**Heart Attack Prediction Using Machine Learning Algorithms**

## MINI PROJECT REPORT

***Submitted by***

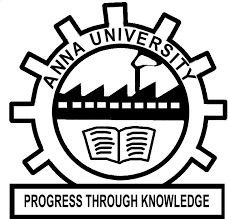
|  |  |
| --- | --- |
| **VISHAL B** | **210701312** |
| **NANDHA KRISHNA R** | **210701518** |

***in partial fulfillment for the award of the degree of***

**BACHELOR OF ENGINEERING**

***in***

**COMPUTER SCIENCE AND ENGINEERING**



**RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI ANNA UNIVERSITY: CHENNAI 600 025**

## APRIL 2024

**RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI**

**BONAFIDE CERTIFICATE**

Certified that this Report titled “**Heart Attack Prediction Using Machine Learning Algorithms**” is the bonafide work of **“VISHAL B (210701312), NANDHA KRISHNA R (210701518)”** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

|  |
| --- |
| **SIGNATURE** |
| **Rahul Chiranjeevi. V** |
| **Assistant Professor,** |
| Department of Computer Science and Engineering, |
| Rajalakshmi Engineering College, |
| Chennai – 602015 |

Submitted to Mini Project Viva-Voce Examination held on

**Internal Examiner External Examiner**

## ABSTRACT

The heart, a vital organ in humans, plays a crucial role in purifying and circulating blood throughout the body. Heart attacks, a leading cause of death worldwide, manifest through symptoms like chest pain, rapid heartbeat, and breathing difficulties. This study presents a comprehensive examination of heart attacks and contemporary prediction techniques. It provides an overview of different machine learning methodologies employed for heart attack prediction, including Decision Tree, Logistic Regression, Support Vector Machines (SVM), Naive Bayes, Random Forest, K-Nearest Neighbours (KNN), and XGBoost Classifier. The comparison of these algorithms is conducted based on their respective features, offering insights into their effectiveness. The purpose of this study is to support ongoing efforts to create precise and effective models of prediction for cardiac diagnosis and prevention.

## ACKNOWLEDGEMENT

Initially we thank the Almighty for being with us through every walk of our life and showering his blessings through the endeavor to put forth this report. Our sincere thanks to our Chairman **Mr. S. MEGANATHAN, B.E, F.I.E.**, our Vice Chairman **Mr. ABHAY SHANKAR MEGANATHAN, B.E., M.S.,** and our

respected Chairperson **Dr. (Mrs.) THANGAM MEGANATHAN**, **Ph.D.,** for providing us with the requisite infrastructure and sincere endeavoring in educating us in their premier institution.

Our sincere thanks to **Dr. S.N. MURUGESAN, M.E., Ph.D.,** our beloved Principal for his kind support and facilities provided to complete our work in time. We express our sincere thanks to **Dr. P. KUMAR, Ph.D.**, Professor and Head of the Department of Computer Science and Engineering for his guidance and encouragement throughout the project work. We convey our sincere and deepest gratitude to our internal guide, **Rahul Chiranjeevi V** Professor, Department of Computer Science and Engineering. Rajalakshmi Engineering College for his valuable guidance throughout the course of the project.

**VISHAL B -210701312**

**NANDHA KRISHNA R-210701518**

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **CHAPTER NO.** | **TITLE** | **PAGE NO.** |
|  | **ABSTRACT** | **iii** |
|  | **ACKNOWLEDGEMENT** | **iv** |
|  | **LIST OF FIGURES** | **vii** |
|  | **LIST OF TABLES** | **viii** |
|  | **LIST OF ABBREVIATIONS** | **ix** |
| **1.** | **INTRODUCTION** | **10** |
|  | 1.1 GENERAL | 10 |
|  | 1.2 OBJECTIVE | 10 |
|  | 1.3 EXISTING SYSTEM | 10 |
|  | 1.4 PROPOSED SYSTEM | 11 |
| **2.** | **LITERATURE SURVEY** | **12** |
| **3.** | **SYSTEM DESIGN** | **14** |
|  | 3.1 DEVELOPMENT ENVIRONMENT | 14 |
|  | 3.1.1 HARDWARE SPECIFICATIONS | 14 |
|  | 3.1.2 SOFTWARE SPECIFICATIONS | 14 |
|  | 3.2 SYSTEM DESIGN | 15 |
|  | 3.2.1 ARCHITECTURE DIAGRAM | 15 |

