Before)	Accuracy (After)	F1 (Before)	F1 (After)	ROC-AUC (Before)	ROC-AUC (After)
DT	0.722	0.759	0.424	0.457	0.698	0.779
LR	0.775	0.826	0.444	0.418	0.800	0.798
XGBoost	0.768	0.839	0.459	0.369	0.799	0.807
RF	0.751	0.800	0.407	0.470	0.753	0.803
KNN	0.702	0.818	0.343	0.295	0.715	0.742
NN	0.781	0.808	0.474	0.430	0.805	0.788
GB	0.779	0.838	0.449	0.377	0.805	0.808
AdaBoost	0.774	0.835	0.447	0.397	0.800	0.804
LightGBM	0.778	0.840	0.446	0.345	0.805	0.808

Model Performance

Before vs. After Threshold Optimization

Model	Accuracy Before	Precision Before	Recall Before	F1 Before	ROC_AUC Before	Accuracy After	Precision After	Recall After	F1 After	ROC_AUC After	Threshold
DT	0.758781	0.369324	0.599298	0.45701	0.7794	0.758781	0.369324	0.599298	0.45701	0.7794	0.492462
LR	0.826059	0.4824	0.368655	0.417926	0.797955	0.784971	0.406189	0.583392	0.478925	0.797955	0.336683
XGBoost	0.83856	0.545892	0.278947	0.369224	0.806689	0.770054	0.39128	0.643392	0.48662	0.806689	0.256281
RF	0.800127	0.426637	0.523392	0.470088	0.802515	0.766349	0.386255	0.644211	0.482946	0.802515	0.427136
KNN	0.818432	0.431038	0.224795	0.295488	0.742102	0.729679	0.334653	0.603041	0.430438	0.742102	0.201005
NN	0.80815	0.432708	0.426433	0.429548	0.787801	0.741882	0.354889	0.640585	0.45674	0.787801	0.321608
GB	0.837966	0.540476	0.289708	0.377218	0.807896	0.789944	0.416266	0.596842	0.490461	0.807896	0.301508
AdaBoost	0.834935	0.520691	0.320819	0.397018	0.804228	0.766131	0.387331	0.654386	0.486628	0.804228	0.351759
LightGBM	0.840264	0.564916	0.247836	0.344525	0.808233	0.773679	0.396186	0.641404	0.489818	0.808233	0.251256

Feature Importance



- GeneralHealth and HighBloodPressure are the most influential predictors of diabetes risk, together contributing over 45% of the model's decisions.
- Other significant features include BMI, Highcholesterol, and AgeCategory, indicating a strong link between general well-being, metabolic indicators, and diabetes outcomes.

Benchmarking

Generally, the performance of GB in this research consistently outperformed previous studies.

Author	Best Model	ROC-AUC	PR-AUC	Accuracy	Recall	Precision	F1-Score
(Chowdhury et al., 2024)	GB	**0.791**	NA	0.703	**0.717**	0.635	NA
(Nguyen & Zhang, 2025)	LR	NA	NA	**0.76**	0.76	0.74	0.75
(Pechprasarn et al., 2025)	SVM	NA	NA	**0.743**	**0.831**	**0.706**	**0.763**
(Horestani, 2024)	LR	**0.7743**	NA	0.7432	0.7169	0.6987	0.7078
(Omoora et al., 2023)	XGBoost	0.768	NA	**0.7403**	0.6897	0.7667	0.726
(Su, 2023)	LR	NA	NA	**0.7448**	NA	NA	NA
This Study	GB	0.8058	0.4494	0.77	0.64	0.39	0.49

