THE CURIOSITY CUP 2023
A Global SAS® Student Competition

Comparing Classification Models of Predicting Diabetes using SAS Viya®

Team Shazam!

INTRODUCTION

Diabetes is a persistent health issue characterized by high levels of glucose in the blood, caused by a lack of insulin production or insensitivity to it. This leads to an accumulation of glucose in the bloodstream, increasing the risk of serious long-term health problems such as cardiovascular disease, kidney disease, vision loss, and nerve damage. Type 2 diabetes, primarily caused by insulin resistance, is mostly found in adults. According to the CDC (2022), over 37 million adults in the US are living with diabetes, with one in five remaining undiagnosed. Diabetes is the seventh leading cause of death and a major contributor to end-stage renal disease, lower limb amputations, and adult blindness. In recent decades, the number of adult diabetes cases has doubled. Early detection and effective management of diabetes are crucial for reducing its harmful effects on health and its associated socioeconomic burden. This project aims to analyze various factors that can aid in early detection through regular check-ups and provide early warning signs of the onset of the disease.

DATA PREPARATION

The data was obtained from Kaggle, and the author indicated that it was sourced from the Behavioral Risk Factor Surveillance System 2015 dataset (Chuks, P., n.d.). The author also performed cleaning on the data before uploading it to Kaggle, resulting in a dataset with 70,692 records and 18 features, which are outlined in Appendix Table 1 The descriptions of the variables were cross-referenced with the calculation description and codebook of the BRFSS 2015 (CDC, n.d.). The original diabetes.csv data file was transformed into a SAS data file using SAS Studio® from SAS OnDemand for Academics®. Appendix Table 1 shows the variables, their values, and summary statistics. This newly formatted data file was then uploaded into SAS Viya for Learners to be used for the study.

PROBLEM STATEMENT

Chronic illnesses such as diabetes are a leading cause of death and disability in the United States, as reported by the CDC (July 21, 2022). The development of this condition is influenced by a few key risk factors, including tobacco use, unhealthy diets, physical inactivity, and excessive alcohol consumption, which greatly contribute to the nation's annual healthcare costs of $4.1 trillion. According to Buttorff et al. (2017), 60% of American adults suffer from at least one chronic condition, with 42% of individuals suffering from multiple chronic conditions, including diabetes. Another study by Ward et al. (2021) found that obesity results in excessive medical costs, highlighting the importance of good health for personal, economic, and social growth. The use of big data analytics and machine learning has become increasingly crucial in healthcare, as it allows for the discovery of correlations and prediction of health risk based on behavioral factors. The purpose of this study is to develop a predictive model for diabetes using SAS Viya® and compare the performance of various machine learning models. The success of this project will be measured by its ability to provide insights and strategies for preventing and managing diabetes, ultimately improving public health, and reducing healthcare costs.

DATA CLEANING AND VALIDATION

On the imported diabetes SAS dataset in the SAS viya variable 'diabetes' was selected as a target variable and data exploration node was run to gain insight of the variables and its values. The 'stroke' variable was rejected, as it functions as a target variable if stroke prediction was the goal. The distribution of variables was viewed using the data exploration node, which showed six variables that were either imbalanced class variables or non-normally distributed interval variables (Figure 1). The severe imbalance for 'CholCheck' (2.47%) and 'HeavyAlcoholConsump' (4.27%) was determined best resolved by rejecting the input.