|  |  |  |
| --- | --- | --- |
| **4.** | **PROJECT DESCRIPTION** | **16** |
|  | 4.1 MODULES DESCRIPTION | 16 |
| **5.** | **IMPLEMENTATION AND RESULTS** | 18 |
|  | 5.1 IMPLEMENTATION | 18 |
|  | 5.2 OUTPUT SCREENSHOTS | 20 |
| **6.** | **CONCLUSION AND FUTURE ENHANCEMENT** | **21** |
|  | 6.1 CONCLUSION | 21 |
|  | 6.2 FUTURE ENHANCEMENT | 21 |
|  | **REFERENCES 22** | |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **S.NO** | **NAME** | **PAGE NO** |
| 3.3.1 | ARCHITECTURE DIAGRAM | 15 |
| 5.2.1 | ACCURACY CURVE | 19 |
| 5.2.2 | WEEKLY MINUTES WATCHED | 19 |
| 5.2.3 | MODELS VS ACCURACY | 19 |

**LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
| **S.NO** | **NAME** | **PAGE NO** |
| 3.2.1 | HARDWARE SPECIFICATIONS | 14 |
| 3.2.2 | SOFTWARE SPECIFICATIONS | 14 |
|  |  |  |

**LIST OF ABBREVIATIONS**

|  |  |
| --- | --- |
| OTT | Over-the-top |
| EDA | Exploratory Data Analysis |
| CR | Classification Report |
| BPNN | Backpropagation Neural Network |
| NLP | Natural Language Processing |
| PSO | Particle Swarm Optimization |

## CHAPTER 1 INTRODUCTION

### GENERAL

The heart, being a vital organ, demands meticulous care to ensure overall well-being. Given its association with various diseases, including heart attacks,

predictive studies in this domain are crucial. Currently, a significant number of individuals succumb to heart attacks, often diagnosed at advanced stages due to the lack of precise predictive tools.

### OBJECTIVE

This research attains approximately 80% accuracy on the testing set during evaluation, although the translation of data into practical use remains time-consuming. To address the challenge of accuracy and efficiency, the Random Forest method emerges as a preferred choice, facilitating more precise outcomes in less time.

**1.3 EXISTING SYSTEM**

Heart attack prediction using machine learning has evolved significantly, leveraging various data sources such as medical databases, electronic health records (EHRs), and wearable devices. These systems utilize a wide range of features, including demographic information, medical history, lifestyle factors, clinical measurements, and lab results. Commonly employed machine learning algorithms include logistic regression, decision trees, random forests, support vector machines, k-nearest neighbors, naive Bayes, neural networks, and gradient boosting machines. Models are evaluated using metrics like accuracy, precision, recall, F1 score, and ROC-AUC. Preprocessing techniques such as data cleaning, normalization, feature selection, and handling imbalanced data are crucial for model performance. Deployment is facilitated through cloud platforms and API integration, with user interfaces designed for both patients and clinicians.

### PROPOSED SYSTEM

The proposed system for heart attack prediction using machine learning aims to enhance accuracy, real-time processing, and user accessibility. It will integrate data from diverse sources, including EHRs, wearable devices, and patient-reported information, ensuring a comprehensive dataset. Advanced algorithms such as deep neural networks and ensemble methods like XGBoost will be employed to capture complex patterns and interactions in the data. The system will feature robust preprocessing techniques, including automated data cleaning, normalization, and feature engineering, to handle large and varied datasets effectively. To address the challenge of interpretability, the system will incorporate explainable AI techniques, providing clear insights into prediction results for clinicians and patients. Deployment will utilize scalable cloud infrastructure, ensuring real-time processing and seamless integration with existing healthcare systems through secure APIs. The user interface will be designed for both desktop and mobile platforms, offering an intuitive experience for end-users. The system will prioritize data privacy and security, complying with relevant regulations such as HIPAA and GDPR. Overall, this proposed system aims to provide a highly accurate, interpretable, and user-friendly solution for predicting heart attacks, ultimately improving patient outcomes and preventive care.

