ISSN: 3065-758X
Research Article

Journal of Infectious Diseases and Viruses Research

Data-driven Approach to Age Prediction on Patients Diabetes and Cardiovascular Diseases using Machine Learning: National Health and Nutrition Health Survey (NHANES)

Irfan Abbas[1*], Syarifuddin Israil[1], Muhammad Bayu[1] and Rahmat Widia Sembiring[2]

[1]*Department of business digital. Faculty of economics and business Muhammadiyah Berau University, Indonesia*
[2]*Departement of computer and information technology, Politeknik negeri medan university, Indonesia*

[*]Corresponding author: Irfan Abbas, Department of business digital. Faculty of economics and business Muhammadiyah Berau University, Indonesia.

Submitted: 19 January 2024 **Accepted:** 25 January 2024 **Published:** 31 January 2024

https://doi.org/10.63620/MKJIDVR.2024.1018

Citation: *Abbas, I., Israil, S., Bayu, M., & Sembiring, R. W. (2024). Data-Driven Approach to Age Prediction on Patients Diabetes and Cardiovascular Diseases Using Machine Learning: National Health and Nutrition Health Survey (Nhanes). J of Infec Dise and Vir Res, 3(1), 01-10.*

Abstract

Background: *Diabetes and cardiovascular disease are two of the main causes of death in the United States. Identifying and predicting these diseases in patients is the first step towards stopping their progression. We evaluate the capabilities of machine learning models in detecting at-risk patients using survey data (and laboratory results) and identify key variables within the data contributing to these diseases among the patients.*

Methods: *Our research explores data-driven approaches which utilize supervised machine learning models to identify patients with such diseases. Using the National Health and Nutrition Examination Survey (NHANES) dataset, we conducted an exhaustive search of all available feature variables within the data to develop models for cardiovascular, prediabetes, and diabetes detection.*

Using different timeframes and feature sets for the data (based on laboratory data), multiple machine learning models (Support vector machines and adaptive boosting) were evaluated on their classification performance. The models were then combined to develop a weighted ensemble model, capable of leveraging the performance of the disparate models to improve detection accuracy. Information gain of tree-based models was used to identify the key variables within the patient data that contributed to the detection of at-risk patients in each of the disease's classes by the data-learned models.

Results: *Diabetes and cardiovascular disease (CVD) are two of the leading causes of death in the United States. Detecting and predicting these diseases in patients is the first step to halting their progression. In this study, it was used Adaptive Boosting (AdaBoost) and Support Vector Machines (SVM) together as prediction. The purpose of this study was to know whether AdaBoost SVM could produce good accuracy. Tests were conducted using 50% data training and 50% data testing. Dot kernels were used to SVM. The highest accuracy value of AdaBoost SVM was accuracy 98.54%. Therefore, it could be that AdaBoost can improve the performance of SVM in prediction of CVD disease severity.*

Conclusion: *We conclude machine learned models based on survey questionnaire can provide an automated identification mechanism for patients at risk of diabetes and cardiovascular diseases. We also identify key contributors to the prediction, which can be further explored for their implications on electronic health records.*

Keywords: Machine Learning, AdaBoost, SVM, NHANES, Diabetes and Cardiovascular Disease (CVD)

Abbreviations

Performance Vector: Accuracy: 98.54%
NHANES: National Health and Nutrition Examination Survey
AdaBoost: Adaptive Boosting
SVM: Support Vector Machine
CVD: Diabetes and Cardiovascular Disease

Introduction

Diabetes and cardiovascular disease (CVD) are two of the most common chronic diseases that cause death in the United States in 2023, approximately 9% of the U.S. population was diagnosed with diabetes, and the remaining 3% were undiagnosed. Additionally, approximately 34% were prediabetic. However, almost 90% of adults with prediabetes were unaware of their condition. Meanwhile, cardiovascular disease is the leading cause of death in the United States every year. Approximately 92.1 million adults in the United States live with some form of cardiovascular disease or stroke, with direct and indirect medical costs estimated to be more than $329.7 [1-5].

There is also a link between cardiovascular disease and diabetes. The American Heart Association reports that at least 68% of people with diabetes over the age of 65 die from heart disease. A systematic literature review by Einarson et al. 4,444 authors concluded that 32.2% of all patients with type 2 diabetes suffer from heart disease [6-8].

In a world of ever-increasing data volumes, where hospitals are gradually implementing big data systems, the use of data analytics in the healthcare system offers great advantages to provide analytical information, improve diagnoses, improve outcomes and reduce costs in particular, the successful implementation of machine learning expands the work of medical professionals and improves the efficiency of the healthcare system [6-8, 5].

Significant improvements in diagnostic accuracy have been demonstrated through the use of machine learning models in collaboration with clinicians. Since then, machine learning models have been used to predict many common diseases, including predictioning diabetes, identifying hypertension in diabetics, and predictioning patients with cardiovascular disease among diabetic patients [9]. Machine learning models can help identify patients with diabetes and heart disease. There are often a number of factors that help identify patients at risk for these common conditions. Machine learning can help identify hidden patterns in these factors that might otherwise be overlooked [5, 9-17].

In this article, we use a supervised machine learning model to predict diabetes and cardiovascular disease [2-4]. Although the association between these diseases is known, we are designing models to independently predict cardiovascular disease and diabetes to benefit a broader patient population [3].

As a result, commonalities between diseases that influence prediction ions can be identified. He also considers prediction prediabetes and undiagnosed diabetes. The National Health and Nutrition Exam Survey (NHANES) dataset is used to train and test multiple models to predict these diseases [18, 19]. This article also describes weighted ensemble models, which combine the results of multiple supervised learning models to improve prediction power [10, 20].

