**NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY**

(AN AUTONOMOUS INSTITUTION, AFFILIATED TO VISVESVARAYA TECHNOLOGICAL UNIVERSITY, BELGAUM, APPROVED BY AICTE & GOVT.OF KARNATAKA

****

**FINAL YEAR PROJECT REPORT**

on

**Heart Disease Percentile Prediction Using ML Algorithm**

*Submitted in partial fulfilment of the requirement for the award of Degree of*

*Bachelor of Engineering*

*in*

*Computer Science and Engineering*

*Submitted by:*

|  |  |
| --- | --- |
| ANUBHAV YADAV | 1NT18CS017 |
| KINSHUK CHATURVEDI | 1NT18CS076 |
| NEHA V M | 1NT18CS106 |
| MUSKAN KHATWANI | 1NT18IS202 |

Under the Guidance of

Mr Vasanthakumar G. U.

DESIGNATION, Dept. of CS&E, NMIT



Department of Computer Science and Engineering

**(Accredited by NBA Tier-1)**

2021-22

**NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY**

(AN AUTONOMOUS INSTITUTION, AFFILIATED TO VISVESVARAYA TECHNOLOGICAL UNIVERSITY, BELGAUM

, APPROVED BY AICTE & GOVT.OF KARNATAKA)

Department of Computer Science and Engineering

**(Accredited by NBA Tier-1)**

****

**CERTIFICATE**

This is to certify that the Final Year Phase 2 Report on **Heart Disease Percentile Prediction Using ML Algorithm** is an authentic work carried out by ANUBHAV YADAV (**1NT18CS017**), KINSHUK CHATURVEDI **(1NT18CS076),** NEHA V M **(1NT18CS106)** and MUSKAN KHATWANI **(1NT18IS202)** bonafide students of **Nitte Meenakshi Institute of Technology**, Bangalore in partial fulfilment for the award of the degree of ***Bachelor of Engineering*** in COMPUTER SCIENCE AND ENGINEERING of Visvesvaraya Technological University, Belagavi during the academic year ***2021-2022.*** It is certified that all corrections and suggestions indicated during the internal assessment has been incorporated in the report.

|  |  |  |  |
| --- | --- | --- | --- |
| **Internal Guide** | | **Signature of the HOD** | |
|  | |  | |
| Dr. Vasanthakumar G. U.  Associate Professor, Dept. CSE,  NMIT Bangalore | | Dr. Saroja Devi.H  Professor & Head, Dept. CSE,  NMIT Bangalore | |
| **Signature of Examiners** | |
|  |  |
|  |  |

**DECLARATION**

We are hereby declare that

(i) The project work is our original work

(ii) This Project work has not been submitted for the award of any degree or examination at any other university/College/Institute.

(iii) This Project Work does not contain other persons’ data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.

(iv) This Project Work does not contain other persons’ writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:

a) their words have been re-written but the general information attributed to them has been referenced;

b) where their exact words have been used, their writing has been placed inside quotation marks, and referenced.

(v) This Project Work does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the thesis and in the References sections.

|  |  |  |
| --- | --- | --- |
| **NAME** | **USN** | **Signature** |
| ANUBHAV YADAV | 1NT18CS017 |  |
| KINSHUK CHATURVEDI | 1NT18CS076 |  |
| NEHA V M | 1NT18CS106 |  |
| MUSKAN KHATWANI | 1NT18IS202 |  |

Date: 21st May 2022

**ACKNOWLEDGEMENT**

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of the people who made it possible, whose constant guidance and encouragement crowned our effort with success. I express my sincere gratitude to our Principal **Dr. H. C. Nagaraj**, Nitte Meenakshi Institute of Technology for providing facilities.

We wish to thank our HoD**, Dr. Saroja Devi H** for the excellent environment created to further educational growth in our college. We also thank him for the invaluable guidance provided which has helped in the creation of a better project.

I hereby like to thank our **Dr. Vasanthakumar G. U., Associate Professor,** Department of Computer Science & Engineering on his periodic inspection, time to time evaluation of the project and help to bring the project to the present form.

