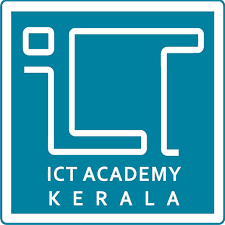
**ICT ACADEMY OF KERALA**

**Summer Internship Report**

**Machine Learning & Artificial Intelligence**



**Team Members: (Group 4) Institution name:**

**Abhishek S Younus College of Engineering**

**Roopak Susap Thomas and Technology (YCET)**

**Aleena Shaji Rekha**

**Fathima Saleem**

**Azeya Y**

**31.04.2023**

# **EXECUTIVE SUMMARY**

# During our internship, we undertook a project focused on cardiovascular disease prediction using a dataset which was provided in our Paatshala accounts. The purpose of this project was to develop a predictive model that could accurately identify individuals at risk of developing cardiovascular disease based on various input factors.

# The model aimed to classify individuals into two categories: those who are at a high risk of developing cardiovascular disease and those who are at a low risk. By identifying individuals at high risk, appropriate interventions and preventive measures could be implemented to mitigate the potential risks and improve overall health outcomes.

# To achieve this objective, we performed several key activities throughout the project. Firstly, we thoroughly explored and preprocessed the dataset to ensure data quality and consistency. This involved handling missing values, removing outliers, and normalizing the features to create a reliable and homogeneous dataset. After all the preprocessing the data was reduced from 70,000 to 62,520.

# Next, we conducted a comprehensive analysis of the dataset, employing various machine learning techniques such as logistic regression, decision trees, random forests, KNN model, and support vector machines. These algorithms were utilized to train and evaluate multiple models, comparing their performance metrics such as accuracy, precision.

# After thorough experimentation and model tuning, we selected the most effective algorithm (XG Boosting algorithm) that yielded the highest predictive accuracy. Once the final model was determined, we utilized it to predict cardiovascular disease for new, unseen data points which was collected from nearest hospital.

# The outcomes of this project were twofold. Firstly, we developed a partially successful predictive model with 73.68% accuracy capable of identifying individuals at risk of cardiovascular disease.

# Secondly, through this internship project, we gained valuable hands-on experience in data preprocessing, exploratory analysis, and machine learning model development. We enhanced my skills in programming, data manipulation, and statistical analysis, while also learning about the challenges and considerations involved in developing a predictive model for cardiovascular disease.

# Overall, this project provided me with an opportunity to contribute to the field of healthcare by leveraging data science techniques to predict and potentially prevent cardiovascular disease, while also fostering my personal and professional growth in the field of machine learning and data analysis.

# **INTRODUCTION**

# The internship project aimed to develop a predictive model for identifying individuals at risk of cardiovascular disease. Cardiovascular disease, including conditions such as heart attacks and strokes, is a leading cause of mortality worldwide. Early identification of individuals at high risk can enable targeted interventions and preventive measures, potentially reducing the burden of cardiovascular disease and improving health outcomes.

# **Background:**

# Cardiovascular disease has long been a major public health concern globally. Numerous studies have explored the risk factors associated with this disease, including age, gender, blood pressure, cholesterol levels, smoking habits, and family history. However, the growing availability of large-scale datasets and advancements in machine learning techniques have opened up new possibilities for developing more accurate and efficient prediction models.

# **Existing Studies:**

# Several studies have utilized machine learning algorithms to predict cardiovascular disease using various datasets. These studies have explored the predictive capabilities of algorithms such as logistic regression, decision trees, random forests, support vector machines, and neural networks. They have also investigated the significance of different risk factors and their contribution to disease prediction. While these studies have achieved promising results, there is still a need for further research and refinement of predictive models to improve accuracy and applicability.

# **Specific Context of the Project:**

# The specific context of the project encompassed data preprocessing, exploratory analysis, feature selection, model training, and evaluation. The project focused on employing various machine learning algorithms to identify the most effective approach for cardiovascular disease prediction. By fine-tuning the model and optimizing its performance metrics, the project sought to contribute to the existing body of research in this field and potentially provide insights for clinical decision-making.

