IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, May 2024

Disease Prevalence Estimation

Lokesh Singhvi, Satyam Pathak, Harvi Patel, Bhoumik Rajput, Prof. Revati Raspayle
Department of Computer Science and Engineering
MIT ADT University Pune, India

Abstract: *Nowadays, disease prevalence estimation is a significant concern, with heart disease being one of the most common ailments. Unfortunately, the treatment of such diseases can be costly, often beyond the means of the average individual. However, we can mitigate this issue to some extent by accurately estimating disease prevalence before it reaches dangerous levels, using techniques such as Machine Learning and Data Mining.*

In the healthcare biomedical field, there's a vast amount of health data available, ranging from text to images. However, much of this data remains unexplored and unmined. Introducing a Disease Prevalence Estimation System could address this gap. Such a system would not only help in reducing costs but also enhance the quality of treatment for patients.

Machine Learning and Data Mining techniques can be employed to construct this Disease Prevalence Estimation System. By analyzing patient profiles including factors like blood pressure, age, sex, cholesterol, and blood sugar levels, the system can predict the likelihood of individuals developing various health issues.

Furthermore, the system can identify complex problems and make intelligent medical decisions, thereby improving overall healthcare outcomes. Performance evaluation can be done using metrics such as the confusion matrix, allowing for the calculation of accuracy, precision, and recall.

In conclusion, a Disease Prevalence Estimation System has the potential to offer high performance and better accuracy, thus significantly contributing to the early detection and management of various diseases..

Keywords: Disease Prevalence Estimation System

I. INTRODUCTION

The healthcare industry accumulates vast amounts of data containing hidden insights crucial for informed decision-making. To harness this information effectively, advanced data mining techniques are indispensable. In this research, a Disease Prevalence Estimation System (DPES) is developed utilizing Naive Bayes and Decision Tree algorithms to predict the risk level of various diseases, including heart disease.

The DPES integrates 15 medical parameters, encompassing factors such as age, sex, blood pressure, cholesterol, and obesity, to forecast the likelihood of patients developing heart disease. By employing these algorithms, the system facilitates the extraction of significant knowledge, including the establishment of relationships between medical factors associated with heart disease and discernible patterns.

Furthermore, a multilayer perceptron neural network with backpropagation serves as the training algorithm, enhancing the system's predictive capabilities. The obtained results underscore the effectiveness of the designed diagnostic system in accurately predicting the risk level of heart diseases, thereby empowering healthcare practitioners with valuable insights for proactive disease management.

Problem Statement

In the context of user service systems facilitating direct interaction with users, accurately predicting heart disease poses a significant challenge. This research aims to develop a robust predictive model integrated within such systems, utilizing user-provided data to deliver precise risk assessments for heart disease. Additionally, the system will incorporate a feature to suggestnearby doctors in the event of a health concern, thereby enhancing user support services, improving health outcomes, and promoting proactive healthcare management.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, May 2024

II. METHODOLOGY

Data Collection

The data utilized in this study were sourced from a publicly available dataset containing medical records of patients, encompassing various demographic and clinical attributes. These attributes include sex, chest pain type, fasting blood sugar, restecg, exercise-induced angina (exang), slope of the peak exercise ST segment, number of major vessels colored by fluoroscopy (CA), thalliumstress test result (thal), resting blood pressure, serum cholesterol level, maximum heart rate achieved during exercise (thalach), ST depression induced by exercise relative to rest (oldpeak), and age.

Data Preprocessing

Before conducting analyses, the dataset underwent rigorous preprocessing to ensure its suitability for machine learning tasks. This involved handling missing values, which were imputed using appropriate strategies such as mean or median imputation. Categorical variables were encoded using one-hot encoding or label encoding as applicable. Numerical features werescaled to a standardized range to prevent any biases during model training

Feature Selection

Feature selection was performed to identify the most relevant attributes for predicting heartdisease. This step involved evaluating the correlation between each feature and the target variable, as well as conducting exploratory data analysis to understand the relationships between variables. Features deemed redundant or non-informative were excluded from further analysis to enhance model performance and interpretability.