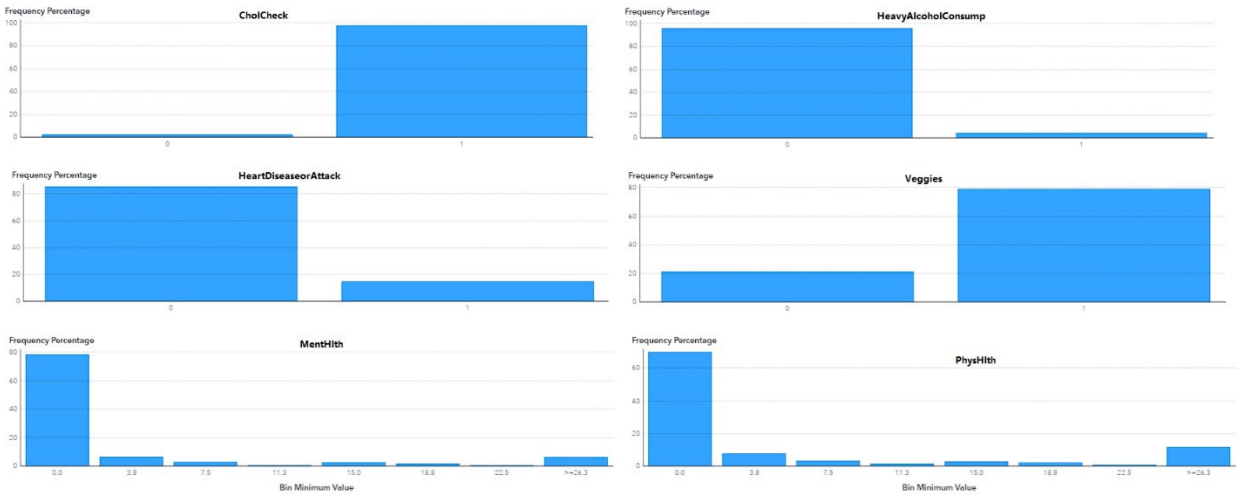

Figure 1: Variables with imbalanced classes or non-normal distribution

Although imbalanced, Veggies (21.1% veggies=0) had a large enough minority to be retained. For both 'MentHlth' and 'PhysHlth', in spite of 78.3% and 69.8% of their data grouped into the 0-3 bin respectively, they were retained because the cross-tabulation shows significant differences between diabetic patients and non-diabetic patients (Figure 2). That is, although most data fit into a single bin, the other bins show significant differences that would be worth exploring as possible predictive variables.

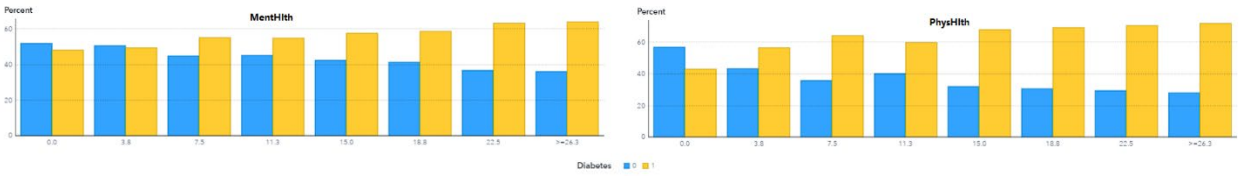

Figure 2: Cross tabulation of MentHlth (left) and PhysHlth (right) with diabetic (yellow) and non-diabetic (blue) patients.

A total of 14 variables were selected as inputs. 'Age' and 'GenHlth' were automatically determined to be nominal, as the original BRFSS data put them into bins. 'MentHlth' and 'PhysHlth' were the only interval variables, as they measured a numerical count. All other variables ('DiffWalk', 'Fruits', 'HeartDiseaseorAttack', 'HighBP', 'HighChol', 'PhysActivity', 'Sex', 'Smoker', and 'Stroke') were binary classifications.

DATA MODELS

This study aims to predict the occurrence of diabetes using SAS viya. The objective was to identify the most effective supervised learning model for predicting diabetes. Despite the limitation of the SAS Viya® for Learners platform, which disables the autotuning option, models were constructed and evaluated using the available auto run option. Six models were constructed to classify the data: Logistic Regression, Neural Network, Support Vector

Machines (SVM), Decision Tree, Random Forest, and Gradient Boost. The Model Comparison node was used to compare the models and determine the champion model for diabetes prediction. The gradient boost max depth was set to 6, while all other parameters were kept default. Data was automatically split into 60:30:10 for training, validation, and test dataset for all these models.