## CHAPTER 2 LITERATURE SURVEY

Numerous studies conducted in medical research focus on developing heart attack prediction systems utilizing a range of algorithms for machine learning.

Santhana Krishnan explored heart attack prediction employing classification techniques. The paper offers comprehensive insights into heart attacks, encompassing their types and risk factors. Naive Bayes and Decision Tree are utilized for prediction, with Naive Bayes demonstrating higher accuracy.

Avinash Golande proposed effective Machine Learning techniques for heart attack prediction. Decision trees, k-nearest neighbour, and Naive Bayes are among the commonly used methodologies, with Decision trees showing superior accuracy.

V.V. Ramalingam recommended machine learning approaches for heart attack prediction, leveraging various algorithms to automate the analysis of extensive medical datasets. In their work, researchers frequently give decision trees, ensemble models, support vector machines, K-nearest neighbours, Naive Bayes, and Random Forest top priority. To assist in the identification of heart-related illnesses, healthcare professionals are advised to make use of supervised learning techniques such ensemble models, Random Forest, K-nearest neighbours, Naive Bayes, support vector machines, and decision trees.

## CHAPTER 3 SYSTEM DESIGN

### DEVELOPMENT ENVIRONMENT

* + 1. **HARDWARE SPECIFICATIONS**

This project uses minimal hardware but in order to run the project efficiently without any lack of user experience, the following specifications are recommended

**Table 3.1.1** Hardware Specifications

|  |  |
| --- | --- |
| **PROCESSOR** | Intel Core i5 |
| **RAM** | 4GB or above (DDR4 RAM) |
| **GPU** | Intel Integrated Graphics |
| **HARD DISK** | 6GB |
| **PROCESSOR FREQUENCY** | 1.5 GHz or above |

