Open Access Library Journal
2024, Volume 11, e12281
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Statistical Analysis of Cardiovascular Diseases Dataset of BRFSS

Sushant Kumar Gupta, Ashank Anshuman, Aakarshit Uppal, Indrajit Mukherjee

Computer Science and Engineering, Birla Institute of Technology, Mesra, India
Email: iamsushantgupta@gmail.com

How to cite this paper: Gupta, S.K., Anshuman, A., Uppal, A. and Mukherjee, I. (2024) Statistical Analysis of Cardiovascular Diseases Dataset of BRFSS. *Open Access Library Journal*, **11**: e12281.
https://doi.org/10.4236/oalib.1112281

Received: September 11, 2024
Accepted: October 28, 2024
Published: October 31, 2024

Abstract

Cardiovascular Diseases (CVDs) remain a leading cause of death in the United States. These diseases, including coronary heart disease, heart attack, and stroke, pose significant health risks. Accurate prediction of CVD probability can aid in prevention and management. To address this challenge, we analyzed data from the Behavioral Risk Factor Surveillance System (BRFSS) spanning 1995-2017. We developed innovative methods to handle missing data and normalize values. Deep learning models were employed to predict risk factors and, subsequently, the likelihood of CVDs. Our models were implemented using TensorFlow and trained on a high-performance computing server. The models accurately predicted risk factors with over 90% accuracy, enabling targeted interventions. We successfully predicted CVD probability with greater than 95% accuracy, providing valuable insights for healthcare providers. An online portal was developed to forecast CVD trends over the next 31 years, facilitating proactive planning and resource allocation.

Subject Areas

Deep Learning, High Performance Computing

Keywords

Deep Learning, Predictive Models, Bioinformatics, Healthcare, Medicine

1. Introduction

Cardiovascular diseases (CVDs) were a leading cause of death globally in 2016, accounting for approximately 17.6 million fatalities. In the United States, CVDs were among the top ten causes of death, affecting 9% of adults (roughly 24.3 million) that year. However, the prevalence of CVDs varied among different age groups and racial populations, with certain demographics experiencing a disproportionate

burden of disease [1]. The Framingham Heart Study [2] developed a risk scoring system to predict the 10-year risk of coronary heart disease. This system assigns weights to various risk factors to determine an individual's overall risk. While initially focused on coronary heart disease, the scoring method was later generalized [3] to estimate the risk of all CVDs. However, a notable limitation of this approach is its failure to consider social and economic factors that can significantly influence cardiovascular health. A recent study [4] found a correlation between resting heart rate and the risk of CVD, particularly among individuals with diabetes. Compared to non-diabetic participants, those with diabetes exhibited higher heart rates, which were associated with an increased likelihood of developing cardiovascular complications.

A relatively older study [5] from 2003 utilized BRFSS data to examine obesity prevalence among U.S. citizens. Given its substantial size and diverse population coverage, the study concluded that BRFSS data could effectively predict statewide and nationwide health-related quality of life (HRQL). More recently, [6] employed big data analytics to predict heart attack risk, drawing from a wide range of clinical, pharmaceutical, behavioral, and sentimental data.

While previous research has explored CVD risk factors, there remains a gap in understanding the impact of social and economic factors. This study aims to address this limitation by utilizing BRFSS data to investigate the influence of socioeconomic factors on CVD risk and to predict the occurrence of these diseases based on identified risk factors.

1.1. Dataset

The American Heart Association has identified seven key ***risk factors*** that significantly influence cardiovascular health [7]. These factors include ***blood pressure***, ***physical activity levels***, ***cholesterol levels***, ***dietary habits***, ***weight***, ***smoking status***, and ***blood glucose levels***. The Behavioral Risk Factor Surveillance System (BRFSS) [8] [9] is a nationally recognized annual survey that collects data on the health status of U.S. residents. Conducted via telephone, BRFSS has been instrumental in numerous research studies due to its reliability and comprehensive data [10]-[13]. For this analysis, we utilized the entire BRFSS dataset spanning 1995 to 2017, consisting of more than 3M rows and 27 columns [14] [15]. Key variables included:

- *Year*: The year the data was collected.
- *Locationabbr*: The state code where the data was collected.
- *Break_Out_Category*: The demographic group surveyed categorized based on age, race/ethnicity, income, education, and gender.
- *Break_Out*: The specific value within the *Break_Out_Category* (e.g., "*White, non-Hispanic*").
- *Question*: The question posed to the respondent [16].
- *Response*: The respondent's answer—binary yes/no or four-level.
- *Topic*: The subject matter of the question.
- *Class*: The survey category to which the topic belongs.

- *Data_value*: The percentage of respondents giving a particular response.
- *Sample_Size*: The number of respondents giving a particular response.

The following unnecessary or redundant columns were removed: *Locationdesc*, *ClassId*, *TopicId*, *Data_value_unit*, *Confidence_limit_Low*, *BreakoutID*, *Display_order*, *Confidence_limit_High*, *QuestionID*, *ResponseID*, *BreakOutCategoryID*, *DataSource*, *LocationID*, *Data_value_type*, *Data_Value_Footnote_Symbol*, *Data_Value_Footnote*, and *GeoLocation*.

Each row represents a specific group of individuals within a given year, location, demographic category, and survey question. For example, consider the question "*Four Level Smoking Status*" and the response "*Smoke everyday*". Focusing on the year 2011, the state of Alabama, and the breakout category "*Gender*" with the value "*Male*", this row would represent the percentage of surveyed males in Alabama during 2011 who smoked daily.

1.2. Analysis

Each risk factor was analyzed individually, followed by an analysis of each disease considering the combined effect of all risk factors. The analysis of the entire dataset was conducted in three phases: raw analysis, subset data analysis, and complete analysis. The results of the first two phases are omitted as they were primarily intended to understand trends and design the analysis model. This paper presents the results of the third and final phase, which analyzed the entire BRFSS data (1996-2017).

1.3. Hardware Stack

The BRFSS dataset, spanning over 3M rows, required significant computational resources. A high-performance computing server [17] was used to handle the demanding preprocessing and model training tasks.

2. Data Preprocessing

Table 1 lists the questions used to extract data for each disease or risk factor. The combination of *Year*, *Locationabbr*, *Break_Out_Category*, and *Break_Out* served as a unique identifier. Each key had exactly one corresponding entry in the dataset for each possible response. Data values were compared after joining the key columns to compare datasets for different questions.

2.1. Choice of Questions

For all categories except "*Physical Activity*", a single question within the dataset directly related to the disease or risk factor, eliminating the need for further preprocessing. However, for *Physical Activity*, six related questions were identified:

1) *Participated in* 150 *minutes or more of Aerobic Physical Activity per week* (asked in 2011, 2013, 2015, & 2017).

2) *During the past month*, *did you participate in any physical activities* (asked in 1996, 1998, & 2000-2017)?

3) *Adults with* 20+ *minutes of vigorous physical activity three or more days per week* (asked in 2001, 2003, 2005, 2007, & 2009).

Table 1. Questions For extracting diseases and risk factors data.

Disease/RF	Question	Possible responses
Coronary heart disease	Ever told you had angina or coronary heart disease?	Yes/No
Heart attack	Ever told you had a heart attack?	Yes/No
Stroke	Ever told you had a stroke?	Yes/No
Hypertension	Adults who have been told they have high blood pressure.	Yes/No
Physical activity	During the past month, did you participate in any physical activities?	Yes/No
Cholesterol	Adults who have had their blood cholesterol checked and have been told it was high.	Yes/No
Nutrition	Adults who have consumed fruits and vegetables five or more times per day**, Consumed fruit less than one time per day*, Consumed vegetables less than one time per day*.	Yes/No
Obesity	Weight classification by Body Mass Index (BMI).	Obese (BMI 30.0 - 99.8)/ Overweight (BMI 25.0 - 29.9)/ Normal weight (BMI 18.5 - 24.9)*/ Underweight (BMI 12.0 - 18.4)*/ Neither overweight nor obese (BMI le 24.9)**
Smoking	Four Level Smoking Status.	Smoke everyday/Smoke some day/ Former smoker/Never smoked
Diabetes	Have you ever been told by a doctor that you have diabetes?	Yes/Yes, pregnancy-related/No, pre-diabetes or borderline diabetes/No

*Indicates that the question/response was available only in the data set post 2010; **Indicates that the question/response was available only in the data set prior to 2011.

