Journal Homepage: - www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI: 10.21474/IJAR01/17783
DOI URL: http://dx.doi.org/10.21474/IJAR01/17783

RESEARCH ARTICLE

PREDICT2PROTECT - MACHINE LEARNING APPLICATION IN THE PREDICTION OF HEART DISEASE

Ankita Mandal

Manuscript Info

Manuscript History
Received: 28 August 2023
Final Accepted: 30 September 2023
Published: October 2023

Abstract

Across the world, there are few universal scenarios, but the pain of losing a loved one to heart disease is an exception and a reality shared by millions every year. Heart disease is the greatest killer in society today, and one prevalent root of this issue is untimely diagnosis, often caused by the unsustainable costs and lack of accessible healthcare for underserved populations. Recognizing these disparities, the goal of this project was to create an easily available application and interface for all that accurately indicate one's risk of heart disease. To address this, a machine learning model, Predict2Protect, was built in Python. An open-source dataset compiled of 1025 patients of diverse backgrounds was scaled, adjusted to include inquiries answerable by patients, and split into 75% train, 15% validation, and 25% test. Four models were tested with the hypothesis that if the RandomForestClassifier was used, it would have the highest validity. This was not supported as the Decision Tree model had a 100% accuracy for training data and 95% for test data. Through the application software Streamlit, this program was processed into a web application now found in browser extensions. The application reports the risk of one having heart disease with a 95% accuracy and describes the risk percentage in their developing heart disease within the next year. With a simple interface and high accuracy, Predict2Protect aims to provide a view into one's health with the goals of accessible heart disease prediction and early treatment for patients around the world.

Introduction:-

Every 34 seconds, another individual's life is claimed by heart disease, a condition characterized by threatening numbers and one that has increasingly become a killer at large. As the leading cause of death throughout the world, heart disease has claimed innumerable lives and continues to at this very moment, a reality faced by families around the world that have been plagued furthermore by acute shortcomings in the form of accessible healthcare resources. The availability and widespread distribution of healthcare in third-world and developing countries is scarce, and the threat of unreliable medical care is only exacerbated by issues of poverty and financial instability. Even within developed countries, healthcare is often infringed upon by the burdens of inadequate scheduling and unaffordable costs for the average citizen (Ratnam et al, 2014) .

Without proper screening and regular medical visits, early detection and treatment of heart disease, one of the prime methods of diminishing the risks the condition carries, the numbers of people who succumb to heart diseases are at a

Corresponding Author:- Ankita Mandal

much higher rate than what they could be with modern technology. Recognizing the potential in providing accessible care for everyone, this project was created with the aim of predicting heart disease using the same indicators identified by healthcare professionals. In this way, those with a family history of heart disease, pre-existing conditions, a lack of proper access to medical care, and even just a yearning to assess their health will have the ability to do so, a measure with the potential of saving one's life.

By utilizing existing datasets containing thousands of patients' data from diverse backgrounds, four different data-fit models were used to create a machine learning program for heart disease prediction and protection. Therefore, the hypothesis was as follows: if four fit-models are used to create a machine learning framework for heart disease prediction, then RandomForestClassifier will be the most accurate and successful. This was believed because of RandomForestClassifier's ability to dissect larger dataset with greater accuracy in similar projects where trial points of up to three thousand participants were utilized. Of all the data-fit models tested in these other projects, which included KNN and SVM that are to be tested in this project as well, RandomForestClassifier was able to take into account the most significant variabilities of the data and respond with the greatest accuracy (Fadnavis et al, 2021).