Model Development

The heart disease prediction model was developed using a combination of Gradient Boosting Classifiers (GBC) and Logistic Regression (LR). The choice of these algorithms was based on their demonstrated effectiveness in handling binary classification tasks and their ability to capture complex relationships within the data.

Gradient Boosting Classifiers (GBC): A gradient boosting ensemble technique was employed to iteratively train a sequence of weak learners, typically decision trees, with each subsequent learner focusing on the residual errors of its predecessors. This approach allows the model to learn from its mistakes and continuously improve predictive performance.

Logistic Regression (LR): Logistic regression was utilized as a baseline model to establish a benchmark for comparison against the more complex GBC ensemble. This linear regression-based algorithm models the probability of the presence of heart disease based onthe input features.

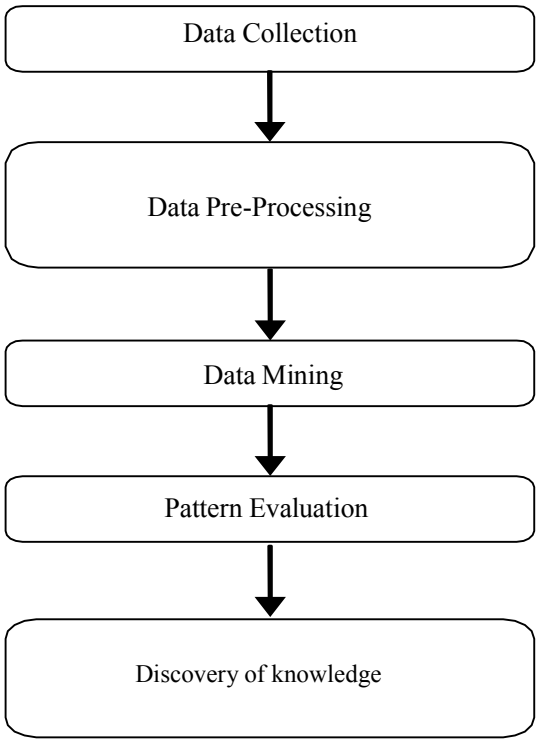
Model Evaluation

The performance of the developed models was assessed using standard evaluation metricssuch as accuracy, precision, recall, and F1-score. The dataset was randomly split into training and testing sets to evaluate model generalization performance. Cross-validation techniques were employed to mitigate overfitting and ensure robustness of the results.