4) *Adults with* 30+ *minutes of any physical activity five or more days per week* (asked in 1996, 1998, & 2000).

5) *Adults with* 30+ *minutes of moderate physical activity five or more days per week, or vigorous physical activity for* 20+ *minutes three or more days per week* (asked in 2001, 2003, 2005, 2007, & 2009).

6) *Participated in enough Aerobic and Muscle Strengthening exercises to meet guidelines* (asked in 2011, 2013, 2015, & 2017).

To analyze the relationship between these questions, the data values were plotted against each other. One variable was plotted in sorted order, while the trend of the other was observed. Figure 1 illustrates several of these plots. The plots revealed a consistent trend across all questions related to physical activity. Increases in the number of individuals participating in 150 minutes of aerobic activity were generally associated with higher levels of physical activity in the past month and greater involvement in muscle strengthening activities. The overall trend indicated a positive correlation between these factors.

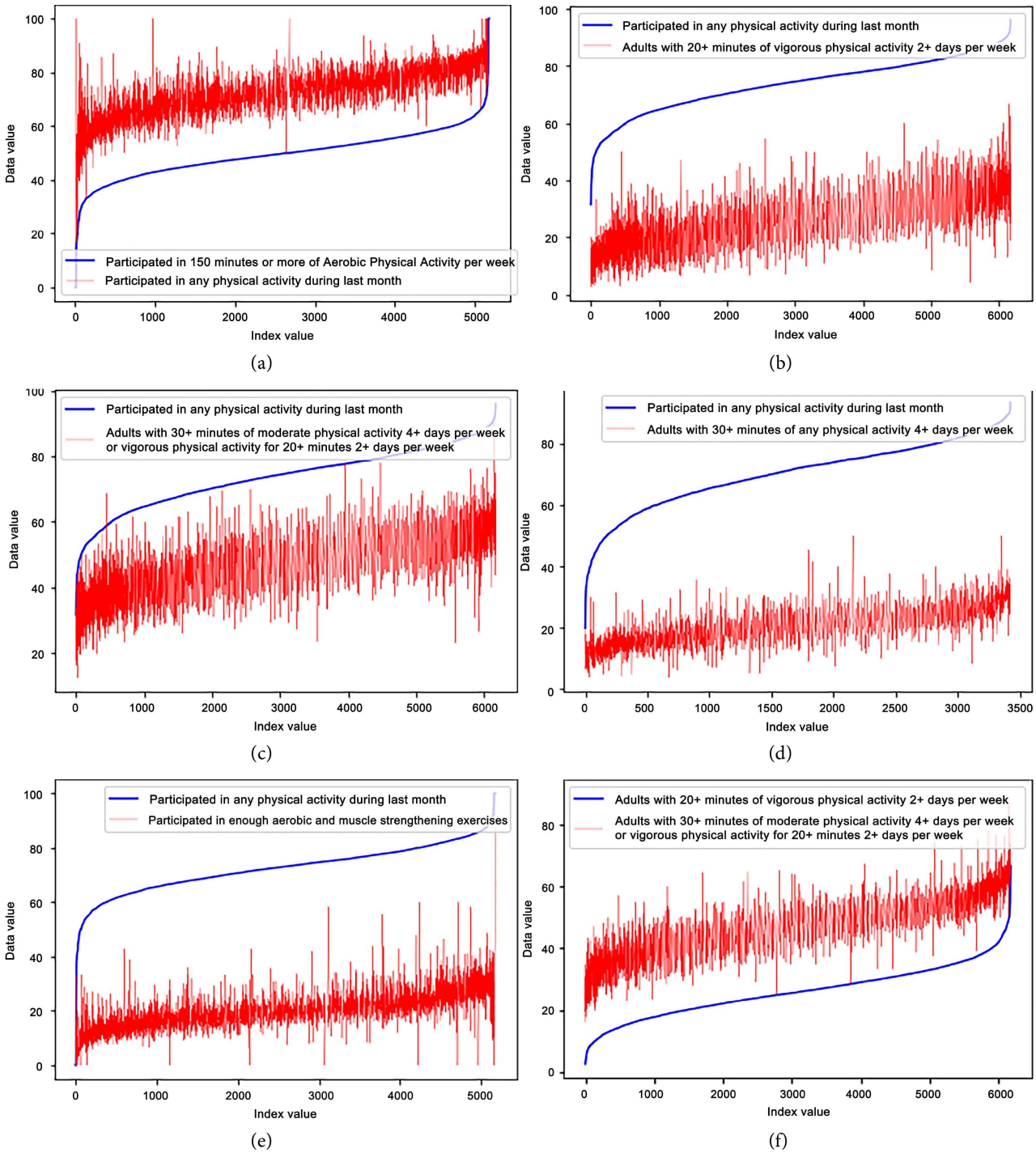


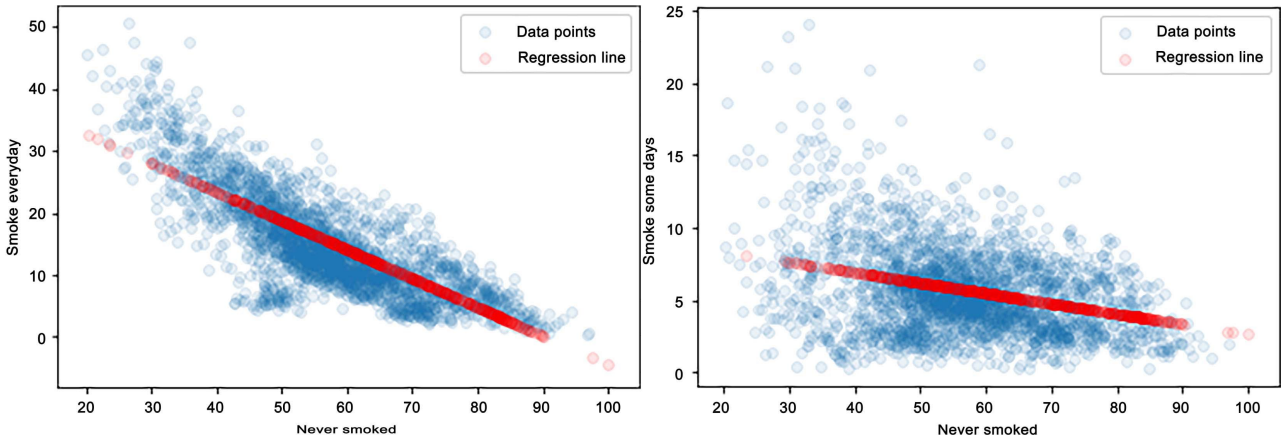
Figure 1. Plot of data values for different questions for the risk factor "physical activity". The blue line represents the sorted variable and the red line indicates the trend of the other variable.

The question "*During the past month, did you participate in any physical activities?*" was selected due to its availability across all surveyed years. Similarly, for Smoking, two related questions were analyzed: "*Adults who are current smokers*" and "*Four Level Smoking Status*". A positive correlation between these measures was observed, leading to the selection of "*Four Level Smoking Status*" due to its

broader applicability.

2.2. Missing Values Fill-Up

For "yes/no" questions, missing values were filled by subtracting the available data value from 100. For questions with four-level responses, a more complex approach was necessary. If three out of four responses were already present, the missing response was calculated by subtracting the sum of the others from 100%. If fewer than three responses were available, the response with the most available data was used as a predictor in a regression model to impute missing values for the other responses. The response that yielded the best regression accuracy was filled first. After filling in missing values, if three responses were complete, the fourth was calculated using the aforementioned method.



(a) (b)

Figure 2. Plot showing the regression line to fill up the missing values for smoking data.

