Unveiling Cardiovascular Health Patterns using Statistical and Predictive Analytics

Abhishek Anumalla
Data Analytics Engineering
George Mason University
Fairfax, USA
aanumall@gmu.edu

Aakash Boenal
Data Analytics Engineering
George Mason University
Fairfax, USA
aboenal@gmu.edu

Pavan Tejavath
Information Systems
George Mason University
Fairfax, USA
ptejavat@gmu.edu

Shashank Yelagandula
Data Analytics Engineering
George Mason University
Fairfax, USA
syelagan@gmu.edu

***Abstract* — Cardiovascular Disease (CVD) continues to be the leading cause of mortality globally, responsible for approximately 31% of all deaths according to the World Health Organization. This research aims to tackle the challenge of early detection and management of CVD by leveraging advanced statistical and predictive analytics techniques. Utilizing a dataset comprising 70,000 patient records, our research employs advanced statistical and predictive analytics to uncover key predictors of cardiovascular health outcomes. We conducted extensive data cleaning, exploratory data analysis, and rigorous testing of multiple machine learning models to determine the most effective predictive approach. Among the various models evaluated, the Gradient Boosting Algorithm was identified as the superior model, demonstrating an accuracy of 74% and an Area Under the Curve (AUC) of 81%. This high level of predictive accuracy highlights the potential of machine learning techniques in transforming CVD diagnostics and management. Our findings not only support the integration of sophisticated analytics into public health strategies but also pave the way for future innovations in cardiovascular health prediction tools, potentially leading to improved preventative measures and treatment methodologies for CVD.**

Keywords — Cardiovascular health, CVD, statistical analytics, predictive analytics, machine learning, heart disease prediction. Prediction tool.

I. INTRODUCTION

Cardiovascular disease (CVD) encompasses a range of conditions affecting the heart and blood vessels, including coronary artery disease, heart failure, and stroke. It remains the most formidable health challenge across the globe, leading to substantial morbidity and mortality. Despite significant advancements in medical science and public health, CVD is responsible for approximately 31% of all global deaths, as reported by the World Health Organization (WHO). This persistent threat underscores the critical need for effective diagnostic tools and intervention strategies that can mitigate the burden of CVD, particularly in aging populations and regions with rising incidence rates due to lifestyle changes.

The global impact of CVD is not only a health concern but also a significant economic burden, affecting the quality of life of individuals and communities, and straining healthcare systems worldwide. Traditional methods of diagnosing and managing CVD rely heavily on clinical assessments and standard statistical methods, which may not adequately capture the complexities of individual patient profiles and their risk factors. In response to these challenges, this study aims to harness the power of statistical and predictive analytics to improve the prediction and management of cardiovascular diseases.

By applying advanced machine learning techniques to a large dataset of 70,000 patient records, our research seeks to identify and analyze the key demographic, clinical, and lifestyle factors that contribute to CVD risk. This approach not only enhances our understanding of cardiovascular health patterns but also contributes to the development of more precise predictive tools, offering significant potential for early detection and personalized treatment strategies.

Through this research, we aim to contribute to the global effort in reducing the prevalence and impact of cardiovascular diseases by integrating innovative data-driven approaches into public health and medical practices.

II. LITERATURE REVIEW

The literature review offers a concise overview of research efforts focused on understanding and managing heart disease using diverse analytical approaches. From data mining and statistical analyses to machine learning and risk assessment tools, each study sheds light on various aspects of cardiovascular health.

By exploring the predictive capabilities of different models and emphasizing the importance of early intervention strategies, this review highlights the value of employing advanced analytics to enhance patient care and inform healthcare decision-making in the domain of heart disease management:

A. Prediction of heart disease using data mining and big data analytics:

This review delves into the utilization of data mining models and techniques to forecast heart disease based on extensive patient datasets. It underscores the pivotal role of data mining in extracting actionable insights from vast volumes of medical data, facilitating early diagnosis and preventive measures for heart disease. Encompassing a spectrum of data mining approaches, including support vector machines, neural networks, and decision trees, the review highlights their collective contribution to enhancing

patient outcomes and diagnostic precision. Moreover, it emphasizes the integration of big data analytics to manage complex medical datasets effectively, thereby optimizing predictive modeling efforts and offering potential strategies for managing heart disease more efficiently. [1]

B. Analysis of heart disease using statistical techniques:

Centering on heart disease, this paper accentuates its diverse manifestations and related risk factors, ranging from age and gender to obesity and smoking history. Employing binary logistic regression as its principal analytical tool, the research methodology aims to elucidate associations between these risk factors and the likelihood of developing heart disease. Noteworthy findings highlight the significance of factors such as depression, obesity, and chest pain as pivotal indicators of heart disease. By leveraging logistic regression's predictive capabilities, the study underscores its suitability for identifying and controlling cardiovascular risk factors, thereby averting adverse outcomes and bolstering patient care protocols. [2]

