Madanapalle Institute of Technology & Science Post Box No: 14, KadiriRoadAngallu (V), Madanapalle-517325,Annamayya District, Andhra Pradesh,India

**PROJECT REPORT**

**ON**

**“ Cardiac Risk Prediction”**

Under the guidance of

* Mr. AKASH V

**CERTIFICATE OF INTERNSHIP**

This is to certify that **Goddumarrili Likitha**, bearing **Registration Number: 23691A0592**, a student of **Madanapalle Institute of Technology & Science**, Kadiri Road, Angallu (V), Madanapalle - 517325, Annamayya District, Andhra Pradesh, India, has successfully completed an internship course from **16/06/2025 to 28/07/2025** at our organization.

During her internship, **Goddumarrili Likitha** worked on a **Machine Learning Project** and gained experience in the following areas:

**STUDENT DECLARATION**

I, **Goddumarrili Likitha**, Register Number: **23691A0592**, hereby declare that this report entitled **“Performance Evaluation of ML Models in Cardiac Risk Prediction”** was completed during the internship period from **16/06/2025 to 28/07/2025** at **Prinston Smart Engineers**, under the supervision and guidance of **Mr. AKASH V**.

**Table of Contents:**

|  |  |  |
| --- | --- | --- |
| CHAPTER NO. | PARTICULARS |  |
| 1 | Executive Summary |  |
| 2 | Introduction |
| 3 | Description of the Organization |
| 4 | Experiential Learning |
| 5 | Internship Outcomes and Conclusion |
| 6 | Bibliography |

### **Brief Overview of the Internship Report**

The **"Cardiac risk Prediction"** project is a machine learning-based application developed with the goal of predicting the likelihood of a person developing heart disease based on key medical and personal features. With heart disease being a leading cause of death globally, early prediction and diagnosis can significantly improve treatment outcomes and preventive healthcare strategies.

The project follows a standard machine learning pipeline—beginning with **data collection, preprocessing, and exploratory data analysis (EDA)**. The dataset includes attributes such as age, gender, blood pressure, cholesterol levels, smoking habits, glucose levels, BMI, and other relevant health indicators. Missing values in the dataset are handled using **KNN imputation**, and class imbalance is addressed using the **SMOTE (Synthetic Minority Over-sampling Technique)** technique to improve model fairness and accuracy.

Feature selection is carried out using **SelectKBest**, helping to focus the training on the most relevant features. Multiple classification algorithms are then implemented and compared, including:

**Logistic Regression**: A fundamental and interpretable binary classification algorithm that models the probability of a target class (presence or absence of heart disease).

**Random Forest Classifier**: A robust ensemble method that builds multiple decision trees and averages the results to improve accuracy and reduce overfitting.

**XGBoost Classifier**: A powerful gradient boosting framework optimized for speed and performance, capable of capturing complex patterns in the data.

Each model is trained on the processed data and evaluated on a holdout test set using metrics like **accuracy, precision, recall, F1-score, and confusion matrix**. Among the models tested, **XGBoost and Random Forest** showed superior performance due to their ability to handle complex interactions between features and reduce overfitting.

Visualization tools like **Seaborn and Matplotlib** were used during EDA to better understand data distributions and class balance. The final model not only achieved a high accuracy score but also demonstrated reliable generalization to unseen data.

### **Outcomes**

This internship project effectively applied machine learning techniques to a real-world healthcare problem. The final model achieved over **85% accuracy** in detecting the risk of heart disease. Insights gained include the importance of preprocessing technique such as **KNN imputation** , and the benefits of ensemble methods like **XGBoost** for improved prediction accuracy.

### **Conclusion**

In conclusion, this internship project highlights the successful application of classification algorithms in predicting heart disease risk using clinical data. It emphasizes the critical role of **data preprocessing, feature selection, and model evaluation** in developing accurate predictive systems. The project provides a valuable foundation for integrating machine learning into medical diagnostics and could be further enhanced through **deep learning**, **time-series analysis**, or integration with **real-time health monitoring systems** in the future.