NHANES Data

The National Health and Nutrition Examination Survey (NHANES) is a program developed by the National Center for Health Statistics (NCHS) to assess the health and nutritional status of the United States population [6, 7, 10]. This dataset is unique in that it combines survey interviews with physical and laboratory tests performed in a medical setting. Survey data consists of socio-economic, demographic, nutritional, and health-related questions. Laboratory tests consist of medical, dental, physical and physiological measurements performed by health care professionals.

NHANES continuous data began in 1999 and is conducted annually with a sample of 4,444 to 5,000 participants. The sample uses a nationally representative sample of civilians obtained through a multistage probability sampling design. In addition to each individual's test results, the frequency of chronic diseases in the population is also recorded. For example, information regarding anemia, cardiovascular disease, diabetes, environmental pollution, eye disease, and hearing loss is collected [6-8, 20].

NHANES provides insightful data that has made important contributions to the American people. This provides researchers with important clues about the causes of disease, based on the distribution of health problems and risk factors within the population. In addition, health planners and government agencies can identify and set policies and plan research and health promotion programs to improve current health conditions and prevent future health problems [21]. For example, data from previous studies have been used to create growth curves for assessing child growth, which have been adapted and adopted as reference standards around the world. Evidence of undiagnosed diabetes, obesity epidemic, high blood pressure, and cholesterol levels has led to intensification of education and prevention programs to raise public awareness with a focus on diet and physical activity.

Material and Methods

Machine Learning Models

Machine Learning Models in our study, we use supervised learning model-based support vector machine (SVM) based Adaptive boosting (AdaBoost) to age prediction diabetes and cardiovascular disease (CVD) [18, 22]. The learning algorithm is provided with training data that includes both recorded observations and labels that correspond to categories of observations. The algorithm uses this information to create a model that can predict which output label to assign to a new observation, given a new observation. The following section briefly describes the model used in this project [13-17].

Supervised Learning

Prediction Using Ad boost Support Vector Machines $\{(\mathcal{X}1, \mathcal{X}1), \ldots, (x_m y_m)\}$ [23-26]. The purpose of this training was to build a model that can provide prediction results according to the target test data. The training data was used to build a predictive model that was applied to the test data. Supervised learning allows you to make discrete predictions called predictions. Prediction is the division of a system into groups or classes according to established rules or criteria [27].

- **Support Vector Machine (SVM)**

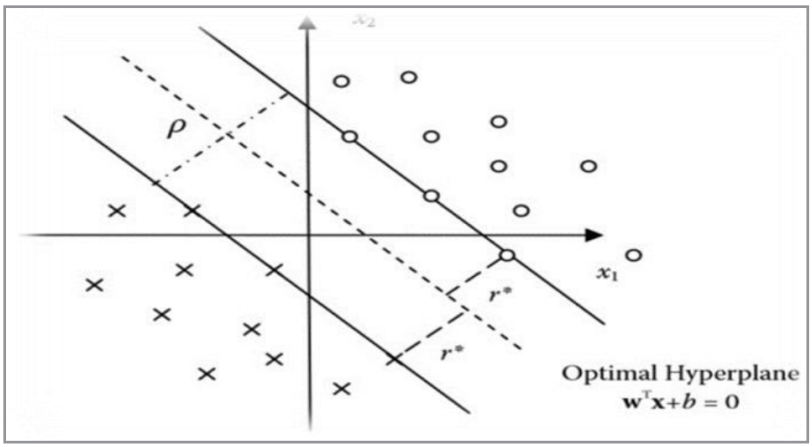


Figure 1: Optimal hyperplane illustration on SVM (Source: [28]

SVM is a supervised learning method developed by Vatnik in 1992 for prediction [29-33]. SVM aimed to solve prediction problems by forming a hyperplane that maximizes the margin. It is done by splitting two data classes. The edge was the minimum distance between the hyperplane and the nearest points (support vectors) for each class. In SVM, the optimal separating hyperplane is determined by specifying the maximum edge distance p between different classes. Given a Dataset $D = \{(x_1, y_1), \ldots, x_m, y_m\}$, $x_i \; E \; X$, $y_i \; E \; Y = \{-1, +1\}$, SVM resolved the following mathematics model:

$$\frac{min}{w,b} \frac{1}{2} \| w \|^2$$

$$s.t \; y_i(w^t x_i + b) \geq 1, i = 1, \ldots, n \qquad (1)$$

For misprediction error cases, it was added C parameter and slack variable so that SVM mathematics model becomes:

$$\frac{min}{w,b} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} \xi_i \qquad (2)$$
$$s.t \; y_i(w^t x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, n$$

With $C > 0$.

For linearly non-separable cases, it was added kernel function. Kernel function could be defined as

$$K(x_i, y_j) = \phi(x_i)^T . \phi(x_j)$$

There were various kernel functions i.e. [28]

1. Linier Kernel $K(x_i, y_j) = x^T x_j$ $_i$
2. Polynomial Kernel: $K(x_i, y_j) = (x^T x_j + \frac{}{j} 1)^d$
3. RBF Kernel $K(x_y, y_j) = \exp(-\gamma \| x_i, x_j \|^2), y > 0$

Adaptive Boosting (AdaBoost)

The AdaBoost (adaptive boosting) algorithm was proposed by Yoav Freund and Robert Shapire in 1995 [28, 34-38]. This method aimed to maintain weight distribution W of base classifier (SVM is a base classifier in this paper) iteratively AdaBoost was an ensemble method that improved the prediction results by constructing a set of classifiers and combining it. Given a dataset:

$D = \{(x_1, y_1), \ldots, x_m, y_m\}, x_i \in X, Y_i \in Y = \{-1, +1\},$

this method performs base classifier training iteratively as many cycles$t = 1, 2 \ldots. T$. Initial weight vector W^1 in this training was arranged the same as follows:

$$w_i^1 = \frac{1}{m}, i = 1,2, \ldots \ldots m$$

At each round, the weight vector would be updated until it obtained the right result. The base classifier worked to find hypothesis $h_t = \{-1 + 1\}$, for W_i the quality of hypothesis $h_t (x_i)$ measured by training error E_t as follow:

$$\varepsilon_t = \sum_{i=1}^{m} y_i \neq h_t(x_i)$$

Training error was calculated from the trained weights vector. If $E_t > 0{,}5$, then the weighting process was stopped, and iteration was not continued.

