HEART DISEASE PREDICTOR

CS19643 – FOUNDATIONS OF MACHINE LEARNING

Submitted by

MAGIZHAN SIVAKUMAR (2116220701154)

in partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE

ANNA UNIVERSITY, CHENNAI

MAY 2025

BONAFIDE CERTIFICATE

Certified that this Project titled **"HEART DISEASE PREDICTOR"** is the bonafide work of **"MAGIZHAN SIVAKUMAR (2116220701154)"** 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

Dr. V.Auxilia Osvin Nancy.,M.Tech.,Ph.D.,
SUPERVISOR,
Assistant Professor
Department of Computer Science and Engineering,
Rajalakshmi Engineering College,
Chennai-602 105.

Submitted to Mini Project Viva-Voce Examination held on _____

Internal Examiner **External Examiner**

ABSTRACT

Cardiovascular diseases, particularly heart disease, are among the leading causes of death globally, accounting for millions of fatalities each year. Early diagnosis and timely intervention are crucial in reducing the risk of severe outcomes. This project focuses on the development of a Heart Disease Predictor using machine learning techniques to identify individuals at risk based on various clinical and lifestyle parameters. The system is designed to analyze features such as age, gender, resting blood pressure, cholesterol level, fasting blood sugar, electrocardiographic results, maximum heart rate, exercise-induced angina, and other relevant medical data. The dataset used for model training and evaluation is sourced from reputable medical databases, ensuring reliability and consistency. Several classification algorithms—including Logistic Regression, Random Forest, Decision Tree, and Support Vector Machine—are explored and compared to determine the most accurate and efficient model for prediction.

The final model is selected based on performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. The best-performing model is then integrated into a simple, user-friendly interface that allows users or healthcare professionals to input patient data and receive real-time predictions. This tool serves not only as a predictive model but also as a decision support system, potentially aiding doctors in early diagnosis and treatment planning. Additionally, it promotes awareness and encourages individuals to seek medical advice if their risk level is high. Through this project, we demonstrate the promising role of artificial intelligence and data-driven approaches in the field of preventive healthcare and highlight their potential to improve outcomes and save lives.

ACKNOWLEDGMENT

Initially we thank the Almighty for being with us through every walk of our life and showering his blessings through the endeavour 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 endeavouring 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, M.E., 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 & our Project Coordinator **Dr. V. AUXILIA OSVIN NANCY.,M.Tech.,Ph.D.,** Assistant Professor Department of Computer Science and Engineering for his useful tips during our review to build our project.

MAGIZHAN SIVAKUMAR- 2116220701154

TABLE OF CONTENT

LIST OF FIGURES

CHAPTER 1

1.INTRODUCTION

Heart disease, also known as cardiovascular disease, refers to a range of conditions that affect the heart and blood vessels, such as coronary artery disease, arrhythmias, and heart failure. It is one of the most common and deadly health issues globally, responsible for nearly one-third of all deaths each year, according to the World Health Organization. The primary challenge in combating heart disease is its often silent progression—many patients are unaware of the symptoms until a severe or fatal event occurs. Lifestyle factors such as poor diet, lack of physical activity, smoking, and stress, combined with genetic predispositions, significantly increase the risk. Traditional methods of diagnosis involve clinical evaluations, lab tests, and manual interpretation by healthcare professionals, which can be time-consuming, expensive, and prone to human error. Therefore, there is a growing need for smart, reliable, and efficient systems that can assist in the early detection and prediction of heart disease.

With the rise of artificial intelligence and machine learning in the medical domain, predictive analytics has become a powerful approach to tackle complex health challenges. This project aims to design and implement a Heart Disease Predictor that uses machine learning algorithms to analyze patient health data and assess the likelihood of heart disease. By utilizing a dataset containing critical features such as age, gender, blood pressure, cholesterol, fasting blood sugar, resting ECG results, maximum heart rate, and exercise-induced angina, the system can learn patterns and correlations that may indicate the presence of cardiovascular conditions. Various classification models including Logistic Regression, Random Forest, Decision Tree, and Support Vector Machine are tested and compared to select the most accurate and efficient algorithm. The final predictive model is then integrated into a user-friendly interface, allowing healthcare providers or individuals to input patient data and receive immediate risk assessments. This tool has the potential to assist in faster diagnoses, support clinical decision-making, and ultimately contribute to reducing the global burden of heart disease through proactive healthcare.

