Heart Disease Prediction

Submitted by:

221FA04107	221FA04160	221FA04221	221FA04225
N M N C Sai Lakshmi	Sk Khaleel Babu	M Anusha	P Sainath Reddy

Under the guidance of

Dr B Suvarna

VIGNAN'S
FOUNDATION FOR SCIENCE, TECHNOLOGY & RESEARCH
(Deemed to be University) - Estd. u/s 3 of UGC Act 1956

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING VIGNAN'S FOUNDATION FOR SCIENCE, TECHNOLOGY AND RESEARCH Deemed to be UNIVERSITY

Vadlamudi, Guntur.

ANDHRA PRADESH, INDIA, PIN-522213.

VIGNAN'S
FOUNDATION FOR SCIENCE, TECHNOLOGY & RESEARCH
(Deemed to be University) - Estd. u/s 3 of UGC Act 1956

CERTIFICATE

This is to certify that the Field Project entitled **"Heart Disease Prediction"** that is being submitted by **221FA04107 (N M N C Sai Lakshmi)**, **221FA04160 (Sk Khaleel Babu)**, **221FA04221 (M Anusha)**,**221FA04225(P Sainath Reddy)** for partial fulfilment of Field Project is a bonafide work carried out under the supervision of **Dr.B.Suvarna, Department of CSE**.

Guide name& Signature

Dr. S. V. Phani Kumar

Dr.K.V. Krishna Kishore

Assistant/Associate/Professor,CSE

HOD CSE

Dean, SoCI

VIGNAN'S

FOUNDATION FOR SCIENCE, TECHNOLOGY & RESEARCH

(Deemed to be University) - Estd. u/s 3 of UGC Act 1956

DECLARATION

We hereby declare that the Field Project entitled **“Heart Disease Prediction”** is being submitted by **221FA04107 (N M N C Sai Lakshmi)**, **221FA04160 (Sk Khaleel Babu)**, **221FA04221(M Anusha)**,**221FA04225(P Sainath Reddy)** in partial fulfilment of Field Project course work. This is our original work, and this project has not formed the basis for the award of any degree. We have worked under the supervision of **Dr.B.Suvarna Assistant Professor, Department of CSE**.

By

221FA04107(N M N C
Sai Lakshmi),
221FA04160 (Sk Khaleel Babu),
221FA04221 (M Anusha), 221FA04225(P
Sainath Reddy)

Date:

ABSTRACT

Heart disease is a leading cause of death globally, making early detection essential for prevention. This study focuses on predicting heart disease using data mining techniques, with a primary focus on logistic regression due to its simplicity and effectiveness in binary classification. We review other methods, such as decision trees and neural networks, and justify logistic regression's selection. Using a heart disease dataset, we preprocess the data, select key features, and develop a logistic regression model to predict the likelihood of heart disease. The model is evaluated using metrics like accuracy, precision.

Results show that logistic regression provides reliable predictions while maintaining interpretability. This makes it suitable for healthcare applications where explainability is crucial. Future work may include integrating advanced machine learning models to enhance prediction accuracy.

Table of contents

List of figures

CHAPTER-1

INTRODUCTION

1. INTRODUCTION

1.1 Overview of Heart Disease

This section introduces heart disease, describing its causes, risk factors, and global prevalence. Heart disease refers to a range of conditions affecting the heart, such as coronary artery disease, heart attacks, and arrhythmias. Highlight the major contributors like high cholesterol, hypertension, smoking, and poor diet. Also, discuss its significant impact on public health.

1.2 Importance of Early Prediction

Early detection is crucial in preventing serious health issues and reducing mortality rates associated with heart disease. Predictive models help in identifying individuals at high risk before symptoms manifest, allowing timely intervention. The goal is to save lives and reduce the cost of long-term treatment by targeting high-risk patients early.

1.3 Role of Data Mining in Healthcare

Data mining techniques have become instrumental in extracting meaningful patterns from large datasets in healthcare. These methods are particularly useful for predictive analytics, allowing the identification of disease patterns and trends. Discuss how data mining helps in diagnosing diseases, personalizing treatment, and improving healthcare efficiency by analyzing historical data.

CHAPTER-2

LITERATURE SURVEY

2.Literature Review

2.1 Data Mining Techniques for Heart Disease Prediction

This section reviews different data mining techniques that have been applied in heart disease prediction. Common techniques include classification models like decision trees, random forests, neural networks, and support vector machines. These methods extract patterns from clinical data to predict the likelihood of heart disease.

