ASEE AMERICAN SOCIETY FOR ENGINEERING EDUCATION

Paper ID #35700

Applying Supervised Machine Learning Algorithms to Detect Cardiac Events

Eileen Deng, Rye Country Day School

Eileen Deng is a junior at Rye Country Day School, Class of 2023. Her areas of interest include many fields within science such as psychology–especially in personality–sociology, and computer science.

Eunice Lee, Townsend Harris High School

Eunice Lee is a senior at Townsend Harris High School, class 2022. She has various interests within engineering and computer science, primarily in machine learning and finite element analysis.

Daniel Shameti, Midwood High School, Brooklyn, NY

Daniel Shameti – Senior at Midwood High School, Brooklyn, NY. He is part of the Medical Science Program/Research track at Midwood High School. His interests are in biochemistry and research in the medical field.

Dr. Yu Wang, New York City College of Technology

Dr. Wang received a doctoral degree in Electrical Engineering from the CUNY Graduate Center and joined the Department of Computer Engineering Technology at New York City College of Technology in 2009. Her research areas of interest are in engineering education, biomedical sensors, optoelectronics, modeling real-time systems, embedded system design, deep neural network and machine learning.

Applying Supervised Machine Learning Algorithms to Detect Cardiac Events

Eileen Deng[1], Eunice Lee[2], Daniel Shameti[3], Yu Wang[4*]

Abstract

Within the realm of machine learning, numerous research advancements have enhanced the understanding of data analytics and prediction models. One of the more recent achievements in artificial intelligence is the rise of machine learning in healthcare, aiding in the development of streamlined treatment and diagnosis. Cardiac focus in this paper is due to an interest in how the pandemic restricted extracurriculars and athletics in school, which led to a decrease in physical activity in adolescents. With a decrease in physical activity, the cardiac systems of students might have weakened thus fostering an interest in applying machine learning to cardiac health in adolescents. By using wearable devices and mobile devices to collect data from participants (mainly adolescents), machine learning algorithms can be applied to the data and then analyzed to get information about the cardiac states of adolescents. Cardiac features were measured using the YAMAY Smart Watch wearable device; a variety of supervised machine learning algorithms (KNN, Naïve Bayes, Random Forest, and Decision Trees) were used to predict the expected data with the target data. Overall, after testing each of the supervised machine learning algorithms, Random Forest had the best prediction accuracy of 75.86%. With these results in mind, research focusing on applying supervised machine learning algorithms to detect cardiac events would benefit from using Random Forest.

Keywords– machine learning algorithms, healthcare, prediction models, cardiac events, physical activity in adolescents

Introduction

Machine Learning is a branch of artificial intelligence that uses computer algorithms with various types of datasets to process, train, and test the algorithm itself. As the algorithm prediction accuracy improves, it is then able to automatedly process data and reach a target conclusion based on the datasets used. Machine learning is especially effective in fields where an abundant amount of data is available; more data allows the algorithm to train itself to have more accurate predictions. This can be seen in healthcare and medicine, where large amounts of medical records and daily collected data are outlets for algorithms to prove themselves effective and have accurate predictions [1]. There are several different subtypes of machine learning: supervised, semi-supervised, unsupervised, and reinforcement, to name a common few. Supervised learning utilizes labeled datasets to train algorithms and then classify the data or predict target data. Some examples of algorithms associated with supervised learning include Support Vector Machine (SVM), Naïve Bayes, K-Nearest Neighbor (KNN), Random Forest, and Decision Tree. In healthcare, machine learning provides for several innovations that enhance present treatments and diagnosis. Currently, focus is on attempting to diagnose diseases and detect abnormalities from data and different types

[1] High school student attending Rye Country Day School, Rye, NY
[2] High school student attending Townsend Harris High School, Queens, NY
[3] High school student attending Midwood High School, Brooklyn, NY
[1, 2, 3] Precollege students are listed alphabetically by last name
[4*] Corresponding author, a faculty member of Computer Engineering Technology Department
New York City College of Technology, Brooklyn, NY

of imaging scans with the additional use of biometrics and deep learning [2]. An example is the use of KNN algorithm in diagnosing heart disease patients at higher accuracy than neural network ensembles [3]. If highly reliable algorithms produce consistent results of high accuracy, it would create automation of a task thus removing the possibility of human error. This automation will allow healthcare workers to have more time to do tasks that require more attention or manual work [4]. Despite the many advancements in machine learning, the goal of automation and official use of algorithms in healthcare is far from being achieved.