After the hypothesis was accepted. AdaBoost would determine the weight of hypothesis h_t, a_t. It was obtained $a_t \geq 0 \; if \; E_t \leq 0{,}5$ and the value of a_t would be increase since ε_t was decreased. Thus, it was formulated as follows:

$$\alpha_t = \frac{1}{2} \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$$

Then, the weight vector W^t_i were updated to.

$$w_i^{t+1} = \frac{w_i^t exp\{-\alpha_t y_i h_t(x_i)\}}{z_t} = \frac{w_i^t}{z_t} \; x \; \{exp\{-a_t\}, y_i\} = h_t(x_i)$$
$$\{exp\{\alpha_t\}, y_i\} \neq h_t(x_i)$$
$$\sum_{i=1}^{m} w_i^{t+1} = 1 \qquad (7)$$

With Z_t was a normalization constant that made

$$\sum_{i=1}^{m} w_i^{t+1} = 1 \; so \; w_i^{t+1}$$

could be distributed. The hypothesis resulted $H(x)$ based on the number of weights of T hypothesis of the base classifier as follows:

$$H(x) \; sign \left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

Table 1: AdaBoost Algoritm [39-42]

1.	Input Dataset $D = \{(x_1 y_1), \ldots, (x_m y_m)\}$, a Base Classifier algorithm, the number of cycles T
2.	**Initialize**: the weights of training samples: $w_i^1 = 1/m$, for all $i = 1, 2, \ldots, m$.
3.	**Do for** $t = 1, \ldots, T$ (a). Use the Base Classifier algorithmto get hypotesis $\boldsymbol{h_t}$ on the weighted training samples. (b). Calculate the training error of $\boldsymbol{h_t}$: $\boldsymbol{\varepsilon_t} = \sum_{i=1}^{m} \boldsymbol{w_i^t, y_1} \neq \boldsymbol{h_t(x_i)}$ (c). *if* $\varepsilon_t => 0.5$ *then stop*. (d). Set weight for the hypotesis h_t: $\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$ (e). Update the weights of training samples: $w_i^{t+1} = \frac{w_i^t exp\{-\alpha_t y_i h_t(x_i)\}}{z_t} = \frac{w_i^t}{x_t}\ X$ $\{exp\{-\alpha_t\}, y_i = h_t(x_i)\}$ Where $z_t$is a normalization constant and $\sum_{i=1}^{m} w_i^{t+1} = 1$ $\{exp\{\ \alpha_t\}, y_i \neq h_t(x_i)\}$
4.	**Output** $\boldsymbol{H}(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$

AdaBoost SVM

AdaBoost SVM was an AdaBoost method that uses SVM as the base classifier [24-26, 39, 40, 43, 44]. The algorithm for this method was similar to that in Table 1. AdaBoost performed hypothesis weighting of the SVM method to achieve higher accuracy. At each cycle, the weight of misprediction errors was increased while the weight of already correctly classified errors was decreased, reducing their potential weight in the next cycle. This process was to predict the class (label) of hypothesis h_t

Table 2: AdaBoost SVM algorithm [45, 41, 42]

1.	Input: Dataset $D = \{(x_1 y_1), \ldots, (x_m y_m)\}$, algorithm, the number of cycles T
2.	Initialize: the weights of training samples: $w_i^1 = {}^1/_m$, for all $i = 1,2, \ldots, m$
3.	Do for $t = 1, \ldots, T$ a) Use SVM algorithmto get hypothesis h_t, on the weighted training samples b) Calculate the training error of $h_t : \varepsilon_t = \sum_{i=1}^{m} w_i^t, y_i \neq h_t(x_i)$. c) If $\varepsilon_t > 0{,}5$; then stop d) Set weight for the hypotesis $h_t : \alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$. e) Update the weights of training samples: $w_i^{t+1} = \frac{w_i^t exp\{-\alpha_t y_i h_t(x_i)\}}{z_t} = \frac{w_i^t}{z_t} x$ $\{exp\{-\alpha_t\}, y_i = h_t(x_i)\}$ Where $z_t$is a normalization constant and $\sum_{i=1}^{m} w_i^{t+1} = 1$ $\{exp\{\ \alpha_t\}, y_i \neq h_t(x_i)\}$
4.	**Output** $\boldsymbol{H}(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$

Result and Discussion

Source Data

National Health and Nutrition Health Survey 2013-2014 (NHANES) [46]. Subset of age prediction conducted by the Centers for Disease Control and Prevention (CDC), collects extensive health and nutrition information from a diverse population in the United States. Datasets are huge, but often too large for specific analytical purposes. In this sub dataset, we focus on predicting the respondent's age by extracting a subset of features from the large NHANES dataset.