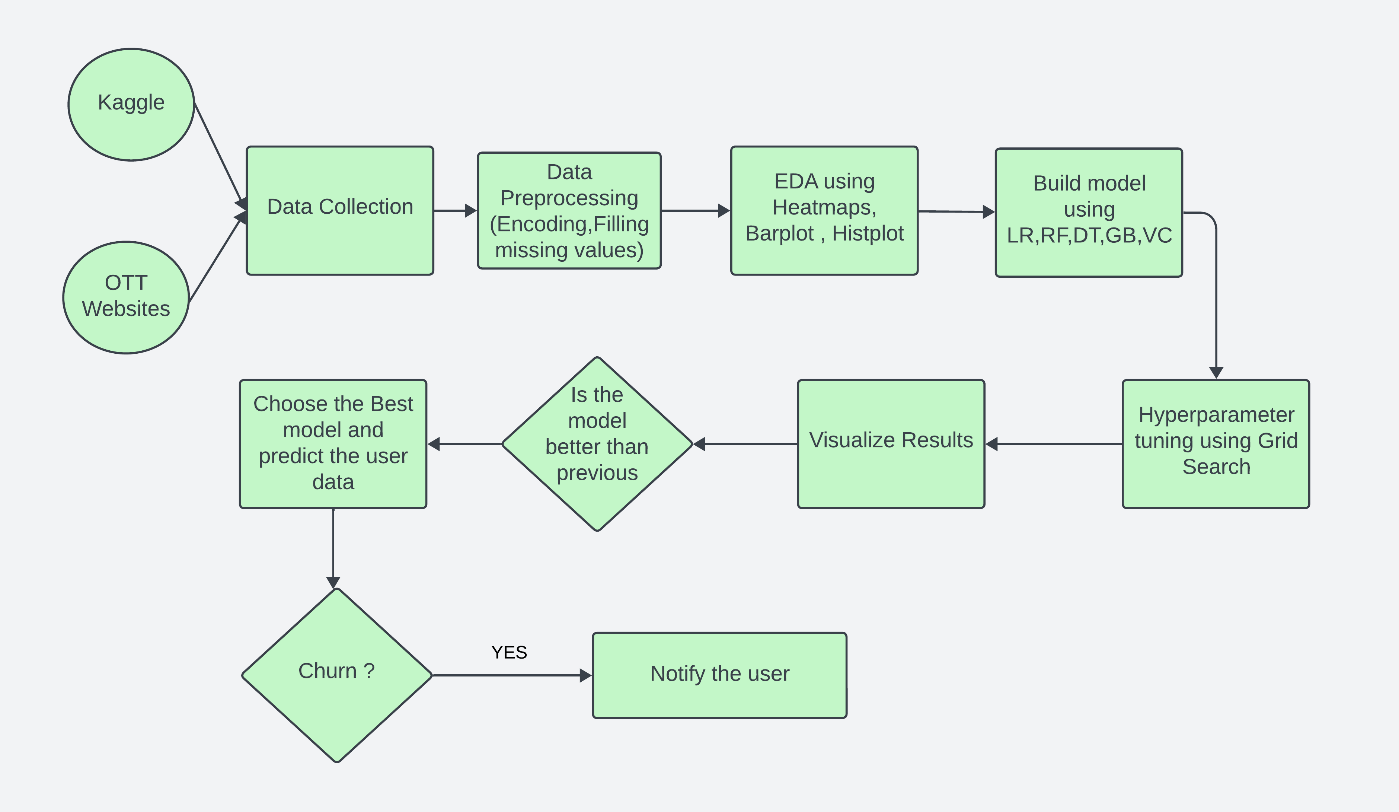
### SOFTWARE SPECIFICATIONS

The software specifications in order to execute the project has been listed down in the below table. The requirements in terms of the software that needs to be pre- installed and the languages needed to develop the project has been listed out below.

**Table 3.1.2** Software Specifications

|  |  |
| --- | --- |
| **BACK END** | Python, Django |
| **LIBRARIES** | Numpy, Pandas, Matplotlib, Seaborn, ScikitLearn |
| **SOFTWARES USED** | Visual Studio, Jupyter Notebook, Microsoft Excel |

* 1. **SYSTEM DESIGN**
     1. **ARCHITECTURE DIAGRAM**

****

**Fig 3.2.1 Architecture Diagram**

## CHAPTER 4 PROJECT DESCRIPTION

### MODULE DESCRIPTION

#### 4.1.1 DATA COLLECTION:

The first phase of the project involves data from trusted sources such as kaggle and Google dataset search. The data set collected should have desired data columns and be able provide better results and the size should be sufficient enough.

#### 4.1.2 DATA PREPROCESSING:

The Data collected won’t be in a state that can be used for training purposes hence, the data should undergo the step of preprocessing in which common problems are eradicated such as missing values, improper spelling in data or incorrectness in data etc. Various python libraries specialized for data analysis can be utilized for this purpose such as Numpy, Pandas. This step is crucial for the project as these may cause inefficiency if they are fed directly to the model.

#### 4.1.3 EDA:

The acquired data is analyzed for its relation within the data. Any outliers or deviation of data can be inferred at this point and also this helps to gain the significance of each data column. Libraries of these are visualization tools commonly used in the project. Through EDA, we concluded that several attributes of users such as phone number, user id etc. are redundant and thus they are dropped. Heatmaps are extensively used to know the correlation between various attributes

#### 4.1.4 BUILD MODELS:

The next step in the project involves building various machine learning models in supervised classification algorithms such as LR, DT, Random Forest, KNN and also several ensemble algorithms such as A-Boosting, G-Boosting, V-Classifier are used to build and keep track of the models’ performance. The library will be helpful in building those models.