DATA ANALYSIS

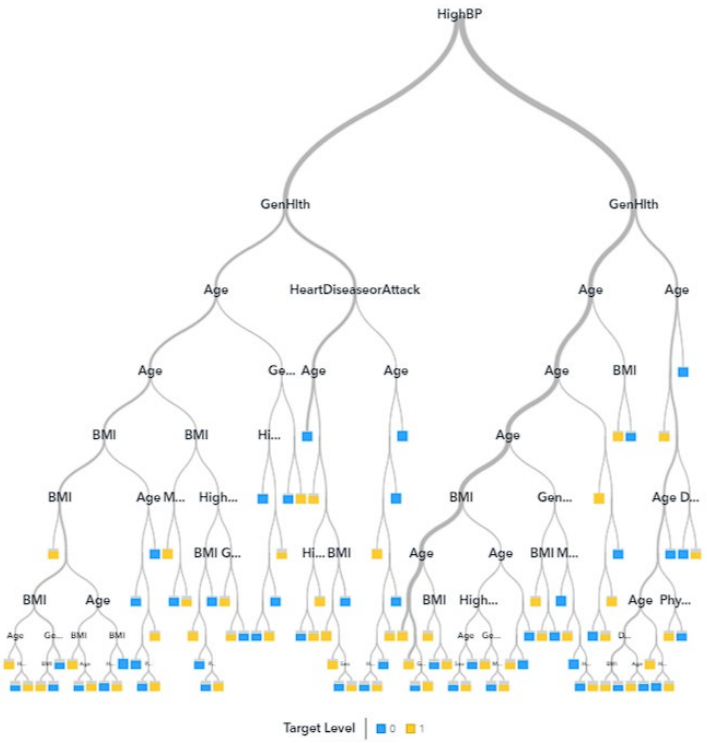

Figure 3: Decision tree. Yellow leaves indicating diabetic and blue leaves indicating non-diabetic patients. Binary variables are represented with 0 on the left branch and 1 on the right branch. At depth of 2, variable GenHlth was split into two groups, one with values of 1 and 2 on the left, and the other with values of 3, 4, and 5 on the right.

Although the decision tree offers the best interpretation of the decision process, it was the worst performing of the six models with a maximum depth of 11. Many leaf nodes were not pure, resulting in only a 73.7% accuracy in the test set. Figure 3 shows that nearly all decision nodes were associated with just four variables: GenHlth, Age, BMI, and HighBP. General health rating often had a split branch with values 1 and 2 on one side, with a value 2 being the cutoff for good health. Furthermore, several BMI branches had a decision point of 40, although it did not necessarily trend towards diabetes.

The neural network, which was the second worst accurate model, was more straightforward. It used only three input variables: GenHlth, BMI, and HighBP (Figure 4). The network consisted of just one hidden layer, containing 50 nodes, leading to one output node. Numerous nodes were linked to BMI, GenHlth 1, and GenHlth 2, indicating that it requires much more subtlety to determine if a patient has diabetes based on their good health and BMI values. Interestingly, a GenHlth rating of 3 was not considered as an input, suggesting that a "neutral" health rating is too indistinct to be used as a decision input.