Thanks to our Departmental Project coordinators. We also thank all our friends, teaching and non-teaching staff at NMIT, Bangalore, for all the direct and indirect help provided in the completion of the project.

|  |  |  |
| --- | --- | --- |
| **NAME** | **USN** | **Signature** |
| ANUBHAV YADAV | 1NT18CS017 |  |
| KINSHUK CHATURVEDI | 1NT18CS076 |  |
| NEHA V M | 1NT18CS106 |  |
| MUSKAN KHATWANI | 1NT18IS202 |  |

Date: 21st May 2022

**ABSTRACT**

Heart disease cases are increasing at an alarming rate, and it's critical and concerning to be able to predict such diseases in advance. This is a difficult diagnosis to make, so it must be done correctly and quickly. The main focus of the research paper is on which patients are more likely to develop heart disease based on various medical characteristics. We created a heart disease prediction system that uses the patient's medical history to predict whether or not the patient will be diagnosed with heart disease. To predict and classify patients with heart disease, we used various machine learning algorithms such as Logistic Regression, Random Forest Classifier, Support Vector Classifier, Decision Tree Classifier, and K-Nearest Neighbors Classifier. To regulate how the model can be used to improve the accuracy of prediction of Heart Attacks in any individual, a very helpful approach was used. The proposed model's strength was quite satisfying, as it was able to predict evidence of heart disease with a minimum score of 86.88 % for Random Forest Classifier, 91.8 % for K-Nearest Neighbors Classifier, 90.16 % for Support Vector Classifier, and 88.52 % for Decision Tree Classifier and Logistic Regression. So, by using the given model to determine the probability of the classifier correctly and accurately identifying heart disease, a significant amount of pressure has been relieved. Given's heart disease prediction system improves medical care while lowering costs.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Chapter. No.** | **Title** | **Page No.** |
|  | PRELIMINARIES | 1 |
|  | Title Page | 1 |
|  | Certificate | 2 |
|  | Declaration | 3 |
|  | Acknowledgement | 4 |
|  | Abstract | 5 |
|  | Table of Contents (ToC) | 6 |
| 1 | Introduction | 8 |
|  | 1.1 ML and Heart Disease Prediction | 8 |
|  | 1.2 Problem Definition | 8 |
|  | 1.3 Objectives of the Project | 9 |
|  | 1.4 Feasibility Study | 9 |
|  | 1.5 Economical Feasibility | 9 |
| 2 | Software Description | 10 |
|  | 2.1 System Requirement Specification | 10 |
|  | 2.1.1 Software Requirements | 10 |
|  | 2.1.2 Hardware Requirements | 10 |
|  | 2.2 Existing System | 10 |
|  | 2.3 Proposed System | 10 |
| 3 | Literature Survey | 11 |
| 4 | Methodology | 12 |
|  | 4.1 System Design & Architecture | 12 |
|  | 4.1.1 Data Collection | 12 |
|  | 4.1.2 Data Pre-processing | 12 |
|  | 4.1.2.1 Understanding Dataset | 12 |
|  | 4.1.2.2 Analyzing Data Through Graph | 14 |
|  | 4.1.2.2.1 Heat Map | 14 |
|  | 4.1.2.2.2 Histogram | 15 |
|  | 4.1.2.2.3 Target Bar Chart | 15 |
|  | 4.1.2.3 Handling Categorical Data & Scaling Data | 16 |
|  | 4.1.3 Model Creation | 16 |
|  | 4.1.3.1 Random Forest Classifier | 16 |
|  | 4.1.3.2 K-Nearest Neighbors Classifier | 17 |
|  | 4.1.3.3 Support Vector Classifier | 18 |
|  | 4.1.3.4 Decision Tree Classifier | 19 |
|  | 4.1.3.5 Logistic Regression | 19 |
| 5 | Conclusion | 20 |
| 6 | References | 21 |

**CHAPTER 1: INTRODUCTION**

* 1. **ML and Heart Disease Prediction**

"Machine Learning" [1] is defined as "a method of manipulating and extracting implicit, previously unknown/known, and potentially useful information about data." Machine Learning is a vast and diverse field, and its scope and application are expanding all the time. Machine learning includes supervised, unsupervised, and ensemble learning classifiers that are used to predict and find the accuracy of a dataset. We can apply that knowledge to our HDPS project, as it will benefit a large number of people. Cardiovascular diseases are a broad term that refers to a variety of conditions that can affect your heart. According to the World Health Organization, (Cardiovascular diseases), CVDs cause 17.9 million deaths worldwide [2]. It is the leading cause of death among adults. With the help of their medical history, our project can help predict who is likely to be diagnosed with heart disease [6]. It recognizes who has any heart disease symptoms, such as chest pain or high blood pressure, and can assist in diagnosing the disease with fewer medical tests and more effective treatments, allowing them to be treated appropriately.

This project focuses on five Machine Learning algorithms: (1) K Nearest Neighbors Classifier, (2) Random Forest Classifier (3) Decision Tree Classifier (4) Support Vector Classifier and (5) Logistic Regression. The highest accuracy of our project comes from the K Nearest Neighbors Classifier (91.8%), which is better than the previous system, which used only one ML Algorithm. As a result, employing more ML Algorithms improved HDPS accuracy and efficiency.