2.2 Overview of Logistic Regression in Medical Predictions

Logistic regression is widely used in medical predictions because of its simplicity and interpretability. This technique predicts binary outcomes, such as whether a patient has heart disease or not, based on clinical risk factors. Highlight the advantages of logistic regression, such as its ability to provide insights into the relationship between risk factors and outcomes.

2.3 Comparative Studies on Machine Learning Models for Heart Disease Prediction

Compare the performance of various machine learning models used for heart disease prediction, such as decision trees, support vector machines (SVM), and neural networks, with logistic regression. Discuss their accuracy, complexity, and interpretability, and highlight why logistic regression remains a preferred choice in healthcare.

CHAPTER-3

Data mining techniques

3. Data Mining Techniques

3.1 Overview of Common Data Mining Techniques

Explain the commonly used data mining techniques in predictive modeling:

- **3.1.1 Decision Trees**: A simple, intuitive model that splits data into branches based on features.
- **3.1.2 Random Forests**: An ensemble method that improves accuracy by combining multiple decision trees.
- **3.1.3 Support Vector Machines (SVM)**: A powerful model that separates data into classes using hyperplanes.
- **3.1.4 Neural Networks**: Complex models capable of learning intricate patterns in large datasets.

3.2 Selecting a Suitable Technique

Criteria for choosing an appropriate technique include model accuracy, interpretability, and computational complexity. Logistic regression is selected for its balance between simplicity and effectiveness, making it well-suited for predicting heart disease based on clinical data.

3.3 Logistic Regression: An Overview

Logistic regression is a statistical method used for binary classification problems. It estimates the probability of a patient having heart disease based on predictor variables, using the logistic function. Discuss its mathematical foundation, where the output is constrained between 0 and 1, representing probabilities.

CHAPTER -4

Logistic regression for heart disease prediction

4 **Logistic Regression for Heart Disease Prediction**

4.1 Theoretical Background of Logistic Regression

Logistic regression is based on the logistic function, which transforms a linear combination of input features into a probability score. Explain the underlying mathematics, including the concept of odds, and how the model uses coefficients to determine the importance of each predictor variable.

4.2 Application of Logistic Regression in Heart Disease Prediction

Logistic regression is applied to predict heart disease by analyzing risk factors such as age, blood pressure, cholesterol levels, and smoking history. It is useful in healthcare due to its ability to provide easily interpretable results that healthcare professionals can use to assess patient risk.

4.3 Assumptions and Limitations of Logistic Regression

Logistic regression assumes a linear relationship between the log odds of the dependent variable and the independent variables. It may not perform well with non-linear relationships or complex interactions between variables. Discuss other limitations, such as sensitivity to outliers and the need for balanced data.

CHAPTER -5

Data collection and preprocessing

5 Data Collection and Preprocessing

5.1 Data Sources (UCI Heart Disease Dataset or others)

Introduce the dataset used for the analysis, such as the UCI Heart Disease Dataset. Describe the data’s structure, including features like age, gender, cholesterol, blood pressure, and target variables indicating the presence or absence of heart disease.

5.2 Data Cleaning and Missing Value Handling

Preprocessing involves cleaning the dataset by removing or imputing missing values, addressing incorrect or inconsistent data, and handling outliers. Data cleaning is essential to ensure the model is trained on high-quality data, improving the reliability of predictions.

5.3 Feature Selection

Feature selection involves choosing the most relevant predictors for the model to improve accuracy and reduce complexity. Discuss methods for identifying significant features, such as correlation analysis or domain knowledge, which may include age, cholesterol levels, and blood pressure.

5.4 Data Normalization and Scaling

Data normalization or scaling ensures that all features are on a similar scale, which is particularly important for distance-based algorithms. Even in logistic regression, scaling can improve the model’s convergence during training.

CHAPTER -6

Implementation pf logistic regression model

6 Implementation of Logistic Regression Model

6.1 Steps in Logistic Regression Model Building

Outline the process of building the logistic regression model, including data splitting (into training and testing sets), training the model on the training data, and evaluating performance on the test data.

6.2 Model Training and Testing

Training involves fitting the logistic regression model to the training data and adjusting the model's coefficients to minimize the prediction error. Testing evaluates the model on unseen data, providing an estimate of how well it generalizes to new data.

6.3 Performance Metrics (Accuracy, Precision, Recall, F1 Score)

Performance is measured using several metrics:

- **Accuracy**: The percentage of correct predictions.
- **Precision**: The proportion of positive predictions that are correct.
- **Recall**: The proportion of actual positives that were correctly predicted.
- **F1 Score**: A balance between precision and recall, especially useful for imbalanced datasets.