Our research that inspired in adolescents and healthcare began with an increasing rise in machine learning for real world applications, especially adolescents in healthcare, as seen the restriction in extracurriculars in school due COVID-19. Cardiac focus is due to an interest in how the pandemic restricted extracurriculars and athletics in school, which led to a decrease of physical activity in adolescents. With a decrease in physical activity, the cardiac systems of students might have weaken thus fostering an interest to find the most accurate algorithm to predicate the cardiac events by using machine learning algorithms to report varying levels of accuracy.

A growing number of fitness devices also prompted this research toward healthcare since gathering and tracking data is more accessible than ever. In our research, by using wearable devices and mobile devices to collect data from participants (mainly adolescents), machine learning algorithms can be applied to the data and then analyzed to get information about the cardiac states of adolescents. Features of the collected participant data include those gathered from surveys (age, biological sex, body mass index) and those monitored through a wearable device (heart rate, blood pressure, blood oxygen). After this collection, the data will be pre-processed and then imported into an integrated development environment (IDE). Then, a variety of supervised learning algorithms (Naïve Bayes, KNN, Random Forest, and Decision Tree) will be applied to the training dataset. After, the algorithm will be analyzed by the testing set to determine which algorithm is most accurate in predicting the cardiac event of adolescents.

Our approach and experiment setup

A. Build a dataset and data visualization

To conduct the experiment, a combination of physical and digital tools is utilized to record, collect, and process the data. Raw data is recorded by participants through the YAMAY SmartWatch SW023 (Fig. 1) and collected by using a shared excel file whose members include those willing to participate in the experiment. Then the algorithms selected predict the targets from the features in the data collected and which are rated for their accuracy.

Figure 1 YAMAY SmartWatch SW023 used in the project

The YAMAY SmartWatch measures the heartrate, blood pressure, and blood oxygen of the participants before and after an activity. Inside the YAMAY SmartWatch, an optical heart rate monitoring (OHRM) uses a photoplethysmorgram (PPG) sensor to detect changes in blood volume by measuring the amount of light that is reflected or absorbed by the blood vessels [5]. The PPG sensor can discern pulsatile blood volume (which is related to blood volume changes in the arteries and heartbeat) and non-pulsatile volume (which is related to basic blood volume, respiration, sympathetic nervous system, and thermoregulation) [6]. Within the YAMAY SmartWatch, a green light (565 nm) is used for the reflective PPG sensors penetrating the tissue to measure heartrate and blood pressure. The YAMAY SmartWatch also uses a red light (610 nm) for its transmissive PPG sensor that can reach further into the tissue to detect blood oxygen [7]. If the values for the features (heartrate, blood pressure, and blood oxygen) are above or below the typical range for adolescents (60-100 resting bpm, <120/80 mmHG, 95%-100% blood oxygen), it may be considered a risk factor that indicates a greater likelihood of developing cardiovascular disease in adulthood [8][9].

Additional features measured for specificity reasons include body mass index (BMI), biological sex, weight (kg), height (cm), type of activity, time interval of activity, time of day, the activity took place, etc. Some variables may or may not have been recorded given circumstances, availability, or willingness. BMI is also considered as an indication of cardiovascular disease risk, along with heartrate, blood pressure, and blood oxygen; however, it is important to consider the distinction of normal cardiac health between sexes as blood pressure in females may be lower than in males [10][11]. With these considerations in mind, a list of features of the dataset to made into targets for classification is included: weight (kg), height (cm), gender, BMI, heartrate, blood pressure, blood oxygen, and activity state (rest, exercise, and eating). A total of 142 samples were include in the dataset, 46.5% female (n=66) and 53.5% male (n=76). In Figure 2, on the horizontal axis, the second number in each label is given a number 0, 1, or 2; 0 represents rest, 1 represent exercise, and 2 represents eating. From the top leftmost to the bottom rightmost, it shows that the number of collected samples vs different weight, height, sex, systolic blood pressure (SYSBP), blood oxygen (BO), and activity state (Target).