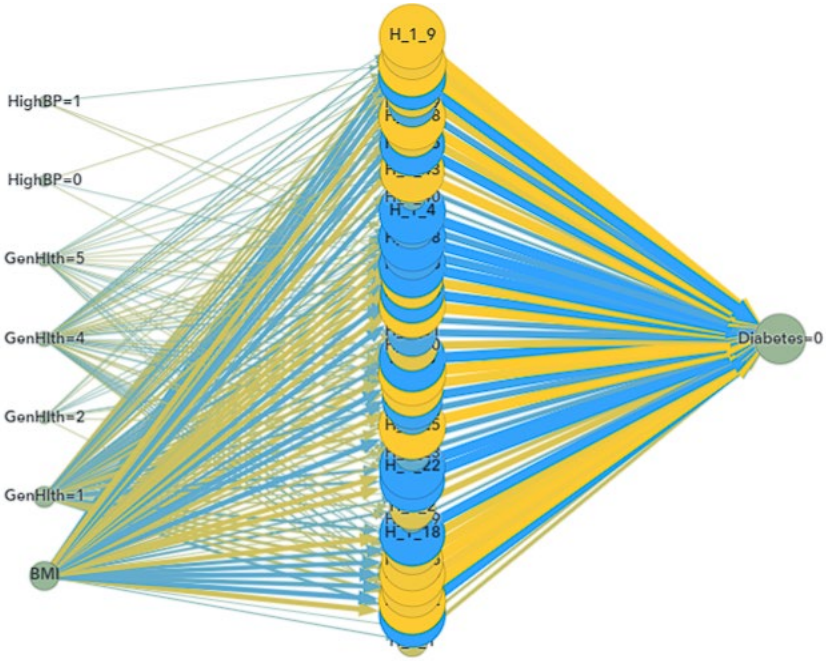

Figure 4. Neural network diagram

The SVM and Random Forest models performed well in terms of accuracy, the random forest being a more complex version of the decision tree. Although it had better accuracy and fit statistics than the decision tree, it exhibited the greatest possibility of overfitting, as the ROC curve for the training set was significantly higher than that for the validation and testing set (Figure 5). On the other hand, SVMs are prone to overfitting, but the SVM model had better ROC curves for both the validation and test datasets. However, the difference was insignificant, with 74.84% and 74.43% accuracy for the validation and test sets respectively, compared to 74.40% for the training set.

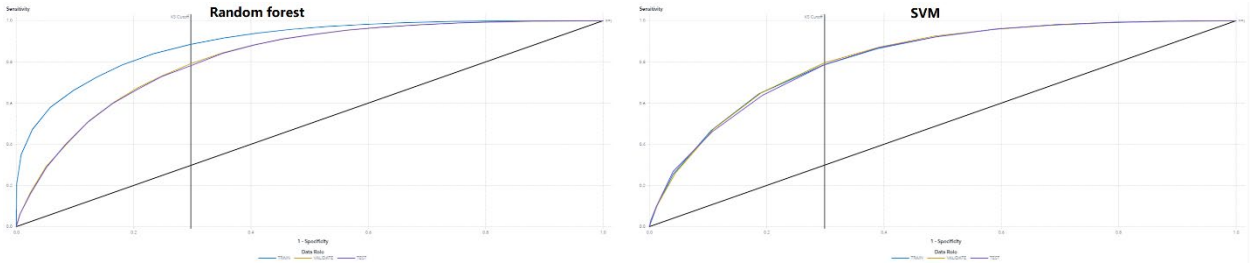

Figure 5: ROC curve of random forest (left) and SVM (right).

The logistic regression, despite having only two interval variables, resulted in 25 terms. Positive coefficients of these terms indicated contributing factors to diabetes, while negative coefficient represented the opposite effect. The analysis revealed that higher BMI values, age of 74 or above, and not consuming vegetables were contributing factors for diabetes, whereas all other factors helped prevent diabetes (Figure 6).

