American Journal of Public Health Research, 2025, Vol. 13, No. 3, 90-102
Available online at http://pubs.sciepub.com/ajphr/13/3/1
Published by Science and Education Publishing
DOI:10.12691/ajphr-13-3-1

Using Statistical Machine Learning to Find Complex Interactions and Important CVD Risk Factors When Predicting General Health in Adults

Peter D. Hart[1,2,*]

[1]Health Promotion Research, Havre, Montana, USA
[2]Kinesmetrics Lab, Tallahassee, Florida, USA
*Corresponding author: pdhart@outlook.com

Received March 28, 2025; Revised April 30, 2025; Accepted May 07, 2025

Abstract **Background**: Cardiovascular disease (CVD) is the leading cause of premature mortality among U.S. adults. Many risk factors for CVD are established and widely used in health promotion and preventive medicine. However, the extent to which the major CVD risk factors interrelate in relation to health outcomes is less understood. The purpose of this study was to use statistical machine learning to identify complex interactions and important variables when predicting general health (GH) with CVD risk factors. **Methods**: The analysis plan included five objectives. First, a decision (regression) tree was built on training data and fine-tuned using validation data. Second, ordinary least squares (OLS) regression was used to confirm terminal splits provided by the decision tree algorithm. Third, new test data were used to evaluate generalization of the decision tree branches. Fourth, a random forest was run and examined for consistency with decision tree fit performance using training, validation, and test data. Fifth, CVD risk factor variable importance was assessed along with a sensitivity analysis to examine stability in rankings. The 2017-2018 NHANES (N = 3,487) was used for training and validation and 2015-2016 NHANES (N = 3,897) for testing. A residualized self-assessed GH T-score with age, race/ethnicity, sex, and income removed, served as the outcome variable (aka., target). Eight CVD risk factors inspired by Life's Essential 8 (LE8) were used as predictors (aka., features or inputs) and included healthy eating index (HEI; 0-100), moderate-to-vigorous-physical activity (MVPA; min/week), nicotine exposure (NE; non-smoker, quit smoker, other nicotine device user, smoker), sleep time (ST; hr/day), body mass index (BMI; kg/m^2), non-high density lipoprotein cholesterol (NHDL; mg/dL), glycohemoglobin (A1C; %), and mean arterial pressure (MAP; mmHg). SAS HPSPLIT and HPFOREST were the primary reporting procedures. The variable importance sensitivity analysis was performed using R (train and randomForest) and Python (DecisionTreeRegressor and RandomForestRegressor). **Results**: The decision tree built on training data and 10-fold cross validation resulted in a 16-leaf tree with a 6-node depth (ASE $_{\text{Training}}$ = 86.3, ASE $_{\text{Validation}}$ = 91.1, Δ = 5.6%). BMI split first with A1C splitting next for high BMI (BMI $\geq$ 30.1) and MVPA splitting next for low BMI (BMI < 30.1). OLS regression confirmed (ps < .05) all terminal splits in the training data. Greatest GH (Mean = 56.2) was observed in those with low BMI (BMI < 30.1), high MVPA (MVPA $\geq$ 137.2 min/day), low NE (non-smoker, quit smoker, other nicotine device user), low A1C (A1C < 6.2%), and high HEI (HEI $\geq$ 37.8). Lowest GH (Mean = 43.8) was observed in those with high BMI (BMI $\geq$ 30.1) and high A1C (A1C $\geq$ 6.8%). The decision tree generalized well (ASE $_{\text{Test}}$ = 93.1, Δ = 8.0%) with OLS regression confirming (ps < .05) majority of terminal splits. Decision tree variable importance rankings were consistent with the random forest (r $_{\text{Spearman}}$ = .83, p = .011) and robust against the sensitivity analysis (avg r $_{\text{Spearman}}$ = .84, p = .009, ICC(3,6) = 0.97). **Conclusion**: This study demonstrated a novel use of machine learning that complements conventional statistical analyses. Decision trees along with random forests can identify extremely complex patterns in data and identify variables that contribute the most to group separation of an outcome variable. BMI, MVPA, A1C, and NE are likely the more important predictors of GH in this population.

Keywords: *Data science, Machine learning, Cardiovascular disease (CVD), Population health*

Cite This Article: Peter D. Hart, "Using Statistical Machine Learning to Find Complex Interactions and Important CVD Risk Factors When Predicting General Health in Adults." *American Journal of Public Health Research*, vol. 13, no. 3 (2025): 90-102. doi: 10.12691/ajphr-13-3-1.

1. Introduction

Heart disease has been the leading cause of mortality in the U.S. since 1919, accounted for over 680,000 deaths in 2023 alone [1,2]. Cardiovascular disease (CVD) mortality includes deaths from coronary heart disease, stroke, hypertension, heart failure, diseases of the arteries, and other minor CVD conditions and was responsible for over 940,000 deaths in 2022 [3]. The negative impacts associated with CVD extends beyond loss of life to include economic and social burdens. For example, in 2020, CVD was responsible for an estimated $393 billion in direct health care costs and $627 billion when additionally considering loss of productivity from related morbidity and mortality [4]. Furthermore, in 2019, CVD in the U.S. accounted for 4,398.1 years of life lost (YLL), 737.8 years lived with disability (YLD), and 5,135.9 disability-adjusted life years (DALYs) per 100,000 population [5]. Given these alarming statistics, it is not surprising that CVD is also linked to declines in health - related quality of life [6,7].

Health promotion and preventive medicine efforts regarding CVD are most often directed toward improving associated risk factors. These risk factors can be addressed individually with medical and/or behavioral intervention or in the aggregate using an overall CVD risk scoring algorithm. The American Heart Association (AHA) has designed such an algorithm using eight individual cardiovascular health (CVH) metrics. The Life's Essential 8 (LE8) is a composite CVH score based on individual subscores of physical activity (PA), diet quality, sleep health, nicotine use and exposure, body mass index (BMI), blood lipids, blood glucose, and blood pressure [8]. The LE8 metric ranges from zero to 100 where higher scores represent better CVH. The Healthy People 2030 program uses this CVH metric with an objective currently set to increase the national mean score in adults from 65.5 (2017-2020) to 72.2 by 2030 [9].

Making CVH a national priority helps researchers and public health professionals justify programs directed toward improving CVD risk factors. Incorporating the LE8 algorithm also allows for useful outcome measures easily linked to program evaluation [10,11,12,13]. CVH metrics can also be as predictors for other health-related outcomes. Two important participant-reported outcome measures that have grown in popularity are self-rated general health (GH) and health-related quality of life (HRQOL) [14,15]. Not surprisingly, LE8 has shown to be strongly correlate with both GH and HRQOL in adult populations [16]. However, the extent to which CVD risk factors interact in relation to participant-reported health outcomes is less understood. In addition, advances in data science have allowed for this type of deeper examination of structure otherwise hidden in health-related datasets. Therefore, the first aim of this study was to use statistical machine learning to identify complex interactions when predicting GH with CVD risk variables. The second aim of this study was to identify the most important CVD risk factors associated with GH.

2. Methods

Study design

The methods used in this study were the result of combining applied statistics with machine learning techniques [17]. Therefore, some terminology should be outlined before describing the design. Outcome variables or dependent variables are often termed *targets* or *outputs* in machine learning nomenclature while predictors or independent variables are often termed *features* or *inputs* [18]. This study will use these terms interchangeably. It is also often the case that machine learning procedures use multiple datasets. Specifically, a *training dataset* is used to build different models, a *validation dataset* is used to fine-tune and select a best-case model, and a *test dataset* is used to make a final evaluation of a model's performance [18].