These selected characteristics include physiological measurements, lifestyle choices, and biochemical markers that are thought to be highly correlated with age [47]. The NHANES dataset was created to assess the health and nutritional status of senior and adult in the United States. Centers for Disease Control and Prevention (CDC), specifically through its National Center for Health Statistics (NCHS). Survey respondents throughout the United States. Data was gathered through interviews, physical examinations, and laboratory tests. Was there any data preprocessing performed? For this subset respondents 65 years old and older were labeled as “senior” and all individuals under 65 years old as “adult.”. Has no missing values. Additional Information The original full dataset can be foundat: https://wwwn.cdc.gov/nchs/nhanes/search/DataPage.aspx?Component=Questionnaire&CycleBeginYear=2013

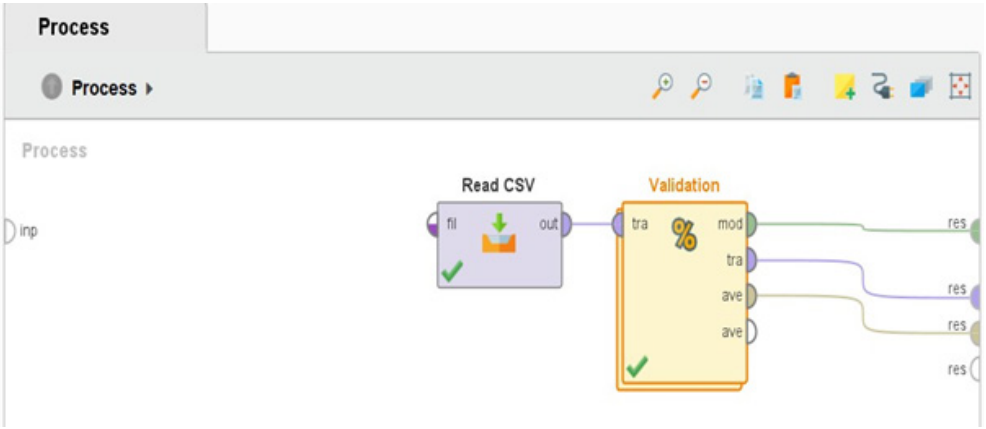
Table 3: Variable Table

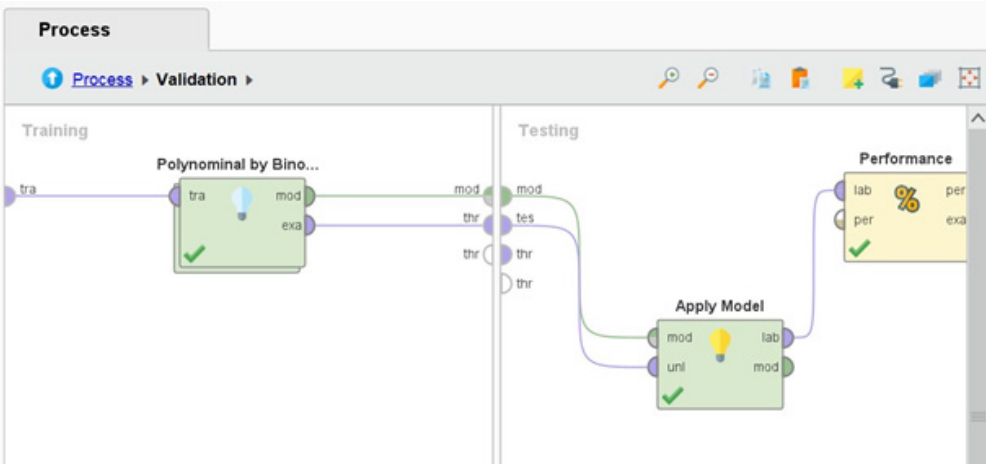
Variable Name	Role	Type	Demographic	Description	Units	Missing Values
SEQN	ID	Continuous		Respondent Sequence Number		no
Variable Name	Role	Type	Demographic	Description	Units	Missing Values
age group	Target	Categorical	Age	Respondent's Age Group (senior/non-senior)		no
RIDAGEYR	Other	Continuous	Age	Respondent's Age		no
RIAGENDR	Feature	Continuous	Gender	Respondent's Gender		no
PAQ605	Feature	Continuous		If the respondent engages in moderate or vigorous-intensity sports, fitness, or recreational activities in the typical week		no
BMXBMI	Feature	Continuous		Respondent's Body Mass Index		no
LBXGLU	Feature	Continuous		Respondent's Blood Glucose after fasting		no
DIQ010	Feature	Continuous		If the Respondent is diabetic		no
LBXGLT	Feature	Continuous		Respondent's Oral		no
LBXIN	Feature	Continuous		Respondent's Blood Insulin Levels		no

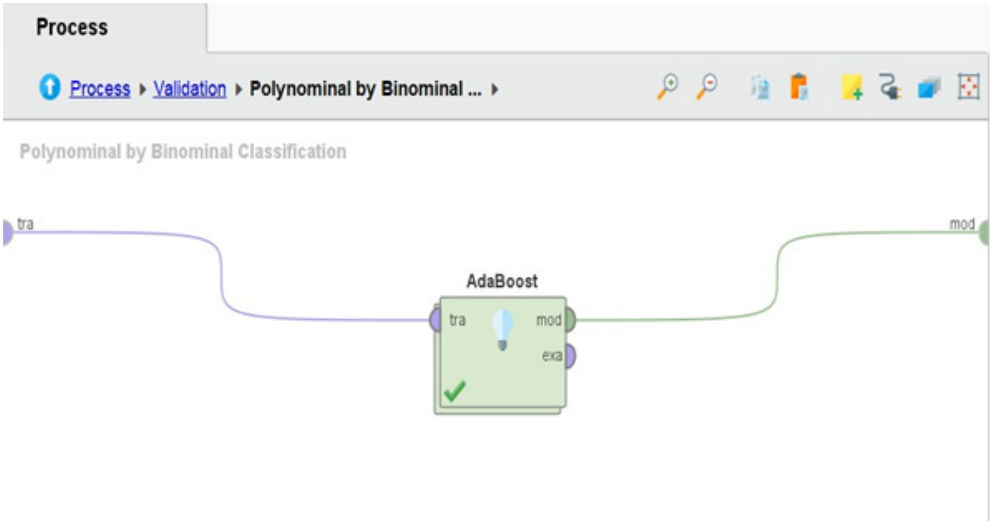
Additional Variable Information Class Labels: RIAGENDR: 1 represents Male and 2 represents Female. PAQ605: 1 represents that the respondent takes part in weekly moderate or vigorous- intensity physical activity and a 2 represents that they do not.