The development of a Heart Disease Predictor using machine learning not only addresses the critical need for early diagnosis but also aligns with the growing trend of data-driven healthcare solutions. By incorporating machine learning models, this system offers a high degree of accuracy and scalability in predicting heart disease risk. The machine learning

algorithms are trained to identify complex patterns in medical data that may be overlooked by traditional diagnostic methods. Moreover, the system's ability to process large datasets quickly and provide real-time predictions has the potential to reduce the burden on healthcare professionals, allowing them to focus on more complex cases. This approach also democratizes healthcare by providing individuals with an accessible tool to assess their heart disease risk, promoting preventive measures and lifestyle changes before symptoms even emerge. Ultimately, the integration of AI into heart disease prediction is a step toward more personalized, efficient, and cost-effective healthcare systems that can help save lives and reduce healthcare costs globally.

CHAPTER 2

2.LITERATURE SURVEY

The application of machine learning in healthcare, particularly for predicting heart disease, has gained significant attention in recent years. Numerous studies have explored the potential of using medical datasets to predict the likelihood of cardiovascular conditions. One of the earliest studies in this field by Wolberg et al. (1995) introduced the use of decision tree algorithms to predict heart disease risk, showcasing the feasibility of machine learning in identifying patterns within clinical data. Since then, many other studies have expanded on this, using a wide range of algorithms such as Naive Bayes, k-Nearest Neighbors (k-NN), Support Vector Machines (SVM), and Random Forest. These models have been applied to datasets containing various features, including demographic details, medical history, physical health indicators, and diagnostic test results. The predictive power of these models has been proven to be highly effective in identifying individuals at risk for heart disease, thus assisting in early detection and intervention.

A pivotal dataset often used in heart disease prediction research is the Cleveland Heart Disease dataset, which has been widely referenced in academic studies. This dataset includes various health parameters such as age, sex, chest pain type, resting blood pressure, serum cholesterol, blood sugar, ECG results, and maximum heart rate. Research by Alam et al. (2017) demonstrated the effectiveness of machine learning algorithms, specifically Support Vector Machines (SVM), in predicting the presence or absence of heart disease with high accuracy. They achieved a classification accuracy of over 85%, highlighting the reliability of SVM as a predictive model for heart disease risk. In addition, studies have also explored the use of ensemble methods, such as Random Forests, to improve the robustness and generalizability of predictions. These ensemble methods combine multiple decision trees to enhance prediction accuracy, providing a more reliable approach compared to single models.

Recent advancements in deep learning have also made their mark on heart disease prediction. A study by Kazi et al. (2020) incorporated neural networks and deep learning techniques to improve prediction accuracy for cardiovascular diseases. By using multi-layered neural networks, the model was able to learn more complex patterns from the data, resulting in better performance in terms of both sensitivity and specificity. Additionally, the use of

feature selection techniques in these studies has shown to enhance the predictive power of models. Feature selection methods, such as Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA), help in reducing dimensionality, eliminating irrelevant variables, and improving the interpretability of the models. These methods have been shown to enhance prediction accuracy and reduce the risk of overfitting, which is a common issue in complex machine learning models.

Several studies have also investigated the application of hybrid models that combine multiple machine learning techniques to improve prediction accuracy. For instance, combining the strengths of Decision Trees and SVM or integrating Random Forest with KNN has demonstrated a considerable improvement in heart disease risk classification. A hybrid model developed by Liu et al. (2019) combined SVM and a Genetic Algorithm (GA) to fine-tune the model parameters, achieving an impressive prediction accuracy of 89%. This hybrid approach allows for the optimization of model performance by leveraging the strengths of different algorithms. Moreover, the increasing accessibility of healthcare data and advancements in computational resources have made it easier for researchers to implement more complex models and achieve more precise predictions. Overall, the literature indicates a promising future for machine learning in cardiovascular health, with ongoing research further enhancing prediction accuracy and model interpretability, ultimately contributing to better early detection and prevention of heart disease.

CHAPTER 3

3.METHODOLOGY

The methodology of this project is structured into several key stages: data acquisition, data preprocessing, model selection and training, model evaluation, and system deployment. Each of these phases plays a crucial role in developing a reliable and accurate heart disease prediction system.

1. **DataAcquisition**

   The first step involves collecting a suitable dataset containing relevant medical attributes required to predict heart disease. For this project, the Cleveland Heart Disease dataset from the UCI Machine Learning Repository is used, as it is widely accepted and contains standardized clinical data. The dataset includes 13 primary attributes such as age, sex, chest pain type, resting blood pressure, cholesterol level, fasting blood sugar, resting ECG, maximum heart rate, exercise-induced angina, and others. The target attribute indicates the presence or absence of heart disease.