Statistical machine learning was defined in this study as the application of statistical procedures with the purpose of discovering patterns, relationships, and structure in data to address a specific research problem. *Statistical machine learning* was also defined as a set of procedures that incorporate training, validation, and test datasets to iteratively learn from data, evaluate models, and ultimately improve generalizations. Furthermore, since this study included a measured outcome variable in the training dataset (i.e., supervised machine learning), which can be predicted using new or future datasets, we also considered this machine learning effort *statistical predictive modeling* [18]. The term *statistical* was intentionally added to the above terms to denote 1) the use of applied statistics, 2) the use of statistical criteria to make decisions, as well as 3) the use of researcher judgement in the decision-making process.

Thus, to apply statistical machine learning techniques, the current study employed training, validation, and test datasets using the same set of inputs to predict the same single target. The 2017-2018 National Health and Nutrition Examination Survey (NHANES) consisting of N = 3,487 adults 20 years of age and older was used for both training and validation datasets. Whereas the 2015-2016 NHANES consisting of N = 3,897 same aged adults was used as the independent test dataset. Both NHANES cycles contained the same variables collected using survey questions, physical examinations, and clinical laboratory tests [19,20]. NHANES study procedures have been approved by the National Center for Health Statistics (NCHS) Ethics Review Board (ERB) and all participants provided consent.

Self-assessed general health (GH)

A residualized self-assessed GH T-score variable with age, sex, race, and income removed, served as the outcome variable (aka., target). The initial self-assessed GH variable was obtained from a single question asking participants how they would rate their general health. The response options for this question included "excellent" (5), "very good" (4), "good" (3), "fair" (2), and "poor" (1) and coded numerically as shown. The GH variable was then residualized by regressing GH onto the participant demographic variables of age, sex, race/ethnicity, and

income. The residual from that regression was output to the dataset and then converted to a T-score (i.e., Mean = 50 and SD = 10) to aid interpretation. GH was also used as a categorical variable where those responding “excellent” or “very good” were placed in a group and those responding “good”, “fair” or “poor” placed in the other group.

Cardiovascular disease (CVD) risk factor variables

Eight CVD risk factors inspired by the LE8 metrics served as predictors (aka., features or inputs) and included healthy eating index (HEI), moderate-to-vigorous-physical activity (MVPA), nicotine exposure (NE), sleep time (ST), body mass index (BMI), non-high density lipoprotein cholesterol (NHDL), glycohemoglobin (A1C), and mean arterial pressure (MAP). Raw risk factor values, in original units, were used instead of LE8 metric values and will be briefly explained.

HEI (0 to 100) was assessed by computing Healthy Eating Index (HEI-2015) individual scores using two days of dietary interview recall and a macro program provided by the National Cancer Institute (NCI). MVPA (min/week) was assessed using survey questions regarding both moderate (MPA) and vigorous PA (VPA). MVPA was computed by adding weekly minutes of MPA with two times weekly minutes of VPA (i.e., MVPA = MPA + 2 × VPA). NE (N, Q, O, S) was assessed by categorizing participants into one of four groups: 1 = 'N' (non-smoker), 2 = 'Q' (quit smoker), 3 = 'O' (other nicotine device user), and 4 = 'S' (cigarette smoker). NE (1 to 4) was also coded as an ordinal-level variable of increasing exposure, as shown. ST (hours/day) was assessed using a survey question asking participants how much sleep they usually get at night on weekdays or workdays. BMI (kg/m^2) was assessed by dividing participant weight in kilograms by their height in meters squared. NHDL (mg/dL) was computed by subtracting participant HDL cholesterol from their total cholesterol using laboratory data. A1C (%) was assessed directly from laboratory data. MAP (mmHg) was assessed by first computing average systolic BP (SBP) and average diastolic BP (DBP) variables from examination data and then computing (SBP + 2 × DBP) / 3.

Covariates

Demographic variables were used both to describe sample participants as well as for statistical adjustment. Age was used as a numeric variable (20 years to 80+ years) as well as a categorical variable (20 to 24 years, 25 to 34 years, 35 to 44 years, 45 to 54 years, 55 to 64 years, and 65+ years). Sex included male and female categories. Race/ethnicity groups included White, Black, Hispanic, and Other. Finally, income was assessed and presented numerically as a ratio of family income to poverty (0 to 5). Income was also categorized into quartiles where larger quartiles contained families with relatively greater income.

Statistical analyses

Sample characteristics were described by GH groups using percentages, 95% confidence intervals (CIs), and the chi-square test of independence. The Cochran-Armitage trend test was also used on demographic variables with order. All study variables were described using means, standard deviations (SDs), and 95% CIs and compared across GH groups using t tests. Additionally, both Pearson and Spearman correlation coefficients were computed to describe the bivariate association between GH and each CVD risk factor variable. The Spearman correlation was added due to minor skewness as well as slight departure from linearity in some predictor variables. As noted above, the original GH was residualized by removing the confounding demographic effects of age, sex, race/ethnicity, and income. The purpose of residualizing GH was to 1) simplify the interpretation of the decision tree by making demographic predictor variables unnecessary and 2) transform GH into a continuous variable with an approximate normal distribution. GH was residualized by regressing GH onto age, sex, race/ethnicity, and income. The residualized GH was then converted to a T-score with mean of 50 and SD of 10.

The primary analysis plan included five objectives. First, a decision (regression) tree was built on training data and fine-tuned using validation data from a 10-fold cross-validation procedure. A decision tree approach was selected because it 1) can analyze a continuous outcome variable using both continuous and categorical predictors, 2) models using a series of if-then statements proving an intuitive interpretation, 3) does not require strict assumptions shared by parametric models, and 4) can find interactions between predictor variables that may otherwise be hidden [21]. The decision tree was grown using residual sum of squares (RSS) as split criteria (aka., ANOVA criterion). That is, a specific predictor is selected with a specific split value that minimizes variability the most in the outcome variable within the newly split nodes. The splitting continues until a full tree is achieved. However, to prevent overfitting, the decision tree was then pruned using the cost-complexity method (CCM). The CCM selects the smallest tree that is within one standard error (1-SE) of the minimum cross-validation average square error (ASE) (aka., 1-SE rule) [22].

Second, ordinary least squares (OLS) regression was used to confirm the terminal splits (i.e., complex interactions) provided by the decision tree algorithm. Since decision tree terminal nodes (aka., leaves) are easily represented by if-then statements, leaf paths, and thus interactions, can be confirmed using the same statements in a regression model. Third, new test data were used to evaluate the generalizability of the decision tree branches. That is, OLS regression from above was repeated but with the independent test dataset to examine the reliability of the complex interactions found during training.

Fourth, a random forest was run to examine consistency with decision tree fit performance using training, validation, and test data. A random forest uses decision tree methods but results in a predictive model averaged from several (hundreds) independent trees [23]. Each tree created in a random forest is built using only a sample of the training data and only a sample of the input variables. Thus, a random forest predictive model is less likely to overfit training data and can result in lower bias than a single decision tree. A 60-to-40 training-to-validation split was employed for the random forest using NHANES 2017-2018 data. Additionally, several fit statistics were used for comparison and included ASE for assessing model bias, percent (%) change in ASE for assessing model variance, and correlations between observed and model predicted values with mean comparisons for additional bias assessment.

Fifth, CVD risk factor variable importance was assessed with a sensitivity analysis implemented to

examine stability in the rankings. The variable importance analysis was conducted using four different hierarchical approaches: 1) a Spearman correlation was computed between the primary decision tree and primary random forest variable importance rankings, 2) an average Spearman correlation was computed between the primary decision tree variable importance rankings and rankings from the other five statistical methods (i.e., N = 5), 3) an average Spearman correlation was computed between all pairs of variable importance rankings across the six statistical methods (i.e., N = 15), and 4) an average variable importance rank was computed for each CVD risk factor for its rankings given by the six statistical methods (i.e., N = 6). SAS HPSPLIT and HPFOREST were the primary statistical procedures used in the study [24,25]. Additionally, the sensitivity analysis used R (train and randomForest) and Python (DecisionTreeRegressor and RandomForestRegressor) functions [26,27,28,29,30].