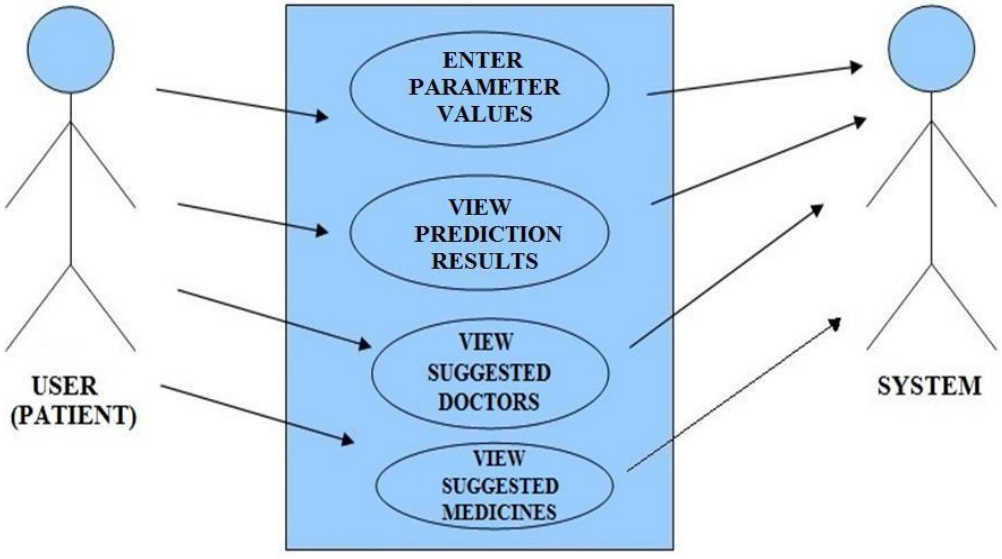
System Implementation

The heart disease prediction system was implemented using Python programming languageand relevant libraries such as Pandas for data manipulation and Scikit-learn for model development and evaluation. A user-friendly interface was designed to allow seamless interaction with the system, enabling users to input their medical parameters and receive personalized risk assessments for heart disease.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, May 2024



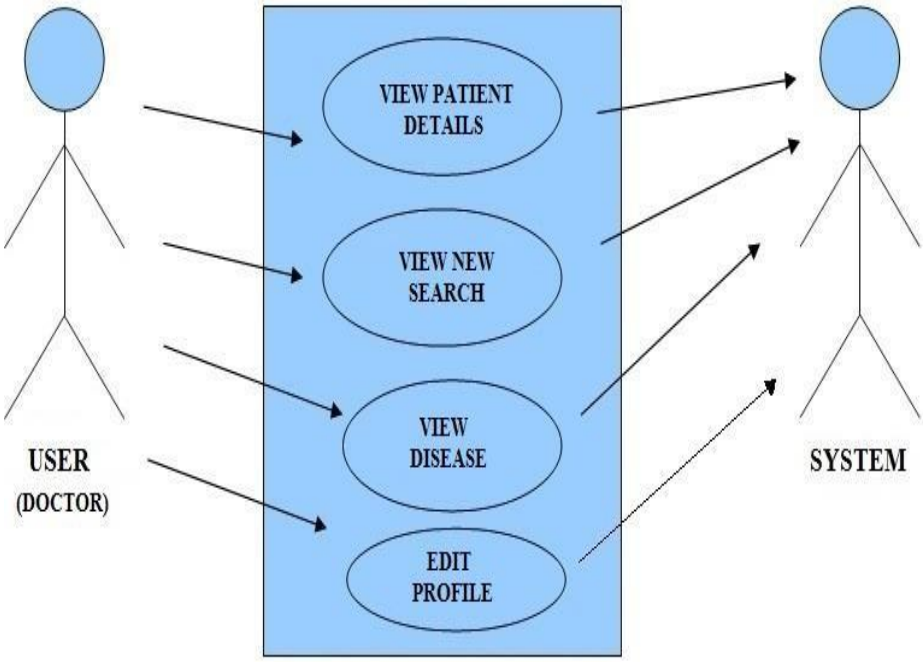
Use case diagram between patient and system –



Use Case Diagram between user and system

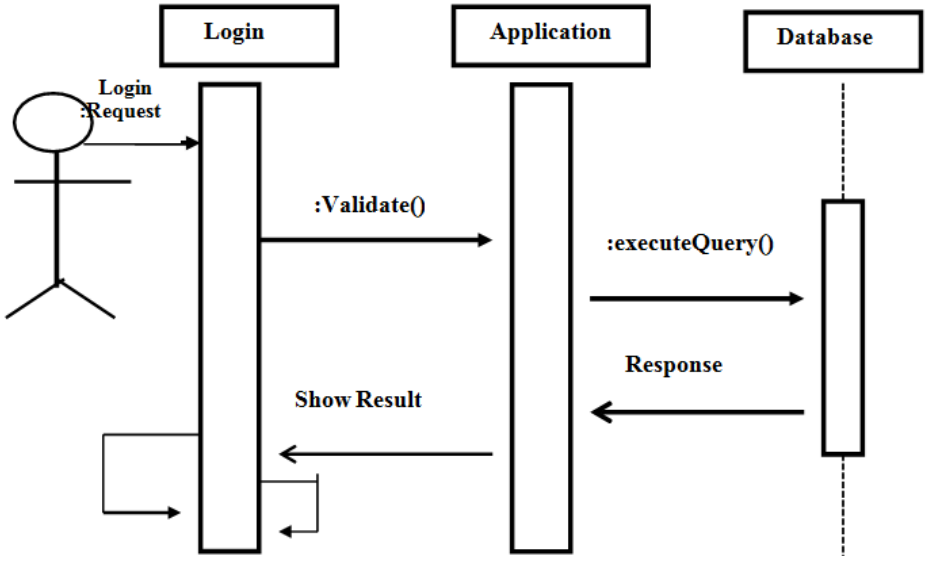
IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, May 2024