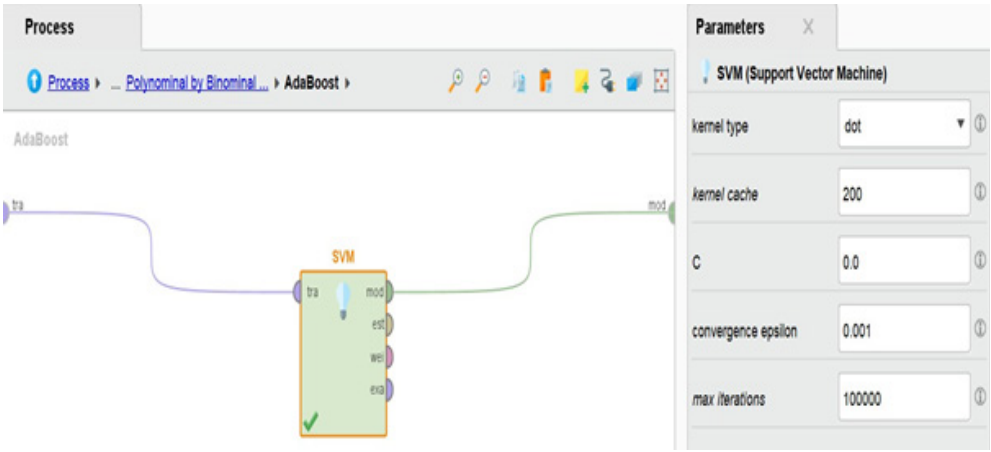
Process

In this experiment we use framework rapid miner tools [48].









Accuracy

Performance Vector:
accuracy: 98.54% Confusion Matrix:
True: Adult Senior Adult: 571 7
Senior: 3 102
classification error: 1.46% Confusion Matrix:
True: Adult Senior Adult: 571 7
Senior: 3 102
kappa: 0.945 Confusion Matrix:
True: Adult Senior Adult: 571 7
Senior: 3 102
weighted_mean_recall: 96.53%, weights: 1, 1 Confusion Matrix:
True: Adult Senior Adult: 571 7
Senior: 3 102
weighted_mean_precision: 97.97%, weights: 1, 1
Confusion Matrix:
True: Adult Senior Adult: 571 7
Senior: 3 102
spearman_rho: 0.945
kendall_tau: 0.945
absolute_error: 0.018 +/- 0.105
relative_error: 1.75% +/- 10.52%
relative_error_lenient: 1.75% +/- 10.52%
relative_error_strict: 20.68% +/- 269.77%
normalized_absolute_error: 0.021
root_mean_squared_error: 0.107 +/- 0.000
root_relative_squared_error: 0.127
squared_error: 0.011 +/- 0.087
correlation: 0.945
squared_correlation: 0.893
cross-entropy: 0.054
margin: 0.016
soft_margin_loss: 0.018
logistic_loss: 0.319