Example: For smoking data, with four possible responses ("*Smoke everyday*", "*Smoke some days*", "*Former smoker*", "*Never smoked*"), "*Never smoked*" had the most available data and was used as the predictor in regression models. Missing values for "*Smoke everyday*" were filled first, followed by "*Smoke some days*". Figure 2 illustrates the regression lines for "*Never Smoked*" vs "*Smoke everyday*" and "*Never Smoked*" vs "*Smoke some days*".

2.3. The Special Case of Obesity

Prior to 2011, the obesity question had three responses: "*Obese*", "*Overweight*", and "*Neither overweight nor obese*". After 2010, it included four responses: "*Obese*", "*Overweight*", "*Normal Weight*", and "*Underweight*". To ensure consistency across the dataset, the following procedure was applied to data prior to 2011:

- The sum of "*Normal Weight*" and "*Underweight*" data was plotted against "*Underweight*" data, and a regression line was fitted.
- Using the regression line, "*Underweight*" data values corresponding to each

value in the "*Neither overweight nor obese*" category were calculated.

- "*Normal Weight*" data values were obtained by subtracting the calculated "*Underweight*" values from the given "*Neither overweight nor obese*" values.
- The given sample size was divided proportionally based on the calculated data values.

N, SS = Data value and sample size for "*Neither overweight nor obese*" respectively.

UW, NW = Calculated "*Underweight*" and "*Normal Weight*" data value respectively.

SSUW, SSNW = Sample size for "*Underweight*" and "*Normal Weight*" respectively.

$$NW = N - UW$$

$$SSUW = SS * UW/N$$

$$SSNW = SS * (NW/N)$$

After distributing the "*Neither overweight nor obese*" category and establishing four-level responses, the missing data imputation procedure described in Section 2.2 was applied. Figure 3 shows the regression line for the sum of "*Normal Weight*" and "*Underweight*" versus "*Underweight*" responses.

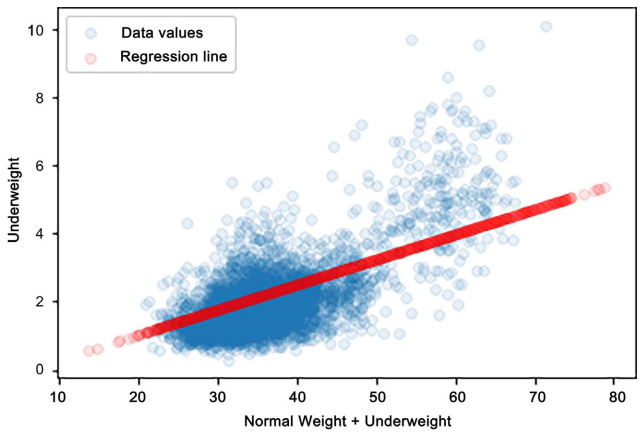


Figure 3. Regression line for calculating data values for the response "*underweight*" from the combination of data values of "*normal weight*" and "*underweight*".

3. Risk Factor Analysis

Following preprocessing, each risk factor was analyzed separately. All models for risk factors and diseases were constructed using neural networks, renowned for their ability to learn complex patterns and generate continuous outputs. To optimize model performance, validation loss monitoring was employed, ensuring the model with the least validation loss was selected. The dataset included six age groups, two genders, eight races, five income levels, four education levels, and one overall category. A total of 25 dummy variables [18] were used to encode the

breakout category. Data was available for 54 states and union territories, requiring 53 dummy variables [18] to encode location.

Proposed Models

For each risk factor, dense sequential models were used with varying numbers of layers, nodes, activation functions, and initialization methods in each layer. All models were trained using the Adagrad optimizer with a learning rate of 0.1. Mean absolute error was used as the loss function during training. Table 2 lists the models and their performance.

Table 2. Neural network models and their performance for risk factors.

Risk factor	Num layers	Num nodes	Activation	Initialization	Loss	Validation loss
Hypertension	3	100	Sigmoid	Uniform	0.0255	0.0298
		50	Sigmoid	Xavier		
		1	Sigmoid	Xavier		
Physical activity	4	100	Tanh	Random normal	0.0278	0.0475
		50	ReLU	Xavier		
		25	Tanh	Xavier		
		1	Sigmoid	Xavier		
Cholesterol	4	100	Tanh	Random normal	0.0367	0.0376
		50	ReLU	Xavier		
		25	Tanh	Xavier		
		1	Sigmoid	Xavier		
Obesity	3	100	ReLU	Uniform	0.0298	0.0369
		50	ReLU	Xavier		
		1	ReLU	Xavier		
Smoking	3	100	Sigmoid	Uniform	0.0336	0.0445
		50	Sigmoid	Xavier		
		1	Sigmoid	Xavier		
Diabetes	3	100	Sigmoid	Uniform	0.027	0.0301
		50	Sigmoid	Xavier		
		1	Sigmoid	Xavier		

3.1. Hypertension

Hypertension is a significant risk factor for CVDs [19]. Year-over-year analysis revealed an increase in hypertension rates from 23% in 1995 to 33% in 2017, as depicted in Figure 4. While location did not exhibit a clear trend with hypertension rates, age and gender played significant roles. Older individuals and males experienced higher rates of hypertension compared to younger individuals and females. Additionally, higher income and education levels were associated with lower hypertension risk, while Black, American Indian, and Native Hawaiian

populations faced a disproportionate burden of hypertension. A study by [20] found that individuals living at higher altitudes had lower blood pressure, possibly due to lower air pressure. This trend is illustrated in Figure 5. By analyzing hypertension rates across states and union territories, a significant shift was observed, with rates decreasing from 36% to 24% at higher elevations.

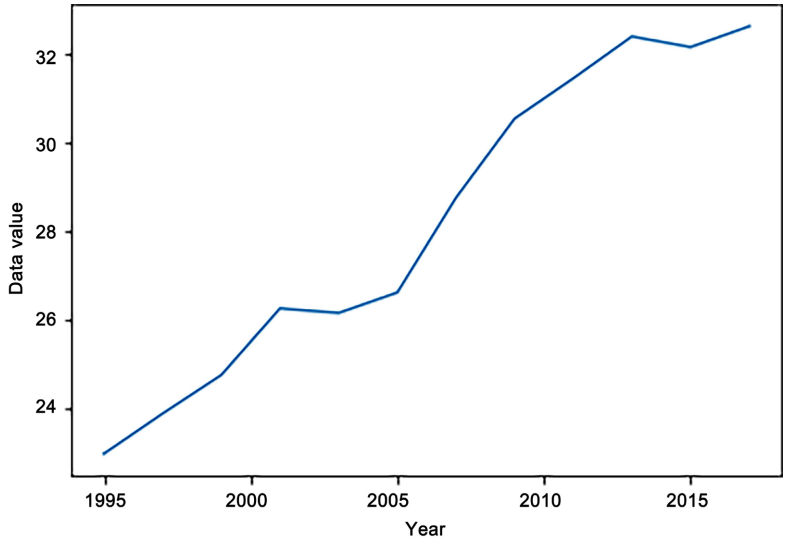


Figure 4. Plot showing the mean percentage of people suffering from hypertension over the years.

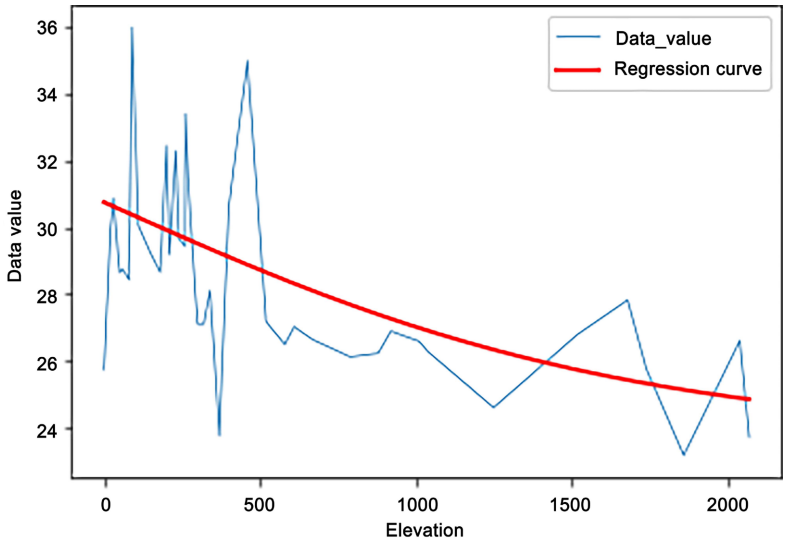


Figure 5. Plot showing mean percentage of people suffering from hypertension by elevation (in meters).