Use Case Diagram between doctor and system -



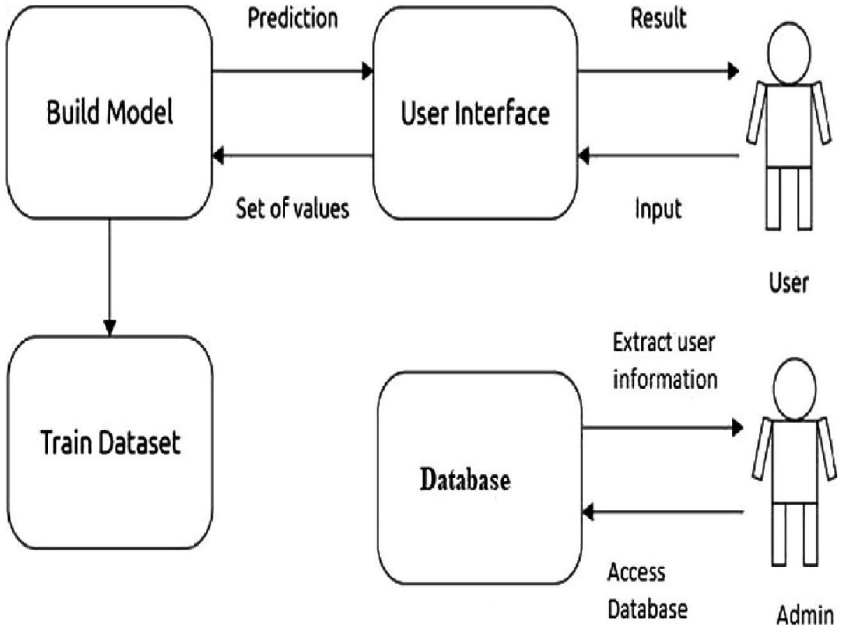
Use Case Diagram between user and system