#### 4.1.5 HYPERPARAMETER TUNING:

Once the basic models are built, the models are then tuned based on their Hyperparameters such as max\_depth, iteration count to improve the performance of the existing models. The tuning of the models will help it to find the best parameters for training.

#### 4.1.6 VISUALIZING RESULTS:

The results of various parameters and also the accuracies along the time are analyzed in this step to get insights of the working of various models. The CR and accuracy scores of the models play a vital role in analyzing them.

#### 4.1.7 CHOOSING BEST MODEL:

       As we have a track of the model’s performance, we can choose the best model among the trained models and can utilize it for the further development of the project which leads to an iterative development process. Among those models, the Enhanced Random Forest algorithm with tuning is concluded as the best model as its performance was at the top. report and accuracy score form the basis of the evaluation of the model.

## CHAPTER 5 IMPLEMENTATION AND RESULTS

### 5.1 IMPLEMENTATION

#### 5.1.1 Random Forest:

#### Random Forest is a versatile and widely-used ensemble learning method for classification, regression, and other tasks. It works by constructing a multitude of decision trees during training and outputting the class that is the mode of the classes (classification) or the mean prediction (regression) of the individual trees. The key idea behind Random Forest is to reduce the overfitting that individual decision trees often exhibit by averaging multiple trees, each trained on a random subset of the data and features. This randomness introduces diversity among the trees, leading to more robust and generalizable predictions.

#### The strength of Random Forest lies in its ability to handle large datasets with higher dimensionality, including datasets with many features and missing values. It provides an estimate of feature importance, which can be useful for understanding the influence of different variables in the predictive model. Additionally, Random Forests are relatively easy to tune and parallelize, making them efficient in practice. Despite its complexity, the method often performs well out-of-the-box and is less sensitive to hyperparameter settings compared to other machine learning algorithms, making it a popular choice for a wide range of applications.

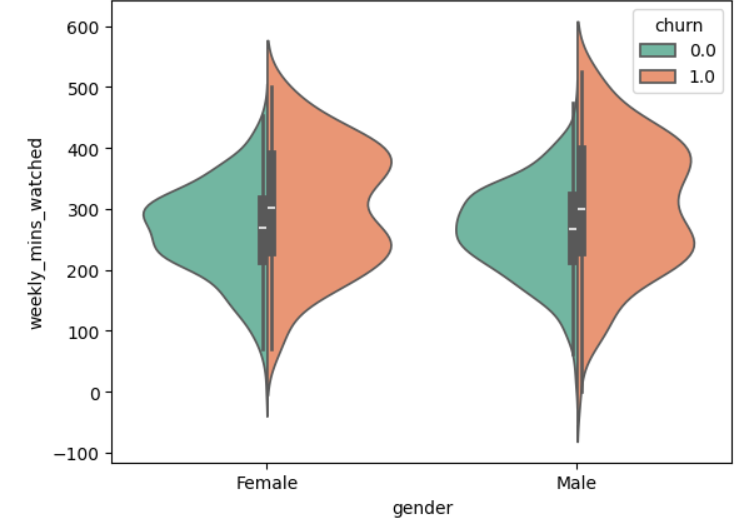
#### 5.1.2 Hyperparameter Tuning:

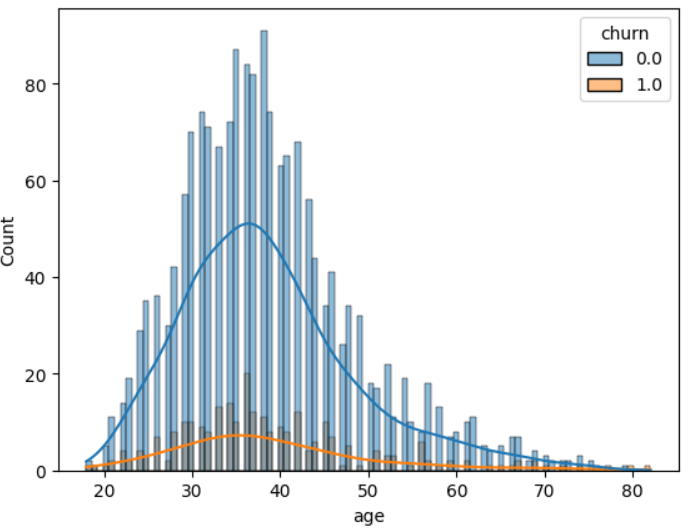
#### Hyperparameter tuning is the process of optimizing the parameters that govern the behavior of a machine learning model but are not learned from the data. These parameters, known as hyperparameters, can significantly affect the performance of a model. Examples include the learning rate in neural networks, the number of trees in a Random Forest, and the regularization strength in logistic regression. Effective hyperparameter tuning can lead to improved model accuracy, robustness, and generalization to new data. The tuning process typically involves searching over a predefined space of hyperparameters and evaluating the model's performance using techniques such as cross-validation.

#### There are several methods for hyperparameter tuning, including grid search, random search, and more advanced techniques like Bayesian optimization and genetic algorithms. Grid search exhaustively searches through a specified subset of hyperparameters, evaluating every possible combination, which can be computationally expensive but thorough. Random search, on the other hand, samples hyperparameters randomly, often leading to better results in less time. Bayesian optimization builds a probabilistic model of the objective function and uses it to select the most promising hyperparameters to evaluate, balancing exploration and exploitation efficiently. These methods help identify the best hyperparameters, ultimately enhancing the predictive performance and reliability of machine learning models.