C. Cardiovascular disease analysis using data mining techniques:

Addressing the pressing issue of cardiovascular diseases (CVDs), this study underscores the imperative of early detection and treatment. Analyzing a dataset encompassing 14 characteristics linked to heart disease diagnosis, the research employs an array of data mining techniques, including decision trees, support vector machines, and Bayesian networks. Notable findings reveal the superior performance of support vector machines, with precision and recall metrics reaching significant levels, thereby affirming its utility as a diagnostic tool for cardiovascular disorders. By showcasing the potential of data mining techniques in healthcare decision-making, the study advances the development of diagnostic tools for CVDs, thereby contributing to improved patient care outcomes and healthcare practices. [3]

D. Machine Learning prediction in cardiovascular diseases: a meta-analysis:

This meta-analysis presents a machine learning (ML) approach to predict mortality following cardiac arrest by analyzing electrocardiogram (ECG) parameters. Through an in-depth exploration of ECG waveform patterns and abnormal signals, the research endeavors to identify predictive indicators associated with post-arrest death. By harnessing machine learning algorithms to evaluate ECG data, the study underscores the potential for enhancing prediction accuracy beyond conventional risk assessment methodologies. The integration of clinical data with computational methods not only augments patient care but also provides insightful guidance for risk prediction and stratification in cardiac emergencies, thereby advancing critical care protocols and bolstering healthcare decision-making frameworks. [4]

E. Primary Prevention of Cardiovascular Disease:

A special report from esteemed authorities such as the American Heart Association and American College of Cardiology underscores the pivotal role of risk assessment tools in guiding primary prevention strategies for atherosclerotic cardiovascular disease (ASCVD). Emphasizing the precise determination of individual cardiovascular risk profiles, the study advocates for tailored preventive measures integrating a comprehensive range of clinical, lifestyle, and demographic variables. By integrating the latest research and expert opinions, the report offers thorough guidance on the selection and utilization of risk assessment techniques, including validated risk scores and pooled cohort equations. This comprehensive approach not only facilitates more accurate risk prediction but also empowers healthcare practitioners to implement targeted preventive measures, ultimately reducing the burden of ASCVD and enhancing population health outcomes. [5]

F. Early prediction of cardiovascular disease using machine learning: Unveiling risk factors from health records:

This paper explores predictive analytics and cardiovascular health trends, with a focus on early cardiovascular disease (CVD) prediction by machine learning. Important genetic variants have been identified by genome-wide investigations and lifestyle factors, clinical indicators, and genetic predispositions all play crucial roles. The risk of CVD is influenced by well-established biomarkers such as blood pressure and unhealthy lifestyle. By utilizing techniques such as logistic regression and neural networks on electronic health records, one can improve predictive accuracy and identify individuals who are at high risk. The ramifications are enormous, allowing for focused therapies and wise resource distribution to lessen the growing worldwide burden of CVD. [6]

G. Machine Learning-Based Predictive Models for Detection of Cardiovascular Diseases:

The worldwide problem of cardiovascular illnesses emphasizes how urgently accurate detection techniques are needed. While previous research offers insightful information, predictive model developments are essential. This work focuses on early myocardial infarction diagnosis using machine learning, filling in gaps such as imbalanced datasets. In order to find efficient methods, it searches the literature and uses seven classifiers, including XGBoost and K-Nearest Neighbors. Results show that XGBoost performs exceptionally well, with solid precision, recall, and F1 scores in addition to 98.50% accuracy. This improved model shows usage of machine learning to improve predictive analytics and reveal complex health trends, which is a major step towards accurately diagnosing cardiovascular disease. [7]

H. Revolutionizing Cardiovascular Health: Integrating Deep Learning Techniques for Predictive Analysis of Personal Key Indicators in Heart Disease:

This review of the literature explores the paradigm change in cardiovascular health that comes from integrating deep learning approaches for personal key indicator prediction in heart disease. The study investigates the revolutionary potential of deep learning in decoding complex cardiovascular data, with an emphasis on revealing nuanced patterns. By utilizing sophisticated algorithms like convolutional and recurrent neural networks, scientists hope to reveal complex connections between specific health metrics and cardiovascular consequences. This review, which marks the beginning of a new age in preventive cardiovascular healthcare measures, highlights the

importance of predictive analytics in anticipatory risk assessment and individualized therapies by combining data from existing research. [8]

