***Heart Disease Prediction***

# **The domain of the Project: Health Care**

# Data Science

# **Team Mentors (and their designation):**

Mr Bhargavesh Dakka Sir

Data Scientist at MuSigma

# **Team Members:**

Ms. Sapna Chauhan

# **Period of the project**

# **May 2025 to December 2025**

***Declaration***

The project titled “Heart Disease Prediction” has been mentored by Mr Bhargavesh Dakka, organised by SURE Trust, from May 2025 to December 2025, for the benefit of the educated unemployed rural youth for gaining hands-on experience in working on industry relevant projects that would take them closer to the prospective employer. I declare that to the best of my knowledge the members of the team mentioned below, have worked on it successfully and enhanced their practical knowledge in the domain.

Team Members:

Ms. Sapna Chauhan Signature

Mentor’s Name

Mr Bhargavesh Dakka

Prof. Radhakumari

Executive Director & Founder

SURE Trust

***Tabel of contents***

1. Executive summary
2. Introduction
3. Project Objectives
4. Methodology & Results
5. Social / Industry relevance of the project
6. Learning & Reflection
7. Future Scope & Conclusion

***Executive Summary***

Heart disease is one of the leading causes of death globally, making early detection and prevention extremely important. This project focuses on building a machine learning–based system to predict the risk of heart disease using patient demographic, clinical, lifestyle, and medical history data. The objective of the project is to assist healthcare professionals by providing early risk assessment and supporting preventive decision-making.

The dataset used in this project includes key features such as age, blood pressure, cholesterol levels, blood sugar, lifestyle habits (smoking, alcohol consumption, physical activity), stress levels, and previous medical history. Data preprocessing techniques were applied to ensure data quality, including handling missing values, encoding categorical variables using one-hot encoding, and scaling numerical features. Feature engineering was performed using medical domain knowledge to create meaningful indicators such as cholesterol ratios, blood pressure metrics, lifestyle risk scores, and metabolic risk flags.

The results indicate that machine learning techniques can effectively predict heart disease risk and provide valuable insights for early intervention. While the model is not intended to replace medical diagnosis, it serves as a supportive tool for clinicians. This project highlights the potential of data-driven approaches in preventive healthcare and lays the foundation for future enhancements using real-time data and advanced predictive models.

***Introduction***

Heart disease is one of the most serious global health challenges and remains a leading cause of death across all age groups. Factors such as unhealthy lifestyle habits, increasing stress levels, poor dietary choices, and lack of physical activity have significantly contributed to the rising number of heart-related illnesses. Early detection of heart disease is critical, as it can help prevent severe complications and reduce mortality through timely medical intervention and lifestyle modification.

With the rapid growth of digital healthcare records, large volumes of patient data are now available for analysis. Machine learning provides powerful techniques to analyze this data and uncover hidden patterns that may not be easily identified through traditional methods. By leveraging demographic information, clinical measurements, lifestyle indicators, and medical history, predictive models can estimate an individual’s risk of developing heart disease.

This project focuses on building a heart disease prediction system using machine learning techniques. The goal is to analyze multiple risk factors such as age, blood pressure, cholesterol levels, blood sugar, smoking habits, stress levels, and previous medical conditions to predict the likelihood of a heart attack. The project demonstrates how data-driven approaches can support preventive healthcare and assist medical professionals in making informed decisions, while emphasizing that such systems are meant to complement, not replace, clinical expertise.

***Project Objectives***

The primary objective of this project is to develop a machine learning–based model capable of predicting the risk of heart disease using patient data. The project aims to analyze various demographic, clinical, and lifestyle-related features to understand their impact on heart health and identify the most influential risk factors.

Another key objective is to perform effective data preprocessing, including handling missing values, encoding categorical variables, and scaling numerical features, to ensure high-quality input for machine learning models. Feature engineering is also an important goal of this project, as it helps create new meaningful features such as cholesterol ratios, blood pressure indicators, and lifestyle risk scores that improve prediction accuracy.

Additionally, the project aims to compare the performance of different machine learning algorithms, such as Logistic Regression and ensemble models like Random Forest, to determine the most suitable model for healthcare prediction tasks. Emphasis is placed on evaluating models using appropriate metrics such as recall, precision, F1-score, and ROC-AUC, rather than relying solely on accuracy.

Finally, the project seeks to demonstrate how predictive analytics can be applied in the healthcare domain to support early diagnosis and preventive care. The overall objective is to build a reliable, interpretable, and practical prediction system that can assist healthcare professionals in identifying high-risk individuals at an early stage.