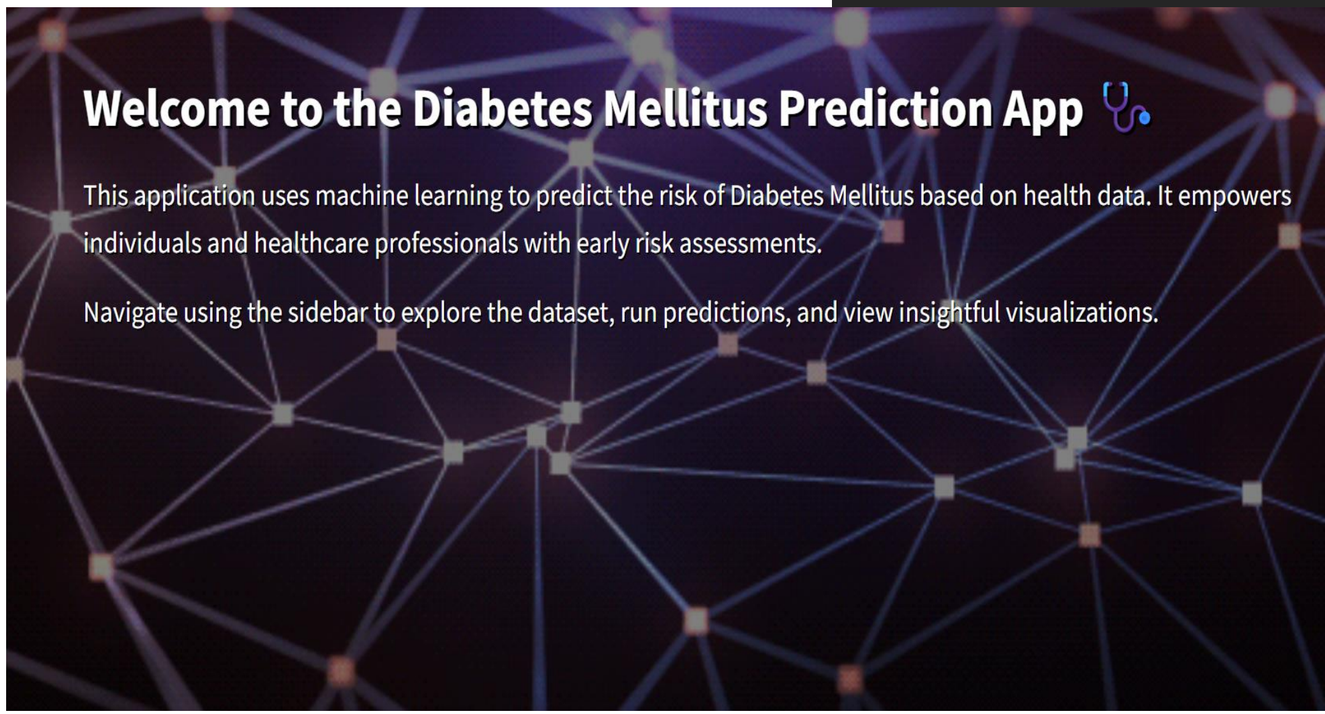
Deployment

User Input

Preprocessing Pipeline

ML Model Prediction

Risk Score



...

Conclusion

Revisiting the Research Questions

Question 1:
What are the factors that contribute to the likelihood of diabetes mellitus?

Feature Selection Strategy:

- Applied variance thresholding (< 0.05) to remove low-variability features
- Conducted statistical tests (t-test, chi-square, Mann-Whitney U)
- Applied model-based selection (RFECV, permutation importance)

Final Predictors:

- GeneralHealth, AgeCategory, HighBloodPressure, BMI, HighCholesterol, AlcoholDrinkers, Gender, Race/Ethnicity, PhysicalActivities, DifficultyWalking, HouseholdIncome

Revisiting the Research Questions

Question 2:
What are the methods that can circumvent the class imbalance in real-world datasets?

Problem:
- Only ~17% of cases were diabetic — leading to skewed predictions toward the majority class
- Imbalance reduces sensitivity in identifying diabetic individuals

Resampling Strategies Tested:
- SVMSMOTE (Best), SMOTE, SMOTEENN, SMOTETomek, BorderlineSMOTE, ADASYN (with SKCV)

Finding:
- SVMSMOTE outperformed others across 9 ML models

Revisiting the Research Questions

Question 3:
How effective are the ML models in diabetes mellitus prediction?

9 ML models tested — including LR, DT, KNN, RF, GB, XGBoost, LightGBM, AdaBoost, NN

- Trained using SVMSMOTE + evaluated with SKCV

Evaluation Metrics:

- Primary: ROC-AUC, with Accuracy, Precision, Recall, F1-score
- Supporting: PR_AUC

Findings:

- GB achieved highest ROC-AUC: 0.8058
- Showed strong performance across all metrics after threshold optimization
- Ensemble models outperformed simpler models overall

Limitations and Future Works

Limitations

Self-reported Data
BRFSS responses may contain recall or social desirability bias.

No Clinical Validation
Diabetes status not confirmed by lab tests or medical records.

Limited Generalizability
Model not validated on external datasets or different populations.

Interpretability Constraints
GB is less transparent for clinical use.

Future Works

External Validation
Test model on EHR data and other demographic groups.

Explainable AI Integration
Use SHAP to enhance model transparency.

System Integration
Expand Streamlit app for mHealth and healthcare system use.

Extended Disease Prediction
Apply framework to comorbid conditions (e.g., hypertension).