The training models incorporated the following normalized input parameters: *Year*, *Breakout* (encoded using dummy variables), *Location* (encoded using dummy variables), *Elevation*, and *Data value*. A total of 80 inputs were fed into a neural network, and the model's error was compared against the given data values.

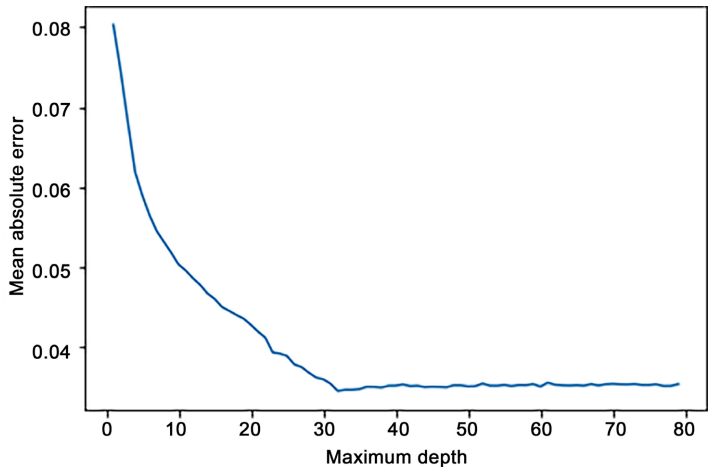


Figure 6. Plot showing the mean absolute error obtained by decision trees of given maximum depth.

Evaluation

The proposed model was evaluated against four other models: Linear Regression, Decision Tree, Random Forest, and Support Vector. For each model, the optimal parameter values were selected based on performance metrics. For example, the mean absolute error was plotted against the maximum depth of the decision tree. The maximum depth corresponding to the lowest mean absolute error was chosen for comparison with the proposed model. Figure 6 illustrates the plot of maximum depth versus mean absolute error for the decision tree model.

The optimal parameters for the other models were determined using similar methods. All models were evaluated using 6-fold cross-validation. Figure 7 illustrates the variation in mean absolute error across the six folds for each model. Figure 8 shows the variation in R2 score across the folds for each model. Based on the plots, it is evident that the proposed neural network model surpasses all other models in performance. The evaluation of other risk factor models followed similar techniques and has been omitted for brevity.

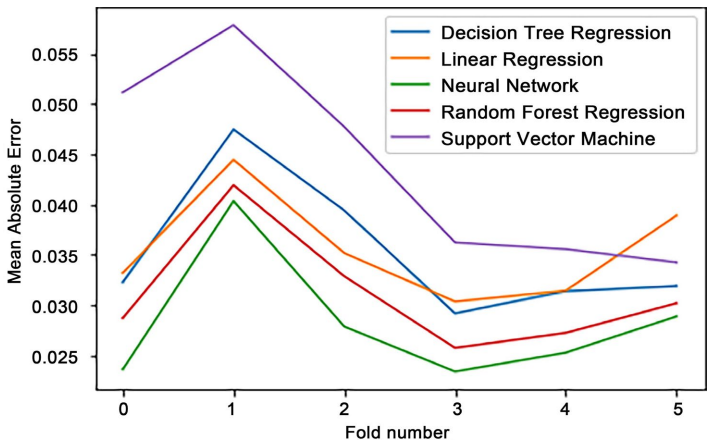


Figure 7. Plot showing the mean absolute error of each model for each fold of cross validation.

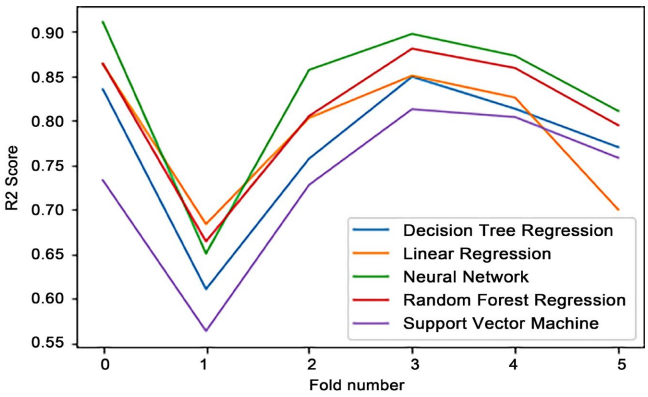


Figure 8. Plot showing the R2 score of each model for each fold of cross validation.

Error Analysis

For about 41.58% of all the data points, the model output deviated by 1%.
For about 16.23% of all the data points, the model output deviated by 3%.
For about 13.41% of all the data points, the model output deviated by 5%.

3.2. Physical Activity

Physical activity was associated with cardiovascular benefits [21]. Year-over-year analysis revealed an increase in physical activity levels during the late 20th century, followed by a period of relative stability and a recent decline, as depicted in Figure 9. Additionally, elevation seemed to influence physical activity levels, with individuals at higher altitudes exhibiting greater physical activity, as illustrated in Figure 10. Age and gender also played roles, with younger individuals and males demonstrating higher levels of physical activity compared to older individuals and females. Furthermore, higher education and income levels were associated with increased physical activity, while Whites, Asians, and Native Hawaiians exhibited higher levels compared to other racial groups. The input parameters for training the physical activity model were identical to those used for hypertension (refer to Section 3.1).

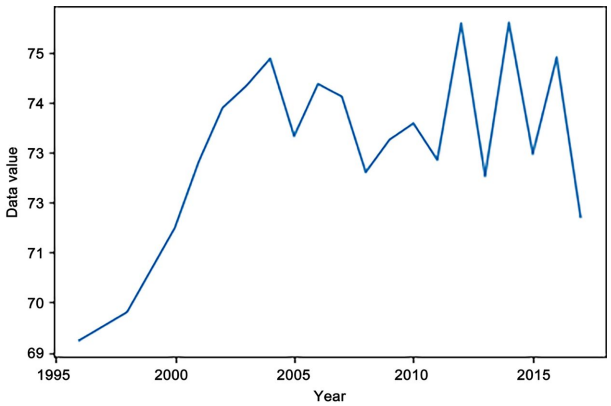


Figure 9. Plot showing mean percentage of people participating in physical activity over the years.

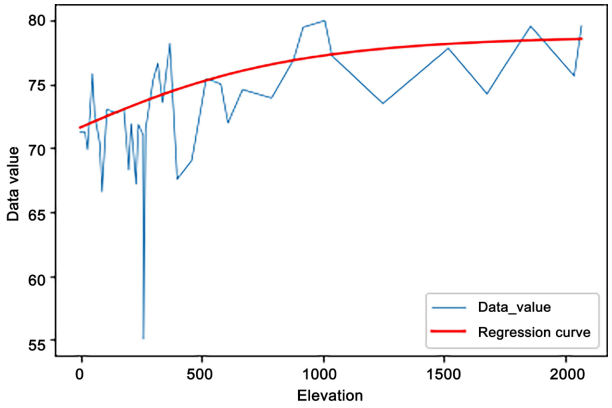


Figure 10. Plot showing mean percentage of people participating in physical activity by elevation (in meters).

Error Analysis

For about 37.32% of all the data points, the model output deviated by 1%.
For about 17.80% of all the data points, the model output deviated by 3%.
For about 17.72% of all the data points, the model output deviated by 5%.