I. Risk prediction of cardiovascular disease using machine learning classifiers:

This review of the literature looks at the necessity of automated and early diagnosis of cardiovascular disease (CVD), which is a vital first step in reducing related deaths and disabilities. Even though this objective has been the focus of many studies, there are still ways to improve performance. This work makes a contribution by using data from the University of California, Irvine repository to apply two strong machine learning techniques: multi-layer perceptron (MLP) and K-nearest neighbor (K-NN). Model performance is maximized by attribute management and outlier elimination, with the MLP model surpassing the K-NN model with an area-under-the-curve value of 86.41% and improved accuracy of 82.47%. With potential applicability in different disease detection situations, this approach shows promise for automated CVD detection and encourages additional validation across a variety of datasets. [9]

J. Unveiling the Influence of AI Predictive Analytics on Patient Outcomes: A Comprehensive Narrative Review:

This thorough review of the literature explores the significant effects of AI predictive analytics on healthcare, especially in terms of improving patient outcomes related to illness progression, treatment effectiveness and recovery. Artificial intelligence (AI) uses machine learning (ML) and deep learning (DL) techniques to examine large datasets, such as genetic data and electronic health records (EHRs), to enable personalized care and early illness identification. When implementing AI, ethical factors like data privacy and bias reduction are crucial. The review highlights how AI has the ability to completely transform clinical decision-making and the provision of healthcare, but it also stresses how ethical standards and continuous validation are necessary to guarantee AI's safe and efficient incorporation into medical practice. [10]

K. Precision Health Analytics With Predictive Analytics and Implementation Research: JACC State-of-the-Art Review:

Using a focus on cardiovascular health trends, this literature review investigates the rapidly developing subject of predictive analytics in precision health. Insights into blood, lung, heart and sleep diseases are provided by emerging data science tools, opening doors to better knowledge and treatment. Examining the advantages, drawbacks and long-term viability of predictive analytics in healthcare delivery requires implementation research. This review emphasizes the integration of predictive analytics into clinical practice and public health initiatives as crucial domains, with a focus on precision medicine and public health. It also emphasizes the need for ongoing research and training to improve cardiovascular disease prevention and management strategies. [11]

L. PREDICTIVE ANALYSIS OF HEART DISEASES WITH MACHINE LEARNING APPROACHES:

This review of the literature looks at the necessity of automated and early diagnosis of cardiovascular disease (CVD), which is a vital first step in reducing related deaths and disabilities. Even though this objective has been the focus of many studies, there are still ways to improve performance. This work makes a contribution by using data from the University of California, Irvine repository to apply two strong machine learning techniques: multi-layer perceptron (MLP) and K-nearest neighbor (K-NN). Model performance is maximized by attribute management and outlier elimination, with the MLP model surpassing the K-NN model with an area-under-the-curve value of 86.41% and improved accuracy of 82.47%. With potential applicability in different disease detection situations, this approach shows promise for automated CVD detection and encourages additional validation across a variety of datasets. [12]

M. Unveiling Insights: Exploring Cardiovascular Health through Data Analytics

This review of the literature explores the relationship between cardiovascular health and data analytics, shedding light on the revolutionary possibilities of using data to better understand and treat cardiovascular disorders. The story opens with an overview of the importance of cardiovascular illness and emphasizes how crucial data analytics is to understanding its intricacies and developing effective interventions. The review sets out on a quest to unearth hidden insights and patterns through examination of a cardiovascular health-focused dataset, highlighting the significant influence of data analytics in influencing the direction of cardiovascular health research and preventive measures in the future. [13]

N. Cardiovascular Disease Prediction combination Using Machine and Deep Learning Model

This review of the literature looks at how deep learning (DL) and machine learning (ML) models work together to predict cardiovascular disease (CVD), with an emphasis on the need for early and precise risk assessment to enhance patient outcomes. The effectiveness of DL models like LSTM, CNN and Naive Bayes, as well as ML methods like Decision Tree, Naive Bayes, Logistic Regression, SVM and Random Forest, is assessed in this study. We present a new ensemble classifier that maximizes predicted accuracy by combining the best features of several models. The findings provide encouraging results, highlighting the significance of classifier selection and supporting the use of ensemble models in proactive healthcare approaches for the treatment of cardiovascular health. [14]

III. ABOUT THE DATASET

The dataset utilized in this study was sourced from the Huggingface platform, comprising 70,000 records with each record containing a range of demographic, clinical, and lifestyle characteristics across 12 distinct features. These features provide a comprehensive view of each patient's cardiovascular health profile, which is crucial for our predictive modeling efforts.