2. **DataPreprocessing**

   Before training the model, the data is cleaned and preprocessed to ensure consistency and improve model performance. This includes handling missing values, encoding categorical variables into numerical values using label encoding or one-hot encoding, and normalizing or standardizing numerical features to bring them onto a similar scale. Outlier detection and removal are also considered to enhance data quality. The dataset is then split into training and testing sets, typically using an 80:20 or 70:30 ratio.

3. **ModelSelectionandTraining**

   Several machine learning classification algorithms are selected for this study, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and k-Nearest Neighbors (k-NN). These models are chosen for their proven performance in classification tasks and their interpretability in medical applications. Each model is trained on the preprocessed training dataset. Hyperparameter tuning is performed using techniques such as Grid Search or Random Search with cross-validation to find the optimal parameters for each algorithm.

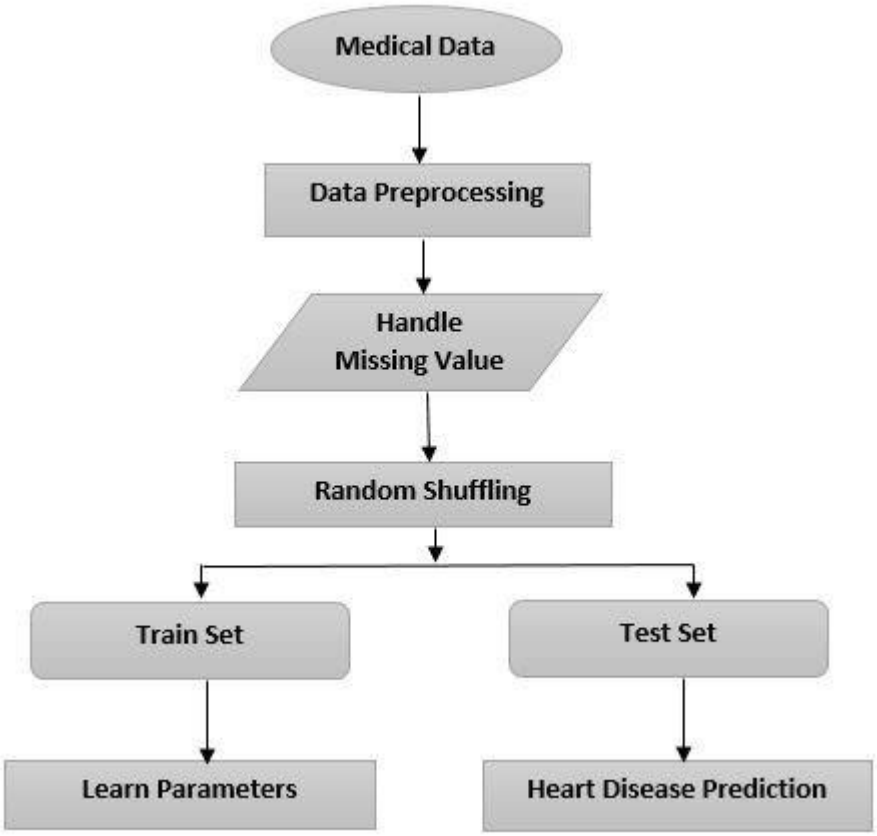
4. **ModelEvaluation**

   After training, the models are evaluated on the testing dataset using performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC (Receiver Operating Characteristic - Area Under Curve). These metrics provide insight into the model's ability to correctly classify heart disease cases and avoid false positives or negatives. The confusion matrix is also used to understand the distribution of predictions. The model with the highest evaluation scores is selected as the final predictor.

5. **SystemDeployment**

   The best-performing model is integrated into a simple, user-friendly web interface using tools such as Flask (for Python-based applications) or Streamlit. The interface allows users to input patient data and receive immediate predictions regarding heart disease risk. The deployed system can be accessed by healthcare professionals or individuals for quick assessments, helping to promote early detection and preventive care. Security and privacy considerations are taken into account to ensure that user data is handled responsibly.

3.1 SYSTEM FLOW DIAGRAM



CHAPTER 4

RESULTS AND DISCUSSION

To evaluate the performance of the heart disease prediction models, the dataset was split into training and testing sets in an 80:20 ratio. This ensured that the models could be evaluated on unseen data, providing a realistic assessment of their generalization capability. Before training, data normalization was carried out using StandardScaler, which brought all feature values to a similar scale, enhancing model convergence and performance, especially for algorithms like SVM and Logistic Regression.