# In summary, the internship project aimed to leverage machine learning techniques and the available Kaggle dataset to develop an accurate and reliable predictive model for cardiovascular disease. By building upon existing studies and considering specific contextual factors, the project sought to contribute to the growing field of cardiovascular disease prediction, ultimately aiming to improve patient outcomes and inform preventive healthcare strategies.

# **OBJECTIVES**

# 1. To develop a predictive model for cardiovascular disease: The main aim of the project was to create a robust predictive model and deploy it as a WebApp that could accurately identify individuals at risk of developing cardiovascular disease.

# 2. To enhance skills in data preprocessing and analysis: Throughout the internship, we aimed to improve my proficiency in data preprocessing techniques. This included handling missing values, removing outliers, and normalizing the dataset to ensure data quality and consistency. Additionally, we sought to enhance my skills in exploratory data analysis, gaining insights into the dataset and identifying relevant patterns and trends.

# 3. To gain experience in machine learning model development: The project provided an opportunity to develop and fine-tune machine learning models for predicting cardiovascular disease. We aimed to explore various algorithms such as logistic regression, decision trees, random forests, support vector machines, and potentially neural networks.

# To foster programming and analytical skills: Throughout the internship, we intended to further develop my programming skills, particularly in languages such as Python. We aimed to implement machine learning algorithms, perform model evaluation, and visualize the results effectively. Additionally, we aimed to enhance my analytical skills by interpreting and communicating the findings from the predictive model accurately.

# **SCOPE AND DELIVERABLES**

# The scope of my project during the internship encompassed the following tasks, activities, and deliverables:

# 1. Dataset exploration and preprocessing: we were responsible for thoroughly exploring the provided dataset and understanding its structure, variables, and potential data quality issues. This involved identifying missing values, outliers, and inconsistencies. We performed data cleaning and preprocessing tasks, including imputation of missing values, handling outliers, and normalizing the features for further analysis.

# 2. Feature selection and engineering: We conducted a comprehensive analysis of the dataset to identify the most relevant features for predicting cardiovascular disease. We performed feature selection techniques, such as correlation analysis and statistical tests, to select the most informative variables. Additionally, we explored the possibility of creating new features through feature engineering, leveraging domain knowledge and insights gained from the dataset.

# 3. Machine learning model development: We implemented various machine learning algorithms, including logistic regression, decision trees, random forests, support vector machines, and potentially neural networks. We trained these models using the preprocessed dataset, fine-tuning their hyper parameters to optimize their performance. Model evaluation techniques, such as cross-validation and performance metrics (accuracy, precision), were applied to assess the models' effectiveness.

# 4. Model selection and prediction: Based on the evaluation results, we selected the most effective machine learning algorithm for predicting cardiovascular disease. This selected model was then used to predict disease outcomes for new, unseen data points. The model's predictions were evaluated based on its accuracy and other relevant performance metrics.

# 5. Documentation and reporting: Throughout the internship, we maintained detailed documentation of the tasks performed, methodologies applied, and results obtained. We prepared reports summarizing the project's progress, including data preprocessing steps, feature selection techniques, model development, and evaluation outcomes. These reports served as deliverables to communicate the project's findings and insights.

# The deliverables of the internship project included a well-preprocessed dataset, a developed machine learning model for cardiovascular disease prediction, evaluation results, and comprehensive documentation and reports detailing the project's methodology and outcomes.

# **METHODOLOGY**

# The project was executed following a divide and conquer methodology, which involved dividing the tasks and responsibilities among team members to efficiently achieve project success. The team members were assigned specific roles based on their expertise and interests. Here is a breakdown of the team's roles and responsibilities:

# 1. **Roopak Susap Thomas (Partly Learner/Coder):** Roopak had a learning focus and contributed to the project by learning and gaining knowledge in various aspects of the project. He actively participated in discussions, collaborated with team members, and provided valuable input based on their learning experience.

# 2. **Abhishek S (Coding and Presentation):** Abhishek was primarily responsible for coding tasks related to the project. He implemented the machine learning algorithms, developed the predictive model, and fine-tuned its performance. Additionally, Abhishek played a key role in creating presentations to showcase the project's progress and outcomes. He was also responsible for front end coding

# 3. **Aleena Shaji Rekha and Azeya Y** **(Error Handling and Correction):** Aleena and Azeya were responsible for error handling and correction tasks. They actively identified and resolved any errors or issues that arose during the project's execution. Their focus was on ensuring the accuracy and reliability of the developed model by addressing any potential errors or discrepancies.