3.3. Cholesterol Level

Initial studies linked elevated cholesterol levels to increased CVD risk [22], but subsequent research presented conflicting evidence [23]. Over time (1995-2015), high cholesterol rates increased in the late 20th century but declined recently (**Figure 11**). Older individuals and males exhibited higher cholesterol risk compared to younger individuals and females. Higher education and income levels were associated with lower cholesterol risk, while Whites, American Indians, and Multiracial Americans faced a disproportionate burden. The training models incorporated normalized *Year*, *Breakout* (encoded), *Location* (encoded), and *Data value* parameters, totaling 79 input nodes.

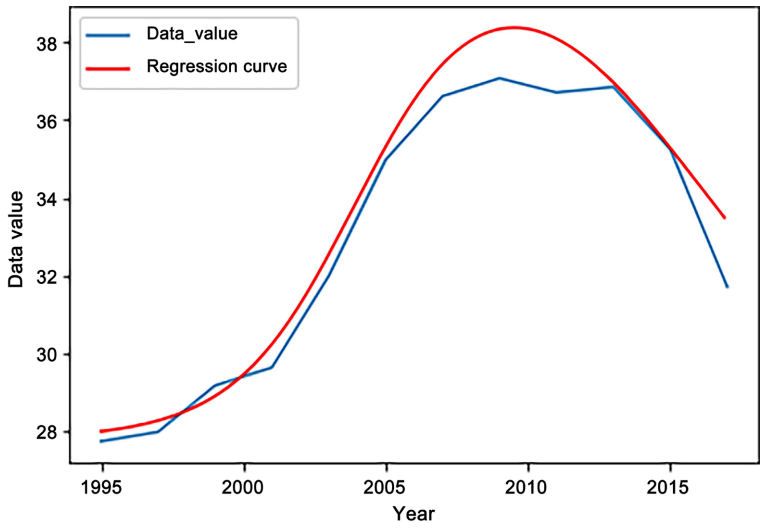


Figure 11. Plot showing mean data values for high cholesterol over the years.

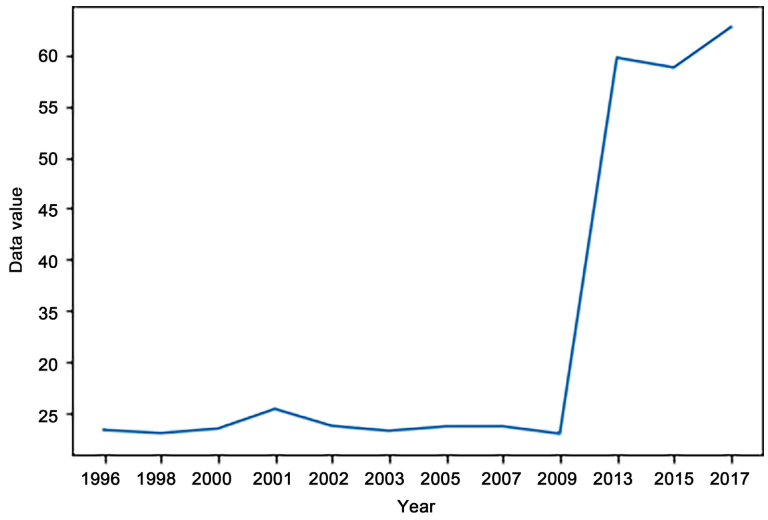


Figure 12. Plot showing data values for nutrition against the years. The data value after 2011 is drastically different from the earlier years.

Error Analysis

For about 35.14% of all the data points, the model output deviated by 1%.
For about 21.22% of all the data points, the model output deviated by 3%.
For about 23.17% of all the data points, the model output deviated by 5%.

3.4. Nutrition

Nutrition data was available only for 2013, 2015, and 2017 after 2010. The questions related to nutrition consumption were "*Consumed fruit less than one time per day*" and "*Consumed vegetables less than one time per day*". Although some correlation was observed between these factors and disease outcomes, the limited data availability hindered definitive conclusions (refer to Sections 4.1, 4.2, and 4.3).

Before 2011, the nutrition question focused on "*Adults who have consumed fruits and vegetables five or more times per day*". This question was significantly different from the post-2010 questions. For example, **Figure 12** illustrates the distinct mean data values for the two periods. Additionally, the data values for nutrition exhibited minimal correlation with disease outcomes, suggesting a negligible impact (refer to Sections 4.1, 4.2, and 4.3). Given these limitations, nutrition data was excluded from further analysis as a risk factor.

3.5. Obesity

Obesity is a significant risk factor for cardiovascular diseases [24]. Over the period 1995-2015, obesity rates increased from 16% to 31%, as shown in **Figure 13**. While obesity generally increased with age, a sharp decline was observed after the age of 65. Males exhibited higher obesity rates than females, while higher education and income levels were associated with lower obesity risk. American Indians and Native Hawaiians had a disproportionately higher prevalence of obesity compared to other groups.

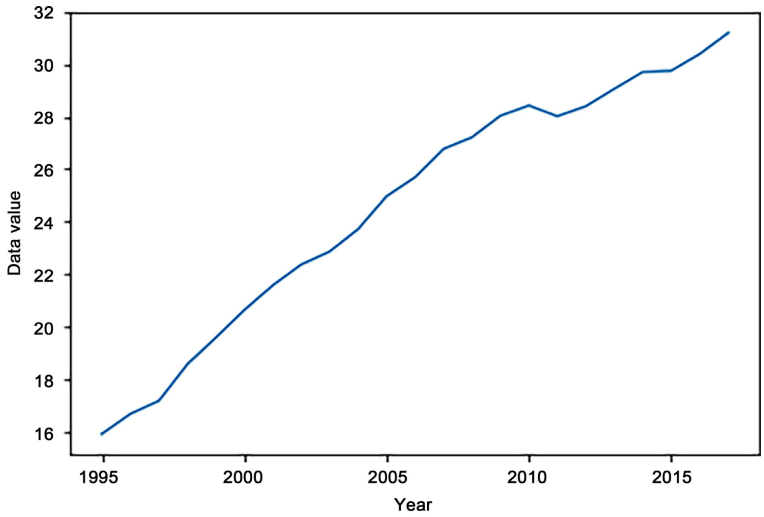


Figure 13. Plot showing mean data values for people suffering from obesity over the years.

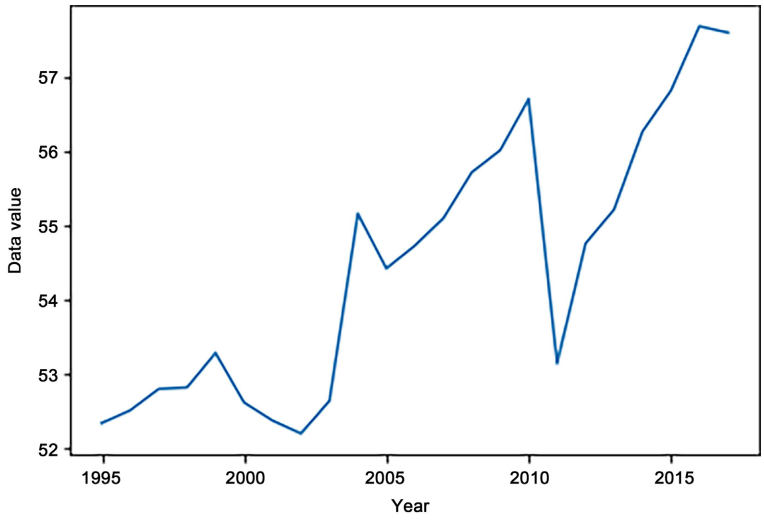


Figure 14. Plot showing mean data values for people who had never smoked over the years.

Although obesity data included four responses, only "*Obese* (*BMI* 30.0 - 99.8)" was considered a representative risk factor (refer to Sections 4.1, 4.2, and 4.3).

The input parameters for training the obesity model were identical to those used for cholesterol analysis (refer to Section 3.3).

Error Analysis

For about 40.62% of all the data points, the model output deviated by 1%.

For about 12.41% of all the data points, the model output deviated by 3%.

For about 5.88% of all the data points, the model output deviated by 5%.