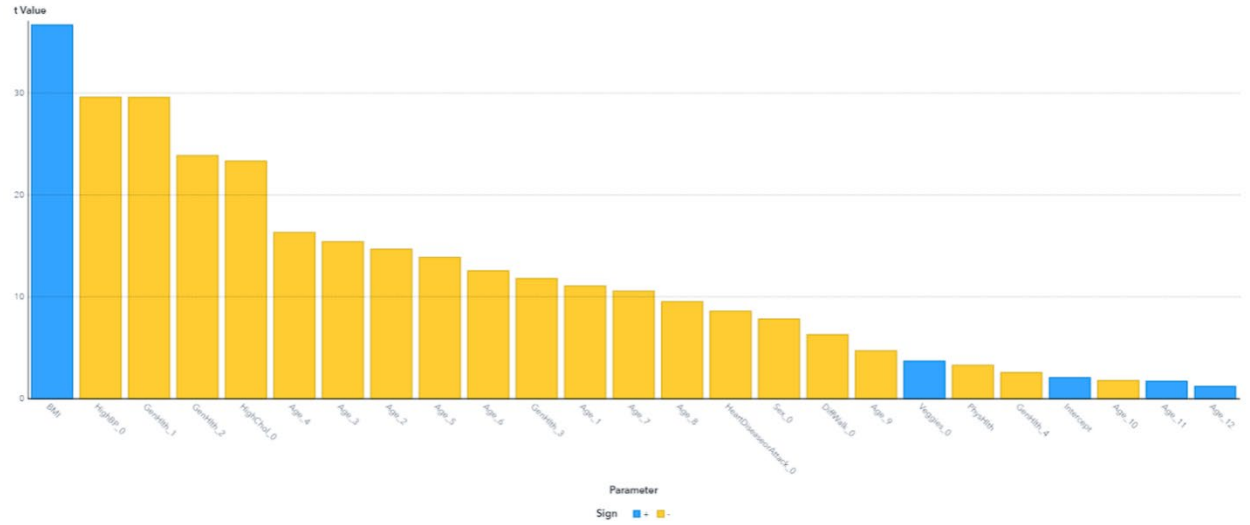

Figure 6: Terms of the logistic regression formula. The yellow bars represent terms that have a preventive effect on diabetes, while the blue terms, including the intercept, have a contributing effect to the likelihood of diabetes.

CHAMPION MODEL

SAS Viya determined the gradient boost model as the most effective model. This model utilized a learning rate of 10% and maximum depth of 6 and generated 84 trees, resulting in just over 75% accuracy for the training, validation, and test datasets. The model identified GenHlth, HighBP, Age, and BMI as the most important variables, while HighChol, PhysHlth, and MentHlth were considered somewhat important (Figure 7).

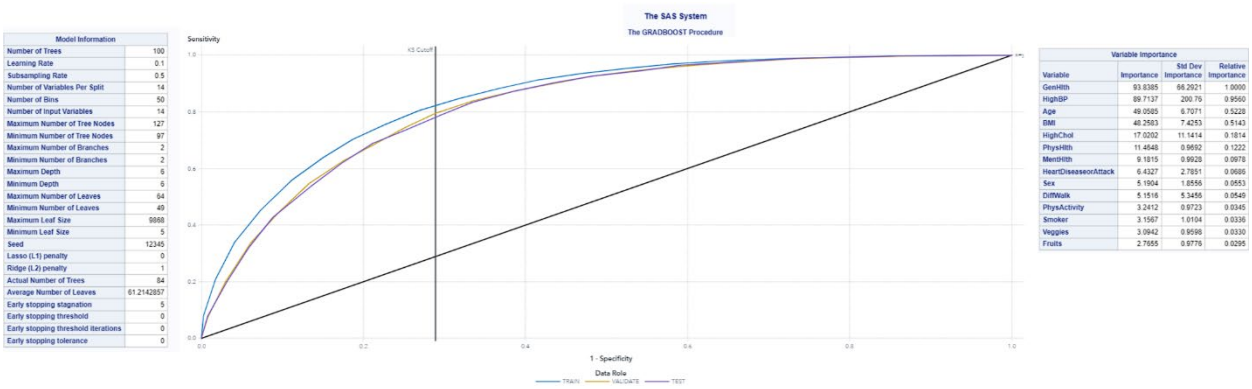

Figure 7: ROC curve with KS Cutoff for Gradient Boost model with its procedure and variable importance table.

CONCLUSION

In conclusion, the analysis showed that all models emphasized the significance of overall health rating, with high blood pressure being the next important variable. BMI and Age were found to be crucial in predicting diabetes, while the presence of high cholesterol only had a moderate impact. The Other variables were found to be of marginal importance, at best.