Sequence diagram for administrator or user–

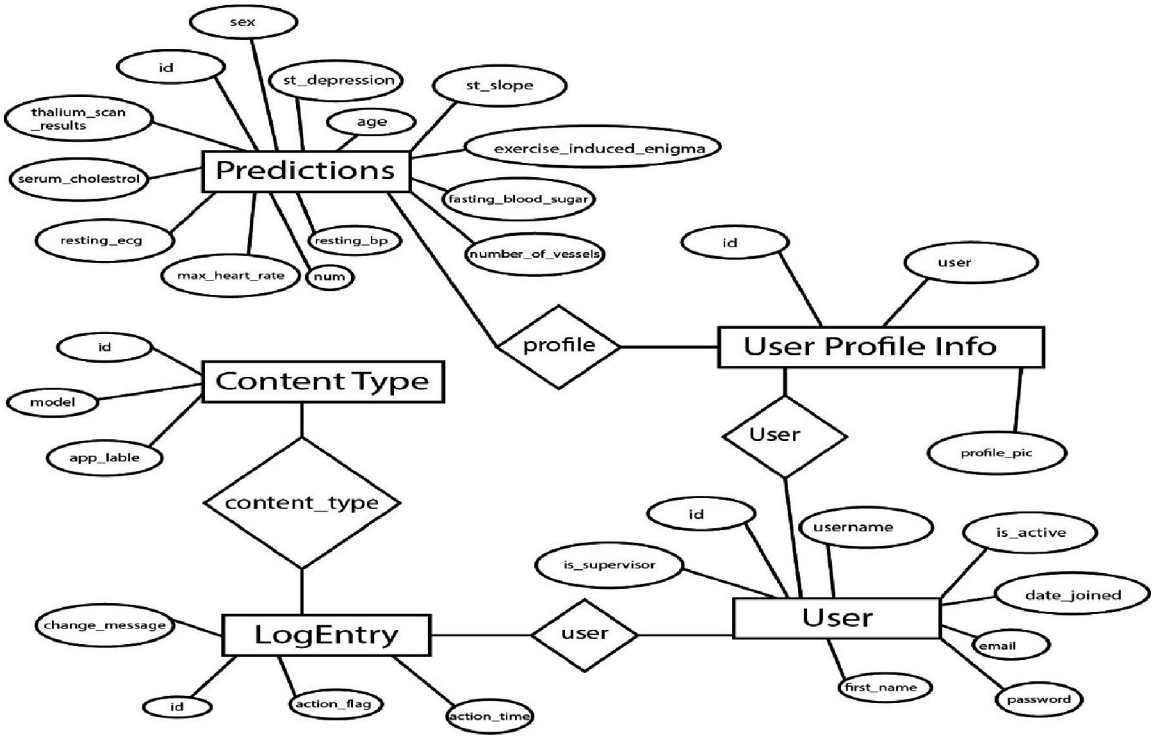


IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, May 2024

Data flow diagram –

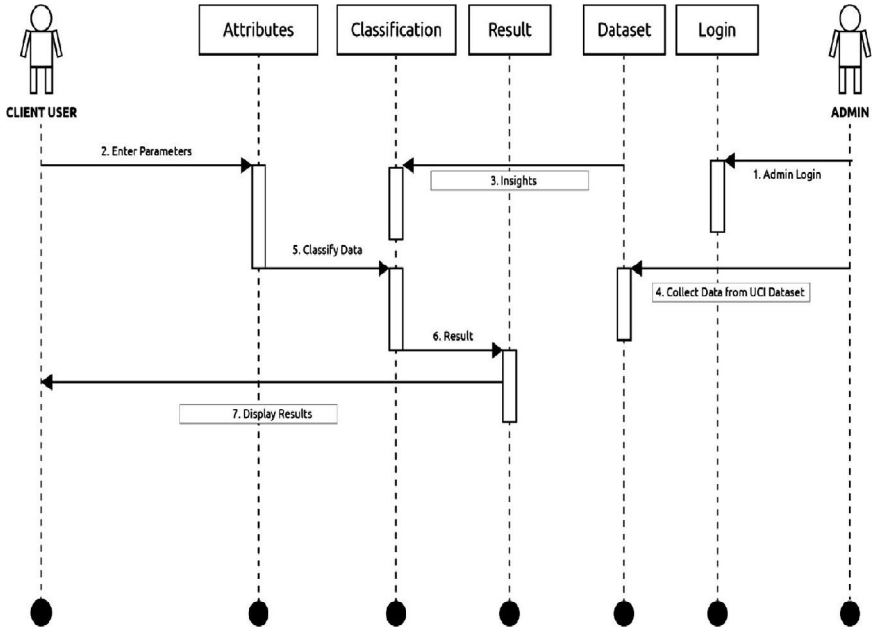


ER Diagram –



IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, May 2024

Activity Diagram –



Our Contributions

Our research contributes significantly to the field of healthcare technology and machine learning by developing a comprehensive heart disease prediction system that integrates machine learningalgorithms with web technologies. By leveraging Python, Django, HTML, CSS, JavaScript, and various machine learning algorithms such as Gradient Boosting Classifiers (GBC), Logistic Regression (LR), Support Vector Machines (SVM), and Random Forests, we created a cohesiveplatform that seamlessly integrates frontend and backend components. Through rigorous testing and evaluation, we identified GBC as the most effective algorithm, achieving around 92% accuracy in heart disease prediction. This system empowers individuals to assess their risk of heart disease based on personal medical parameters and facilitates proactive healthcare management by providing personalized risk assessments and suggesting nearby doctors for consultation. Our research also opens up avenues for future exploration, including enhancing model performance, incorporating additional data sources, implementing real-time monitoring capabilities, expanding to predict other chronic diseases, and integrating with electronic health record systems. These potential areas for research and development have the potential to further advance healthcare technology and improve patient outcomes. Overall, our research aims to make a positive impact on the well-being of individuals and communities by leveraging innovative technologies and methodologies in preventive healthcare

III. LITERATURE SURVEY

1.Year -2022

Author - Prasannavenkatesan Theerthagiri

Title - Predictive Analysis of CardiovascularDisease using Gradient Boosting based Learning and Recursive Feature Elimination Technique

Description - Heart disease remains one of the most prevalent chronic illnesses affecting individuals worldwide. Timely identification of risk factors and early intervention can significantly reduce mortality rates by preventing or mitigating the severity of cardiovascular disease (CVD). In recent years, machine learning algorithms have emerged as a promising approach for detecting risk indicators associated with heart disease.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, May 2024

To address the need for accurate cardiac disease prediction, this study introduces a novel approach known as recursive feature elimination-based gradient boosting (RFE-GB). This method aims to leverage the power of machine learning algorithms to analyze patients' health records and identify crucial characteristics related to CVD. By iteratively selecting and prioritizing features that contribute most to predictive accuracy, the RFE-GB approach seeks to improve the effectiveness of heart disease prediction models.