Reference

Akter, S. B., Akter, S., & Pias, T. S. (2023). Stroke Risk Prediction from Medical Survey Data: AI-Driven Risk Analysis with Insightful Feature Importance using Explainable AI (XAI). *medRxiv*, 2023.2011. 2017.23298646.

American Diabetes Association. (2024). Standards of medical care in diabetes—2024. *Diabetes Care, 47*(Supplement_1).

Arslan, N. N., & Özdemir, D. (2023). A comparison of traditional and state-of-the-art machine learning algorithms for type 2 diabetes prediction. *Journal of Scientific Reports-C*(006), 1-11.

Budhathoki, N., Bhandari, R., Bashyal, S., & Lee, C. (2023). Predicting asthma using imbalanced data modeling techniques: Evidence from 2019 Michigan BRFSS data. *Plos one, 18*(12), e0295427.

Burch, A. E., Elliott, S. K., & Harris, S. T. (2024). Associations between social determinants of health and diabetes self-care behaviors among insured adult patients. *Diabetes Research and Clinical Practice, 207*, 111048.

Chang, V., Ganatra, M. A., Hall, K., Golightly, L., & Xu, Q. A. (2022). An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators. *Healthcare Analytics, 2*, 100118. doi:https://doi.org/10.1016/j.health.2022.100118

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321-357.

Chowdhury, M. M., Ayon, R. S., & Hossain, M. S. (2024). An investigation of machine learning algorithms and data augmentation techniques for diabetes diagnosis using class imbalanced BRFSS dataset. *Healthcare Analytics, 5*, 100297.

Collins, G. S., Mallett, S., Omar, O., & Yu, L.-M. (2011). Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC medicine, 9*, 1-14.

Dong, W. (2023). Research on Type 2 Diabetes Risk based on Lifestyle Factors. *Highlights in Science, Engineering and Technology, 54*, 461-473. doi:10.54097/hset.v54i.9826

Fleitman, O. (2024). Prominent Risk Factors in Diabetes.

Hama Saeed, M. A. (2023). Diabetes type 2 classification using machine learning algorithms with up-sampling technique. *Journal of Electrical Systems and Information Technology, 10*(1), 8.

Hossain, M. J., Al-Mamun, M., & Islam, M. R. (2024). Diabetes mellitus, the fastest growing global public health concern: Early detection should be focused. *Health Sci Rep, 7*(3), e2004. doi:10.1002/hsr2.2004

Hotz, N. (2024, 16/11/2024). What is SEMMA? *Data Science PM.* Retrieved from https://www.datascience-pm.com/semma/

Lakshmi, H., Reddy, A. S., & Naidu, K. (2023). *Analysis of Diabetic Prediction Using Machine Learning Algorithms on BRFSS Dataset.* Paper presented at the 2023 7th International Conference on Trends in Electronics and Informatics (ICOEI).

Lee, D. C., Ross, L., Quintero Arias, C., Rony, M., Patel, R., Jensen, E., . . . Anthopolos, R. (2024). Demographic and geographic distribution of diabetes and pre-diabetes risk in rural settings: results from a cross-sectional, countywide rural health survey in Sullivan County, New York. *BMJ Open, 14*(8). doi:10.1136/bmjopen-2023-080831

Akter, S. B., Akter, S., & Pias, T. S. (2023). Stroke Risk Prediction from Medical Survey Data: AI-Driven Risk Analysis with Insightful Feature Importance using Explainable AI (XAI). *medRxiv*, 2023.2011. 2017.23298646.

American Diabetes Association. (2024). Standards of medical care in diabetes—2024. *Diabetes Care, 47*(Supplement_1).

Reference

Arslan, N. N., & Özdemir, D. (2023). A comparison of traditional and state-of-the-art machine learning algorithms for type 2 diabetes prediction. *Journal of Scientific Reports-C*(006), 1-11.

Budhathoki, N., Bhandari, R., Bashyal, S., & Lee, C. (2023). Predicting asthma using imbalanced data modeling techniques: Evidence from 2019 Michigan BRFSS data. *Plos one, 18*(12), e0295427.

Burch, A. E., Elliott, S. K., & Harris, S. T. (2024). Associations between social determinants of health and diabetes self-care behaviors among insured adult patients. *Diabetes Research and Clinical Practice, 207*, 111048.

Chang, V., Ganatra, M. A., Hall, K., Golightly, L., & Xu, Q. A. (2022). An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators. *Healthcare Analytics, 2*, 100118. doi:https://doi.org/10.1016/j.health.2022.100118

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321-357.

Chowdhury, M. M., Ayon, R. S., & Hossain, M. S. (2024). An investigation of machine learning algorithms and data augmentation techniques for diabetes diagnosis using class imbalanced BRFSS dataset. *Healthcare Analytics, 5*, 100297.