***Methodology and Results***

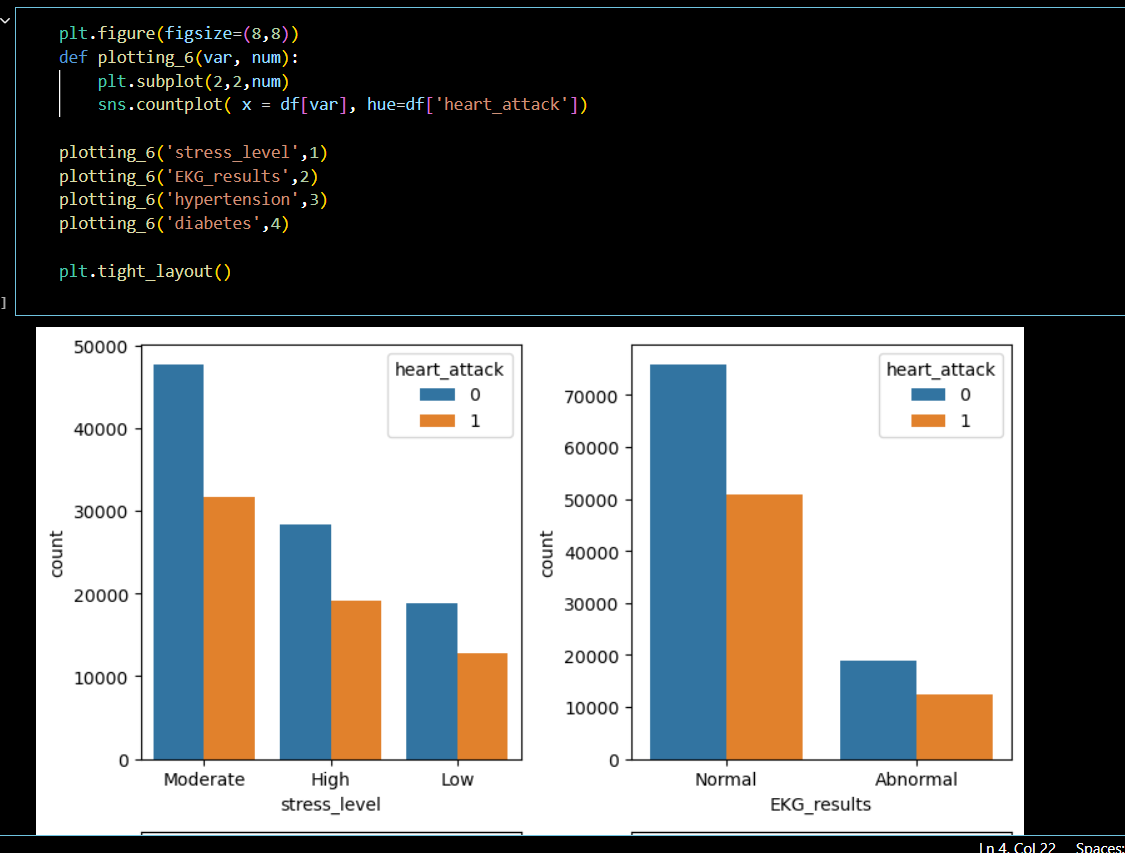
The methodology of this project follows a structured machine learning workflow. Initially, the healthcare dataset was explored to understand the distribution of features and identify potential data quality issues. Data preprocessing was performed by handling missing values, correcting inconsistent entries, and removing irrelevant noise from the dataset. Categorical variables such as gender, smoking status, and dietary habits were converted into numerical format using one-hot encoding, while numerical features were scaled to maintain uniformity.