In addition to the RFE-GB approach, this study explores various other machine learning techniques for building predictive models of cardiac disease. These models are evaluated and compared to the proposed RFE-GB method to assess their performance and efficacy in predicting CVD risk.

The findings of this study indicate that the combined RFE-GB approach achieves the highest accuracy among the evaluated models, with an accuracy rate of 88.8%.

Furthermore, the RFE-GB method demonstrates superior performance compared to previous strategies, as evidenced by its area under the curve (AUC) value of 0.84.

In conclusion, the developed RFE-GB method holds significant promise as a reliable model for predicting and managing cardiovascular disease. By leveraging advanced machine learning techniques and analyzing comprehensive health records, this approach has the potential to enhance early detection and treatment of CVD, ultimately improving patient outcomes and reducing the burden of heart disease worldwide.

2. Dataset Research –

Title - Cardiovascular disease dataset, retrieved from Kaggle repository

Description - Heart disease remains a significant public health concern worldwide, with early identification of risk factors crucial for effective prevention and management. In recent years, machine learning algorithms have emerged as promising tools for detecting risk indicators associated with cardiovascular diseases (CVDs). Datasets play a crucial role in cardiovascular research, providing researchers with the necessary information to develop and evaluate predictive models for CVD diagnosis and risk assessment. One such dataset of interest is the "Cardiovascular Disease Dataset" available on Kaggle, curated by Sulianova in 2021. The Cardiovascular Disease Dataset offers a comprehensive collection of health-related data, encompassing a wide range of patient attributes and clinical measurements. Compiled from over 70,000 records spanning multiple years, the dataset includes demographic factors, lifestyle habits, and clinical measurements such as blood pressure, cholesterol levels, and glucose levels. Binary labels indicating the presence or absence of cardiovascular disease enable researchers to develop predictive models for disease diagnosis and risk assessment. By leveraging datasets such as the Cardiovascular Disease Dataset, researchers can employ data-driven approaches to explore the complex interplay between patient characteristics and cardiovascular health outcomes. These approaches facilitate the identification of novel biomarkers, validation of existing risk prediction models, and development of data-driven strategies for CVD prevention and management. In conclusion, the Cardiovascular Disease Dataset serves as a valuable resource for cardiovascular research, providing researchers with the necessary data to advance our understanding of CVDs and improve patient care. By incorporating datasets into their analyses, researchers can enhance the transparency and reproducibility of their studies while contributing to the broader efforts aimed at combating cardiovascular diseases.