# 4. **Fathima Saleem (Reference Gathering):** Fathima had the responsibility of gathering relevant references and resources related to cardiovascular disease prediction and machine learning techniques. She conducted literature reviews, collected scholarly articles, and compiled a comprehensive list of references to support the project's findings and methodologies.

# In terms of tools and frameworks, the team utilized Google Colab for collaborative coding and data analysis tasks. VS Code was used as the primary integrated development environment (IDE) for coding and model development. For deploying the model as a web application, the team employed the Flask framework, which allowed them to create a user-friendly interface to interact with the predictive model.

# By following the divide and conquer methodology and leveraging tools such as Google Colab, VS Code, Flask, and an HTML renderer, the team effectively collaborated, implemented the necessary functionalities, and successfully executed the project in a structured and organized manner.

# **PROJECT ACTIVITIES**

# During my internship, we undertook a series of activities that involved various tasks, tools, and technologies. Here is a detailed breakdown of the activities we performed, the tools or technologies we used, and the challenges or obstacles encountered:

# 1. **Dataset Exploration and Preprocessing:**

# - **Task:** we thoroughly explored the dataset, analyzing its structure, variables, and potential data quality issues. we identified missing values, outliers, and inconsistencies.

# - **Tools/Technologies:** Python, Pandas, NumPy

# - **Challenges:** One challenge we encountered was errors in the dataset, such as inconsistent formatting or invalid entries. It required careful handling and preprocessing techniques to ensure data quality.

# 2. **Feature Selection and Engineering:**

# - **Task:** we performed a comprehensive analysis of the dataset to identify relevant features for predicting cardiovascular disease. we applied feature selection techniques and explored the possibility of creating new features through feature engineering.

# - **Tools/Technologies:** Python, Scikit-learn

# - **Challenges:** One challenge was determining the most informative features while avoiding overfitting or including irrelevant variables. It required careful analysis and experimentation to strike the right balance.

# 3**. Machine Learning Model Development:**

# - **Task:** we implemented various machine learning algorithms such as logistic regression, decision trees, random forests, and support vector machines (SVM). We trained and evaluated these models using the preprocessed dataset.

# - **Tools/Technologies:** Python, Scikit-learn

# - **Challenges:** One major challenge was limited computational power, particularly when working with complex algorithms like SVM. We had to optimize the model parameters and potentially reduce the dataset size to overcome this obstacle.

# 4. **Model Evaluation and Performance Metrics:**

# - **Task:** We evaluated the trained models using cross-validation techniques and assessed their performance using metrics such as accuracy, precision.

# - **Tools/Technologies:** Python, Scikit-learn

# - **Challenges:** The challenge here was selecting the most appropriate evaluation metrics based on the project's objectives and interpreting the results accurately.

# 5. **Deployment as a Web Application:**

# - **Task:** We deployed the developed predictive model as a web application using the Flask framework. This allowed users to interact with the model and obtain predictions through a user-friendly interface.

# - **Tools/Technologies:** Python, Flask, HTML

# - **Challenges:** A challenge encountered during deployment was creating a seamless integration between the machine learning model and the web application. Ensuring smooth functionality and handling any compatibility issues required careful debugging and troubleshooting.

# Overall, while working on the internship project, We faced challenges such as errors in the dataset and limited computational power for complex algorithms. However, by employing appropriate data preprocessing techniques, optimizing model parameters, and carefully selecting evaluation metrics, we were able to overcome these obstacles. The use of Python, Pandas, Scikit-learn and Flask provided the necessary tools and technologies to execute the project effectively.

# **RESULTS & FINDINGS**

# The internship project resulted in the successful deployment of a predictive model for cardiovascular disease, achieving an accuracy of 73.68%. The findings from the project revealed interesting insights about the attributes that are related to cardiovascular disease prediction. Here are the key accomplishments, insights, and data analysis conducted:

# 1. **Model Deployment and Accuracy:**

# - The developed predictive model was deployed as a web application, allowing users to input their data and obtain predictions about their risk of cardiovascular disease.