A simple input of personal data from a patient is incredible in its capacity to help one determine their medical standing in terms of heart disease and decide what steps they would like to take since many manifestations of early heart disease are detectable through symptom description. When given a platform to voice these symptoms, detection becomes much simpler. As such, the variables to be considered in this platform encompass the extent of the issue by considering the independent variable as the fit-model used to frame the program. Specifically, KNN, SVM, DecisionTree, and RandomForestClassifier were tested as they were found to be the most accurate and well-fitting models for data with specific characteristics and varied scales as applied in the utilized dataset. Each model has its own rate for success and includes varying degrees of fit. Based on these measures, the program will interpret user input in varying manners and create models of different accuracies. It is also vital to consider the true meaning of each of the result diagnostics from testing different fit-models since, even if they are high, they may not be accurate in representing the true nature of the model. For example, one of the issues of a supposedly high-percentage fit model is over-fitness, or when the program is too specific to a data set, and when this occurs, a false sense of accuracy can also occur. To prevent such issues, the data will be split into test and train sets with an extra validation set for confirmation of the initial train data. The result of the independent variable changing is a greater overall difference in accuracy of the model, and this is considered the dependent variable. As the accuracy of the model fluctuates in the presence of various models, the program can be evaluated for its applicability in daily life to understand if it has the potential to properly assess one's heart health. If the percentage for accuracy is low, then the model will indicate lesser implementability as it will not be as informative to a patient and may incorrectly skew self-perceptions of one's health with false interpretations of data.

A control in this project could be considered the results of each model's validation set of data as this set the basis of comparison between the original train data and testing data to further shape the model and increase its fit, but there was not a true control group because no model could be considered a plain fit model that could be used against all others. However, to counteract this deficiency, 1025 repeated trials were conducted for maximum accuracy and allowance for the greatest fit of the data and the factors considered, number of repeated trials, scaling of data, and amount of data separated into test, train, and validation were all kept constant to maintain consistency in the models. These data points were collected with diversity in race, gender, and age, which are represented as user input options that will allow the model to take into account historical trends for varying backgrounds. This data will be kept confidential and can be left blank with a notification that this may change the accuracy of their results.

As a web application, this software will be accessible and recommend medical help to patients if necessary. Heart disease has taken the lives of many, but as technology evolves, humans' response to it must as well, and this project takes a stride in that direction. With the development of a smart model that will only become more accurate and efficient with further use, this project can be used well into the future as countries and people grow into more equitable places where the true needs of healthcare are met system wide. This application will facilitate a greater approach by people to healthcare as they will receive medical support without the many present-day burdens that it brings. The opportunities of new technology in healthcare are beginning to change medicine for the general public, and this project strives for this vision.

Methods and Materials:-

To create a mobile application with easily imputable patient data, first, a dataset had to be acquired. Using the platform Kaggle, a dataset was found from user David Lapp, and this dataset was selected with certain characteristics that indicated reliability. Firstly, the data was compiled from four widely varying areas': Cleveland, Hungary, Switzerland, and Long Beach. This allowed for variation in the data since concentrating results to a certain group would limit the data. This dataset had over 365,000 views, 60,000 downloads and was published just four years ago, indicating a high efficacy rate. Libraries such as NumPy, Matplotlib, pandas, and Scikit-learn were imported into Jupyter Notebook as the program was written in Python.

Image 1:- Correlation Matrix for All Factors.

The dataset of 1025 patients of various ages, genders, and backgrounds was then separated into training data and testing data, from which the information was fed into a machine learning model that incorporated 13 factors referenced in the image above could be measured, 4 of which can be found using an ECG and were therefore excluded from the final product due to lack of access to proper equipment in the majority of areas. The remaining factors are age, sex, chest pain type, resting blood pressure, cholesterol levels, blood sugar, resting heart rate pain levels, maximum heart rate pain levels, and exercise-induced pain levels. The pain levels were all standardized to a scale of 10 based on possible responses, and this was reflected in the overall scaling process. For example, moderate heart pain in response to exercise-inducedation could be selected in a dropdown menu, and in the program, this was interpreted as 5 on a scale of 1-10. With the use of EDA, or Exploratory Data Analysis, the data was separated and the model was run. The above correlation matrix was created to assess the validity of each factor, and it was found that each attribute contributed with validity since matrix scores varied significantly.

Then, the data was scaled with the code shown in the image below to allow the model to understand the significance of different values, such as when Boolean values were entered versus standard integer inputs. To determine it was as sufficient and indicative as possible, the data was split into X and Y sets first, where X contained all the attributes tested upon to find Y, the data containing whether or not one had heart disease. From here, the data was further split into 75% training and 25% testing data to ensure that the model would have new inputs to test once the model had been trained. To ensure the training data was also adequately tested, this X set was also split so as to contain a validation set with which the model could be checked against once more after the initial training.