3. Results

A total of N = 3,487 adults with complete data from the NHANES 2017-2018 were included in the training dataset (Table 1). Approximately 35% of adults self-rated their GH as "Very good" or better. More younger adults than older adults (p for trend = .002) and more with greater income than with less income (p for trend < .001) rated their GH as "Very good" or better. Racial disparities were also observed in terms of GH (p < .001). A total of N = 3,897 adults from the NHANES 2015-2016 were included in the test dataset and had similar characteristics to the training data (Table S1).

The residualized GH outcome variable standardized to T-score units approximated a normal distribution with mean of 50.0 and standard deviation (SD) of 10 (Figure 1). A similar distribution was seen for GH in the test dataset (Figure S1). As expected, GH was significantly greater (p < .001) for those with "Excellent" or "Very good" GH (Mean = 60.9) compared to those with "Good," "Fair," or "Poor" GH (Mean = 44.3) (Table 2). Additionally, all CVD risk factor variables were superior for adults with "Excellent" or "Very good" GH, compared to their counterparts. Specifically, HEI, MVPA, and ST variables were greater (ps < .05) among those with better GH while NE, BMI, NHDL, A1C, and MAP variables were lower (ps < .05). Similar findings were observed, less ST (p = .674), in the test dataset (Table S2). The bivariate correlations indicated significant (ps < .01) associations between GH and all CVD risk factor variables. Specifically, positive associations were observed between GH and HEI, MVPA, and ST and negative correlations between GH and NE, BMI, NHDL, A1C, and MAP. Once more, similar findings were observed, less ST (p = .517), with the test dataset (Table S3).

The decision tree analysis with 10-fold cross validation resulted in a 16-leaf tree with a 6-node depth ($ASE_{Training}$ = 86.3, $ASE_{Validation}$ = 91.1, Δ = 5.6%) (Figure 2). BMI split first with A1C splitting next for high BMI (BMI ≥ 30.1) and MVPA splitting next for low BMI (BMI < 30.1) (Figure 3). NE was the next important predictor of GH among those with low BMI whereas BMI was again the next important predictor among those with high BMI (Figure 3). GH was greatest (Mean = 56.2) in those with low BMI (BMI < 30.1), high MVPA (MVPA ≥ 137.2 min/week), low NE (non-smoker, quit smoker, other nicotine device user), low A1C (A1C < 6.2%), and high HEI (HEI ≥ 37.8) (Table 4). GH was lowest (Mean = 43.8) in those with high BMI (BMI ≥ 30.1) and high A1C (A1C ≥ 6.8%) (Table 4). OLS regression confirmed (ps < .05) all interactions in the training data (Table 5). The decision tree generalized well using test data (ASE_{Test} = 93.1, Δ = 8.0%) with OLS regression confirming (ps < .05) majority of training data terminal splits (Table 5).

GH predictions on test data using random forest (ASE = 89.9, Mean GH = 50.2, r = .32) were consistent with that of the decision tree (ASE = 93.1, Mean GH = 50.1, r = .28) (Table 6). Decision tree CVD variable rankings were considered consistent with the random forest ($r_{Spearman}$ = .83, p = .011) (Table 7). The decision tree variable rankings were also considered consistent with the other five machine learning methods (avg $r_{Spearman}$ = .88, p = .004) (Figure 4). Finally, average rank of CVD risk factor variable importance across all six machine learning methods were considered reliable (avg $r_{Spearman}$ = .84, p = .009, ICC(3,6) = 0.97) with BMI, MVPA, A1C, and NE ranking as the top predictors of GH (Figure 4) (Figure 5).

Table 1. Sample characteristics by general health (GH) status, NHANES 2017-2018

		Excellent or Very good GH			Good, Fair, or Poor GH			
Characteristic	*N*	*%*	*LL*	*UL*	%	*LL*	*UL*	*p*
Overall	3,487	34.5	32.9	36.0	65.5	64.0	67.1	<.001
Sex								.551
Female	1,762	34.0	31.8	36.2	66.0	63.8	68.2	
Male	1,725	35.0	32.7	37.2	65.0	62.8	67.3	
Age Group (yr)								<.001
20 to 24	246	40.2	34.1	46.4	59.8	53.6	65.9	
25 to 34	547	41.3	37.2	45.4	58.7	54.6	62.8	
35 to 44	529	33.6	29.6	37.7	66.4	62.3	70.4	
45 to 54	543	29.3	25.5	33.1	70.7	66.9	74.5	
55 to 64	762	33.2	29.9	36.5	66.8	63.5	70.1	
65+	860	33.4	30.2	36.5	66.6	63.5	69.8	
Trend								.002

		Excellent or Very good GH			Good, Fair, or Poor GH			
Characteristic	*N*	*%*	*LL*	*UL*	%	*LL*	*UL*	*p*
Race/Ethnicity								<.001
White	1,352	40.8	38.2	43.4	59.2	56.6	61.8	
Black	764	29.7	26.5	33.0	70.3	67.0	73.5	
Hispanic	742	24.8	21.7	27.9	75.2	72.1	78.3	
Other	629	38.0	34.2	41.8	62.0	58.2	65.8	
Income Quartile								<.001
Q1 (low)	861	21.4	18.6	24.1	78.6	75.9	81.4	
Q2 (low-middle)	885	28.6	25.6	31.6	71.4	68.4	74.4	
Q3 (high-middle)	874	38.1	34.9	41.3	61.9	58.7	65.1	
Q4 (high)	867	49.8	46.5	53.2	50.2	46.8	53.5	
Trend								<.001

Note. General health (GH) is overall self-rated general health. Group p values are for χ^2 test of independence testing for association between each characteristic and GH groups. Trend p values are for Cochran-Armitage trend test.

Table S1. (Supplement). Sample characteristics by general health (GH) status, NHANES 2015-2016

		Excellent or Very good GH			Good, Fair, or Poor GH			
Characteristic	*N*	*%*	*LL*	*UL*	%	*LL*	*UL*	*p*
Overall	3,897	33.8	32.3	35.3	66.2	64.7	67.7	<.001
Sex								.785
Female	1,987	33.6	31.5	35.7	66.4	64.3	68.5	
Male	1,910	34.0	31.9	36.2	66.0	63.8	68.1	
Age Group (yr)								<.001
20 to 24	298	42.3	36.7	47.9	57.7	52.1	63.3	
25 to 34	719	41.3	37.7	44.9	58.7	55.1	62.3	
35 to 44	651	32.6	29.0	36.2	67.4	63.8	71.0	
45 to 54	666	32.7	29.2	36.3	67.3	63.7	70.8	
55 to 64	697	25.5	22.3	28.8	74.5	71.2	77.7	
65+	866	33.1	30.0	36.3	66.9	63.7	70.0	
Trend								<.001
Race/Ethnicity								<.001
White	1,408	43.1	40.5	45.7	56.9	54.3	59.5	
Black	767	31.7	28.4	35.0	68.3	65.0	71.6	
Hispanic	1,206	20.7	18.4	23.0	79.3	77.0	81.6	
Other	516	42.2	38.0	46.5	57.8	53.5	62.0	
Income Quartile								<.001
Q1 (low)	975	21.4	18.9	24.0	78.6	76.0	81.1	
Q2 (low-middle)	973	28.6	25.7	31.4	71.4	68.6	74.3	
Q3 (high-middle)	976	38.2	35.2	41.3	61.8	58.7	64.8	
Q4 (high)	973	47.1	43.9	50.2	52.9	49.8	56.1	
Trend								<.001

Note. General health (GH) is overall self-rated general health. Group p values are for χ^2 test of independence testing for association between each characteristic and GH groups. Trend p values are for Cochran-Armitage trend test.