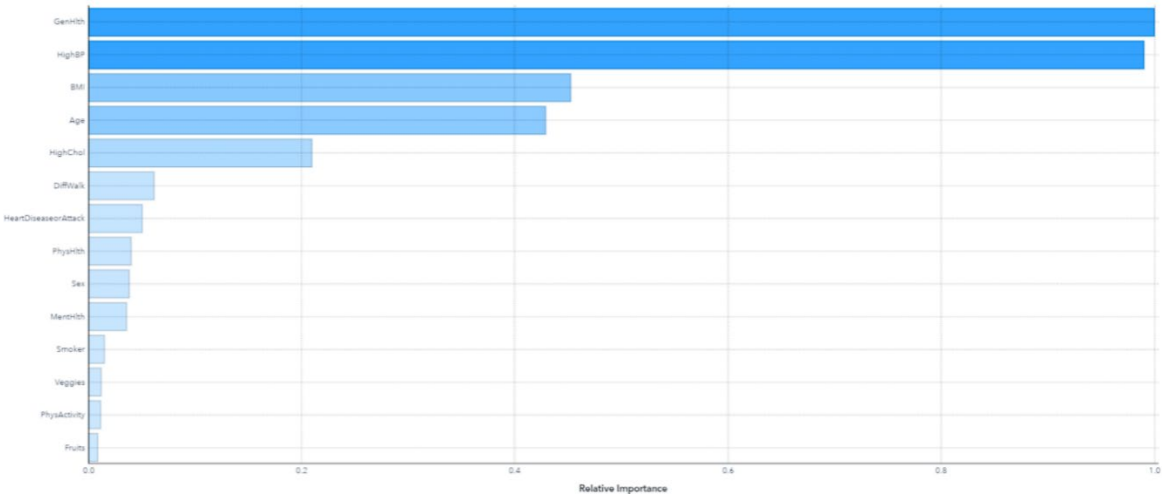

Figure 8: Variables listed in order of their importance.

Despite overall health being important, the study found that most controllable activities did not significantly affect the presence of diabetes. However, given the low benchmarks for these variables, such as consuming only one vegetable a day, the low impact was to be expected. Ultimately, the models did not provide any insights beyond what is already known – maining a healthy weight, being mindful of age, and keeping overall health high are key factors in preventing diabetes.

SUGGESTIONS FOR FUTURE STUDIES

Future diabetes studies for machine learning model development need to improve the questionnaire by incorporating important diet factors. They should also increase the sample size, balance variables like BMI and Mental Health, and utilize the full version of SAS Viya software to improve the model's performance through autotuning capabilities. Additionally, creating subsets of the data using specific variables could be useful in determining if different relationships could be found.

REFERENCES

Buttorff, C., Ruder, T., & Bauman, M. (2017). Multiple Chronic Conditions in the United States. Santa Monica, CA: *Rand Corporation*. https://www.rand.org/content/dam/rand/pubs/tools/TL200/TL221/RAND_TL221.pdf

Chuks, P. n.d. Diabetes, Hypertension and Stroke Prediction. Kaggle Accessed January 18, 2023. https://www.kaggle.com/datasets/prosperchuks/health-dataset

CDC. 2016. Behavioral Risk factor Surveillance System 2015 Codebook Report Accessed January 28, 2023. https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf

Gaurav. 2021. An introduction to gradient boosting decision trees. Accessed February 13, 2023. https://www.machinelearningplus.com/machine-learning/an-introduction-to-gradient-boosting-decision-trees/

Ward, Z. J., Bleich, S. N., Long, M. W., & Gortmaker, S. L. 2021. Association of body mass index with health care expenditures in the United States by age and sex. *PLOS ONE*. https://doi.org/10.1371/journal.pone.0247307

SAS Institute Inc. 2018. SAS® Studio 3.8: User's Guide. Cary, NC: SAS Institute Inc Accessed February 8, 2023. https://documentation.sas.com/doc/en/sasstudiocdc/3.8/webeditorcdc/webeditorug/titlepage.htm