The models were then run on the training, validation, and testing set chronologically and analyzed for the highest accuracy rate. Refining of each model was performed accordingly, such as post-pruning for the DecisionTree model. Then, the model to be used in the program was decided to be the DecisionTree model due to its high accuracy rate from classification reports run within the program. T-test statistical analysis was performed in support, taking into account accuracy averages of all the models and proving the statistical significance of DecisionTree.

Image 2:- Standard Scaling Code with Initial Data Set for Factor 4.

```
import seaborn as sns
import numpy as np
dataset = pd.read_excel("heart_disease.xlsx")
sort_dataset = np.sort(dataset)

Q1 = np.percentile(sort_dataset, 25, interpolation = 'midpoint')
Q2 = np.percentile(sort_dataset, 50, interpolation = 'midpoint')
Q3 = np.percentile(sort_dataset, 75, interpolation = 'midpoint')

print('Q1 25 percentile of the given data is, ', Q1)
print('Q1 50 percentile of the given data is, ', Q2)
print('Q1 75 percentile of the given data is, ', Q3)

IQR = Q3 - Q1
print('Interquartile range is', IQR)
```

```
Q1 25 percentile of the given data is,  0.0
Q1 50 percentile of the given data is,  1.0
Q1 75 percentile of the given data is,  56.0
Interquartile range is 56.0
```

Following this, an extra measure to transform this model from a program to a widespread application was made since this model was meant for more accessible use. With this considered, a mobile application was decided upon as it can be opened through any browser rather than reliant on a mobile device, which people may not have access to. Whether one has a mobile device, tablet, or even laptop accessible through a shared space, this application can be reached in diverse areas, as is the purpose of this study.

Using a pickle file, the program containing the machine learning model was then imported to Spyder for processing into a mobile application. The interface of the application was designed with the software Streamlit, which was imported into Spyder as well. From here, a series of inquiries requesting user input were formulated based on each attribute from the dataset, and a user's input of their health will determine their outcomes. An additional probability function was added to the program in order to process the development of heart disease within a year of input. This was based on the use of the time progression model from StatsModels library that used the correlation of age, symptoms and positive heart disease outcomes, and by considering one's age one year after their inputted age combined with symptoms, the probability for their risk for positive heart disease was indicated with a percentage. Input was designed to be as basic and universally understood as possible with only integer and dropdown inputs, as well as captions describing further detail where needed. All inputs were only saved in the program to enhance its accuracy, but were not saved in any other cache or capacity. Personal data, such as name, email, or further details to easily identify a user were not inquired at all in this program to maintain maximum privacy.

Results:-

When each of the decision techniques for the data were run, there was a slight difference between the average accuracy rate of the model between training and testing data, and this is to be expected as the data fluctuates and increases the expanse of the model's capabilities. However, there was a clear forerunner despite this variation. Referencing figure 1 below, the Decision Tree model was observed to have the highest validity of all the models with a 100% accuracy with training and validation data sets, and when the testing set was factored in, the accuracy dipped ever so slightly to 95%. SVM, KNN, and RandomForestClassifier had accuracy rates of 69%, 84%, and 77% respectively with only the training data considered, which clearly revealed the Decision Tree model as most reliable.

Figure 1:- Classification Report of Decision Tree Model for Training and Validation Data.

Decision Tree Model Analysis			
	Precision	Recall	F1-Score
	1	1	1
	1	1	1
Accuracy			1
Macro avg.	1	1	1
Weighted avg	1	1	1

From this analysis, the Decision Tree model was able to determine the requirements to qualify for a person who was at risk for heart disease and differentiate these results from one who was not at the same risk. There is an issue of over-fit present with the validation data factored in to the above table's results, but this was counteracted by the consideration of testing data that reduced the model's accuracy to the aforementioned 95%, a probability level that is based on thousands of trials and relays a high degree of reliability. However, the work could not stop here as this model was meant for more accessible use than an encoded program, and with this considered, the process of converting this program into a usable mobile application ensued. As for the other models, there were consistent discrepancies in the classification reports conducted for data analysis that lowered accuracy.