Table 2. Cardiovascular disease (CVD) variables and covariates by general health (GH) status, NHANES 2017-2018

	Excellent or Very good GH (N=1,202)				Good, Fair, or Poor GH (N=2,285)				
Variables	*Mean*	*SD*	*LL*	*UL*	*Mean*	*SD*	*LL*	*UL*	*p*
GH (adj T-score)	60.91	5.91	60.58	61.25	44.26	6.21	44.00	44.51	<.001
HEI (0 - 100)	55.63	14.13	54.83	56.43	51.56	13.83	50.99	52.12	<.001
MVPA (min/week)	317.32	498.36	289.12	345.52	155.11	337.89	141.25	168.98	<.001
NE (1 - 4)	1.69	1.04	1.63	1.75	1.99	1.19	1.94	2.03	<.001
ST (hrs/day)	7.79	1.32	7.71	7.86	7.68	1.54	7.62	7.74	.031
BMI (kg/m^2)	27.52	5.66	27.20	27.84	31.06	7.50	30.76	31.37	<.001
NHDL (mg/dL)	134.17	39.66	131.93	136.42	137.08	43.00	135.31	138.84	.046
A1C (%)	5.61	0.78	5.57	5.66	5.97	1.21	5.92	6.02	<.001
MAP (mmHg)	89.69	11.60	89.04	90.35	92.45	12.91	91.92	92.98	<.001
Age (yr)	49.74	18.09	48.72	50.77	51.51	16.82	50.82	52.20	.005

	Excellent or Very good GH (N=1,202)				Good, Fair, or Poor GH (N=2,285)				
Variables	*Mean*	*SD*	*LL*	*UL*	*Mean*	*SD*	*LL*	*UL*	*p*
Sex (1=Male, 0=Female)	0.50	0.50	0.47	0.53	0.49	0.50	0.47	0.51	.551
Income (0 - 5)	3.14	1.62	3.05	3.23	2.35	1.55	2.28	2.41	<.001
White (1=White, 0=Other)	0.46	0.50	0.43	0.49	0.35	0.48	0.33	0.37	<.001

Note. N = 3,487. GH is general health T-score adjusted for age, race, sex, and income. HEI is healthy eating index (HEI-2015). MVPA is moderate-to-vigorous-physical activity. NE is nicotine exposure where 1 = 'N' (non-smoker), 2 = 'Q' (quit smoker), 3 = 'O' (other nicotine device user), and 4 = 'S' (cigarette smoker). ST is sleep time. BMI is body mass index. NHDL is non-high density lipoprotein cholesterol. A1C is glycohemoglobin. MAP is mean arterial pressure. t statistic p-values are testing for mean difference between GH groups.

Table 3. Correlations between general health (GH) and study variables, NHANES 2017-2018

Variable	*Mean*	*SD*	*LL*	*UL*	r_P	*p*	r_S	*p*
GH (adj T-score)	50.00	10.00	49.67	50.33	-	-	-	-
HEI (0 - 100)	52.96	14.06	52.49	53.43	.138	<.001	.139	<.001
MVPA (min/week)	211.03	407.82	197.49	224.57	.162	<.001	.188	<.001
NE (1 - 4)	1.88	1.15	1.84	1.92	-.115	<.001	-.118	<.001
ST (hrs/day)	7.72	1.47	7.67	7.77	.068	<.001	.059	.001
BMI (kg/m^2)	29.84	7.12	29.60	30.08	-.249	<.001	-.250	<.001
NHDL (mg/dL)	136.07	41.89	134.68	137.47	-.053	.002	-.052	.002
A1C (%)	5.85	1.09	5.81	5.89	-.161	<.001	-.138	<.001
MAP (mmHg)	91.50	12.54	91.09	91.92	-.088	<.001	-.101	<.001
Age (yr)	50.90	17.29	50.32	51.47	.000	1.000	.009	.588
Sex (1=Male, 0=Female)	0.49	0.50	0.48	0.51	.000	1.000	-.008	.643
Income (0 - 5)	2.62	1.61	2.57	2.67	.000	1.000	.015	.373
White (1=White, 0=Other)	0.39	0.49	0.37	0.40	.000	1.000	.012	.463

Note. N = 3,487. GH is general health T-score adjusted for age, race, sex, and income. HEI is healthy eating index (HEI-2015). MVPA is moderate-to-vigorous-physical activity. NE is nicotine exposure where 1 = 'N' (non-smoker), 2 = 'Q' (quit smoker), 3 = 'O' (other nicotine device user), and 4 = 'S' (cigarette smoker). ST is sleep time. BMI is body mass index. NHDL is non-high density lipoprotein cholesterol. A1C is glycohemoglobin. MAP is mean arterial pressure. r_P are Pearson correlations. r_S are Spearman correlations.

Table 4. Decision tree leaf paths predicting general health (GH) with cardiovascular disease (CVD) risk factor variables, NHANES 2017-2018

Leaf	*N*	Depth = 1	GH	Depth = 2	GH	Depth = 3	GH	Depth = 4	GH	Depth = 5	GH	Depth = 6	GH
	40	BMI <	52.	MVPA <		NE =	49.						
D	5	30.1	0	137.2	50.5	O,Q,S	1	A1C < 5.8	**50.0**				
	17	BMI <	52.	MVPA <		NE =	49.						
E	6	30.1	0	137.2	50.5	O,Q,S	1	A1C ≥ 5.8	**47.0**				
		BMI <	52.	MVPA <			51.						
G	30	30.1	0	137.2	50.5	NE = N	8	A1C ≥ 8.0	**46.9**				
		BMI <	52.	MVPA <			51.						
N	69	30.1	0	137.2	50.5	NE = N	8	A1C < 8.0	52.1	ST < 6.3	**48.8**		
	11	BMI <	52.	MVPA <			51.					NHDL <	
T	6	30.1	0	137.2	50.5	NE = N	8	A1C < 8.0	52.1	ST ≥ 6.3	52.5	101.01	**50.4**
	42	BMI <	52.	MVPA <			51.					NHDL ≥	
U	0	30.1	0	137.2	50.5	NE = N	8	A1C < 8.0	52.1	ST ≥ 6.3	52.5	101.01	**53.1**
	11	BMI <	52.	MVPA ≥			**48.**						
9	4	30.1	0	137.2	54.3	NE = S	**9**						
		BMI <	52.	MVPA ≥		NE =	55.						
I	72	30.1	0	137.2	54.3	N,O,Q	2	A1C ≥ 6.2	**50.1**				
		BMI <	52.	MVPA ≥		NE =	55.			HEI <			
P	60	30.1	0	137.2	54.3	N,O,Q	2	A1C < 6.2	55.7	37.8	**51.4**		
	56	BMI <	52.	MVPA ≥		NE =	55.			HEI ≥			
Q	9	30.1	0	137.2	54.3	N,O,Q	2	A1C < 6.2	55.7	37.8	**56.2**		
	22	BMI ≥	47.										
6	3	30.1	2	A1C ≥ 6.8	**43.8**								
	14	BMI ≥	47.			BMI <	48.	HEI <					
J	7	30.1	2	A1C < 6.8	47.8	36.0	8	39.5	**46.0**				
	14	BMI ≥	47.			BMI ≥	46.	HEI <					
L	7	30.1	2	A1C < 6.8	47.8	36.0	2	42.9	**44.5**				
	32	BMI ≥	47.			BMI ≥	46.	HEI ≥					
M	7	30.1	2	A1C < 6.8	47.8	36.0	2	42.9	**46.9**				
	38	BMI ≥	47.			BMI <	48.	HEI ≥		MAP <			
R	2	30.1	2	A1C < 6.8	47.8	36.0	8	39.5	49.4	94.5	**50.5**		
	23	BMI ≥	47.			BMI <	48.	HEI ≥		MAP ≥			
S	0	30.1	2	A1C < 6.8	47.8	36.0	8	39.5	49.4	94.5	**47.6**		

Note. N = 3,487. GH is general health T-score adjusted for age, race, sex, and income. HEI is healthy eating index (HEI-2015). MVPA is moderate-to-vigorous-physical activity. NE is nicotine exposure where 'N' = non-smoker, 'Q' = quit smoker, 'O' = other nicotine device user, and 'S' = cigarette smoker. ST is sleep time. BMI is body mass index. NHDL is non-high density lipoprotein cholesterol. A1C is glycohemoglobin. MAP is mean arterial pressure. GH values in bold are the terminal node (aka., leaf) means.