* 1. **Problem Definition**

The main aim of developing this project is to achieve and overcome the below mentioned problems:

1. The problem is aiming to automate the process of heart disease identification.
2. The problem is aiming to resolve the human error done by medical professionals.
   1. **Objectives of the Project**

Objectives of our system are:

● To identify which ML algorithms are used in Heart disease

● Automate the process of Heart disease identification

● To resolve the issue of human error done by medical professionals

* 1. **Feasibility Study**

A feasibility study is carried out to select the best system that meets performance requirements. The system/application prospers only if it is feasible. The main aim of the feasibility study is to determine whether it would be financially and technically feasible to develop the product.

* 1. **Economical Feasibility**

Economic analysis is the most frequently used technique for evaluating the effectiveness of the proposed system. We have conducted economic feasibility studies to determine the operational and maintenance cost. The project requires an android device to run the app and a PC/Laptop to run the system.

**CHAPTER 2: SOFTWARE DESCRIPTION**

* 1. **System Requirement Specification**
     1. **Software Requirements**

1. Programming Language: Python

2. Machine Learning: Sklearn

3. Data Preprocessing: Numpy, Pandas

4. Data Visualization: Seaborn, Matplotlib

* + 1. **Hardware Requirements**

1. Cloud Computing (Kaggle, Google Collab)
   1. **Existing System**

To date, heart diseases are diagnosed by medical professionals. Medical professionals consume a lot of time for identifying the disease and there is always the issue of human error.

* 1. **Proposed System**

A machine learning-based system is proposed in this project. The model will predict heart disease in less time and the issue of human error will not occur. This system can be implemented in any pandemic situation as well.

**CHAPTER 3: LITERATURE SURVEY**

This work was motivated by a significant amount of work on the diagnosis of Cardiovascular Heart Disease using Machine Learning algorithms. This paper includes a brief review of the literature. Various algorithms, such as Logistic Regression, KNN, Random Forest Classifier, and others, have been used to make an effective cardiovascular disease prediction. Each algorithm has the strength to register the defined objectives, as shown in the Results [7]. Using the previous and new models of machine learning and deep learning, the model incorporating IHDPS was able to calculate the decision boundary. It made the most basic and important factors/knowledge, such as family history, associated with any heart disease, easier to access.

However, the accuracy of such an IHDPS model was far less than that of a new upcoming model for detecting coronary heart disease using artificial neural networks and other machine and deep learning algorithms. McPherson et al.,[8] identified the risk factors of coronary heart disease or atherosclerosis using an inbuilt implementation algorithm that employs some Neural Network techniques and was only able to accurately predict whether the test patient was suffering from the disease or not.

R. Subramanian et al.,[24], introduced the use of neural networks to diagnose and predict heart disease and blood pressure, as well as other attributes. A Deep Neural Network was built using the given disease attributes to produce an output that was carried out by the output perceptron and almost included 120 hidden layers, which is the most basic and relevant technique for ensuring an accurate result of having heart disease if the model is used for the Test Dataset. The use of a supervised network for heart disease diagnosis has been recommended [16]. When the testing of the model was done by a doctor using unfamiliar data, the model was used and trained from the previously learned data and predicted the result thereby calculating the accuracy of the given model.

**CHAPTER 4: METHODOLOGY**

* 1. **System Design & Architecture**

This section illustrates the entire concept of the research work which will aid to understand the whole notion of the paper. At first, we collected the data and preprocessed it. After preprocessing the data. Then the model is created and trained. The model is evaluated using a test dataset and further optimized. This process can be seen in the diagram 4.1.

**Diagram

Description automatically generated**

Figure 4.1: System architecture of chronic kidney disease prediction system.

* + 1. **Data Collection**

Data is gained from Kaggle heart disease Competition.

* + 1. **Data Pre-processing**
       1. **Understanding Dataset**

Before training the model, proper analysis of data needs to be done. As shown in fig 4.2.

**Table

Description automatically generated**

Figure 4.2: Dataset Sample

**Graphical user interface, text

Description automatically generated**

Figure 4.3: Dataset Information

From Figure 4.3, it can be seen that there is no NULL value in the dataset. So, there is no need for imputing NaN values.

* + - 1. **Analyzing Data Through Graph**
         1. **Heat Map**

**A picture containing text, crossword puzzle, scoreboard, cabinet

Description automatically generated**

Figure 4.4 (a): Heatmap

From Figure 4.4 (a), it can be seen from the correlation matrix that there is no single feature that has a very high correlation with the target. Similarly, the correlation between independent variables is also not high.