Figure 2:- Classification Report of RandomForestClassifier for Training and Validation Data.

RandomForestClassifier Model Analysis			
	Precision	Recall	F1-Score
	0.79	0.84	0.83
	0.74	0.62	0.76
Accuracy			0.77
Macro avg.	0.75	0.78	0.8
Weighted avg	0.77	0.77	0.81

As identified in the classification report of Figure 4, the RandomForestClassifier model, which is recognized for its ability to analyze and conform to data, was not nearly as accurate as the Decision Tree model. The accuracy rate of 77% indicates that the model struggled in understanding how often to predict heart disease based on the given systems and how correct these predictions were. The model's confusion was also present in the KNN and SVM models, which were both also weighed down by a lack of precision that, in turn, contributed to a lack of success in identifying heart disease. Verifying that these results were consistent with the data was made possible through a t-test.

Using Python and the imported libraries, a t-test was run between the average accuracies of the data using NumPy, pandas, and SciPy Stats for an alpha of 0.05. The null hypothesis for this condition was as follows: there is no significant difference between the averages of accuracy for each fit-model. Inputting the average accuracies and applying the t-test function, the test statistic was found to be -133.334 and the p value was calculated to be 0.0035, therefore rejecting the null hypothesis and indicating a high statistical significance for the data. This supports the conclusion that the DecisionTree model is highly effective and suitable for the analysis of patient input in conjunction with heart disease detection and predictions.

Discussions and Conclusions:-

Understanding the results of this project are critical to successfully implementing this web application and allowing patients to gain access to views of their health and pursue proper medical attention. The final conclusion to be drawn from testing the different fit-models is that the DecisionTree model is the evidently best-fitting model for Predict2Protect's program with an accuracy rate of 95%. Therefore, the hypothesis stating RandomForestClassifier would presumably be of the best fit is not supported by this project. Despite the high accuracy of the DecisionTree model, there are several factors taken into account to caution against historical principles of the model itself, and keeping these factors in consideration is what creates the strength in this application's program.

With the data set used, the model is able to create a successful program because the nature of machine learning is reliant on training the model with data that proves a trend for any ascertained cause. In this case, each model is trained to detect symptoms for heart disease and relate them back to whether the symptoms were indicative of heart disease in the future or present and at what percent probability. The DecisionTree model is recognized for its overfitting tendencies as the model can often become too focused on the training data to recognize the patterns present and instead conforms to what the majority of the data exhibits. In anticipation of this, several measures were taken (Vijayashree, 2016). The data was split into not only training, but validation data, and this combined with the testing data allowed the model multiple opportunities to relearn the data and recognize patterns. Post-pruning, or generating the decision tree and later removing branches that became too specific rather than pattern-oriented, was also utilized to reduce overfitting. Historically, this is a technique that has verified the significance of DecisionTree models (Thenmozhi, 2014). The program's maintenance of its accuracy rate even with these measures displays the strength of the program and its accuracy in prediction.

There was a nuance addressed earlier in the making of this application that allowed for further accessibility regardless of resources; several of the inputs requested were to be gathered from a recent ECG, or electrocardiograph, screening, and this was included in the original dataset for those who may have had this screening done but did not approach further medical aid or were not able to due to reasons of finance, accessibility, etc. However, this project is aimed particularly towards groups with populations who may not have available results from an ECG as these tests can range from $150-$300, and considering this, these questions were removed (Limbitote, 2020). The program maintained its accuracy rate for the DecisionTree model, but this could be a potential source of error and difficulty for the other models as they may have functioned better with respect to more specific ECG data. This would not be practical for this specific application given its intended audience, but this can be improved by testing other models. Improvements can also be made to the procedure of this project by further eliminating the risks of overfitting and even underfitting by dividing the data more to train the model continuously. This may allow the model to adapt further and gain higher accuracy.