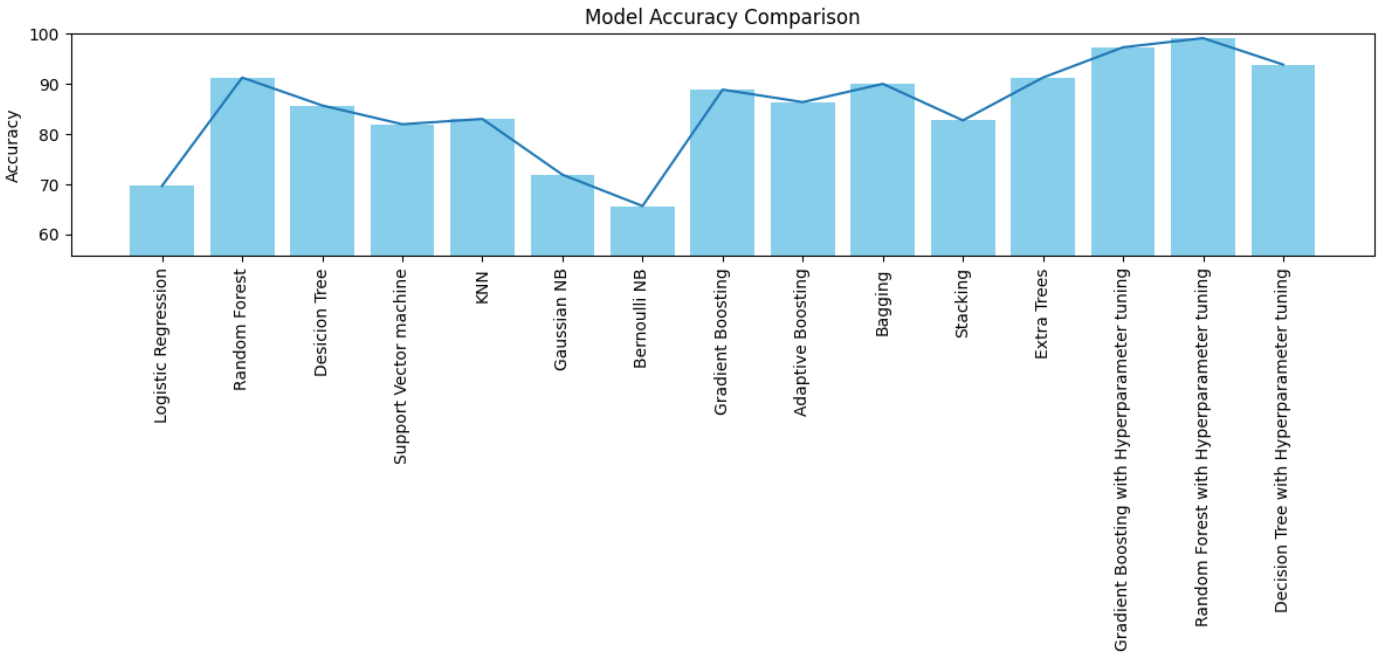
### 5.2 OUTPUT SCREENSHOTS

After analyzing various user criteria, the system successfully predicted whether a user is likely to retain their subscription for the platform or to churn from the subscription plan. Additionally, the system was able to identify the age category and gender of users who are more likely not to retain their plan. Furthermore, the system notifies users who are likely to churn. These results were achieved with an accuracy of 95%.





**Fig 5.2.1 Distribution plot (age vs churn) Fig 5.2.2 Weekly minutes watched**



**Fig 5.2.3 Models vs Accuracy**

## CHAPTER 6

**CONCLUSION AND FUTURE ENHANCEMENTS**

### CONCLUSION

One particularly strong collaborative educational framework that may be used for classification as well as regression tasks is the Random Forest method. To produce a final prediction, it generates many decision trees and aggregates their outputs. This method allows the algorithm to quickly reach a high prediction accuracy, especially in the initial phases. Utilizing this technology for medical records, especially those connected to heart health, has enormous promise for promoting early identification of heart-related problems and, ultimately, saving lives.

The prediction of heart attacks is a major difficulty in the modern world. By entering pertinent parameters from their medical reports, people can use these programs to predict the risk of a heart attack in situations where they do not have instant access to medical personnel. Equipped with this knowledge, individuals can decide whether to seek medical attention, potentially reducing the severity of the crisis.

### FUTURE ENHANCEMENTS

In the future, this platform holds potential for augmentation through the integration of new features. For instance, upon the prediction of a heart attack, a feature could be implemented to dispatch notifications to all family members of the patient. Simultaneously, pertinent information would be relayed to the nearest hospital for immediate attention. Moreover, an additional functionality could facilitate online consultations between physicians, enhancing collaborative medical discussions.

It's crucial to emphasize that machine learning (ML) technologies are used for more than only predicting and analysing heart attacks. Additionally, ML algorithms play a key role in the improvement of several medical sectors, including radiology, bioinformatics, and medical imaging analysis.

### REFERENCES

[1] [W. Sullivan, *Decision Tree and Random Forest: Machine Learning and Algorithms: The Future Is Here!* Createspace Independent Publishing Platform, 2018.](http://paperpile.com/b/ivjfOj/qOTM)

[2] [L. Owen, *Hyperparameter Tuning with Python: Boost your machine learning model’s performance via hyperparameter tuning*. Packt Publishing Ltd, 2022.](http://paperpile.com/b/ivjfOj/qvLr)

[3] [Kalorth and Nithin, *Exploring the Impact of OTT Media on Global Societies*. IGI Global, 2024.](http://paperpile.com/b/ivjfOj/vfsw)

[4] [L. Harte, *OTT Business Opportunities: Streaming TV, Advertising, TV Apps, Social TV, and TCommerce*. Discovernet, 2020.](http://paperpile.com/b/ivjfOj/DVjR)

[5] [G. Sheetrit, *Over The Top SEO (OTT) is a digital marketing & Professional SEO Agency: Professional SEO OTT*. BookRix, 2023.](http://paperpile.com/b/ivjfOj/tRzf)

[6] [H. A. Mengash, N. Al Ruwais, F. Kouki, C. Singla, E. S. Abd Elhameed, and A. Mahmud, “Archimedes Optimization Algorithm-Based Feature Selection with Hybrid Deep-Learning-Based Churn Prediction in Telecom Industries,” *Biomimetics*, vol. 9, no. 1, Dec. 2023, doi:](http://paperpile.com/b/ivjfOj/F5Xe) [10.3390/biomimetics9010001.](http://dx.doi.org/10.3390/biomimetics9010001.)