# - The model achieved an accuracy of 73.68% in predicting cardiovascular disease, which indicates its ability to classify individuals into high-risk and low-risk categories.

# 2. **Findings:**

# - The analysis revealed that certain attributes have a positive relationship with cardiovascular disease prediction. These attributes include age, weight, systolic and diastolic blood pressure, cholesterol levels, glucose levels, and activity level.

# - Surprisingly, attributes such as gender, height, smoking habits, and alcohol consumption showed a negative relationship with cardiovascular disease. This observation deviates from the typical assumption that these factors are strongly associated with the disease.

# 3. **Insights:**

# - The identified attributes with a positive relationship provide valuable insights into the risk factors associated with cardiovascular disease. This information can help healthcare professionals and individuals understand the importance of maintaining a healthy lifestyle and managing their weight, blood pressure, and cholesterol levels.

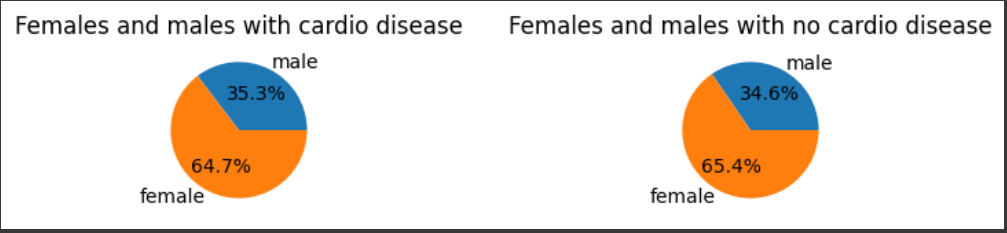
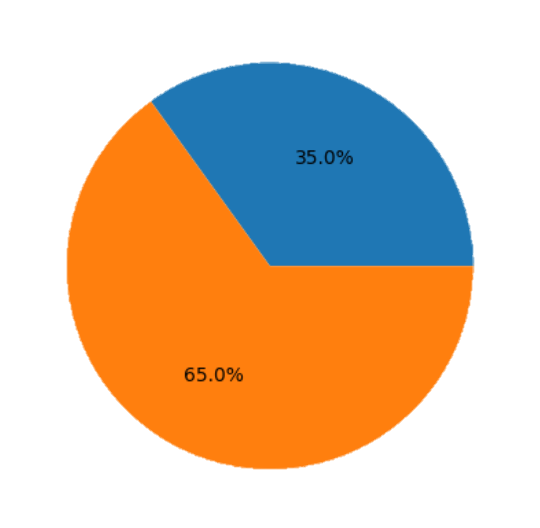
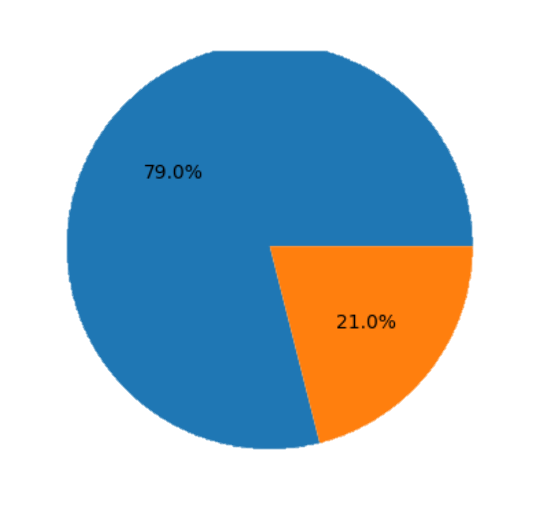
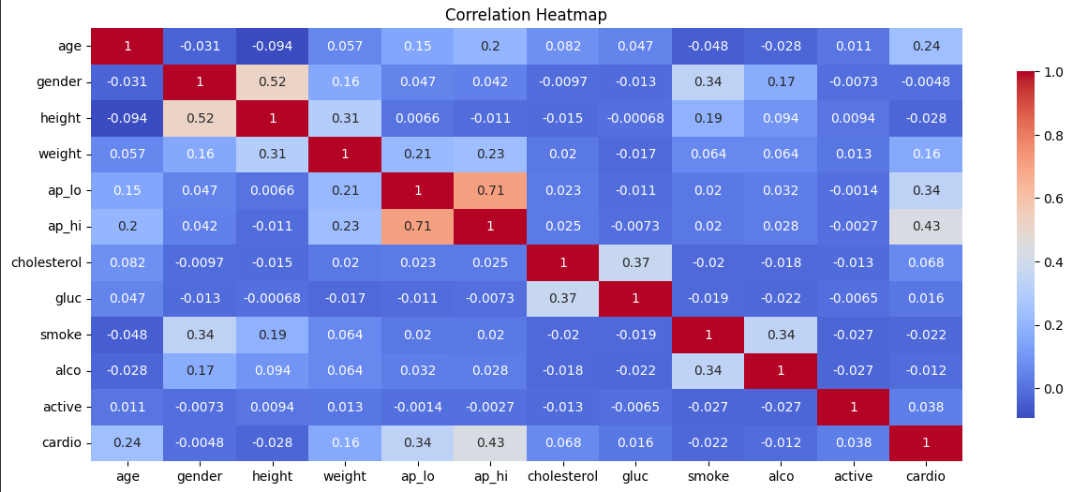
# - The unexpected negative relationship observed for attributes like gender, height, smoking habits, and alcohol consumption suggests that these factors may not play a significant role in predicting cardiovascular disease in the given dataset. This highlights the importance of considering context-specific factors and the limitations of the dataset.