- **Visualizing of plot view the model performance figure. 2**

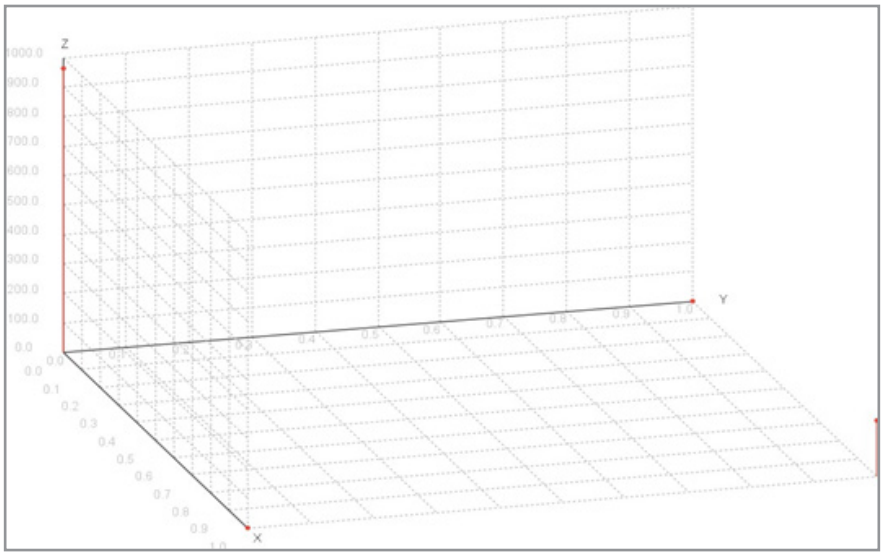


Figure 2: Plot View model performance

- **Visualizing of AUC, the model performance figure. 3**

Figure 3: AUC Optimistic

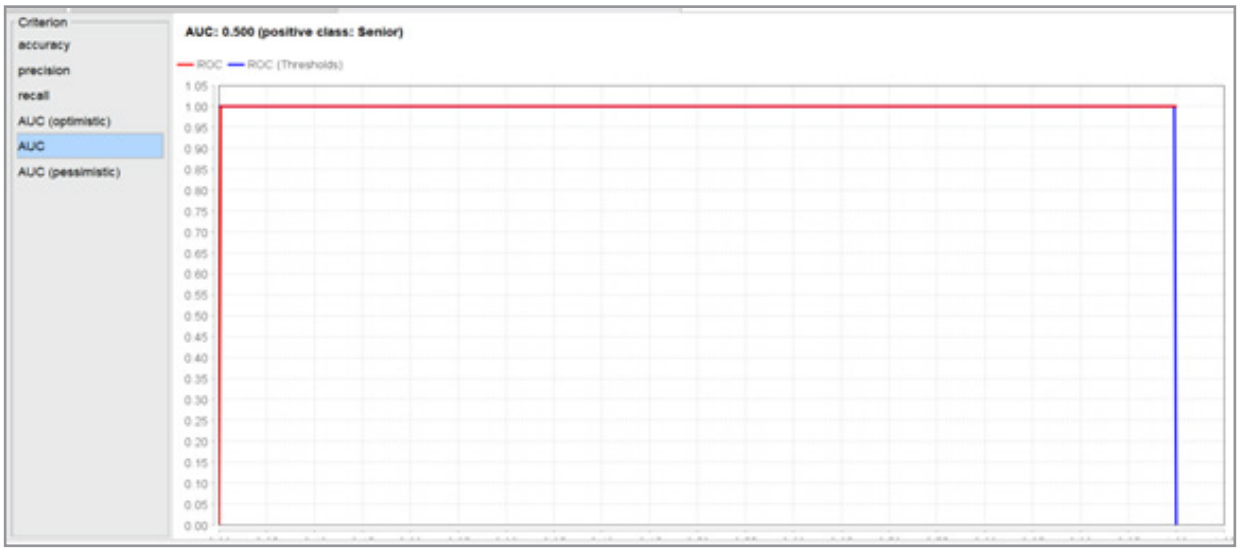


Figure 4: AUC

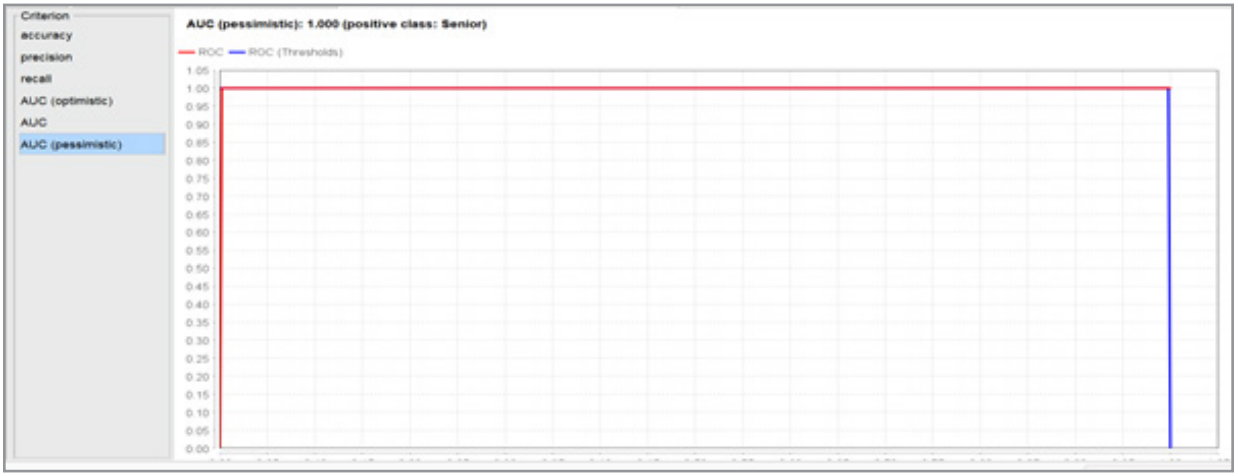


Figure 5: AUC Pessimistic

Conclusion

We conclude machine learned models based on survey questionnaires can provide an automated identification mechanism for patients at risk of diabetes and cardiovascular diseases. We also identify key contributors to the prediction, which can be further explored for their implications on electronic health records.

Acknowledgements

We are thankful for the support of University of Muhammadiyah Berau for providing funding support to conduct the study.

Authors' Contributions

IA (Irfan Abbas) conceived the research study and mentored SI (Syarifuddin Israil), MB (Muhammad Bayu) and RWS (Rahmat Widia Sembiring), IA worked on the data pre-processing and development of the machine learning model for cardiovascular diseases. SI and RWS developed SVM models for diabetes and pre-diabetes patients. MB and RWS develop AdaBoost models for diabetes and pre-diabetes patients. IA and RWS developed the framework of the paper and contributed to introduction, methodology, results/discussion, and conclusion.

Funding

The research study was supported by the Muhammadiyah Berau University – Indonesia and Politeknik Negeri Medan University - Indonesia

Availability of Data and Materials

The National Health and Nutrition Examination Survey (NHANES) continuous data used in the study is available publicly at Center Disease Control (CDC)

website at: https://www.cdc.gov/nchs/tutorials/nhanes/Preparing/Download/ intro.htm.