Chart, waterfall chart

Description automatically generated

Figure 4.4 (b): Correlation with ‘Target’ Column

Even from the figure 4.4 (b) we can observe that there is no single attribute which has a major impact on determining the outcome.

* + - * 1. **Histogram**

**Chart, box and whisker chart

Description automatically generated**

Figure 4.5: Histogram

The figure 4.5 depicts how each feature and label is spread throughout multiple ranges, confirming the requirement for scaling. Following that, the discrete bars indicate that each of them is a categorical variable. Before implementing Machine Learning, these categorical variables must be addressed.

* + - * 1. **Target Bar chart**

**Chart, bar chart

Description automatically generated**

Figure 4.6: Target Bar Chart

An unbalanced dataset is one in which the distribution of the classes in the dataset is highly skewed. This indicates that a dataset is skewed towards one of the dataset's classes. If the dataset favors one class, an algorithm trained on the same data will favor the same class. From Figure 4.6, it can be seen that the target is balanced.

* + - 1. **Handling Categorical Data & Scaling Data**

****

Figure 4.7: Code for Handling Categorical data and Scaling data.

Figure 4.7, shows which features are categorical and these features are broken down. Standard Scaler is used to standardize the data. It normalizes the data so that the mean is 0 and the standard deviation is 1. Finally, the data is split into the train(80%) and test(20%). Various machine learning models are employed in this data.

* + 1. **Model Creation**
       1. **Random Forest Classifier**

A random forest is a meta estimator that employs averaging to increase predicted accuracy and control over-fitting by fitting several decision tree classifiers on various sub-samples of the dataset.

Chart, bar chart

Description automatically generated

Figure 4.8: Random Forest Graph

The figure 4.8 shows the accuracy value of the classifier for different estimator values. Taking a look at the bar graph, we can see that the maximum score of 86.88% was achieved for 100, 200,500, and 1000 trees.

* + - 1. **K-Nearest Neighbors Classifier**

The K-Nearest Neighbour method is based on the Supervised Learning approach and is one of the most basic Machine Learning algorithms. The K-NN method assumes that the new case/data and existing cases are comparable and places the new case in the category that is most similar to the existing categories. The K-NN method saves all available data and classifies a new data point based on its similarity to the existing data. This implies that fresh data may be quickly sorted into a well-defined category using the K-NN method. The K-NN approach may be used for both regression and classification, however, it is more commonly utilized for classification tasks.

**Graphical user interface, chart, application, line chart

Description automatically generated**

Figure 4.9: K-Neighbors Classifier

From the figure 4.9, it is clear the model achieved the maximum score of 91.8% when the number of neighbors was 8.

* + - 1. **Support Vector Classifier**

By altering the distance between the data points and the hyperplane, SVM seeks to construct a hyperplane that can distinguish the classes as much as feasible. The hyperplane is determined depending on the number of kernels. In the current study, four kernels were used: linear, poly, RBF, and sigmoid.

**Chart, bar chart

Description automatically generated**

Figure 4.10: Support Vector Classifier

It can be seen from the figure 4.10, the RBF kernel performed the best for this dataset and achieved a score of 90.16%.

* + - 1. **Decision Tree Classifier**

The Decision Tree Classifier generates a decision tree and assigns class values to each data point depending on it. We may change the maximum number of characteristics to consider while building the model here.

**Chart, line chart

Description automatically generated**

Figure 4.11: Decision Tree Classifier

From Figure 4.11, it can see that the maximum score is 88.52% and is achieved for maximum features being selected to be 4.

* + - 1. **Logistic Regression**

This classifier predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

The accuracy we got from this model was 88.52 %, which was equal to Decision Tree classifier.

**CHAPTER 5: CONCLUSION**

The project involved analysis of the heart disease patient dataset with proper data processing.

Five models were trained and tested with maximum scores as follows:

* Random Forest Classifier (86.88%)
* K Nearest Neighbors Classifier (91.8%)
* Support Vector Classifier (90.16%)
* Decision Tree Classifier (88.52%)
* Logistic Regression (88.52%)

Chart, bar chart

Description automatically generated

If we compare the results with our base papers:

Table

Description automatically generated

We can see that among these models from all papers, K Nearest Neighbors Classifier performs the best.

**CHAPTER 6: REFERENCES**

[1] Soni J, Ansari U, Sharma D & Soni S (2011). Predictive data mining for medical diagnosis: an overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43-8.

[2] Dangare C S & Apte S S (2012). Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications, 47(10), 44-8.