CHAPTER -7

Evaluation and results

7. **Evaluation and Results**

7.1 Model Evaluation Criteria

Explains the criteria used to evaluate the performance of the logistic regression model, focusing on statistical significance and model reliability.

7.2 Model Results and Interpretations

Presents the results of the logistic regression model and interprets the outcomes, such as identifying key risk factors and their influence on heart disease prediction.

7.3 Comparison with Other Data Mining Techniques

Compares the logistic regression results with other data mining techniques to see which approach performs better for heart disease prediction.

7.4 Confusion Matrix and ROC Curve Analysis

Discusses the confusion matrix to evaluate the model’s performance on the test data and uses the ROC (Receiver Operating Characteristic) curve and AUC (Area Under the Curve) to measure classification performance.

CHAPTER -8

Discussion

8. Discussion

8.1 Key Findings

Summarizes the main insights gained from the study, including the logistic regression model’s effectiveness in predicting heart disease.

8.2 Challenges Faced During Model Implementation

Outlines the challenges encountered, such as imbalanced datasets, multicollinearity, or computational limitations, and how they were addressed.

8.3 Potential Improvements and Future Work

Suggests areas for future research or improvements, like incorporating other machine learning models, using larger datasets, or combining logistic regression with other techniques for better predictions.

CHAPTER -9

conclusion

9. Conclusion

9.1 Summary of Findings

Concludes by summarizing the main results of the study, including the importance of logistic regression in heart disease prediction.

9.2 Implications for Healthcare

Discusses the broader implications of the findings, including how predictive models can be implemented in healthcare systems for early detection and prevention strategies.

9.3 Recommendations for Future Research

Provides recommendations for future research, such as exploring other machine learning methods, deep learning models, or improving the model by incorporating additional clinical data.

CHAPTER -10

References

10. References

1.Han, j. and M. Kamber, Data Mining Concepts and Techniques. 2006: Morgan Kaufmann Publishers.

2. Lee, I.-N., S.-C. Liao, and M. Embrechts, Data mining techniques applied to medical information. Med. inform, 2000.

3. Obenshain, M.K., Application of Data Mining Techniques to Healthcare Data. Infection Control and Hospital Epidemiology, 2004.

4. Sandhya, J., et al., Classification of Neurodegenerative Disorders Based on Major Risk Factors Employing Machine Learning Techniques. International Journal of Engineering
and Technology, 2010. Vol.2, No.4.

5. Thuraisingham, B., A Primer for Understanding and Applying Data Mining. IT Professional IEEE, 2000.

6. Ashby, D. and A. Smith, The Best Medicine? Plus Magazine - Living Mathematics., 2005.

7. Liao, S.-C. and I.-N. Lee, Appropriate medical data categorization for data mining classification techniques. MED. INFORM., 2002. Vol. 27, no. 1, 59–67, .

8. Ruben, D.C.J., Data Mining in Healthcare: Current Applications and Issues. 2009.

9. Porter, T. and B. Green, Identifying Diabetic Patients: A Data Mining Approach. Americas Conference on Information Systems, 2009.

10. Mohd, H., Mohamed, S. H. S.: “Acceptance Model of Electronic Medical Record”, Journal of Advancing Information and Management Studies. 2(1), 75-92, 2005.

11. Microsoft Developer Network (MSDN). http://msdn2.microsoft.com/en-us/virtuallabs/aa740409.aspx, 2007.

12. Obenshain, M.K: “Application of Data Mining Techniques to Healthcare Data”, Infection Control and Hospital Epidemiology, 25(8), 690–695, 2004.

13. Sellappan, P., Chua, S.L.: “Model-based Healthcare Decision Support System”, Proc. Of Int. Conf. on Information Technology in Asia CITA’05, 45-50, Kuching, Sarawak, Malaysia, 2005

14. Tang, Z. H., MacLennan, J.: “Data Mining with SQL Server 2005”, Indianapolis: Wiley, 2005.

15. Thuraisingham, B.: “A Primer for Understanding and Applying Data Mining”, IT Professional, 28-31, 2000.

16. Weiguo, F., Wallace, L., Rich, S., Zhongju, Z.: “Tapping the Power of Text Mining”, Communication of the ACM. 49(9), 77-82, 2006.

17. Wu, R., Peters, W., Morgan, M.W.: “The Next Generation Clinical Decision Support: Linking Evidence to Best Practice”, Journal Healthcare Information Management. 16(4), 50-55, 2002.