Collins, G. S., Mallett, S., Omar, O., & Yu, L.-M. (2011). Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC medicine, 9*, 1-14.

Dong, W. (2023). Research on Type 2 Diabetes Risk based on Lifestyle Factors. *Highlights in Science, Engineering and Technology, 54*, 461-473. doi:10.54097/hset.v54i.9826

Fleitman, O. (2024). Prominent Risk Factors in Diabetes.

Hama Saeed, M. A. (2023). Diabetes type 2 classification using machine learning algorithms with up-sampling technique. *Journal of Electrical Systems and Information Technology, 10*(1), 8.

Hossain, M. J., Al-Mamun, M., & Islam, M. R. (2024). Diabetes mellitus, the fastest growing global public health concern: Early detection should be focused. *Health Sci Rep, 7*(3), e2004. doi:10.1002/hsr2.2004

Hotz, N. (2024, 16/11/2024). What is SEMMA? *Data Science PM.* Retrieved from https://www.datascience-pm.com/semma/

Lakshmi, H., Reddy, A. S., & Naidu, K. (2023). *Analysis of Diabetic Prediction Using Machine Learning Algorithms on BRFSS Dataset.* Paper presented at the 2023 7th International Conference on Trends in Electronics and Informatics (ICOEI).

Lee, D. C., Ross, L., Quintero Arias, C., Rony, M., Patel, R., Jensen, E., . . . Anthopolos, R. (2024). Demographic and geographic distribution of diabetes and pre-diabetes risk in rural settings: results from a cross-sectional, countywide rural health survey in Sullivan County, New York. *BMJ Open, 14*(8). doi:10.1136/bmjopen-2023-080831

Mamun, M., Uddin, M. M., Tiwari, V. K., Islam, A. M., & Ferdous, A. U. (2022). *Mlheartdis: Can machine learning techniques enable to predict heart diseases?* Paper presented at the 2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON).

Martin, S. S., Aday, A. W., Almarzooq, Z. I., Anderson, C. A. M., Arora, P., Avery, C. L., . . . Palaniappan, L. P. (2024). 2024 Heart Disease and Stroke Statistics: A Report of US and Global Data From the American Heart Association. *Circulation, 149*(8), e347-e913. doi:10.1161/cir.0000000000001209

Reference

Olayeye, T. O., Bodunwa, O. K., & Adewole, A. I. (2024). PREVALENCE AND RISK FACTORS OF DIABETES MELLITUS AMONG WOMEN USING THE MULTINOMIAL LOGISTIC REGRESSION MODEL. *FUDMA JOURNAL OF SCIENCES, 8*(1), 195-200.

Prasetyo, S. Y., Izdihar, Z. N., & Nabiilah, G. Z. (2024). *Analyzing Machine Learning Approaches for Diabetes Risk Prediction: Comparative Performance Assessment Using BRFSS Data.* Paper presented at the 2024 7th International Conference on Informatics and Computational Sciences (ICICoS).

Sah, M., Kulkarni, P., Sehgal, P., & Victor, A. Preprocessing and Detection of Diabetes Mellitus from Physiological Data Using Deep Learning. In *Deep Learning in Diabetes Mellitus Detection and Diagnosis* (pp. 26-37): CRC Press.

Shinde, A., & Singh, A. (2023, 5-7 April 2023). *Enhancing Diabetes Detection using Machine Learning: A Focus on Optimizing Recall Performance.* Paper presented at the 2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT).

Talebi Moghaddam, M., Jahani, Y., Arefzadeh, Z., Dehghan, A., Khaleghi, M., Sharafi, M., & Nikfar, G. (2024). Predicting diabetes in adults: identifying important features in unbalanced data over a 5-year cohort study using machine learning algorithm. *BMC Medical Research Methodology, 24*(1), 220.

Ullah, Z., Saleem, F., Jamjoom, M., Fakieh, B., Kateb, F., Ali, A. M., & Shah, B. (2022). Detecting High-Risk Factors and Early Diagnosis of Diabetes Using Machine Learning Methods. *Computational Intelligence and Neuroscience, 2022*(1), 2557795. doi:https://doi.org/10.1155/2022/2557795

Wu, S. (2024). Multi-factors correlation to diabetes using Machine Learning: findings from BRFSS. *Transactions on Materials, Biotechnology and Life Sciences, 3*, 143-149.

Xie, Z., Nikolayeva, O., Luo, J., & Li, D. (2019). Building risk prediction models for type 2 diabetes using machine learning techniques. *Preventing Chronic Disease, 16*(9). doi:10.5888/pcd16.190109

Thank You