[7] [K. Peng, Y. Peng, and W. Li, “Research on customer churn prediction and model interpretability analysis,” *PLoS One*, vol. 18, no. 12, p. e0289724, Dec. 2023.](http://paperpile.com/b/ivjfOj/j6yE)

[8] [J. Faritha Banu, S. Neelakandan, B. T. Geetha, V. Selvalakshmi, A. Umadevi, and E. O. Martinson, “Artificial Intelligence Based Customer Churn Prediction Model for Business Markets,” *Comput. Intell. Neurosci.*, vol. 2022, p. 1703696, Sep. 2022.](http://paperpile.com/b/ivjfOj/tLaB)

[9] [I. B. Brandusoiu, *BAYESIAN NETWORKS FOR CHURN PREDICTION IN THE MOBILE TELECOMMUNICATIONS INDUSTRY*. GAER Publishing House, 2020.](http://paperpile.com/b/ivjfOj/L1qK)

[10] [T. Agrawal, *Hyperparameter Optimization in Machine Learning: Make Your Machine Learning and Deep Learning Models More Efficient*. Apress, 2020.](http://paperpile.com/b/ivjfOj/pBJy)

[11] [Y. Zhou, W. Chen, X. Sun, and D. Yang, “Early warning of telecom enterprise customer churn based on ensemble learning,” *PLoS One*, vol. 18, no. 10, p. e0292466, Oct. 2023.](http://paperpile.com/b/ivjfOj/rrvG)

[12] [I. AlShourbaji, N. Helian, Y. Sun, A. G. Hussien, L. Abualigah, and B. Elnaim, “An efficient churn prediction model using gradient boosting machine and metaheuristic optimization,” *Sci. Rep.*, vol. 13, no. 1, p. 14441, Sep. 2023.](http://paperpile.com/b/ivjfOj/aSMk)

[13] [T. Xu, *A New CRM System for Telecoms Customer Churn Analysis Based on Ensemble Learning and RFM*. Overseas Chinese Press, 2023.](http://paperpile.com/b/ivjfOj/xYDC)

[14] [E. Crincoli *et al.*, “Deep learning for automatic prediction of early activation of treatment naïve non-exudative MNVs in AMD,” *Retina*, Mar. 2024, doi:](http://paperpile.com/b/ivjfOj/b98D) [10.1097/IAE.0000000000004106.](http://dx.doi.org/10.1097/IAE.0000000000004106.)

[15] [G. Fatima, S. Khan, F. Aadil, D. H. Kim, G. Atteia, and M. Alabdulhafith, “An autonomous mixed data oversampling method for AIOT-based churn recognition and personalized recommendations using behavioral segmentation,” *PeerJ Comput Sci*, vol. 10, p. e1756, Jan. 2024.](http://paperpile.com/b/ivjfOj/tRxj)

[16] [M. Zheng, *Spatially Explicit Hyperparameter Optimization for Neural Networks*. Springer Nature, 2021.](http://paperpile.com/b/ivjfOj/00UG)

[17] [I. Vasilev, D. Slater, G. Spacagna, P. Roelants, and V. Zocca, *Python Deep Learning: Exploring deep learning techniques and neural network architectures with PyTorch, Keras, and TensorFlow*. Packt Publishing Ltd, 2019.](http://paperpile.com/b/ivjfOj/SdzO)

[18] [S. Vajjala, B. Majumder, A. Gupta, and H. Surana, *Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems*. O’Reilly Media, 2020.](http://paperpile.com/b/ivjfOj/kqSd)

[19] [S. Liu *et al.*, “Interaction Pattern Disentangling for Multi-Agent Reinforcement Learning,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PP, May 2024, doi:](http://paperpile.com/b/ivjfOj/9bBh)