**Introduction**

This project focuses on predicting heart disease using the Framingham Heart Study dataset. It uses important health features like age, blood pressure, cholesterol, diabetes, and smoking habits. Machine learning algorithms such as ***Logistic Regression, Random Forest, and XGBoost*** are used to build models that predict the risk of heart disease. The project follows steps like data cleaning, analysis, model training, and performance evaluation. The aim is to support early detection and better decision-making in healthcare through data science.

**Objectives of the Project**

The main objective of this project is to develop a reliable machine learning model that can accurately predict whether an individual is at risk of developing heart disease. By analyzing the Framingham Heart Study dataset, the project aims to identify key health indicators such as age, cholesterol level, blood pressure, diabetes status, and smoking habits that contribute to heart disease. The goal is to assist healthcare professionals in early diagnosis, promote preventive care, and improve patient outcomes using data-driven decision-making.

**Scope**

This project is intended to support:

**Early diagnosis** of heart disease.

**Assist doctors** in making data-driven decisions.

**Raise awareness** about heart health risk factors.

**Provide individuals** with predictive insights to take preventive measures.

The project can be further extended to work with real-time hospital data, develop web-based tools for healthcare professionals, and integrate with wearable devices for continuous health monitoring.

**Data Collection and Preparation**

#### **Data Collection:**

The dataset used in this project is derived from the **Framingham Heart Study**, which is a long-term, ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. It contains medical history and health-related information of individuals collected over several years. The dataset is publicly available on platforms like [Kaggle](https://www.kaggle.com/" \t "_new), making it accessible for research and educational purposes.

**Source**: Framingham Heart Study Dataset

**Format**: CSV file

**Total Records**: Approximately 4,240 instances.

**Features**: 15 input features + 1 target variable (TenYearCHD)

**Target**: Predict the 10-year risk of coronary heart disease (CHD)

#### **🛠️ Data Preparation:**

To ensure accurate predictions, the raw dataset was cleaned and prepared through the following steps:

**Handling Missing Values**:

Identified columns with missing data (e.g., BPMeds, glucose, totChol)

Imputed missing values using mean or median strategies to maintain data integrity.

**Encoding Categorical Data**:

Converted categorical variables like gender into numeric format if needed for the model.

**Feature Selection**:

Selected relevant features that impact heart disease prediction such as age, cigsPerDay, sysBP, BMI, glucose, etc.

**Feature Scaling**:

Applied normalization or standardization (e.g., using StandardScaler) to bring features to a similar range, improving model performance.

**Splitting the Data**:

Divided the dataset into **training (80%)** and **testing (20%)** sets for model evaluation.

This data preparation step was crucial to ensure the dataset was clean, well-structured, and ready for machine learning model training.

**Description of the Organization**



**PRINSTON SMART ENGINEERS** is a leading MEP (Mechanical, Electrical, and Plumbing) design consultancy specializing in designing, installing, and servicing electro-mechanical systems and utilities. Established in **2016**, the company is known for its full-cycle project involvement—covering engineering, documentation, commissioning, and as-built drawings—strictly following project timelines and consultant specifications.

In addition to core engineering services, the company also focuses on **skill development for engineering students** by offering live, instructor-led online training. They have impacted over **350 engineering colleges** across India, addressing the employability gap among graduates.

Following the **AICTE 2018 internship mandate**, PRINSTON began providing expert-led internships to help students gain industrial exposure. In **2020**, they partnered with **Wedir-tech Trading Contracting & Services W.L.L** in **Doha, Qatar**, further expanding their global presence in electro-mechanical services.

Their **mission** is to offer innovative engineering and educational solutions, while their **vision** is to become a global leader in MEP consultancy and technical training, committed to quality, safety, and building a skilled future workforce.

**About the Organization:**

**PRINSTON SMART ENGINEERS** is a reputed **MEP (Mechanical, Electrical, and Plumbing) Design Consultancy** established in **2016**, offering specialized services in the design, installation, servicing, and upgrading of electro-mechanical systems and utilities. The organization handles all project phases—from engineering documentation to commissioning and as-built drawings—while strictly adhering to project schedules and consultant specifications.