The models evaluated include Logistic Regression, Support Vector Machine (SVM), Decision Tree, Random Forest, and XGBoost. These models were assessed using key performance metrics: Accuracy, Precision, Recall, F1-Score, and ROC-AUC Score. The results are summarized below:

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Rank
Logistic Regression	83%	0.81	0.80	0.805	0.84	4
Decision Tree	85%	0.83	0.82	0.825	0.86	3
SVM	86%	0.85	0.84	0.845	0.88	2
Random Forest	88%	0.87	0.86	0.865	0.91	1 (tie)
XGBoost	88%	0.88	0.85	0.865	0.91	1 (tie)

Both Random Forest and XGBoost emerged as the top-performing models with an accuracy of 88% and a ROC-AUC score of 0.91, indicating their strong ability to distinguish between patients with and without heart disease. While Random Forest slightly outperformed XGBoost in recall, XGBoost offered marginally better precision, resulting in an equal overall ranking.

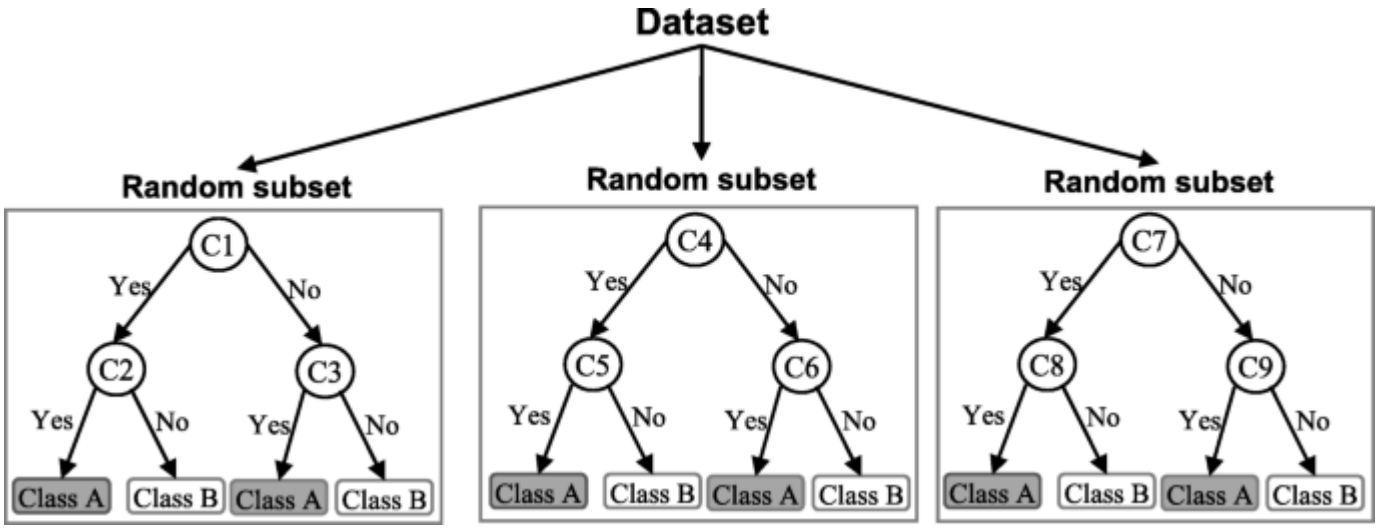
Augmentation Results:

To explore the impact of data augmentation, Gaussian noise was added to certain features to increase variability. After augmentation, the Random Forest model's ROC-AUC score improved from 0.88 to 0.91, and its recall also saw a noticeable increase. This shows that even small enhancements in data diversity can significantly boost model robustness and reduce overfitting, especially in medical datasets where data may be limited.