3.Year – 2023

Title - Implementation of a Heart Disease Risk Prediction Model Using Machine Learning.

Authors - K. Karthick, S. K. Aruna, R. Samikannu, R. Kuppusamy, Y. Teekaraman, and A. R.Thelkar

Description - The research proposes a hybrid model for heart disease prediction, utilizing a fitness function, genetic operators, and a rule encoding method [23]. The study evaluates the classification accuracy of several machine learning algorithms, including Support Vector Machine (SVM) with Radial Basis Function (RBF) kernel, Gaussian Naive Bayes, Logistic Regression, LightGBM, XGBoost, and Random Forest, using the

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, May 2024

Cleveland Heart Disease dataset from the UCI Machine Learning repository [24]. The dataset consists of 303 data instances with a subset of 13 attributes selected based on their significance in predicting heart disease. Attributes such as gender, chest pain category, and fasting blood sugar level are visualized to illustrate their relationship with cardiovascular disease risk [25]. Statistical tests, including the chi-square test, are employed for feature selection, resulting in the selection of 13 best features for model development. Data visualization techniques, such as heat maps and distribution plots, provide insights into attribute correlations and the prevalence of cardiovascular disease across different age groups.

4. Year - 2017
Title – Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques
Authors - Animesh Hazra, 2Subrata Kumar Mandal, 3Amit Gupta, 4Arkomita Mukherjee and Asmita Mukherjee
Description - the literature review explores several key algorithms commonly used in decision-making and classification tasks within machine learning. Decision trees, represented as tree-like graphs, are fundamental tools utilized in operations research and decision analysis. They offer a visual representation of algorithms and aid in identifying strategies to achieve desired outcomes. Among the algorithms discussed, the C4.5 algorithm stands out as a decision tree classifier known for its speed, popularity, and easily interpretable output.
Additionally, the K-means algorithm is highlighted for its effectiveness in partitioning datasets into clusters based on similarities, particularly useful for large datasets where the number of clusters is unknown. The ID3 algorithm is presented as a top-down decision tree building algorithm, emphasizing its ability to partition datasets into homogeneous subsets based on classification criteria.
Furthermore, Support Vector Machines (SVMs) are introduced as powerful supervised learning methods for binary classification tasks, leveraging the concept of maximizing the margin between data points. Naive Bayes classifiers are discussed for their simplicity and effectiveness in constructing classifiers based on probabilistic principles. Lastly, Artificial Neural Networks (ANNs) are described as computational models inspired by biological neural networks, capable of capturing complex relationships between inputs and outputs.
Overall, the literature survey provides an insightful overview of various machine learning algorithms and their applications in decision-making and classification tasks, laying the groundwork for further exploration and implementation in diverse domains.

Learning's –
Introduction to Technologies Used
The development of the heart disease prediction system encompassed both machine learning algorithms and frontend technologies, with Python, Django, HTML, CSS, and JavaScript playing pivotal roles.

Machine Learning Algorithms
The heart disease prediction model served as the cornerstone of the system, leveraging various machine learning algorithms to analyze patient data and make accurate predictions. These algorithms were implemented using Python's scikit-learn library, allowing for efficient model development and evaluation.

Backend Development
Python, in conjunction with the Django web framework, facilitated the implementation of the backend logic for the heart disease prediction system. Django provided a robust foundation for building web applications, enabling backend developers to handle data processing, model integration, and system functionality seamlessly.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, May 2024

Frontend Development

HTML, CSS, and JavaScript formed the frontend stack for the heart disease prediction system. HTML provided the structure of the user interface, CSS facilitated styling and layout design, while JavaScript enhanced interactivity and user experience. The frontend components were designed to be intuitive and user-friendly, ensuring a seamless experience for patients interacting with the system.