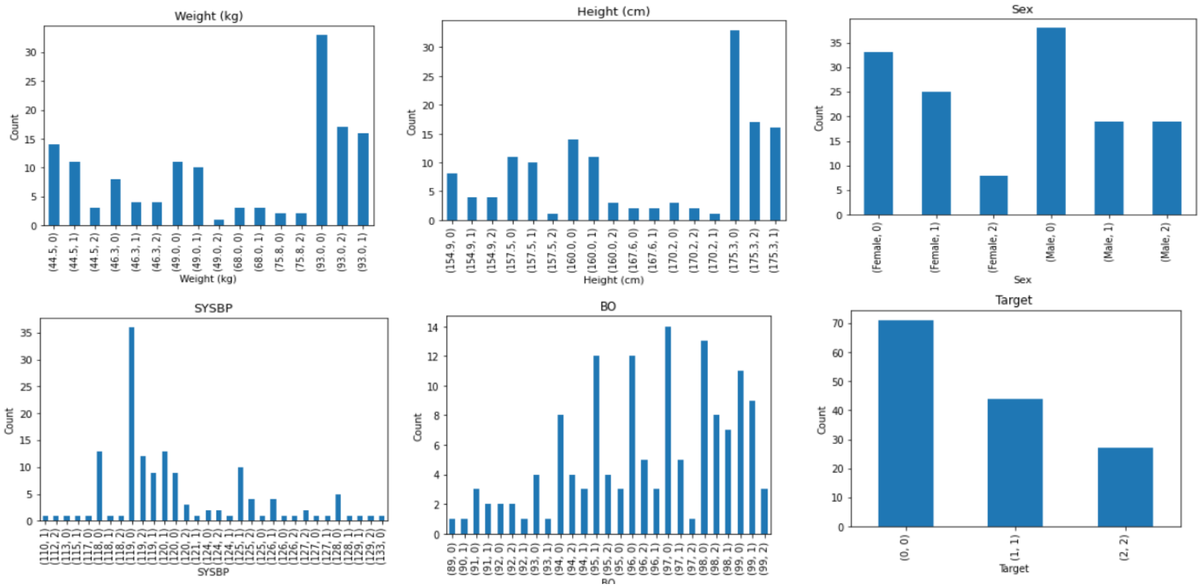

Figure 2 Partial data visualization from our dataset

B. The procedure to applying supervised algorithms

A total of four supervised algorithms were studied to compare the accuracy score of each: Naïve

Bayes, Decision Tree, Random Forest, and k-Nearest Neighbors. The performances of these algorithms are summarized in [12]. The Naive Bayes classifier is a probabilistic model that predicts a probability distribution from inputs based on the Bayes theorem. When the Bayes theorem uses independence assumptions in which the predictors and features are independent, this situation is called Naive. The algorithm performs well with categorical data, but poorly with numerical data in the training set. In classification models, Decision Tree is used in data mining to glean information from a large dataset. It has a tree structure consisting of nodes and branches, representing features in a category to be classified and the outcomes that lead to additional nodes, respectively. The Random Forest algorithm is a probabilistic classifier that has an ensemble of random trees, which is a decision tree drawn at random from a set of possible trees. Regarding classification error, high correlation between the trees in the forest leads to a higher error rate. A particular algorithm that will be used is KNN, K-Nearest Neighbor, a classification algorithm for supervised learning. This algorithm categorizes the samples by comparing the data points by a k number of its nearest data point neighbors. The k number of neighbors vote for a category in which the data point will be categorized into. The Table 1 shows how to apply these algorithms to our dataset to detect cardiac events.

Table 1 The steps of the experiment

Collect data from participants.
Data Preprocessing (eliminating data samples that were outside of the age range, reducing number of features).
Upload preprocessed data csv into Jupyter local host server cache. Open a new notebook. (Jupyter is a browser-based IDE- integrated development environment).
Import libraries.
Import dataset.
Preprocessing and data visualization (Setting target value to activity state, in which 0=resting, 1=activity, 2=eating).
Label encoding string values to float values (Transforming string categorical data to numerical data, in which female=0, male=1).
The data is split and scaled into 80% of the data being trained and the remaining 20% being tested
Algorithms and multiple models trained (Naive Bayes, Random Forest, Decision Tree, KNN)
Predicating and accuracy comparation of Naive Bayes, Random Forest, Decision Tree, KNN

C. Result

It is difficult to know in advanced which machine learning model(s) will perform better for a given dataset. After testing the four algorithms, the accuracy score of each is indicated in Table 2 to show which algorithm has the highest accuracy (highlighted green) and the lowest accuracy (highlighted orange). Overall, after testing out four different supervised machine learning algorithms, the random forest algorithm was the most accurate in determining the activity event of adolescents based on their cardiac data with an accuracy score of 75.86%. The accuracy scores of other algorithms were the similar to the random forest algorithm, with the decision tree algorithm and KNN algorithm (k=3) having the same accuracy score of 72.41%. The worst performing algorithm was the Naive Bayes, with an accuracy score of 62.07%.