3.6. Smoking

Smoking cessation is associated with a lower risk of CVDs [25]. From 1995 to 2015, the proportion of non-smokers increased from 52% to 57%, although the

trend was not consistently upward (Figure 14). While smoking rates generally increased with age, a decline was observed after the age of 65. Males exhibited higher smoking rates than females, while higher education and income levels were associated with lower smoking rates. Asians had a lower prevalence of smoking compared to other groups.

Only the "*Never smoked*" response was considered representative of smoking data (refer to Sections 4.1, 4.2, and 4.3).

The input parameters for training the smoking model were identical to those used for cholesterol analysis (refer to Section 3.3).

Error Analysis

For about 38.51% of all the data points, the model output deviated by 1%.

For about 18.56% of all the data points, the model output deviated by 3%.

For about 15.53% of all the data points, the model output deviated by 5%.

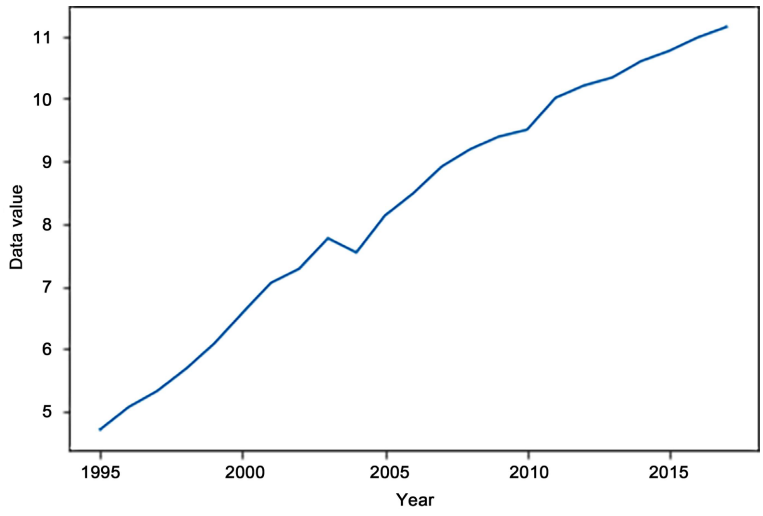


Figure 15. Plot showing mean data values for people suffering from diabetes over the years.

3.7. Diabetes

Both type 1 and type 2 diabetes are closely linked to cardiovascular diseases [26]. From 1995 to 2015, diabetes rates increased from 5% to 11% (Figure 15). Older individuals and males exhibited higher diabetes risk compared to younger individuals and females. Higher education and income levels were associated with lower diabetes risk, while American Indians and Native Hawaiians had a disproportionately higher prevalence.

Only the "*Yes*" response was considered representative of diabetes data (refer to Sections 4.1, 4.2, and 4.3). The input parameters for training the diabetes model were identical to those used for cholesterol analysis (refer to Section 3.3).

Error Analysis

For about 29.95% of all the data points, the model output deviated by 1%.

For about 5.03% of all the data points, the model output deviated by 3%.

For about 1.54% of all the data points, the model output deviated by 5%.

4. Diseases Analysis

Each disease was analyzed independently, along with the impact of each risk factor on that disease. As previously mentioned, *Year*, *Locationabbr*, *Break_Out_Category*, and *Break_Out* served as the primary key for each data point. To create the dataset for disease analysis, the risk factor data and disease data were merged using an inner join. All accuracies mentioned below are calculated using mean absolute error.

Proposed Models

All models employed the same architecture: a three-layered dense sequential neural network with 5 (sigmoid activated and uniformly initialized), 3 (sigmoid activated and Xavier initialized), and 1 (sigmoid activated and Xavier initialized) nodes, respectively. All models were trained using the Adagrad optimizer with a learning rate of 0.1. Mean absolute error was used as the loss function during training. Table 3 lists their performance. Table 4 presents the model losses for predicting disease and risk factor data values.

Table 3. Neural network model performance for diseases.

Disease	Loss	Validation loss
Coronary heart disease	0.0423	0.04622
Heart attack	0.0433	0.04084
Stroke	0.0372	0.03232

Table 4. Individual risk factor model loss for diseases.

Risk factor	Coronary heart disease		Heart attack		Stroke	
	Loss	Validation loss	Loss	Validation loss	Loss	Validation loss
Hypertension	0.01243	0.01766	0.01409	0.01545	0.04136	0.03818
Physical activity	0.01886	0.01984	0.01971	0.02016	0.05341	0.05054
Cholesterol	0.01363	0.0146	0.01539	0.01425	0.04861	0.04194
Nutrition 1	0.09788	0.10288	0.10281	0.10726	0.06457	0.06195
Nutrition 2	0.02166	0.02387	0.02408	0.02652	0.06662	0.07748
Obesity: Obese	0.08614	0.10098	0.09314	0.10052	0.05973	0.05895
Obesity: Overweight	0.09317	0.10226	0.10156	0.10712	0.07296	0.07172
Obesity: Normal weight	0.08508	0.09783	0.09172	0.09931	0.06787	0.06778
Obesity: Underweight	0.09338	0.10187	0.10213	0.1051	0.07279	0.07047
Smoking: Smoke everyday	0.0869	0.09783	0.08956	0.09838	0.0601	0.06381
Smoking: Smoke some days	0.09469	0.10509	0.10249	0.11039	0.06711	0.06791
Smoking: Former smoker	0.07373	0.0821	0.08289	0.09838	0.06106	0.05907
Smoking: Never smoked	0.07278	0.08115	0.07324	0.0796	0.05468	0.05769
Diabetes: No	0.0635	0.07868	0.0652	0.0743	0.03993	0.04057

Continued

Diabetes: Yes	0.05865	0.07485	0.06008	0.10149	0.03857	0.03959
Diabetes: Pre-diabetes	0.08332	0.10246	0.08885	0.10246	0.05708	0.05932
Diabetes: Yes, pregnancy-related	0.091	0.09963	0.09901	0.10298	0.06797	0.06585

4.1. Coronary Heart Diseases

Individual risk factors were analyzed to understand their impact on coronary heart disease. A gradual decline in heart disease rates was observed over time. Figure 16 visualizes the relationship between risk factors and disease data values. Analysis revealed the following trends.

- *Hypertension*: 98% accuracy, positive correlation with disease.
- *Physical Activity*: 98% accuracy, negative correlation with disease.
- *Cholesterol*: 98% accuracy, positive correlation with disease.
- *Nutrition*: Data from before and after 2010 showed limited or no correlation with disease. Discarded from analysis due to insufficient data.
- *Obesity*: "Obese" response performed best (89% accuracy), positive correlation.
- *Smoking*: "Never Smoked" response performed best (91% accuracy), negative correlation.
- *Diabetes*: "Yes" response performed best (92% accuracy), positive correlation.

The final model included six input parameters, excluding nutrition.

Evaluation

A comparison with other similar models was conducted. The proposed neural network model demonstrated superior performance, as evidenced by the lower mean absolute error and higher R2 score in Table 5. The evaluation of other diseases followed a similar methodology and has therefore been omitted.

Table 5. Mean absolute error and R2 score for different models.

Fold	Linear regression		Decision tree		Random forest		Support vector		Neural network	
	Error	R2	Error	R2	Error	R2	Error	R2	Error	R2
1	0.051	0.767	0.041	0.762	0.045	0.791	0.052	0.766	**0.042**	**0.821**
2	0.051	0.745	0.044	0.763	**0.042**	**0.807**	0.051	0.759	0.042	0.794
3	0.048	0.773	0.045	0.773	0.042	0.801	0.048	0.783	**0.041**	**0.811**
4	0.048	0.778	0.046	0.778	**0.042**	**0.813**	0.048	0.784	**0.042**	0.802
5	0.047	0.805	0.046	0.805	0.044	0.824	0.047	0.817	**0.042**	**0.832**
6	0.055	0.675	0.063	0.675	0.058	0.724	0.058	0.742	**0.048**	**0.790**

Error Analysis

For about 29.91% of all the data points, the model output deviated by 1%.

For about 3.53% of all the data points, the model output deviated by 3%.

For less than 1% of all the data points, the model output deviated by 5%.