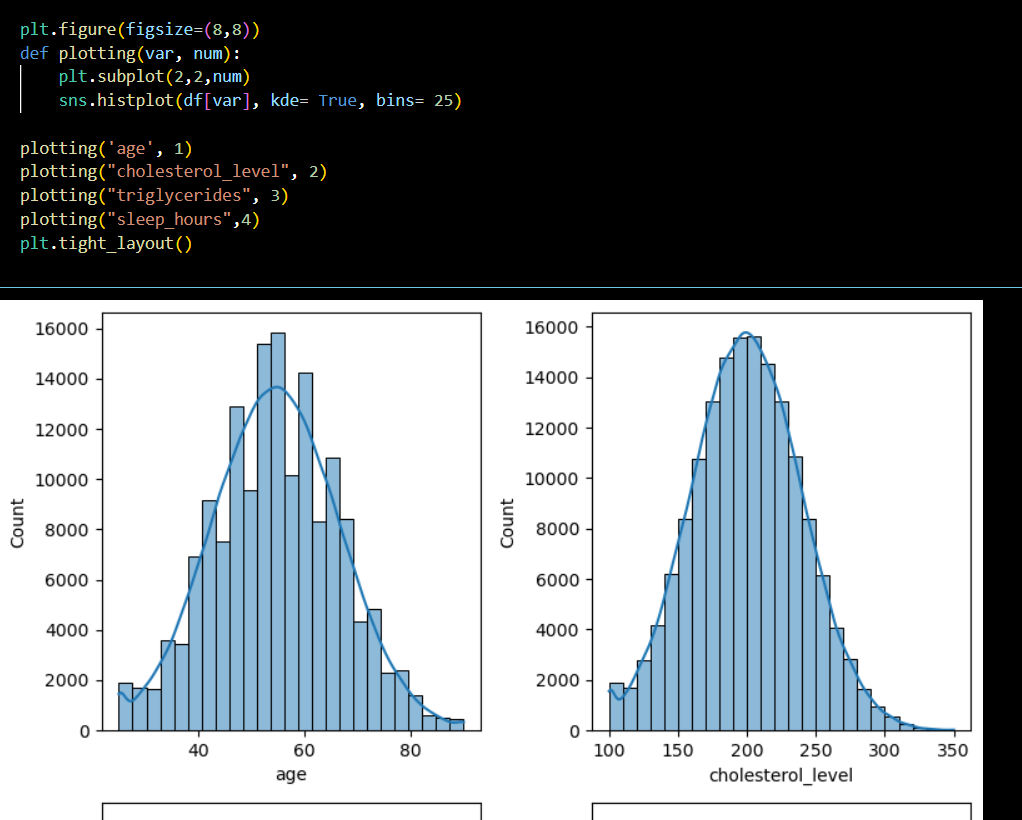
Feature engineering played a crucial role in improving model performance. New features such as cholesterol ratios, pulse pressure, metabolic risk indicators, and lifestyle risk scores were created using medical domain knowledge. These engineered features helped the models better capture underlying health patterns.

Multiple machine learning models were trained and tested, including Logistic Regression and Random Forest. The dataset was split into training and testing sets to ensure unbiased evaluation. Model performance was assessed using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. The results showed that ensemble models like Random Forest performed better due to their ability to handle non-linear relationships and feature interactions.

Overall, the final model achieved satisfactory performance, demonstrating that machine learning techniques can effectively predict heart disease risk when combined with proper preprocessing and feature engineering.







***Learning and Reflection***

This project provided valuable hands-on experience in applying machine learning techniques to a real-world healthcare problem. One of the key learnings was the importance of data preprocessing, as model performance heavily depends on the quality of the input data. Handling missing values, encoding categorical features correctly, and scaling numerical variables were essential steps in building an effective model.

Another important learning was the role of feature engineering. Creating new features based on medical knowledge significantly improved the predictive capability of the model. This highlighted that domain understanding is just as important as technical skills in data science projects. The project also emphasized that accuracy alone is not a sufficient evaluation metric in healthcare applications, where false negatives can have serious consequences. Metrics such as recall and ROC-AUC are more meaningful in such scenarios.

Working with different machine learning models helped me understand the strengths and limitations of various algorithms. Ensemble methods proved to be more robust compared to simpler models. This project also improved my problem-solving skills, debugging ability, and understanding of end-to-end machine learning pipelines.

Overall, this project strengthened my confidence in applying data science techniques to sensitive domains like healthcare and reinforced the importance of ethical and responsible use of predictive models.

***Conclusion and Future Scope***

In conclusion, this project successfully demonstrates the application of machine learning techniques for predicting heart disease risk using healthcare data. By analyzing demographic, clinical, lifestyle, and medical history features, the model was able to identify individuals who may be at higher risk of experiencing a heart attack. The project highlights the importance of early risk prediction in reducing the burden of heart disease and improving patient outcomes.