Table 5. Decision tree leaf/node differences using linear regression with training and test datasets

	Leaf and Leaf/Node comparison						
	Leaf/Node	*N*	*Mean*	Leaf/Node	*N*	*Mean*	*p*
Training data	D [a]	405	50.0	E [a]	176	47.0	<.001
(*N* = 3,487)	G	30	46.9	F	605	52.1	.003
	N	69	48.8	O	536	52.5	.003
	T [a]	116	50.4	U [a]	420	53.1	.008
	9	114	48.9	A	701	55.2	<.001
	I	72	50.1	H	629	55.7	<.001
	P [a]	60	51.4	Q [a]	569	56.2	<.001
	6	223	43.8	5	1233	47.8	<.001
	J	147	46.0	K	612	49.4	<.001
	L [a]	147	44.5	M [a]	327	46.9	.012
	R [a]	382	50.5	S [a]	230	47.6	<.001
Test data	D [a]	421	49.2	E [a]	238	47.5	.035
(*N* = 3,897)	G	28	43.2	F	695	51.9	<.001
	N	90	54.3	O	605	51.5	.011
	T [a]	129	50.8	U [a]	476	51.7	.318
	9	140	52.3	A	796	54.2	.024
	I	67	49.5	H	729	54.7	<.001
	P [a]	55	54.4	Q [a]	674	54.7	.819
	6	251	44.2	5	1328	48.2	<.001
	J	156	48.3	K	653	49.4	.187
	L [a]	159	46.6	M [a]	360	46.7	.961
	R [a]	167	50.0	S [a]	486	49.2	.316

Note. The [a] superscript indicates a leaf (aka., terminal node). An above node is a location on the decision tree in opposition to its leaf but is not terminal. p values are for group differences using ordinary least squares regression.

Table 6. Performance statistics for decision tree and random forest models predicting general health (GH)

	Dataset		
Statistic	Training	Validation	Test
Decision Tree			
N	3,487	3,487	3,897
ASE (bias)	86.27	91.12	93.14
% Δ (variance)	-	5.6	8.0
r	.370	.298	.279
p	<.001	<.001	<.001
Predicted GH Mean	50.00	-	50.10
p	.999	-	.499
Random Forest			
N	2,338	1,559	3,897
ASE (bias)	77.44	88.46	89.91
% Δ (variance)	-	14.2	16.1
r	.508	.339	.318
p	<.001	<.001	<.001
Predicted GH Mean	49.99	-	50.15
p	.992	-	.328

Note. ASE is average square error. Training dataset statistics are for NHANES 2017-2018 data. Validation dataset statistics are from the 10-fold cross-validation with decision tree and 40% hold-out from random forest NHANES 2017-2018 data. Test dataset statistics are for the NHANES 2015-2016 scored data. Decision tree cross-validation resulted in a 16-leaf tree with a 6-node depth. r values are Pearson correlations between observed and model predicted GH. Predicted GH mean p-values are for paired t-tests against observed GH.

Table 7. Variable rankings for decision tree and random forest models predicting general health (GH), NHANES 2017-2018

	Decision Tree		Random Forest	
Rank	Variable	RSS	Variable	MSE
1st	BMI	148.60	BMI	3.33
2nd	A1C	83.25	MVPA	0.84
3rd	MVPA	83.16	NE	0.42
4th	NE	77.91	A1C	0.11
5th	HEI	56.12	ST	-0.82
6th	MAP	33.88	HEI	-0.96
7th	ST	28.48	MAP	-1.02
8th	NHDL	25.66	NHDL	-1.23

Note. Random forest ranking uses the validation data change in mean square error (MSE). Decision tree importance statistic uses the change in residual sum of squares (RSS). The importance of a variable is proportional to the sum of the reduction in node impurity (RSS), summed over nodes that the variable splits. Spearman correlation between the two model variable rankings is $r = .83$, $p = .011$.

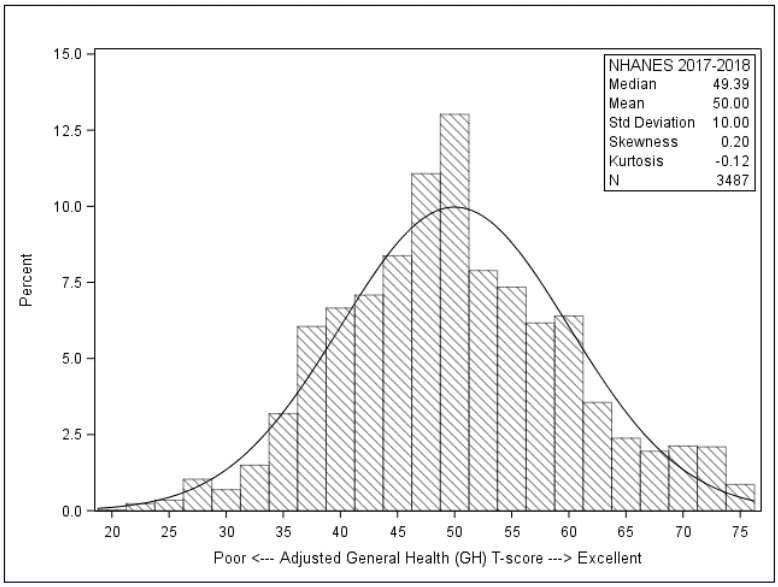


Figure 1. Distribution of new general health (GH) T-scores adjusted for age, sex, race, and income, NHANES 2017-2018

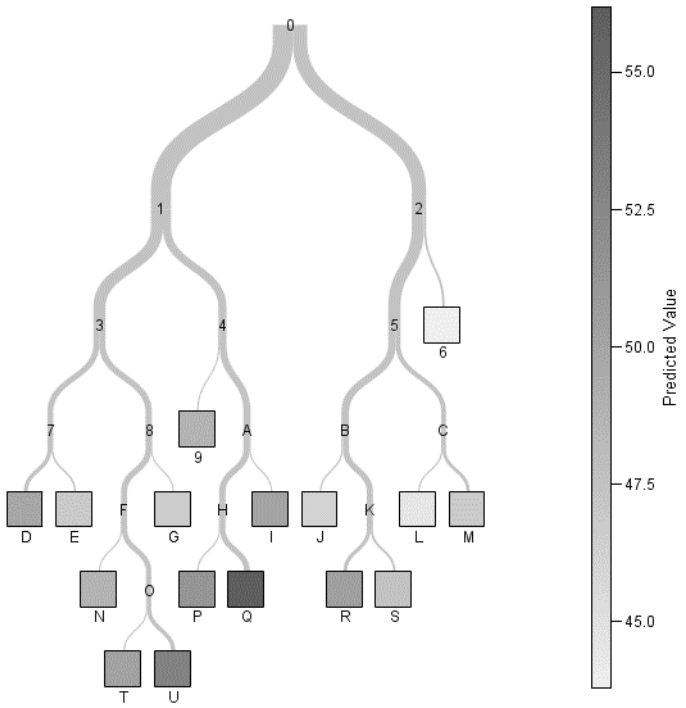


Figure 2. Final decision tree predicting general health (GH) with cardiovascular disease (CVD) variables, NHANES 2017-2018