Hands-on Experience with Technologies

The development process involved practical experience with both machine learning algorithms and frontend technologies. Machine learning engineers experimented with various algorithms, including Gradient Boosting Classifiers (GBC), Logistic Regression (LR), Support Vector Machines (SVM), and Random Forests, among others. Through rigorous testing and evaluation, it was found that the Gradient Boosting Classifier achieved the highest accuracy, reaching around 92%.

Frontend developers honed their skills in HTML, CSS, and JavaScript, mastering the design and implementation of user interfaces that are visually appealing and responsive. They collaborated closely with backend developers to ensure seamless integration of frontend and backend components, facilitating a cohesive user experience.

Collaboration and Integration

Collaboration among team members, including machine learning engineers, frontend developers, backend developers, and domain experts, was instrumental in the success of the project. Regular communication and knowledge sharing sessions facilitated collaboration and ensured alignment with project goals and requirements.

The seamless integration of machine learning algorithms and frontend components required close coordination between different teams. Agile development methodologies were employed to iteratively build and refine the heart disease prediction system, incorporating feedback and addressing challenges as they arose.

In this study, we developed a heart disease prediction system leveraging machine learning algorithms and web technologies. Through the integration of Python, Django, HTML, CSS, and JavaScript, we created a user-friendly interface for patients to input their medical parameters and receive personalized risk assessments for heart disease.

Our investigation into various machine learning algorithms revealed that Gradient Boosting Classifiers (GBC) achieved the highest accuracy, reaching around 92%. This demonstrates the effectiveness of ensemble learning techniques in accurately predicting heart disease risk based on patient data.

The successful development and implementation of the heart disease prediction system have significant implications for healthcare management. By providing patients with timely risk assessments and suggesting nearby doctors for consultation, the system can facilitate proactive healthcare interventions and improve patient outcomes.

Looking ahead, there are several avenues for future research and development:

1. Enhancing Model Performance:

Further optimization of machine learning algorithms and exploration of ensemble techniques could lead to even higher accuracy levels in heart disease prediction.

2. Incorporating Additional Data Sources:

Integration of additional data sources, such as genetic information or lifestyle factors, could enhance the predictive capabilities of the system and provide more comprehensive risk assessments.

3. Implementing Real-time Monitoring:

Development of functionality for real-time monitoring of patient health parameters could enable early detection of cardiovascular risk factors and timely intervention.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 3, May 2024

4. Expanding to Other Diseases:
Extending the system to predict other chronic diseases, such as diabetes or hypertension, would broaden its utility and impact on preventive healthcare.
5. Integration with Electronic Health Records (EHR):
Integration with existing electronic health record systems could streamline data collection and provide healthcare providers with valuable insights for personalized patient care.

In conclusion, the heart disease prediction system developed in this study represents a significant advancement in leveraging technology to address healthcare challenges. By harnessing the power of machine learning and web technologies, we aim to empower individuals to take proactive steps towards better heart health and ultimately improve overall well-being.

REFERENCES

[1] Prasannavenkatesan Theerthagiri , Predictive Analysis of CardiovascularDisease using Gradient Boosting based Learning and Recursive Feature Elimination Technique, Intel-ligent Systems with Applications (2022), doi: https://www.sciencedirect.com/science/article/pii/S266730532200059X?via%3Dihub
[2] Cardiovascular disease dataset, retrieved from Kaggle repository, https://www.kaggle.com/sulianova/cardiovascular-disease-dataset, 2021.
[3] Hindawi Computational and Mathematical Methods in Medicine Volume 2022, Article ID 6517716, 14 pages https://doi.org/10.1155/2022/6517716
[4] Advances in Computational Sciences and Technology ISSN 0973-6107 Volume 10, Number 7 (2017) pp. 2137-2159 © Research India Publications http://www.ripublication.com