[3] Ordonez C (2006). Association rule discovery with the train and test approach for heart disease prediction. IEEE Transactions on Information Technology in Biomedicine, 10(2), 334-43.

[4] Shinde R, Arjun S, Patil P & Waghmare J (2015). An intelligent heart disease prediction system using k-means clustering and Naïve Bayes algorithm. International Journal of Computer Science and Information Technologies, 6(1), 637-9.

[5] Bashir S, Qamar U & Javed M Y (2014, November). An ensemble-based decision support framework for intelligent heart disease diagnosis. In International Conference on Information Society (i-Society 2014) (pp. 259-64). IEEE.

[6] Jee S H, Jang Y, Oh D J, Oh B H, Lee S H, Park S W & Yun Y D (2014). A coronary heart disease prediction model: the Korean Heart Study. BMJ Open, 4(5), e005025.

[7] Ganna A, Magnusson P K, Pedersen N L, de Faire U, Reilly M, Ärnlöv J & Ingelsson E (2013). Multilocus genetic risk scores for coronary heart disease prediction. Arteriosclerosis, thrombosis, and vascular biology, 33(9), 2267-72.

[8] Jabbar M A, Deekshatulu B L & Chandra P (2013, March). Heart disease prediction using lazy associative classification. In 2013 International Multi-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s) (pp. 40- 6). IEEE.

[9] Dangare Chaitrali S and Sulabha S Apte. "Improved study of heart disease prediction system using data mining classification techniques." International Journal of Computer Applications 47.10 (2012): 44-8.

[10] Soni Jyoti. "Predictive data mining for medical diagnosis: An overview of heart disease prediction." International Journal of Computer Applications 17.8 (2011): 43-8.

[11] Chen A H, Huang S Y, Hong P S, Cheng C H & Lin E J (2011, September). HDPS: Heart disease prediction system. In 2011 Computing in Cardiology (pp. 557-60). IEEE.

[12] Parthiban, Latha and R Subramanian. "Intelligent heart disease prediction system using CAN FIS and genetic algorithm." International Journal of Biological, Biomedical and Medical Sciences 3.3 (2008).

[13] Wolgast G, Ehrenborg C, Israelsson A, Helander J, Johansson E & Manefjord H (2016). Wireless body area network for heart attack detection [Education Corner]. IEEE antennas and propagation magazine, 58(5), 84-92.

[14] Patel S & Chauhan Y (2014). Heart attack detection and medical attention using the motion-sensing device -Kinect. International Journal of Scientific and Research Publications, 4(1), 1-4.

[15] Zhang Y, Fogoros R, Thompson J, Kenknight B H, Pederson M J, Patangay A & Mazar S T (2011). U.S. Patent No. 8,014,863. Washington, DC: U.S. Patent and Trademark Office.

[16] Raihan M, Mondal S, More A, Sagor M O F, Sikder G, Majumder M A & Ghosh K (2016, December). Smartphone-based ischemic heart disease (heart attack) risk prediction using clinical data and data mining approaches, a prototype design. In 2016 19th International Conference on Computer and Information Technology (ICCIT) (pp. 299-303). IEEE.

[17] Buechler K F & McPherson P H (1999). U.S. Patent No. 5,947,124. Washington, DC: U.S. Patent and Trademark Office.

[18] Takci H (2018). Improvement of heart attack prediction by the feature selection methods. Turkish Journal of Electrical Engineering & Computer Sciences, 26(1), 1-10.

[19] Worthen W J, Evans S M, Winter S C & Balding D (2002). U.S. Patent No. 6,432, 124. Washington, DC: U.S. Patent and Trademark Office.

[20] Acharya U R, Fujita H, Oh S L, Hagiwara Y, Tan J H & Adam M (2017). Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals. Information Sciences, 415, 190-8.

[21] Brown N, Young T, Gray D, Skene A M & Hampton J R (1997). Inpatient deaths from acute myocardial infarction, 1982-92: analysis of data in the Nottingham heart attack register. BMJ, 315(7101), 159-64.

[22] Piller L B, Davis B R, Cutler J A, Cushman W C, Wright J T, Williamson J D & Haywood L J (2002). Validation of heart failure events in the Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT) participants assigned to doxazosin and chlorthalidone. Current controlled trials in cardiovascular medicine, 3(1), 10.

[23] Folsom A R, Prineas R J, Kaye S A & Soler J T (1989). Body fat distribution and self-reported prevalence of hypertension, heart attack, and other heart diseases in older women. International journal of epidemiology, 18(2), 361-7.

[24] Kiyasu J Y (1982). U.S. Patent No. 4,338,396. Washington, DC: U.S. Patent and Trademark Office