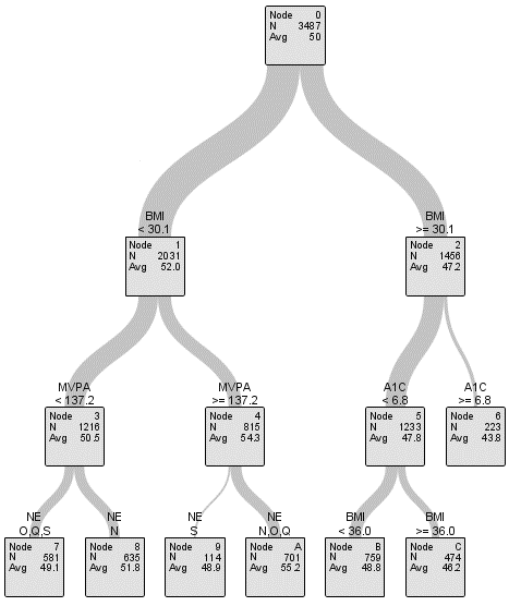
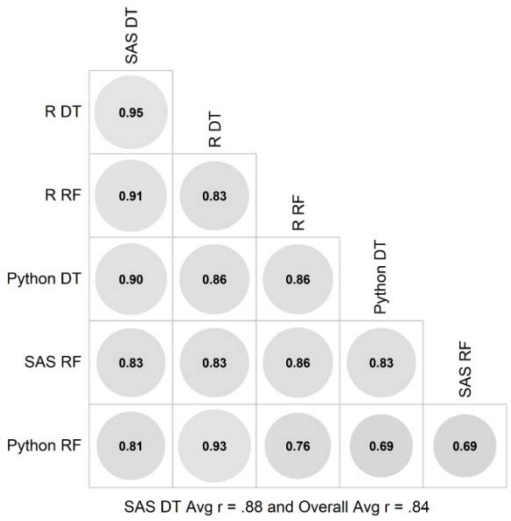


Figure 3. Zoomed top portion of the final decision tree predicting general health (GH) with cardiovascular disease (CVD) variables, NHANES 2017-2018



Note. DT represents decision tree. RF represents random forest. SAS DT Avg correlation (r) has an N = 5. Overall Avg correlation (r) has an N = 15.

Figure 4. Sensitivity analysis using Spearman correlations between variable importance ranking across software analyses, NHANES 2017-2018

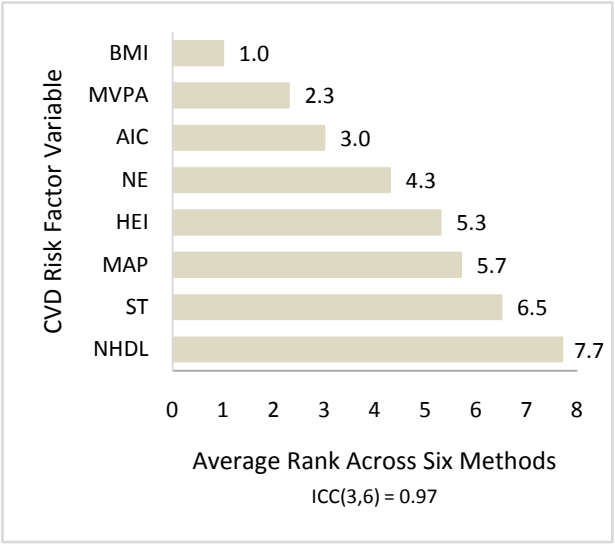


Figure 5. Average rank of CVD risk factor variable importance across all six machine learning methods, NHANES 2017-2018

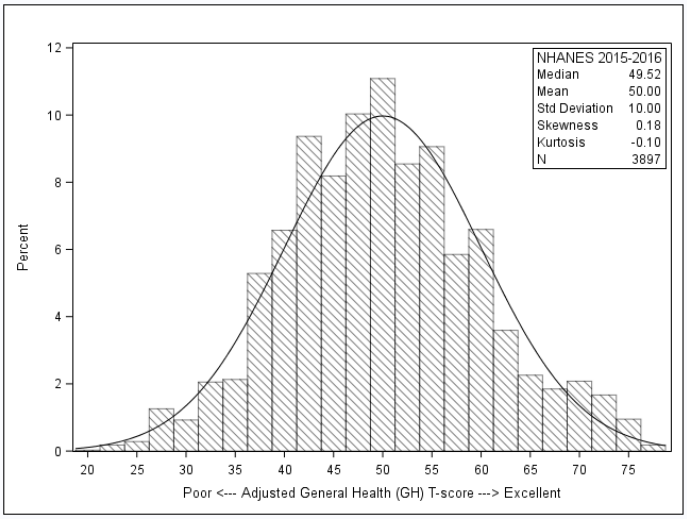


Figure S1. (Supplement). Distribution of new general health (GH) T-scores adjusted for age, sex, race, and income, NHANES 2015-2016

Table S2. (Supplement). Cardiovascular disease (CVD) variables and covariates by general health (GH) status, NHANES 2015-2016

	Excellent or Very good GH (N=1,318)				Good, Fair, or Poor GH (N=2,579)				
Variables	*Mean*	*SD*	*LL*	*UL*	*Mean*	*SD*	*LL*	*UL*	*p*
GH (adj T-score)	*60.88*	5.94	60.56	61.21	*44.44*	6.45	44.19	44.69	<.001
HEI (0 - 100)	54.95	14.02	54.19	55.71	51.95	13.45	51.43	52.47	<.001
MVPA (min/week)	322.34	478.34	296.50	348.19	155.75	333.53	142.87	168.63	<.001
NE (1 - 4)	1.74	1.08	1.68	1.79	1.97	1.20	1.93	2.02	<.001
ST (hrs/day)	7.72	1.37	7.64	7.79	7.70	1.58	7.63	7.76	.674
BMI (kg/m2)	27.46	5.80	27.15	27.78	30.87	7.34	30.59	31.16	<.001
NHDL (mg/dL)	132.17	40.04	130.01	134.34	139.04	43.74	137.35	140.73	<.001
A1C (%)	5.54	0.79	5.50	5.58	5.99	1.32	5.93	6.04	<.001
MAP (mmHg)	86.78	10.69	86.20	87.36	88.71	11.98	88.25	89.17	<.001
Age (yr)	47.28	17.97	46.31	48.25	49.88	16.84	49.23	50.53	<.001
Sex (1=Male, 0=Female)	0.49	0.50	0.47	0.52	0.49	0.50	0.47	0.51	.785
Income (0 - 5)	2.95	1.61	2.86	3.04	2.23	1.53	2.17	2.29	<.001
White (1=White, 0=Other)	0.46	0.50	0.43	0.49	0.31	0.46	0.29	0.33	<.001

Note. Note. N = 3,897. GH is general health T-score adjusted for age, race, sex, and income. HEI is healthy eating index (HEI-2015). MVPA is moderate-to-vigorous-physical activity. NE is nicotine exposure where 1 = 'N' (non-smoker), 2 = 'Q' (quit smoker), 3 = 'O' (other nicotine device user), and 4 = 'S' (cigarette smoker). ST is sleep time. BMI is body mass index. NHDL is non-high density lipoprotein cholesterol. A1C is glycohemoglobin. MAP is mean arterial pressure. t statistic p-values are testing for mean difference between GH groups.

Table S3. (Supplement). Correlations between general health (GH) and study variables, NHANES 2015-2016

Variable	*Mean*	*SD*	*LL*	*UL*	r_P	*p*	r_S	*p*
GH (adj T-score)	50.00	10.00	49.69	50.31	-	-	-	-
HEI (0 - 100)	52.96	13.71	52.53	53.40	.096	<.001	.098	<.001
MVPA (min/week)	212.09	396.45	199.64	224.54	.179	<.001	.198	<.001
NE (1 - 4)	1.89	1.16	1.86	1.93	-.130	<.001	-.124	<.001
ST (hrs/day)	7.70	1.51	7.66	7.75	.003	.836	.010	.517
BMI (kg/m^2)	29.72	7.04	29.50	29.94	-.241	<.001	-.241	<.001
NHDL (mg/dL)	136.72	42.64	135.38	138.06	-.062	<.001	-.058	<.001
A1C (%)	5.84	1.19	5.80	5.87	-.185	<.001	-.171	<.001
MAP (mmHg)	88.06	11.59	87.69	88.42	-.069	<.001	-.076	<.001
Age (yr)	49.00	17.28	48.46	49.54	.000	1.000	-.005	.753
Sex (1=Male, 0=Female)	0.49	0.50	0.47	0.51	.000	1.000	-.005	.766
Income (0 - 5)	2.47	1.60	2.42	2.52	.000	1.000	.021	.182
White (1=White, 0=Other)	0.36	0.48	0.35	0.38	.000	1.000	.011	.502

Note. N = 3,897. GH is general health T-score adjusted for age, race, sex, and income. HEI is healthy eating index (HEI-2015). MVPA is moderate-to-vigorous-physical activity. NE is nicotine exposure where 1 = 'N' (non-smoker), 2 = 'Q' (quit smoker), 3 = 'O' (other nicotine device user), and 4 = 'S' (cigarette smoker). ST is sleep time. BMI is body mass index. NHDL is non-high density lipoprotein cholesterol. A1C is glycohemoglobin. MAP is mean arterial pressure. r_P are Pearson correlations. r_S are Spearman correlations.