4.2. Heart Attack

The risk of heart attacks has remained relatively constant over time. Figure 17 illustrates the relationship between various risk factors and the disease. Analysis revealed the following trends:

- *Hypertension*: 98% accuracy, positive correlation with disease.
- *Physical Activity*: 98% accuracy, negative correlation with disease.
- *Cholesterol*: 98% accuracy, positive correlation with disease.
- *Nutrition*: Data showed similar anomalies as previously observed and was discarded.
- *Obesity*: "*Obese*" response performed best (89% accuracy), positive correlation.
- *Smoking*: "*Never Smoked*" response performed best (92% accuracy), negative correlation.
- *Diabetes*: "*Yes*" response performed best (92% accuracy), positive correlation.

Error Analysis

For about 28.00% of all the data points, the model output deviated by 1%.
For about 2.81% of all the data points, the model output deviated by 3%.
For less than 1% of all the data points, the model output deviated by 5%.

4.3. Stroke

Stroke rates have increased gradually over time. Figure 18 illustrates the relationship between various risk factors and the disease. Analysis revealed the following trends:

- *Hypertension*: 96% accuracy, positive correlation with disease.
- *Physical Activity*: 94% accuracy, negative correlation with disease.
- *Cholesterol*: 93% accuracy, positive correlation with disease.
- *Nutrition*: Data showed similar anomalies as previously observed and was discarded.
- *Obesity*: "Obese" response performed best (94% accuracy), positive correlation.
- *Smoking*: "*Never Smoked*" response performed best (94% accuracy), negative correlation.
- *Diabetes*: "*Yes*" response performed best (96% accuracy), positive correlation.

Error Analysis

For about 20.48% of all the data points, the model output deviated by 1%.
For about 1.64% of all the data points, the model output deviated by 3%.
For less than 1% of all the data points, the model output deviated by 5%.

5. Final Predictor

5.1. Architecture

The final predictor is a combination of seven individual predictors: hypertension risk, physical activity level, cholesterol risk, obesity risk, smoking level, diabetes risk, and disease risk. As illustrated in Figure 19, the first six predictors receive

Year, *Locationabbr*, *Break_Out_Category*, and *Break_Out* as inputs and generate corresponding rates as outputs. These outputs are then fed into the disease predictor to obtain the final disease rate.

5.2. Software Implementation

The software implementation, a web application, expected the user to enter their age, gender, race, income, education and location and used the model to calculate the risk for each disease, for each provided breakout. The output was the mean of risks predicted by all the breakouts for each disease. The variation of mean risk over the coming years was also displayed to the users. The application could thus be used to gather recommendations for a given demographic, by state, to take the necessary actions to mitigate risk of CVDs.

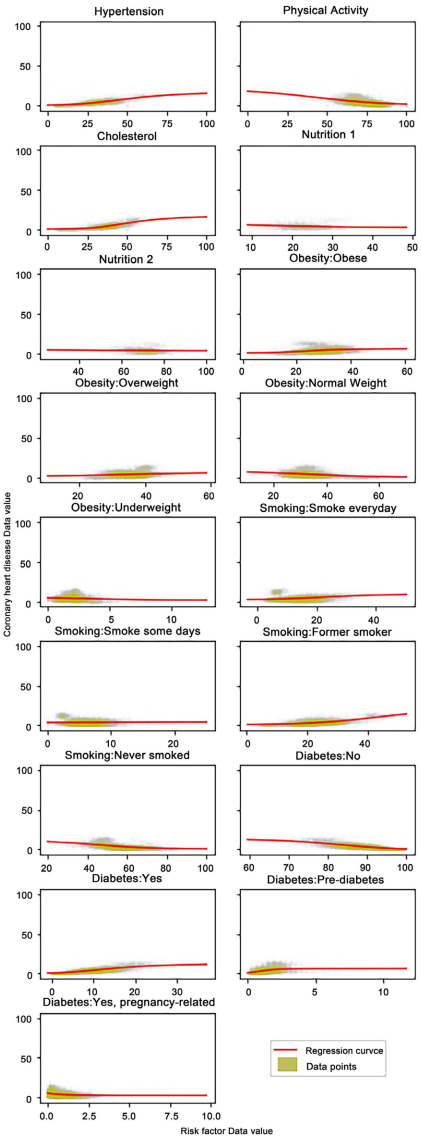


Figure 16. Plot of data value of different risk factors against data value of coronary heart disease.

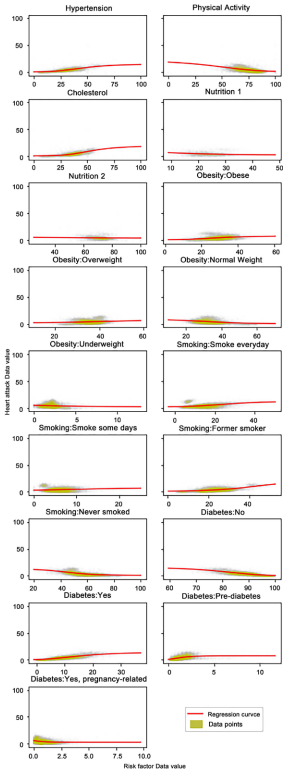


Figure 17. Plot of data value of different risk factors against data value of heart attack.

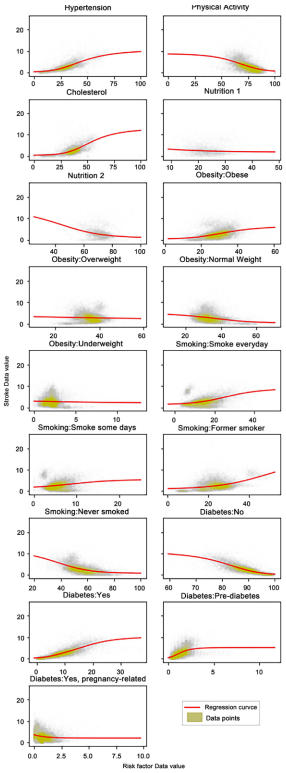


Figure 18. Plot of data value of different risk factors against data value of stroke.

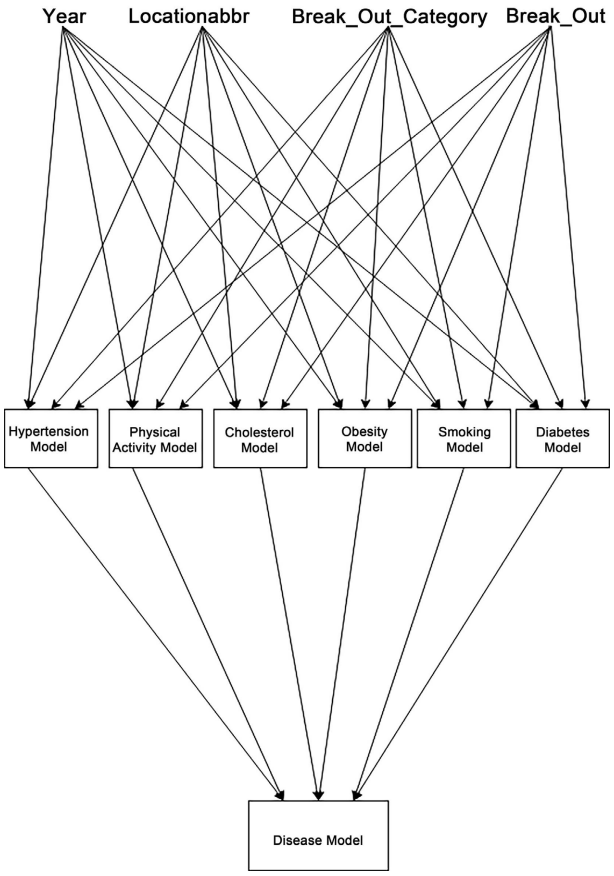


Figure 19. Model for prediction of risk for diseases.

6. Conclusions

Several notable trends emerged during the statistical analysis of the dataset. Risk factors such as hypertension, obesity, and diabetes exhibited a clear upward trajectory over the years. Other factors, including smoking, cholesterol risk, and physical activity, did not show a consistent increase or decrease. All risk factors demonstrated a correlation with age, with higher risks observed in males compared to females. Additionally, there was a general trend of decreasing risk with increasing income and education levels.