Table 2 Accuracy Comparation of Supervised Learning Algorithms

Algorithms	Accuracy Score
Naive Bayes	62.07%
Decision Tree	72.41%
Random Forest	75.86%
K-Nearest Neighbor (k=3)	72.41%

With these results in mind, research focusing on the cardiac health of adolescents would benefit the most in using the random forest algorithm when doing supervised machine learning classification studies. While there are a few modified versions of each algorithm studied that may perform with a greater accuracy with the given data, the outcome of the random forest algorithm proving the highest accuracy nevertheless shows how future adolescent cardiac studies can delve into modifying the random forest algorithm to perform at an even higher level.

Discussion and Reflections

Much knowledge on the basics and fundamentals of machine learning and how it works was gained through the execution of this project. Specific knowledge on how machine learning is a type of artificial intelligence that splits datasets into a training data (which improves the algorithm) and testing data (which tests the prediction accuracy of the algorithm). To add on, we learned there are many types of algorithms that might use statistical or mathematical techniques based on their coding. Each algorithm has an area where it is best suited for, therefore no algorithm is fundamentally better than another. Along the way, we also learned specific parts and processes of Python since machine learning algorithms are mainly done through coding. Also, as our research and dataset focused on healthcare event data, we learned a bit about the healthcare system and sensor technology. Lastly, we learned machine learning works very well with healthcare data and how this relationship can be incorporated into diagnosis of diseases. Learning all this new information is useful because machine learning can be applied to all sorts of fields. As expertise in machine learning algorithms and Python increase, autonomous self-diagnosis may one day be a reality. For us, knowledge in Python and machine learning is especially beneficial for research projects we may work on in the future. Through research papers, articles, videos, and lectures, we gained much knowledge and information which helped in the creation of this research project. With the help of past research studies on applying machine learning in healthcare data, we had a better understanding in how to apply machine learning.

Acknowledgements

We would like to thank Matthew Brittain, Abiola Olatunde, Fadimatou Mossi, and Kadidjatou Mossi for their dedication and great efforts. This work is partially supported by PSC-CUNY Award # 63382-00 51, New York State Collegiate Science and Technology Entry Program (CSTEP), and the College Now STEM Research Academy at City Tech.

References

[1] G. Sasubilli and A. Kumar, "Machine Learning and Big Data Implementation on Health Care data," 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), 2020, pp. 859-864

[2] M. Ferdous, J. Debnath and N. R. Chakraborty, "Machine Learning Algorithms in Healthcare: A Literature Survey," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020, pp. 1-6

[3] Shouman, Mai & Turner, Timothy & Stocker, Rob. (2012). Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients. International Journal of Information and Education Technology

[4] Y. Verma and S. Tayeb, "Evaluation of Machine Learning Architectures in Healthcare," 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), 2021, pp. 1377-1382

[5] Elgendi, M., Fletcher, R., Liang, Y. et al. The use of photoplethysmography for assessing hypertension. npj Digit. Med. 2, 60 (2019)

[6] Utami, N., Setiawan, A. W., Zakaria, H., Mengko, T. R. & Mengko, R. Extracting blood flow parameters from Photoplethysmograph signals: A review. In The 3rd International Conference on Instrumentation, Communications, Information Technology, and Biomedical Engineering 403–407

[7] Cui, W., Ostrander, L. E. & Lee, B. Y. In vivo reflectance of blood and tissue as a function of light wavelength. IEEE Trans. Biomed. Eng. 37, 632–639 (1990)

[8] Riley M, Bluhm B. High blood pressure in children and adolescents. Am Fam Physician. 2012 Apr 1;85(7):693-700. PMID: 22534345.

[9] Kobayashi, Masaru MD*; Fukuda, Shinya MD; Takano, Ken-ichi MD, PhD; Kamizono, Junji MD; Ichikawa, Kotaro MD Can a pulse oxygen saturation of 95% to 96% help predict further vital sign destabilization in school-aged children?, Medicine: June 2018 - Volume 97 - Issue 25 - p e11135

[10] Syme C, Abrahamowicz M, Leonard GT, et al. Sex Differences in Blood Pressure and Its Relationship to Body Composition and Metabolism in Adolescence. Arch Pediatr Adolesc Med. 2009;163(9):818–825

[11] R. Katarya and P. Srinivas, "Predicting Heart Disease at Early Stages using Machine Learning: A Survey," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020, pp. 302-305