4. Discussion

The first aim of this research was to build a decision tree predicting GH using CVD risk factor variables. Examining each leaf in the tree in essence identified an interaction among the CVD predictors occurring at the previous node splits. For example, the simplest leaf in this study (i.e., leaf 6) was seen at a node depth of two and resulted from a BMI×A1C interaction. This interaction revealed a significant difference in GH between those with low A1C (GH Mean = 47.8 for A1C < 6.8) and those with high A1C (GH Mean = 43.8 for A1C ≥ 6.8) among those with high BMI (BMI ≥ 30.1). A total of eleven (11) terminal splits resulting in sixteen (16) leaves were observed in the final decision tree across a tree depth of six (6) nodes. The two most complicated leaves in the tree (i.e., leaf T and leaf U) resulted from a six-way BMI×MVPA×NE×A1C×ST×NHDL interaction. This interaction ultimately found a significant difference in GH between those with low NHDL (GH Mean = 50.4 for NHDL < 101.01) and those with high NHDL (GH Mean = 53.1 for NHDL ≥ 101.01). Identifying a six-variable interaction would be difficult without a decision tree but identifying each variable's split value would be nearly impossible. These interactions highlight the novel use of machine learning in this study.

Although this research discovered many complex associations between the predictors unlikely found otherwise, decision tree models are often criticized for their instability and tendency to overfit data [21]. These limitations, however, were addressed with the use of four (4) safeguards. First, the decision tree was pruned during a 10-fold cross-validation to ensure that the simplest tree was found that also provided an acceptable amount of bias. Second, an equivalent random forest was also employed and resulted in comparable fit statistics, predicted values, and variable importance rankings. Third, an independent test dataset was used to provide a final evaluation of the decision tree and indicated that the model experienced an acceptable amount of variance. Fourth, and final, OLS regression was used to confirm the complex interactions in both the training and test datasets. As expected, OLS regression confirmed all group differences in the training data. In the test data, majority of the group differences were confirmed. Even though not every interaction was confirmed in the test dataset, all non-significant group differences trended toward their training data difference. Thus, the complex interactions found from this statistical machine learning effort appeared to generalize well to new data.

The specific split values found using the decision tree algorithm are also worth discussing. The first variable used to split the data and maximize GH separation was BMI and its split point was 30.1. This decision is noteworthy not just because it likely contributed to the high ranking of BMI but also because its split value is almost exactly at the conventional cutoff value for obese classification of 30.0 [31]. The next two features used to split the data were MVPA and A1C and both ranked high in importance. Although slightly lower than the commonly used guideline, the MVPA cutoff of 137.2 is a close approximation to the 150 min/week [32]. Similarly, the cutoff decision used for A1C of 6.8 is also a close approximation to the 6.5 criteria used for diagnosing diabetes [33]. Interestingly, the next feature used to split the data was NE with a split of smokers versus all others. Although seeing a model direct smokers to a lower GH node makes sense, directing consumers of nicotine via other device modes to a higher GH node makes less sense. However, some evidence suggests that e-cigarette use may be associated with improved health status when compared to combustible cigarette use [34]. Finally, the feature selected to split opposite NE was once again BMI. This selection underscores the importance of BMI in predicting GH and provides face validity for its split value of 36.0 approximating the class II obese cutoff criteria of 35.0 [35].

The second aim of this research was to identify the most important CVD risk factors for predicting GH. Since decision tree variable importance rankings can also be volatile across different models, a sensitivity analysis was performed. The findings of which were considered robust with strong agreement in decision tree rankings to other machine leaning models. Results from the sensitivity analysis also found strong agreement in feature importance rankings across all six machine learning models. This resulted in stable rankings with BMI, MVPA, A1C, and NE placing first, second, third, and fourth, respectively, for their ability to predict GH. Few studies to date, if any, have examined CVD risk factors importance rankings using an ensemble approach when predicting GH. Research has been conducted on feature importance for other health outcomes using similar risk factors, however, rankings from these studies were reported either subjectively or reported only for the best fitting machine learning model [36,37,38].

A limitation regarding this study needing mention is its cross-sectional design. Cross-sectional studies do not allow for the criteria of temporality required to suggest a cause-and-effect relationship [39]. Therefore, inferences from this study should not be interpreted as CVD risk factors causing changes in GH. Another limitation is the use of a single ordinal-level GH item for assessing perceived health. Although GH was residualized and transformed to a continuous and approximate normal variable, and despite the item's reported validity, it still is crude in its ability to cover an entire perceived health trait [40,41]. A final limitation regarding this study is its use of participant-reported predictors of MVPA, ST, and NE. Although the self-reporting of health behaviors such as smoking, inactivity, and poor sleep can be criticized for bias, NHANES survey questions have an established reputation as being psychometrically sounds [42,43,44]. A strength regarding this study is its use of large samples needed for machine learning algorithms, its use of an independent test dataset for final model evaluations, and its use of objectively measured predictors of height and weight for BMI, blood pressure for MAP, and blood tests for NHDL and A1C.

5. Conclusions

This study combined machine learning techniques with applied statistics to predict GH using CVD risk factors as model features. The use of statistical machine learning allowed for the identification of several complex multi-

risk factor patterns that would not likely have been found using conventional methods. Additionally, these results identified variables that contributed the most to group separation of GH. BMI, MVPA, A1C, and NE were found to be the more important predictors of GH in this population. These findings can provide health professionals specialized segments of the population to target as well as provide the most influential CVD risk factors to focus on for improving quality of life in adults.

ACKNOWLEDGMENTS

The author appreciates the staff at the National Center for Health Statistics (NCHS) and National Health and Nutrition Examination Survey (NHANES) for allowing the public use of NHANES data.

Author Contributions

PDH performed all research, analysis, and writing related to this manuscript.

Disclosure Statement

The author reports that there are no competing interests to declare regarding this research.

Funding

This research was not supported by any grant funding.

ORCID

Peter D Hart: https://orcid.org/0000-0002-4947-9310

References

[1] Bastian B, Tejada Vera B, Arias E, et al. Mortality trends in the United States, 1900–2018. National Center for Health Statistics. 2020.

[2] Ahmad FB, Cisewski JA, Anderson RN. Mortality in the United States — Provisional Data, 2023. MMWR Morb Mortal Wkly Rep 2024; 73: 677–681.

[3] Martin SS, Aday AW, Allen NB, et al. 2025 Heart Disease and Stroke Statistics: A Report of US and Global Data From the American Heart Association. Circulation. Published online January 27, 2025.

[4] Kazi DS, Elkind MSV, Deutsch A, et al. Forecasting the Economic Burden of Cardiovascular Disease and Stroke in the United States Through 2050: A Presidential Advisory From the American Heart Association. Circulation. 2024; 150(4): e89-e101.

[5] Cardiovascular disease burden in the Region of the Americas, 2000-2019. ENLACE data portal. Pan American Health Organization. 2021.

[6] Lui JNM, Williams C, Keng MJ, et al. Impact of New Cardiovascular Events on Quality of Life and Hospital Costs in People With Cardiovascular Disease in the United Kingdom and United States. J Am Heart Assoc. 2023; 12(19): e030766.