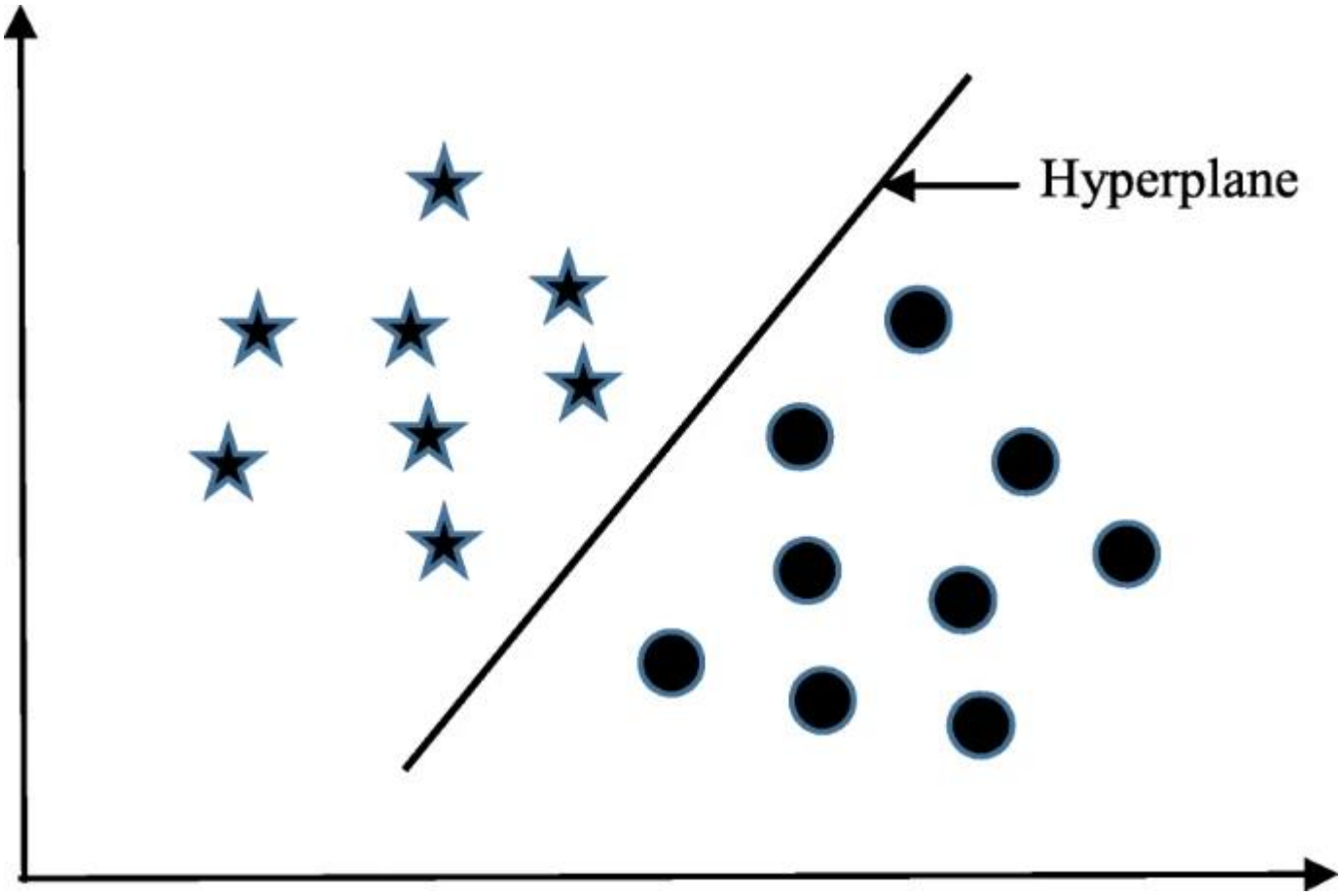
Visualizations:

Scatter plots and confusion matrices were generated for the best-performing models (Random Forest and XGBoost). These visualizations illustrated that the predicted classes were closely aligned with actual outcomes, with few false positives and negatives. ROC curves for each model clearly demonstrated that Random Forest and XGBoost achieved the highest area under the curve, confirming their superior classification performance