To bridge the skill gap among engineering graduates in India, PRINSTON also provides **industry-oriented training** and **internship programs** through live, instructor-led sessions. They have collaborated with over **350 engineering colleges** across India, making quality education and hands-on training accessible and affordable.

Following the **AICTE mandate in 2018** for mandatory internships, PRINSTON began offering internships to help students gain industrial exposure and improve their technical and managerial skills.

In **2020**, they formed a strategic alliance with **Wedir-tech Trading Contracting & Services W.L.L, Qatar**, expanding their operations and expertise in the Middle East's electro-mechanical market.

Their **mission** is to deliver top-tier engineering solutions and educational services while maintaining excellence in quality, safety, and sustainability. Their **vision** is to emerge as a **global leader in MEP design and training**, empowering the next generation of skilled engineers

### **Mission and Vision**

PRINSTON SMART ENGINEERS aims to deliver **innovative engineering solutions** and **educational services** that promote growth.

Their mission is to **exceed client expectations** through **quality**, **safety**, and **sustainability**.

Vision: To be a **global leader** in MEP design and training, building a **skilled workforce** for the future.

### **1. Skill Development and Training Programs**

Offers **live, instructor-led online training sessions** focused on practical engineering skills.

Training modules cover **fundamental to advanced technical topics**, ensuring **industry readiness**.

Programs are designed to provide **hands-on experience** and **real-world exposure** to students.

### **2. Strategic Partnerships and Collaborations**

Formed a key partnership with **Wedir-tech Trading Contracting & Services W.L.L** in **Doha, Qatar**.

This collaboration enhances **service delivery**, **technical expertise**, and access to **international markets**.

It demonstrates PRINSTON’s commitment to meeting **global standards** in MEP consultancy.

### **3. Future Plans and Growth**

Plans to **expand globally** and increase their presence in **international markets**.

Committed to the **continuous development** of their training programs to align with **industry trends**.

Focused on delivering **sustainable**, **cost-effective**, and **client-centric** engineering solutions.

### **Conclusion**

PRINSTON SMART ENGINEERS stands for **excellence, innovation, and sustainability**.

Their strategic initiatives and focus on **skill development** ensure they remain a **trusted engineering partner**.

With a foundation rooted in **expertise** and **continuous improvement**, they are well-positioned for future growth.

**Experiential Learning**

* Gained hands-on experience in handling real-world datasets and performing data preprocessing using Python.
* Learned how to implement and evaluate machine learning algorithms such as Simple Linear Regression (SLR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Decision Tree.
* Understood the importance of feature selection and data transformation techniques for improving model performance.
* Explored various evaluation metrics like R² score, Mean Squared Error (MSE), and classification accuracy.
* Developed the ability to compare models based on both predictive accuracy and interpretability.
* Strengthened proficiency in libraries such as pandas, scikit-learn, matplotlib, and seaborn.
* Improved problem-solving, analytical thinking, and coding skills through iterative experimentation and validation.
* Gained insight into how machine learning can be applied to environmental and automotive challenges in real-world scenarios.

### **Internship Outcomes and Conclusion**

#### **Introduction**

The internship at **Prinston Smart Engineers** was designed to offer practical exposure in the field of **healthcare analytics**, with a specific focus on **predicting heart disease** using **machine learning algorithms** and the **Framingham Heart Study dataset**. The aim was to build predictive models capable of identifying individuals at risk of developing heart disease based on health indicators. This chapter highlights the outcomes, findings, and key takeaways from the internship experience.

#### **Outcomes of the Internship**

Successfully conducted **data preprocessing** on the Framingham dataset, handling missing values and scaling features.

Implemented various **classification algorithms** including **Logistic Regression**, **KNN**, **SVM**, and **Decision Tree**.

Identified key features contributing to heart disease prediction such as **cholesterol level, blood pressure, glucose**, and **smoking status**.

Evaluated models using metrics such as **accuracy, precision, recall**, and **F1-score** to assess performance.

Found that models like **Logistic Regression** and **SVM** provided better generalization and interpretability for medical datasets.