The BRFSS data revealed a gradual decline in the percentage of the population suffering from coronary heart disease and heart attacks over time. However, the percentage of individuals experiencing heart strokes showed an upward trend.

It's important to note that the model's predictions are currently limited to the United States. Other countries, including Australia, Brazil, Canada, China, and others, are developing similar surveillance systems and have sought technical assistance from the BRFSS. This collaboration will enable the models to be applied in other regions as well.

Acknowledgements

The authors would like to thank Dr. Gautam Sarkhel of the Department of Chemical

Engineering at the Birla Institute of Technology, Mesra, India, for providing access to the high-performance computing server at the Birla Institute of Technology, Mesra. This server was used to train all the models.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Benjamin, E.J., *et al.* (2019) Heart Disease and Stroke Statistics—2019 Update: A Report from the American Heart Association. *Circulation*, **10**, e56-e528.

[2] D'Agostino, R.B., Vasan, R.S., Pencina, M.J., Wolf, P.A., Cobain, M., Massaro, J.M., *et al.* (2008) General Cardiovascular Risk Profile for Use in Primary Care. *Circulation*, **117**, 743-753. https://doi.org/10.1161/circulationaha.107.699579

[3] Wilson, P.W.F., D'Agostino, R.B., Levy, D., Belanger, A.M., Silbershatz, H. and Kannel, W.B. (1998) Prediction of Coronary Heart Disease Using Risk Factor Categories. *Circulation*, **97**, 1837-1847. https://doi.org/10.1161/01.cir.97.18.1837

[4] Böhm, M., Schumacher, H., Teo, K.K., Lonn, E.M., Mahfoud, F., *et al.* (2018) Resting Heart Rate and Cardiovascular Outcomes in Diabetic and Non-Diabetic Individuals at High Cardiovascular Risk Analysis from the ONTARGET/TRANSCEND Trials. *European Heart Journal*, **41**, 231-238. https://doi.org/10.1093/eurheartj/ehy808

[5] Hassan, M.K., Joshi, A.V., Madhavan, S.S. and Amonkar, M.M. (2003) Obesity and Health-Related Quality of Life: A Cross-Sectional Analysis of the US Population. *International Journal of Obesity*, **27**, 1227-1232. https://doi.org/10.1038/sj.ijo.0802396

[6] Alexander, C.A. and Wang, L. (2017) Big Data Analytics in Heart Attack Prediction. *Journal of Nursing & Care*, **6**, Article 393. https://doi.org/10.4172/2167-1168.1000393

[7] Effoe, V.S., Carnethon, M.R., Echouffo-Tcheugui, J.B., Chen, H., Joseph, J.J., Norwood, A.F., *et al.* (2017) The American Heart Association Ideal Cardiovascular Health and Incident Type 2 Diabetes Mellitus among Blacks: The Jackson Heart Study. *Journal of the American Heart Association*, **6**, e005008. https://doi.org/10.1161/jaha.116.005008

[8] BRFSS (2019) CDC—About BRFSS. https://www.cdc.gov/brfss/about/index.htm

[9] Tiura, M.L. (2018) Impact of Behavioral Risk Factor Surveillance System Data on Public Health Outcomes Within States in the United States. Central Michigan University.

[10] Nelson, D.E., *et al.* (2001) Reliability and Validity of Measures from the Behavioral Risk Factor Surveillance System (BRFSS). *Sozial-und Praventivmedizin*, **46**, S3-S42.

[11] Schneider, K.L., Clark, M.A., Rakowski, W. and Lapane, K.L. (2010) Evaluating the Impact of Non-Response Bias in the Behavioral Risk Factor Surveillance System (BRFSS). *Journal of Epidemiology and Community Health*, **66**, 290-295. https://doi.org/10.1136/jech.2009.103861

[12] Stein, A.D., Lederman, R.I. and Shea, S. (1993) The Behavioral Risk Factor Surveillance System Questionnaire: Its Reliability in a Statewide Sample. *American Journal of Public Health*, **83**, 1768-1772. https://doi.org/10.2105/ajph.83.12.1768

[13] Yore, M.M., Ham, S.A., Ainsworth, B.E., Kruger, J., Reis, J.P., Kohl, H.W., *et al.* (2007) Reliability and Validity of the Instrument Used in BRFSS to Assess Physical Activity. *Medicine & Science in Sports & Exercise*, **39**, 1267-1274. https://doi.org/10.1249/mss.0b013e3180618bbe

[14] CDC (2019) Prevalence Data (2010 and Prior). https://chronicdata.cdc.gov/Behavioral-Risk-Factors/Behavioral-Risk-Factor-Surveillance-System-BRFSS-P/y4ft-s73h

[15] CDC (2019) Prevalence Data (2011 to Present). https://chronicdata.cdc.gov/Behavioral-Risk-Factors/Behavioral-Risk-Factor-Surveillance-System-BRFSS-P/dttw-5yxu

[16] CDC (2019) BRFSS Questionnaires. https://www.cdc.gov/brfss/questionnaires/index.htm

[17] Mesra, B. (2019) Computer Science Lab Equipments. https://www.bitmesra.ac.in/edudepartment/content/12/317/623

[18] Alkharusi, H. (2012) Categorical Variables in Regression Analysis: A Comparison of Dummy and Effect Coding. *International Journal of Education*, **4**, Article 202. https://doi.org/10.5296/ije.v4i2.1962

[19] Kjeldsen, S.E. (2018) Hypertension and Cardiovascular Risk: General Aspects. *Pharmacological Research*, **129**, 95-99. https://doi.org/10.1016/j.phrs.2017.11.003

[20] Narvaez-Guerra, O., Herrera-Enriquez, K., Medina-Lezama, J. and Chirinos, J.A. (2018) Systemic Hypertension at High Altitude. *Hypertension*, **72**, 567-578. https://doi.org/10.1161/hypertensionaha.118.11140

[21] Lear, S.A., Hu, W., Rangarajan, S., Gasevic, D., Leong, D., Iqbal, R., *et al.* (2017) The Effect of Physical Activity on Mortality and Cardiovascular Disease in 130 000 People from 17 High-Income, Middle-Income, and Low-Income Countries: The Pure Study. *The Lancet*, **390**, 2643-2654. https://doi.org/10.1016/s0140-6736(17)31634-3

[22] Gordon, T., Castelli, W.P., Hjortland, M.C., Kannel, W.B. and Dawber, T.R. (1977) High Density Lipoprotein as a Protective Factor against Coronary Heart Disease. *The American Journal of Medicine*, **62**, 707-714. https://doi.org/10.1016/0002-9343(77)90874-9

[23] Ravnskov, U., Diamond, D.M., Hama, R., Hamazaki, T., Hammarskjöld, B., Hynes, N., *et al.* (2016) Lack of an Association or an Inverse Association between Low-Density-Lipoprotein Cholesterol and Mortality in the Elderly: A Systematic Review. *BMJ Open*, **6**, e010401. https://doi.org/10.1136/bmjopen-2015-010401

[24] Mandviwala, T., Khalid, U. and Deswal, A. (2016) Obesity and Cardiovascular Disease: A Risk Factor or a Risk Marker? *Current Atherosclerosis Reports*, **18**, Article No. 21. https://doi.org/10.1007/s11883-016-0575-4

[25] Liu, G., Hu, Y., Zong, G., Pan, A., Manson, J.E., Rexrode, K.M., *et al.* (2020) Smoking Cessation and Weight Change in Relation to Cardiovascular Disease Incidence and Mortality in People with Type 2 Diabetes: A Population-Based Cohort Study. *The Lancet Diabetes & Endocrinology*, **8**, 125-133. https://doi.org/10.1016/s2213-8587(19)30413-9

[26] Tannus, L.R.M., Cobas, R.A., Palma, C.C.S., *et al.* (2013) Impact of Diabetes on Cardiovascular Disease: An Update. *International Journal of Hypertension*, **2013**, 1-15. https://doi.org/10.1155/2013/653789