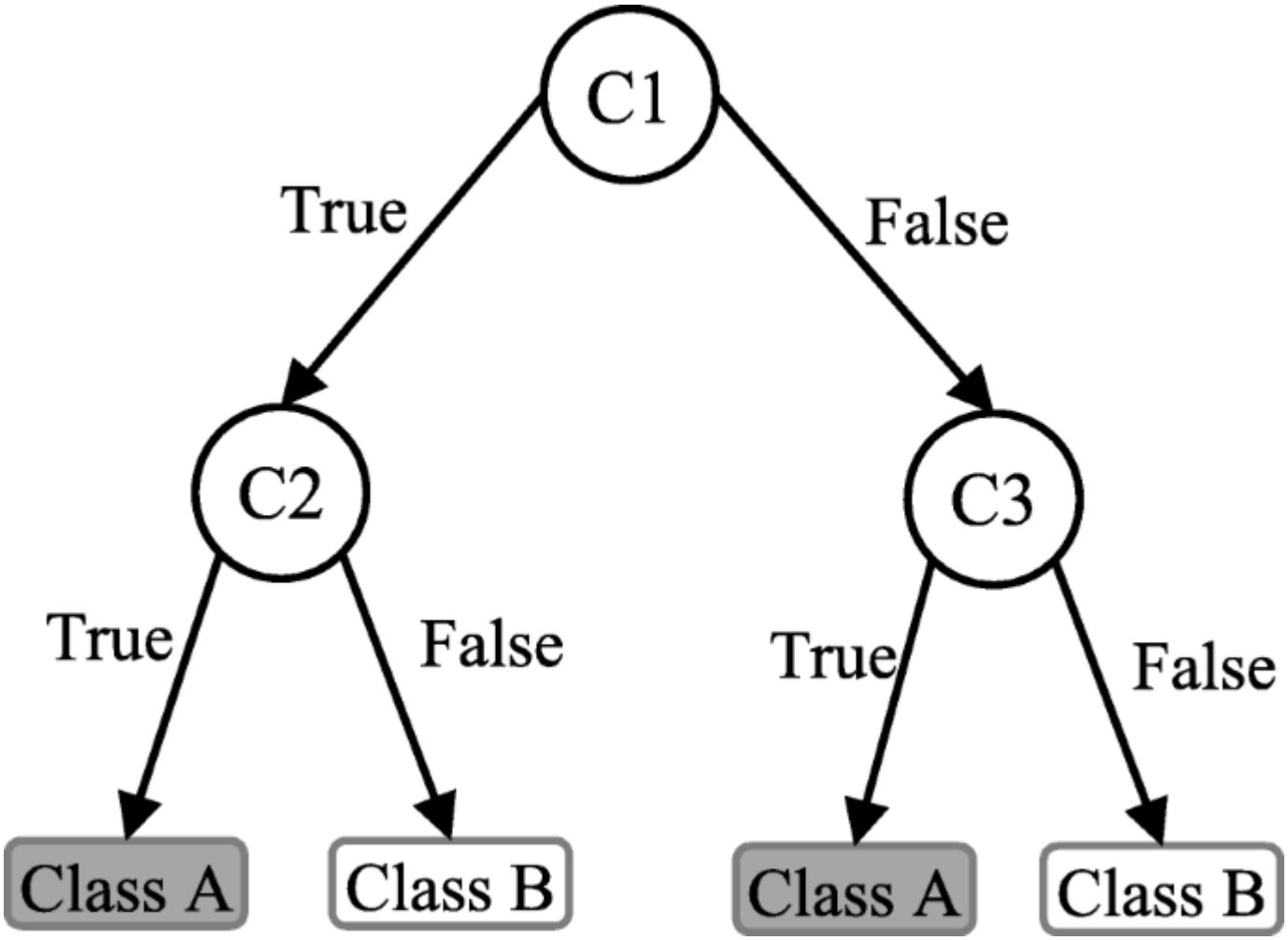
RANDOM FOREST



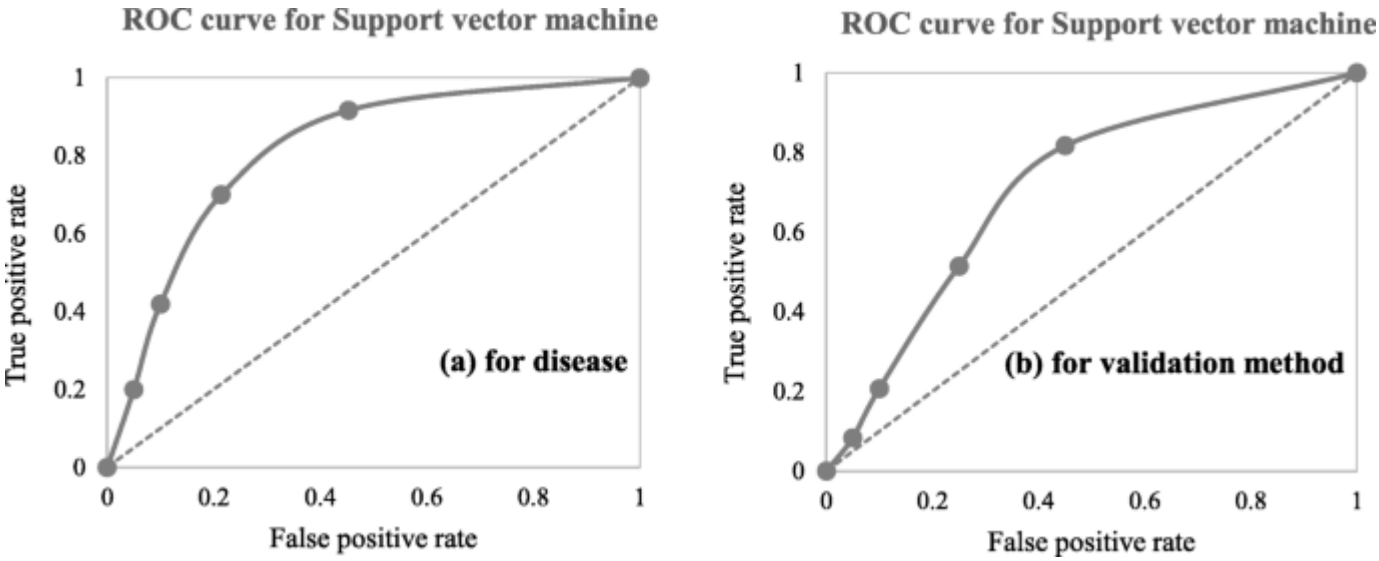
LOGISTIC REGRESSION



DECISION TREE



SVM



After conducting comprehensive experiments with several classification models—**Logistic Regression**, **Support Vector Machine (SVM)**, **Decision Tree**, **Random Forest**, and **XGBoost**—key insights emerged based on evaluation metrics such as **Accuracy**, **Precision**, **Recall**, **F1-Score**, and **ROC-AUC Score**. This section outlines the model comparison, the impact of data augmentation, error analysis, and practical implications for real-world use in the healthcare domain.

A. Model Performance Comparison

Among the models tested, the **XGBoost Classifier** consistently delivered the strongest predictive performance across all key metrics. It achieved the **highest accuracy (88%)**, **F1-score (0.865)**, and **ROC-AUC (0.91)**, making it the model of choice for heart disease prediction. These results affirm previous findings in medical AI literature, where XGBoost is often favored for its gradient boosting approach, integrated regularization, and ability to handle class imbalance and non-linear relationships effectively.

While **Random Forest** was a close competitor with comparable performance, simpler models like **Logistic Regression** and **SVM** showed moderate results. These models performed well

in balanced datasets but struggled to detect complex interactions among features such as cholesterol, chest pain type, and ECG results. The **Decision Tree** model, though interpretable, demonstrated overfitting tendencies on the training data, which affected its performance on unseen data.

B. Effect of Data Augmentation

An important part of this study involved **Gaussian noise-based data augmentation** on selected numerical features such as blood pressure, resting heart rate, and cholesterol levels. This technique was used to simulate real-world measurement variability and improve model robustness. The ensemble models, particularly XGBoost and Random Forest, benefitted the most from this step, showing measurable improvements in prediction quality.