# 4. **Dataset Issues:**

# - The analysis revealed potential problems with the collected data, suggesting the need for further investigation and validation. It is crucial to ensure the accuracy and reliability of the dataset used for training the predictive model.

# - The identified dataset issues could include inconsistencies, errors, or biases in the collected data. Addressing these issues is essential for improving the model's performance and reliability.

# Visualizations, such as graphs and charts, are provided in the ICT\_final\_project.py file. Some sample visualizations are given below

# **CONCLUSION**

# The main outcomes of the internship project on cardiovascular disease prediction were as follows:

# 1. **Developed Predictive Model:** A predictive model for cardiovascular disease was successfully developed, achieving an accuracy of 73.68%. The model was deployed as a web application, allowing users to obtain personalized predictions based on their input data.

# 2. **Insights into Risk Factors:** The project provided valuable insights into the risk factors associated with cardiovascular disease. Attributes such as age, weight, blood pressure, cholesterol levels, glucose levels, and activity level were identified as having a positive relationship with the disease. This information can help individuals and healthcare professionals better understand and address the risk factors.

# 3. **Unexpected Findings:** Surprisingly, attributes such as gender, height, smoking habits, and alcohol consumption showed a negative relationship with cardiovascular disease in the dataset. This suggests that these factors may not be strong predictors of the disease within the specific context of the dataset.

# 4. **Dataset Limitations and Issues:** The project highlighted potential problems with the collected dataset, such as inconsistencies, errors, or biases. Addressing these dataset issues is crucial for improving the model's performance and ensuring the accuracy and reliability of future predictions.

# Reflecting on the objectives of the project, the internship project successfully achieved its goals. The objectives included developing a predictive model, enhancing skills in data preprocessing and analysis, gaining experience in machine learning model development, and fostering programming and analytical skills.

# The project met its objective of developing a predictive model, which was successfully deployed as a web application. Skills in data preprocessing were enhanced through the exploration and cleaning of the dataset. Experience in machine learning model development was gained through implementing and evaluating various algorithms. Programming skills were honed through coding tasks and deployment using tools like Python, Scikit-learn and Flask

# Overall, the internship project provided valuable outcomes and learnings, contributing to the field of cardiovascular disease prediction. It not only achieved the objectives set at the beginning but also provided insights and findings that can inform preventive healthcare strategies and improve patient outcomes.

# **APPENDIX**

Additional supporting documents, samples of your work, project artifacts are given below in the zip file

**Link to Github account with all files are provided here:** <https://github.com/theunknownadorable/Internship-on-ML-AI/tree/master>