APPENDIX

Variables	Variable Level	Variable Label	Variable Values	Diabetes (Target Variable) (0 = No) N = 35346	Diabetes (Target Variable) (1 = Yes) N = 35346	Total N =70692
BMI	Interval	Body Mass Index	mean	27.77	31.94	29.86
			Standard Deviation	6.19	7.36	7.11
PhysHlth	Interval	Days per month of Poor Physical Health	mean	3.04	4.46	3.75
			Standard Deviation	7.2	8.95	8.15
MentHlth	Interval	Days per months of Poor Mental Health	mean	3.67	7.95	5.81
			Standard Deviation	8.1	11.3	10.06
Age	Nominal	Age Group	1 = 18-24 Years	901 (2.55%)	78 (0.22%)	979 (1.38%)
			2 = 25-29 Years	1256 (3.55%)	140 (0.40%)	1396 (1.97%)
			3 = 30-34 Years	1735 (4.91%)	314 (0.89%)	2049 (2.90%)
			4 = 35-39 Years	2167 (6.13%)	626 (1.77%)	2793 (3.95%)
			5 = 40-44 Years	2469 (6.99%)	1051 (2.97%)	3520 (4.98%)
			6 = 45-49 Years	2906 (8.22%)	1742 (4.93%)	4648 (6.58%)
			7 = 50-54 Years	3784 (10.71%)	3088 (8.74%)	6872 (9.72%)
			8 = 55-59 Years	4340 (12.28%)	4263 (12.06%)	8603 (12.17%)
			9 = 60-64 Years	4379 (12.39%)	5733 (16.22%)	1E4 (14.30%)
			10 = 65-69 Years	4298 (12.16%)	6558 (18.55%)	11E3 (15.36%)
			11 = 70-74 Years	2903 (8.21%)	5141 (14.54%)	8044 (11.38%)
			12 = 75-79 Years	1991 (5.63%)	3403 (9.63%)	5394 (7.63%)
			13 = >=80 Years	2217 (6.27%)	3209 (9.08%)	5426 (7.68%)
CholCheck	Binary	Cholesterol Checked in 5 years?	0 = No	1508 (4.27%)	241 (0.68%)	1749 (2.47%)
			1 = Yes	33838 (95.73%)	35105 (99.32%)	68943 (97.53%)
Fruits	Binary	Consumes one Fruit per day	0 = No	12790 (36.19%)	14653 (41.46%)	27443 (38.82%)
			1 = Yes	22556 (63.81%)	20693 (58.54%)	43249 (61.18%)
HvyAlcoholConsump	Binary	Consumes Heavy Alcohol?	0 = No	33158 (93.81%)	34514 (97.65%)	67672 (95.73%)
			1 = Yes	2188 (6.19%)	832 (2.35%)	3020 (4.27%)
Veggies	Binary	Consumes one Vegetable per day	0 = No	6322 (17.89%)	8610 (24.36%)	14932 (21.12%)
			1 = Yes	29024 (82.11%)	26736 (75.64%)	55760 (78.88%)
DiffWalk	Binary	Difficulty Walking or Climbing Stairs?	0 = No	30601 (86.58%)	22225 (62.88%)	52826 (74.73%)
			1 = Yes	4745 (13.42%)	13121 (37.12%)	17866 (25.27%)
Stroke	Binary	Ever had Stroke?	0 = No	34219 (96.81%)	32078 (90.75%)	66297 (93.78%)
			1 = Yes	1127 (3.19%)	3268 (9.25%)	4395 (6.22%)
GenHlth	Nominal	General Health Status	1 = Excellent	7142 (20.21%)	1140 (3.23%)	8282 (11.72%)
			2 = Very Good	13491 (38.17%)	6381 (18.05%)	19872 (28.11%)
			3 = Good	9970 (28.21%)	13457 (38.07%)	23427 (33.14%)
			4 = Fair	3513 (9.94%)	9790 (27.70%)	13303 (18.82%)
			5 = Poor	1230 (3.48%)	4578 (12.95%)	5808 (8.22%)
HeartDiseaseorAttack	Binary	Had CHD or MI?	0 = No	32775 (92.73%)	27468 (77.71%)	60243 (85.22%)
			1 = Yes	2571 (7.27%)	7878 (22.29%)	10449 (14.78%)
HighBP	Binary	Have High Blood Pressure?	0 = No	22118 (62.58%)	8742 (24.73%)	30860 (43.65%)
			1 = Yes	13228 (37.42%)	26604 (75.27%)	39832 (56.35%)
HighChol	Binary	High Cholesterol Level?	0 = No	21869 (61.87%)	11660 (32.99%)	33529 (47.43%)
			1 = Yes	13477 (38.13%)	23686 (67.01%)	37163 (52.57%)
PhysActivity	Binary	Performs Physical Activities?	0 = No	7934 (22.45%)	13059 (36.95%)	20993 (29.70%)
			1 = Yes	27412 (77.55%)	22287 (63.05%)	49699 (70.30%)
Sex	Binary	Sex	0 = Female	19975 (56.51%)	18411 (52.09%)	38386 (54.30%)
			1 = Male	15371 (43.49%)	16935 (47.91%)	32306 (45.70%)
Smoker	Binary	Smoked 100 Cigarettes?	0 = No	20065 (56.77%)	17029 (48.18%)	37094 (52.47%)
			1 = Yes	15281 (43.23%)	18317 (51.82%)	33598 (47.53%)