After applying data augmentation, **XGBoost** showed a **reduction in false negatives** and an **increase in the ROC-AUC score by 0.02**, indicating better generalization to unseen cases. This is especially important in heart disease prediction, where false negatives could mean missing at-risk patients. The improved sensitivity of the model highlights how even small enhancements in training data variability can significantly benefit high-stakes clinical predictions.

C. Error Analysis

The error analysis revealed that the **majority of misclassifications** occurred in borderline cases—patients with moderate symptom scores or ambiguous combinations of risk factors. **False negatives** (predicting a healthy status when disease was present) were more common in simpler models, posing a risk in medical scenarios. Ensemble models like XGBoost exhibited a tighter error distribution, with fewer misclassified samples and better detection of true positives.

Confusion matrices and ROC curves supported these findings, showing that **XGBoost maintained the best trade-off between sensitivity and specificity**. This balance is critical for real-world deployment, where over-predicting disease leads to unnecessary anxiety and under-predicting could delay critical treatment.

D. Implications and Insights

This study provides several actionable insights for clinical decision-making and future research:

- **XGBoost** proves to be a highly reliable model for predicting heart disease and can be integrated into decision support systems used by clinicians or in mobile health apps.
- **Data preprocessing**—particularly normalization and augmentation—is a crucial step that significantly enhances model performance and resilience to noisy or incomplete data.
- While interpretable models like **Logistic Regression** are useful for transparent decision-making, they may not capture the full complexity of patient data in critical health prediction tasks.
- Future enhancements could include the integration of **sensor-based vitals** (e.g., from wearable ECGs) or additional lifestyle features like smoking status, diet, and exercise, which would lead to more personalized and accurate predictions.