The documentation on how to download and use the data is provided at: https://www.cdc.gov/nchs/tutorials/NHANES/index_continuous. htm

Ethics Approval and Consent to Participate

The NHANES survey operates under the approval of the National Center for Health Statistics Research Ethics Review Board (Protocols #2005-06, and #2011- 17), available in www.cdc.gov/nchs/nhanes/irba98.htm. All the NHANES data meet the conditions described in Research Using Publicly Available Datasets (Secondary Analysis) - Policy #39 - for use without application to Institutional Review Board. All study participants provided written informed consent.

Consent for Publication

Not applicable.

Competing Interests

The authors declare that they have no competing interests.

Author Details

Department of Business Digital, University of Muhammadiyah Berau - Indonesia. Politeknik Negeri Medan - Indonesia

References

1. Centers for Disease Control and Prevention. (n.d.). National Diabetes Statistics Report.
2. Adler, A. (2021). Using machine learning techniques to identify key risk factors for diabetes and undiagnosed diabetes.
3. Imtiaz, N., & Haque, A. (2020). Predicting type-2 diabetes using machine learning and feature selection techniques. Advancement of Computer Technology and Its Applications, 3(3).
4. Abdalrada, A. S., Abawajy, J., Al-Quraishi, T., & Islam, S. M. S. (2022). Machine learning models for prediction of co-occurrence of diabetes and cardiovascular diseases: A retrospective cohort study. Journal of Diabetes & Metabolic Disorders, 21, 251–261. https://doi.org/10.1007/s40200-022-00987-5
5. Javaid, A., Zghyer, F., Kim, C., Spaulding, E. M., Isakadze, N., Ding, J., ... & Marvel, F. A. (2022). Medicine 2032: The future of cardiovascular disease prevention with machine learning and digital health technology. American Journal of Preventive Cardiology, 12, 100379. https://doi.org/10.1016/j.ajpc.2022.100379
6. Centers for Disease Control and Prevention. (n.d.). National

Diabetes Statistics Report.
7. National Center for Health Statistics. (n.d.). About the National Health and Nutrition Examination Survey.
8. American Heart Association. (n.d.). Heart attack and stroke symptoms. https://www.heart.org/idc/groups/ahamah-public/@wcm/@sop/@smd/documents/downloadable/ucm_491265.pdf
9. Joo, G., Song, Y., Im, H., & Park, J. (2020). Clinical implication of machine learning in predicting the occurrence of cardiovascular disease using big data (Nationwide Cohort Data in Korea). IEEE Access, 8, 157643–157653. https://doi.org/10.1109/ACCESS.2020.3019398
10. Dinh, A., Miertschin, S., Young, A., & Mohanty, S. D. (2019). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Medical Informatics and Decision Making, 19, 211. https://doi.org/10.1186/s12911-019-0918-5
11. Imtiaz, N., & Haque, A. (2020). Predicting type-2 diabetes using machine learning and feature selection techniques. Advancement of Computer Technology and Its Applications, 3(1), 1–10.
12. Li, J., Xu, Z., Xu, T., & Lin, S. (2022). Predicting diabetes in patients with metabolic syndrome using machine-learning model based on multiple years' data. Diabetes, Metabolic Syndrome and Obesity, 15, 2951–2960. https://doi.org/10.2147/DMSO.S361259
13. Abdalrada, A. S., Abawajy, J., Al-Quraishi, T., & Islam, S. M. S. (2022). Machine learning models for prediction of co-occurrence of diabetes and cardiovascular diseases: A retrospective cohort study. Journal of Diabetes & Metabolic Disorders, 21, 251–261. https://doi.org/10.1007/s40200-022-00987-5
14. Mayya, A., & Solieman, H. (2022). Machine learning system for predicting cardiovascular disorders in diabetic patients. Journal of the Russian Universities. Radio Electronics, 25(2), 116–122.
15. Kibria, H. B., & Matin, A. (2022). The severity prediction of the binary and multi-class cardiovascular disease: A machine learning-based fusion approach. arXiv. https://doi.org/10.48550/arXiv.2203.04921
16. Adler, A. (2021). Using machine learning techniques to identify key risk factors for diabetes and undiagnosed diabetes.
17. Yu, W., Liu, T., Valdez, R., Gwinn, M., & Khoury, M. J. (2010). Application of support vector machine modeling for prediction of common diseases: The case of diabetes and pre-diabetes. BMC Medical Informatics and Decision Making, 10(1), 16. https://doi.org/10.1186/1472-6947-10-16
18. Klados, G. A., Politof, K., Bei, E. S., Moirogiorgou, K., Anousakis-Vlachochristou, N., Matsopoulos, G. K., & Zervakis, M. (2021, November). Machine learning model for predicting CVD risk on NHANES data. In 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) IEEE, 1749-1752 https://doi.org/10.1109/EMBC46164.2021.9629575
19. Dinh, A., Miertschin, S., Young, A., & Mohanty, S. D. (2019). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Medical Informatics and Decision Making, 19(1), 211. https://doi.org/10.1186/s12911-019-0918-5
20. UCI Dataset. (2023). National Health and Nutrition Health Survey 2013-2014 (NHANES) age prediction subset.
21. Mayya, A., & Solieman, H. (2022). Machine learning system for predicting cardiovascular disorders in diabetic patients. Journal of the Russian Universities. Radioelectronics, 25(2), 116–122.
22. Yu, W., Liu, T., Valdez, R., Gwinn, M., & Khoury, M. J. (2010). Application of support vector machine modeling for prediction of common diseases: The case of diabetes and pre-diabetes. BMC Medical Informatics and Decision Making, 10(1), 16. https://doi.org/10.1186/1472-6947-10-16
23. Xu, B. (2019). Proceedings of 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference. Institute of Electrical and Electronics Engineers (IEEE), Beijing Section.
24. Charfi, I., Miteran, J., Dubois, J., Atri, M., & Tourki, R. (2013). Optimized spatio-temporal descriptors for real-time fall detection: Comparison of support vector machine and Adaboost-based classification. Journal of Electronic Imaging, 22(4), 041106.
25. Li, J., Sun, L., & Li, R. (2020). Nondestructive detection of frying times for soybean oil by NIR-spectroscopy technology with Adaboost-SVM (RBF). Optik, 206, 164356.
26. Mehmood, Z., & Asghar, S. (2021). Customizing SVM as a base learner with AdaBoost ensemble to learn from multi-class problems: A hybrid approach AdaBoost-MSVM. Knowledge-Based Systems, 217, 106845.
27. Hitam, N. A., Ismail, A. R., & Saeed, F. (2019). An optimized support vector machine (SVM) based on particle swarm optimization (PSO) for cryptocurrency forecasting. Procedia Computer Science, 163, 427–433.
28. Adyalam, T. R., Rustam, Z., & Pandelaki, J. (2018). Classification of osteoarthritis disease severity using Adaboost support vector machines. Journal of Physics: Conference Series, 1108(1), 012062.
29. Wu, Q. (2011). Hybrid forecasting model based on support vector machine and particle swarm optimization with adaptive and Cauchy mutation. Expert Systems with Applications, 38(7), 9070–9075.
30. Wan, S., Li, X., Yin, Y., & Hong, J. (2021). Milling chatter detection by multi-feature fusion and Adaboost-SVM. Mechanical Systems and Signal Processing, 156, 107671.
31. Mehmood, Z., & Asghar, S. (2021). Customizing SVM as a base learner with AdaBoost ensemble to learn from multi-class problems: A hybrid approach AdaBoost-MSVM. Knowledge-Based Systems, 217, 106845.
32. Harish, N., Mandal, S., Rao, S., & Patil, S. G. (2015). Particle swarm optimization-based support vector machine for damage level prediction of non-reshaped berm breakwater. Applied Soft Computing, 27, 313–321.
33. Wang, R. (2012). AdaBoost for feature selection, classification and its relationship with SVM: A review. Physics Procedia, 25, 800–807.
34. Lan, Y., Zhang, Y., & Lin, W. (2023). Diagnosis algorithms for indirect bridge health monitoring via an optimized AdaBoost-linear SVM. Engineering Structures, 275, 115239.
35. Belghit, A., Lazri, M., Ouallouche, F., Labadi, K., & Ameur, S. (2023). Optimization of One versus All-SVM using AdaBoost algorithm for rainfall classification and estimation from multispectral MSG data. Advances in Space Research, 71(4), 946–963.
36. Li, R., Zhang, H. (2022). State of health and charge es-