Table 1: Summary statistics of all variables

Models	Parameters
Decision tree	Splitting options: Class target Criterion as Information gain ratio and Interval target criterion as Variance; Branches 2, Depth 10, Min leaf 2, Binning: 50 max, quantile method, Pruning: cost complexity method and selection: Automatic
Logistic regression	Binary targe link function: Complementary log-log; Nominal target link function: Generalized logit; Selection options: Stepwise; Optimization options: Newton-Raphson with ridging;
Gradient boost	Trees: 100 Max, learning rate 0.1, subsample rate 0.5, interval target distribution: Normal; Splitting: Max branches 2, Max depth 6, Min leaf 5, binning: Max 50, quantile method. Early stopping, Class target metric as Misclassification rate, early stopping method: stagnation 5
Random forest	Number of trees: 100; Voting method: probability; Splitting: branches 2, depth 20, class target: information gain ratio, interval: variance; Binning: 50 max, quantile method, in bag proportion: 0.6
Neural network	1 hidden layer, midrange standardization, number of neurons per hidden layer: 50, activation function: Tanh; Automatic Optimization, 300 maximum iterations, early stopping stagnation of 5
SVM	Linear kernel, Penalty: 1, tolerance: 0.000001, Max iterations: 25

Table 2: Description of parameter used during building models.

Model	Dataset	Accuracy	Sensitivity	Specificity	Precision	Recall	F1
Random Forest	Training	.8030	.7824	.8270	.8397	.7664	.8100
	Validation	.7470	.7269	.7710	.7912	.7028	.7577
	Test	.7423	.7236	.7644	.7842	.7004	.7526
Gradient Boosting	Training	.7544	.7348	.7776	.7962	.7126	.7642
	Validation	.7517	.7303	.7776	.7984	.7051	.7628
	Test	.7501	.7284	.7763	.7975	.7027	.7614
Neural Network	Training	.7401	.7290	.7523	.7642	.7159	.7462
	Validation	.7419	.7287	.7567	.7707	.7131	.7491
	Test	.7395	.7286	.7514	.7632	.7157	.7455
Logistic Regression	Training	.7452	.7339	.7577	.7695	.7210	.7513
	Validation	.7498	.7346	.7670	.7821	.7175	.7576
	Test	.7465	.7325	.7624	.7768	.7163	.7540
Decision Tree	Training	.7451	.7369	.7539	.7624	.7278	.7495
	Validation	.7391	.7289	.7502	.7613	.7169	.7448
	Test	.7372	.7292	.7458	.7547	.7197	.7417
SVM	Train	.7440	.7260	.7651	.7838	.7042	.7538
	Validation	.7484	.7270	.7744	.7957	.7012	.7598
	Test	.7443	.7243	.7681	.7887	.6999	.7551

Table 3: Fit statistics for all models

$$\begin{aligned} Y = 0.1951 &+ 0.0736BMI - 0.7472HighBP_0 - 2.1258GenHlth_1 - 1.445GenHlth_2 \\ &- .5654HighChol_0 - 1.2968Age_4 - 1.5066Age_3 - 1.9264Age_2 - 0.9725Age_5 \\ &- 0.7966Age_6 - 0.6792GenHlth_3 - 1.7764Age_1 - 0.5925Age_7 - 0.5045Age_8 \\ &- .3111HeartDiseaseorAttack_0 - .1871Sex_0 - 0.2056DiffWalk_0 \\ &+ 0.1079Veggies_0 - 0.0052PhysHlth - 0.1444GenHlth_4 - 0.0917Age_{10} \\ &+ 0.0932Age_{11} + 0.0723Age\_12 \end{aligned}$$

$$Logistic\ Regression = {}^{1}/_{1 + e^{-(Y)}}$$

Formula: Logistic regression model