[7] Allen NB, Badon S, Greenlund KJ, Huffman M, Hong Y, Lloyd-Jones DM. The association between cardiovascular health and health-related quality of life and health status measures among U.S. adults: a cross-sectional study of the National Health and Nutrition Examination Surveys, 2001-2010. Health Qual Life Outcomes. 2015; 13: 152. Published 2015 Sep 22.

[8] Lloyd-Jones DM, Allen NB, Anderson CAM, et al. Life's Essential 8: Updating and Enhancing the American Heart Association's Construct of Cardiovascular Health: A Presidential Advisory From the American Heart Association. Circulation. 2022; 146(5): e18-e43.

[9] Office of Disease Prevention and Health Promotion. (n.d.). Heart Disease and Stroke. Healthy People 2030. U.S. Department of Health and Human Services. https:// odphp.health.gov/ healthypeople/objectives-and-data/browse-objectives/heart-disease-and-stroke.

[10] Cuccia AF, DiPietro L, Hayman LL, Whiteley JA, Napolitano MA. Longitudinal Changes in Cardiovascular Health among Young Adults with Overweight and Obesity. J Cardiovasc Nurs. Published online December 31, 2024.

[11] Brewer LC, Jenkins S, Hayes SN, et al. Community-Based, Cluster-Randomized Pilot Trial of a Cardiovascular Mobile Health Intervention: Preliminary Findings of the FAITH! Trial. Circulation. 2022; 146(3): 175-190.

[12] Gall SL, Feigin V, Thrift AG, et al. Personalized knowledge to reduce the risk of stroke (PERKS-International): Protocol for a randomized controlled trial. Int J Stroke. 2023; 18(4): 477-483.

[13] Krishnamurthi R, Hale L, Barker-Collo S, et al. Mobile Technology for Primary Stroke Prevention: A Proof-of-Concept Pilot Randomized Controlled Trial. Stroke. 2019; 50(1): 196-198.

[14] Dramé M, Cantegrit E, Godaert L. Self-Rated Health as a Predictor of Mortality in Older Adults: A Systematic Review. Int J Environ Res Public Health. 2023; 20(5): 3813. Published 2023 Feb 21.

[15] Tanaka T, Morishita S, Nakano J, et al. Relationship between patient-reported health-related quality of life as measured with the SF-36 or SF-12 and their mortality risk in patients with diverse cancer type: a meta-analysis. Int J Clin Oncol. 2025; 30(2): 252-266.

[16] Herraiz-Adillo Á, Ahlqvist VH, Daka B, et al. Life's Essential 8 in relation to self-rated health and health-related quality of life in a large population-based sample: the SCAPIS project. Qual Life Res. 2024; 33(4): 1003-1014.

[17] Ratner, B. (2017). Statistical and machine-learning data mining: Techniques for better predictive modeling and analysis of big data. Chapman and Hall/CRC.

[18] Pinheiro, Carlos Andre Reis and Mike Patetta. 2021. Introduction to Statistical and Machine Learning Methods for Data Science. Cary, NC: SAS Institute Inc.

[19] Akinbami LJ, Chen TC, Davy O, Ogden CL, Fink S, Clark J, et al. National Health and Nutrition Examination Survey, 2017–March 2020 prepandemic file: Sample design, estimation, and analytic guidelines. National Center for Health Statistics. Vital Health Stat 2(190). 2022.

[20] Chen TC, Clark J, Riddles MK, Mohadjer LK, Fakhouri THI. National Health and Nutrition Examination Survey, 2015-2018: Sample Design and Estimation Procedures. Vital Health Stat 2. 2020; (184): 1-35.

[21] Flom P. An introduction to classification and regression trees with PROC HPSPLIT. In Midwest SAS Users Group (MWSUG) Conference Proceedings. Paper AA-42, 2018.

[22] Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J. (1984). Classification and Regression Trees.

[23] Nord C, Keeley J. An Introduction to the HPFOREST Procedure and its Options. In Midwest SAS Users Group (MWSUG) Conference Proceedings; Paper AA20, 2016.

[24] Belmont, CA: Wadsworth SAS Institute Inc. 2015. The HPSPLIT Procedure. SAS/STAT® 14.1 User's Guide. Cary, NC: SAS Institute Inc.

[25] SAS Institute Inc. 2016. SAS® Enterprise Miner™ 14.2: High-Performance Procedures. Cary, NC: SAS Institute Inc.

[26] Liaw A, Wiener M (2002). Classification and Regression by randomForest. R News, 2(3), 18-22. https://CRAN.R-project.org/doc/Rnews/.

[27] Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software, 28(5), 1–26.

[28] Manual AB. An introduction to statistical learning with applications in R.

[29] Liu YH. Python machine learning by example: unlock machine learning best practices with real-world use cases. Packt Publishing Ltd; 2024 Jul 31.

[30] VanderPlas J. Python data science handbook: Essential tools for working with data. "O'Reilly Media, Inc."; 2016 Nov 21.

[31] Flegal, K. M., Kit, B. K., Orpana, H., & Graubard, B. I. (2013). Association of all-cause mortality with overweight and obesity using standard body mass index categories: a systematic review and meta-analysis. JAMA, 309(1), 71–82.

[32] Piercy, K. L., Troiano, R. P., Ballard, R. M., Carlson, S. A., Fulton, J. E., Galuska, D. A., George, S. M., & Olson, R. D. (2018). The Physical Activity Guidelines for Americans. JAMA, 320(19), 2020–2028.

[33] American Diabetes Association Professional Practice Committee (2022). 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes-2022. Diabetes care, 45(Suppl 1), S17–S38.

[34] Cao, Y., Zhang, X., Fearon, I. M., Li, J., Chen, X., Xiong, Y., Zheng, F., Zhang, J., Sun, X., & Liu, X. (2024). The effects of electronic cigarette use patterns on health-related symptom burden and quality of life: analysis of US prospective longitudinal cohort study data. Frontiers in public health, 12, 1433678.

[35] Weir, C. B., & Jan, A. (2023). BMI Classification Percentile and Cut Off Points. In StatPearls. StatPearls Publishing.

[36] Deng J, Ji W, Liu H, et al. Development and validation of a machine learning-based framework for assessing metabolic-associated fatty liver disease risk. BMC Public Health. 2024; 24(1): 2545. Published 2024 Sep 18.

[37] Ma X, Wu Y, Zhang L, et al. Comparison and development of machine learning tools for the prediction of chronic obstructive pulmonary disease in the Chinese population. J Transl Med. 2020; 18(1): 146. Published 2020 Mar 31.

[38] Hu X, Yang Z, Ma Y, et al. Development and validation of a machine learning-based predictive model for secondary post-tonsillectomy hemorrhage. Front Surg. 2023; 10: 1114922. Published 2023 Feb 7.

[39] Hill AB. The environment and disease: association or causation? Proc R Soc Med. 58:295-300, 1965.

[40] DeSalvo KB, Fisher WP, Tran K, Bloser N, Merrill W, Peabody J. Assessing measurement properties of two single-item general health measures. Qual Life Res. 2006; 15(2): 191-201.

[41] Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory (3rd ed.). New York: McGraw-Hill.

[42] Yeager DS, Krosnick JA. The validity of self-reported nicotine product use in the 2001-2008 National Health and Nutrition Examination Survey. Med Care. 2010; 48(12): 1128-1132.

[43] Cleland CL, Hunter RF, Kee F, Cupples ME, Sallis JF, Tully MA. Validity of the global physical activity questionnaire (GPAQ) in assessing levels and change in moderate-vigorous physical activity and sedentary behaviour. BMC Public Health. 2014; 14: 1255. Published 2014 Dec 10.

[44] Lee PH. Validation of the National Health and Nutritional Survey (NHANES) single-item self-reported sleep duration against wrist-worn accelerometer. Sleep Breath. 2022; 26(4): 2069-2075.