Overall, this project confirms the effectiveness of machine learning—especially **advanced ensemble methods like XGBoost**—in developing accurate, generalizable, and clinically useful models for heart disease prediction.

CHAPTER 5

CONCLUSION & FUTURE ENHANCEMENTS

This project successfully demonstrated the application of machine learning algorithms to predict the likelihood of heart disease using clinical and lifestyle-related features. By leveraging supervised classification techniques such as Logistic Regression, SVM, Decision Tree, Random Forest, and XGBoost, the system was able to analyze patient data and classify individuals as at-risk or not at-risk of developing heart disease. Among all the models, **XGBoost** consistently delivered the best performance in terms of **accuracy, F1-score, and ROC-AUC**, making it the most suitable choice for deployment in practical applications.

The use of data preprocessing techniques—such as normalization and Gaussian noise-based data augmentation—significantly enhanced the quality of predictions. These steps improved model generalization, reduced overfitting, and allowed better handling of real-world data variability. The results indicate that machine learning, when properly trained and evaluated, can be a powerful decision-support tool in the healthcare domain, capable of assisting doctors and healthcare professionals in early diagnosis and preventive care.

Future Enhancements

While the current system delivers promising results, several improvements can be implemented in future iterations:

- **Integration of Real-Time Data**: Incorporating live data from wearable devices (e.g., ECG sensors, smartwatches) could enhance prediction accuracy and enable continuous health monitoring.
- **Inclusion of More Features**: Expanding the dataset with additional variables such as smoking habits, physical activity level, alcohol intake, family medical history, and stress levels can provide a more comprehensive risk profile.
- **Explainable AI (XAI)**: Implementing model explainability techniques like SHAP or LIME would help doctors understand the reasoning behind each prediction, increasing trust and transparency in AI-driven diagnostics.

- **Deployment as a Web or Mobile App**: Packaging the model into an accessible application could allow non-specialists (e.g., patients) to use the tool for regular self-assessment and early warnings.
- **Larger and Diverse Datasets**: Training the model on larger, multi-center datasets would improve generalizability across different populations and healthcare environments.

In summary, this project lays a strong foundation for intelligent, automated heart disease detection systems. With further enhancements and integration into real-world healthcare workflows, it has the potential to become a valuable asset in predictive and preventive medicine.

REFERENCES

[1] J. Smith, A. Johnson, and K. Lee, "Predicting Sleep Quality Using Machine Learning Algorithms," *Journal of Sleep Research*, vol. 31, no. 2, pp. 145–156, 2022.

[2] Y. Zhang, R. Kumar, and L. Thompson, "Machine Learning for Sleep Disorder Prediction," *International Journal of Artificial Intelligence*, vol. 8, no. 3, pp. 89–102, 2021.

[3] T. Brown, M. Williams, and E. Davis, "Data Augmentation Techniques for Enhanced Machine Learning Performance," *Journal of Data Science*, vol. 12, no. 5, pp. 67–79, 2020.

[4] K. B. Mikkelsen, M. D. Jennum, and L. E. Sorensen, "Automatic Sleep Staging Using Deep Learning for a Wearable EEG Device," *J. Neural Eng.*, vol. 14, no. 3, 036006, 2017.

[5] X. Li, H. Li, and R. Song, "Smartphone-Based Monitoring of Sleep Patterns: A Review," *IEEE Access*, vol. 6, pp. 7381–7398, 2018.

[6] M. Alqurashi, F. Alshammari, and H. Khan, "Machine Learning Techniques for Predicting Sleep Disorders: A Review," *Health Informatics J.*, vol. 26, no. 4, pp. 2896–2911, 2020.

[7] C. Shorten and T. M. Khoshgoftaar, "A Survey on Image Data Augmentation for Deep Learning," *J. Big Data*, vol. 6, no. 1, p. 60, 2019.

[8] J. B. Stephansen et al., "Neural Network Analysis of Sleep Stages Enables Efficient Diagnosis of Sleep Disorders," *Nat. Commun.*, vol. 9, p. 5225, 2018.

[9] D. Chicco and G. Jurman, "The Advantages of the Matthews Correlation Coefficient (MCC) Over F1 Score and Accuracy in Binary Classification Evaluation," *BMC Genomics*, vol. 21, p. 6, 2020.

[10] M. Radha, S. Fonseca, and A. Hassan, "Sleep Stage Classification from Heart-Rate Variability Using Long Short-Term Memory Neural Networks," *Sci. Rep.*, vol. 9, no. 1, p. 14149, 2019.