A study recently established a heart-disease prediction program with an extremely similar dataset but instead employed SVM, Gaussian Naive Bayes, logistic regression, and RandomForestClassifier. Here, the accuracy rates produced from running each model over the refined data set were 80.32%, 78.68%, 80.32%, and 88.5% respectively. In this scenario, RandomForestClassifier was found to be the highest ranking in terms of accuracy and represented an overfit of 100% at first when run on only the training data, which is identical to how DecisionTree reacted in this project (Karthick et al, 2022). However, the difference lies in DecisionTree's maintenance of the accuracy rate, and although RandomForestClassifier remains highest for the dataset used in this particular study, DecisionTree offers a more accurate view of the same prediction goal. The difference in RandomForestClassifier's accuracy rate between studies can be attributed to disparities in the datasets used.

On the user interface of this application, after submitting, the user is informed on the likelihood of them having heart disease within less than ten seconds, and this output is framed by the Decision Tree model itself as well as an additional probability function. This function assesses the likelihood of developing heart disease within the next

year, and this is a step towards prediction that allows users to plan ahead given their personal health histories. The patient is then informed to reach out for proper medical care if concerns arise from indicated results. This is an aspect of the program that can be expanded on in future research; this project can be expanded to include greater projections for the future, such as extending this time of projection to predict one's chances of heart disease in the next five years and beyond. This could be an influential field as people will gain further understanding of what to expect and what precautions to take. Similarly, another extension of this project to allow in-depth prediction could be adding more factors to the data itself. This could include more social and environmental factors, such as one's daily anxiety level and even financial situation as research continues to build the connection between economic class-related stress and risk for heart disease (Glenn, 2020).

Creating Predict2Protect serves to urge a user towards implying their outcome in their daily lives and pursuing the correspondingly necessary lifestyle based on their personal knowledge of how they treat their wellbeing. Predict2Protect is a step towards providing this personal assessment of one's health to people across the globe, regardless of their background. Ensuring that one's life can be lived to the fullest for the greatest amount of time is made possible by their health, and this is the goal this application accomplishes, to ensure that everyone has knowledge of their medical needs and risks without the burden of unsustainable costs and unavailable resources. As those around the world without available medical care gain such resources, it is accessible technology like Predict2Protect that have the means of changing lives forever.

Works Cited:-

Peer Reviewed

1. D. Ratnam, P. Himabindu, V. M. Sai, S. P. R. Devi, and P. R. Rao, "Computer-Based Clinical Decision Support System for Prediction of Heart Diseases Using Naïve Bayes Algorithm," Int. J. Comput. Sci. Inf. Technol., vol. 5, no. 2, pp. 2384–2388, 2014.
2. Glenn. (2020, March 18). Social, Financial Factors Critical to Assessing Cardiovascular Risk - American College of Cardiology. American College of Cardiology. https://www.acc.org/about-acc/press-releases/2020/03/18/11/10/social-financial-factors-critical-to-assessing-cardiovascular-risk#:~:text=People%20with%20a%20high%20degree,who%20considered%20themselves%20financially%20secure.
3. H., Karthick, K., Aruna, S. K., Samikannu, R., Kuppusamy, R., Teekaraman, Y., &Thelkar, A. R. (2022, May 2). Implementation of a Heart Disease Risk Prediction Model Using Machine Learning. Implementation of a Heart Disease Risk Prediction Model Using Machine Learning. https://doi.org/10.1155/2022/6517716
4. J. Vijayashree and N. C. S. N. Iyengar, "Heart disease prediction system using data mining and hybrid intelligent techniques: A review," Int. J. Bio-Science Bio-Technology, vol. 8, no. 4, pp. 139–148, 2016, doi: https://doi.org/10.14257/ijbsbt.2016.8.4.16.
5. K. Thenmozhi and P. Deepika, "Heart Disease Prediction Using Classification with Different Decision Tree Techniques," Int. J. Eng. Res. Gen. Sci., vol. 2, no. 6, pp. 6–11, 2014, [Online]. Available: www.ijergs.org.
6. Mangesh Limbitote. (2020). A Survey on Prediction Techniques of Heart Disease using Machine Learning. International Journal of Engineering Research And, V9(06). https://doi.org/10.17577/ijertv9is060298
7. R. Fadnavis, K. Dhore, D. Gupta, J. Waghmare, and D. Kosankar, "Heart disease prediction using data mining," J. Phys. Conf. Ser., vol. 1913, no. 1, 2021, doi: https://doi.org/10.1088/1742-6596/1913/1/012099.