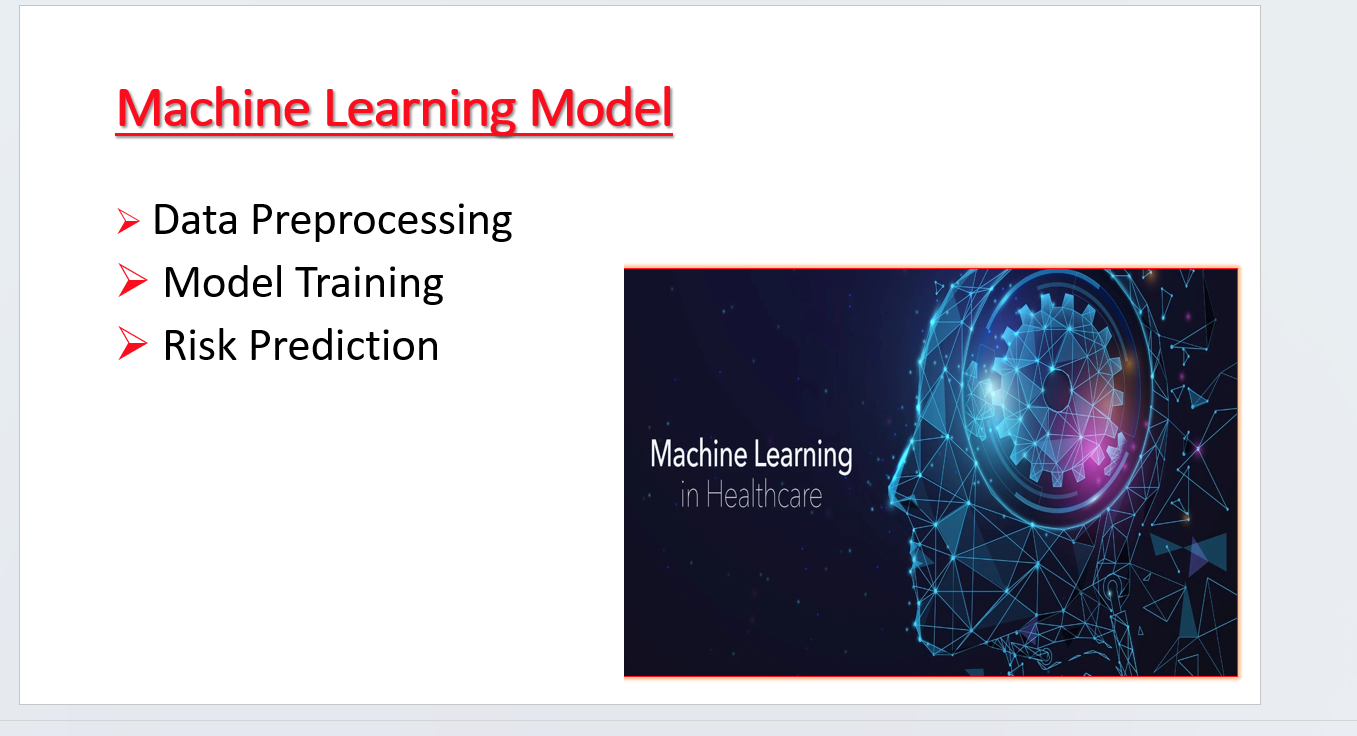
Through effective data preprocessing, feature engineering, and model selection, the performance of the prediction system was significantly enhanced. The use of ensemble models and appropriate evaluation metrics ensured reliable and meaningful results. Rather than relying solely on accuracy, the project focused on recall and overall model robustness, which are critical in medical decision-support systems.

The findings of this project indicate that machine learning can serve as a powerful tool in preventive healthcare by providing early warnings and supporting clinical decision-making. However, it is important to note that such systems are not meant to replace doctors but to assist them by offering data-driven insights.

Overall, this project demonstrates how data science can contribute positively to the healthcare domain and showcases the potential of predictive analytics in addressing real-world medical challenges.

Although the current model shows promising results, there is significant scope for future improvement and expansion. One major area of enhancement is the use of larger and more diverse datasets, including real-time patient data collected from wearable devices and electronic health records. This would help improve model generalization and prediction accuracy.

Another important future direction is model interpretability. Integrating explainable AI techniques can help doctors understand why a particular prediction was made, increasing trust and usability in clinical settings. Additionally, the model can be deployed as a web or mobile application to provide easy access for healthcare providers and patients.



Clinical validation and collaboration with medical professionals will be essential to ensure the reliability and ethical use of the system. With these enhancements, the project can evolve into a comprehensive decision-support tool that contributes meaningfully to preventive healthcare and early disease management.