timation based on adaptive boosting integrated with particle swarm optimization/support vector machine (AdaBoost-PSO-SVM) model for lithium-ion batteries. International Journal of Electrochemical Science, 17, 220212.
37. Li, R., Li, W., & Zhang, H. (2022). State of health and charge estimation based on adaptive boosting integrated with particle swarm optimization/support vector machine (AdaBoost-PSO-SVM) model for lithium-ion batteries. International Journal of Electrochemical Science, 17, 220212.
38. Wei, F. S., Wang, W. J., Miao, Y., Tu, J., & Liu, C. L. (2009). Particle swarm optimization-based support vector machine for forecasting dissolved gases content in power transformer oil. Energy Conversion and Management, 50(6), 1604–1609.
39. Wan, S., Li, X., Yin, Y., & Hong, J. (2021). Milling chatter detection by multi-feature fusion and Adaboost-SVM. Mechanical Systems and Signal Processing, 156, 107671.
40. Li, X., Wang, L., & Sung, E. (2008). AdaBoost with SVM-based component classifiers. Engineering Applications of Artificial Intelligence, 21(5), 785–795.
41. Adyalam, T. R., Rustam, Z., & Pandelaki, J. (2018). Classification of osteoarthritis disease severity using Adaboost support vector machines. Journal of Physics: Conference Series, 1108(1), 012062. https://doi.org/10.1088/1742-6596/1108/1/012062
42. Wang, R. (2012). AdaBoost for feature selection, classification and its relationship with SVM: A review. Physics Procedia, 25, 800–807. https://doi.org/10.1016/j.phpro.2012.03.141
43. Wei, F. S., Wang, M. J., Miao, Y. B., Tu, J., & Liu, C. L. (2009). Particle swarm optimization-based support vector machine for forecasting dissolved gases content in power transformer oil. Energy Conversion and Management, 50(6), 1604–1609. https://doi.org/10.1016/j.enconman.2009.02.007
44. Zhang, X., & Ren, F. (2008). Improving SVM learning accuracy with Adaboost. In Proceedings of the 4th International Conference on Natural Computation (ICNC) (pp. 221–225). IEEE. https://doi.org/10.1109/ICNC.2008.855
45. Amami, R., Ben Ayed, D., & Ellouze, N. (2013). Adaboost with SVM using GMM supervector for imbalanced phoneme data. ResearchGate, Article ID 6577843.
46. National Center for Health Statistics. (2014). 2013-2014 Questionnaire Data - Continuous NHANES. Centers for Disease Control and Prevention (CDC). https://www.cdc.gov/nchs/nhanes/index.htm
47. UCI Machine Learning Repository. (2023). National Health and Nutrition Health Survey 2013-2014 (NHANES) Age Prediction Subset. https://archive.ics.uci.edu/dataset/887/national+health+and+nutrition+health+survey+2013-2014+(nhanes)+age+prediction+subset
48. RapidMiner. (n.d.). RapidMiner framework tools. https://rapidminer.com/

Copyright: *©2024 Irfan Abbas, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.*