Understood the impact of **data quality**, **feature selection**, and **balancing class distribution** on predictive accuracy.

Gained experience in **using Python libraries** such as **pandas**, **seaborn**, **matplotlib**, and **scikit-learn**.

#### **Conclusion**

This internship project reinforced the significance of **machine learning in healthcare**, especially in early diagnosis and prevention of chronic illnesses like **heart disease**. The application of predictive models on the Framingham dataset demonstrated the ability of ML algorithms to support clinical decision-making.  
Through this experience, essential technical and analytical skills were strengthened, and a deeper understanding of **health data science** was gained. Future work could focus on enhancing model accuracy with larger and more diverse datasets, incorporating **deep learning**, or developing **real-time prediction tools** for use in clinical environments.

### **Bibliography (Tailored for Heart Disease Prediction Project)**

#### 🔹 **Machine Learning Foundations**

These foundational books offer core ML concepts and statistical approaches crucial for building predictive healthcare models.

**Bishop, C. M. (2006).** Pattern Recognition and Machine Learning. Springer.  
➤ A fundamental guide to probabilistic models used in disease prediction.

**Hastie, T., Tibshirani, R., & Friedman, J. (2009).** The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.  
➤ Explains key supervised learning techniques like logistic regression and decision trees used in health datasets.

#### 🔹 **Data Preprocessing and Feature Selection**

These works provide methods to clean and select the most relevant features from clinical datasets such as age, cholesterol, blood pressure, etc.

**Guyon, I., & Elisseeff, A. (2003).** An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, 3, 1157–1182.  
➤ Useful for identifying impactful features like cholesterol level or heart rate.

**Liu, H., & Motoda, H. (1998).** Feature Selection for Knowledge Discovery and Data Mining. Springer.  
➤ Helps in understanding methods to reduce irrelevant or noisy medical features.

#### 🔹 **Machine Learning Algorithms**

These references focus on algorithm selection and implementation, helping choose the best models (like Random Forest, KNN, SVM) for heart disease classification.

**James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013).** An Introduction to Statistical Learning: with Applications in R. Springer.  
➤ Includes practical examples of ML algorithms used in health diagnostics.

**Müller, A. C., & Guido, S. (2016).** Introduction to Machine Learning with Python: A Guide for Data Scientists. O’Reilly Media.  
➤ A beginner-friendly approach using Python libraries like scikit-learn for disease prediction tasks.

#### 🔹 **Model Evaluation and Validation**

These sources guide how to assess the accuracy and reliability of your prediction model using metrics like confusion matrix, ROC-AUC, and cross-validation.

**Raschka, S., & Mirjalili, V. (2019).** Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2. Packt Publishing.  
➤ Helps in evaluating model performance using Python-based tools.

**Kohavi, R., & Provost, F. (1998).** Glossary of Terms. Machine Learning, 30(2–3), 271–274.  
➤ Defines evaluation metrics and validation terms important for clinical prediction tasks.

#### 🔹 **Interpretability and Explainability**

These works help in interpreting the output of complex models (like Random Forests) so that healthcare professionals can trust the system’s predictions.

**Ribeiro, M. T., Singh, S., & Guestrin, C. (2016).** "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In KDD '16 Proceedings (pp. 1135–1144).  
➤ Presents techniques like LIME to make model outputs understandable.

**Lipton, Z. C. (2016).** The Mythos of Model Interpretability. arXiv preprint arXiv:1606.03490.  
➤ Discusses the need for transparency in clinical AI systems.

#### 🔹 **Ethical Considerations in ML Projects**

These references emphasize the importance of fairness, privacy, and bias control in healthcare AI applications.

**Jobin, A., Ienca, M., & Vayena, E. (2019).** The Global Landscape of AI Ethics Guidelines. Nature Machine Intelligence, 1(9), 389–399.  
➤ Outlines global AI ethics for responsible use in healthcare systems.

**Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016).** The Ethics of Algorithms: Mapping the Debate. Big Data & Society, 3(2), 2053951716679679.  
➤ Addresses concerns of algorithmic bias and discrimination in medical prediction systems.