The demographic information within the dataset includes age, which was originally recorded in days and converted to

years for our analysis; gender; and height measured in centimeters. Clinical parameters featured in the dataset consist of systolic and diastolic blood pressure readings (noted as ap_hi and ap_lo), along with cholesterol and glucose levels. These latter two parameters are categorized into normal, above normal, and well above normal based on standard medical benchmarks. Lifestyle variables captured include data on smoking status, alcohol intake, and physical activity levels, which are essential for assessing the impact of lifestyle on cardiovascular health.

Each record in the dataset also includes a target variable, 'Cardio', indicating the presence (1) or absence (0) of cardiovascular disease. This binary outcome aids in training the predictive models to distinguish between affected and unaffected individuals based on the provided features. In terms of data quality and preprocessing, an initial examination revealed that the dataset was robust with no missing values, ensuring a high level of completeness.

However, some discrepancies such as outliers in blood pressure measurements were noted. These outliers, including clinically implausible systolic values like 1000 mmHg, were addressed by removing records containing such abnormal values. This step is critical to prevent skewed analysis and to ensure the models trained are robust and reflective of realistic clinical scenarios. The comprehensive nature of the dataset and the rigorous preprocessing steps taken to ensure its quality form the foundation for the exploratory data analysis and modeling phases of the study. Fig. 1 and Fig. 2 displays a snippet of the dataset from the HuggingFace platform. [15]

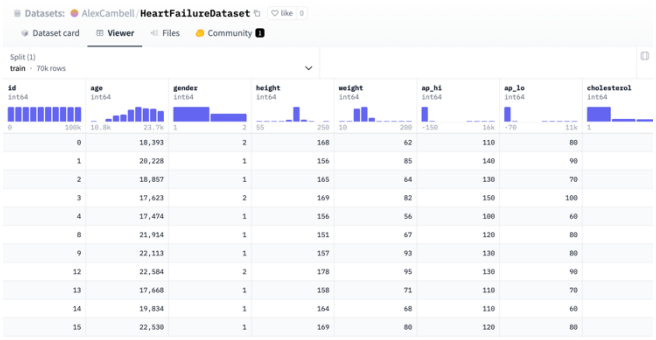

Fig. 1. Snippet of the dataset (1)

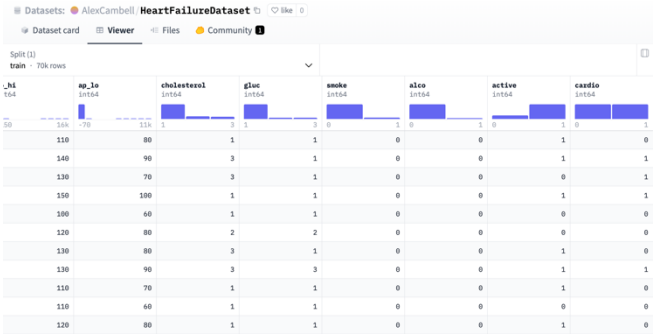

Fig. 2. Snippet of the dataset (2)

IV. PROPOSED APPROACH

Our project is dedicated to advancing the prediction and management of cardiovascular diseases (CVD) using cutting-edge machine learning technologies. Our comprehensive approach is designed to develop, refine, and implement a predictive model that effectively assesses CVD risks from a variety of demographic, clinical, and lifestyle factors.

Overview of Key Phases:

1) Data Acquisition and Preparation: We begin by acquiring a robust dataset from the Huggingface platform, which includes over 70,000 records encompassing a wide range of health indicators. The data is meticulously cleaned and prepared, transforming raw data into a reliable foundation for in-depth analysis. This step is critical for ensuring the accuracy and reliability of our predictive models, as it involves correcting any inconsistencies and standardizing data formats across multiple variables.

2) Model Development: In this phase, we explore a variety of statistical techniques and machine learning algorithms to identify the most effective method for predicting CVD. Our focus is on testing and comparing different models, such as Logistic Regression, Support Vector Machines, and Gradient Boosting, among others, to determine which offers the best balance of accuracy, reliability, and applicability in clinical settings. The development process is iterative, involving continuous refinement and optimization based on performance metrics like accuracy, precision, and the area under the ROC curve.

3) Tool Development: The culmination of our project is the creation of an intuitive, user-friendly predictive tool that integrates our best-performing model. This tool is designed to be used by healthcare professionals in real-world scenarios, allowing them to input patient data and receive immediate risk assessments. The tool's interface is crafted to ensure ease of use, with clear inputs and outputs that support rapid decision-making in clinical environments. This phase also involves rigorous testing and user feedback sessions to refine the tool's functionality and ensure it meets the practical needs of end-users.

The overarching goal of our approach is to bridge the gap between complex analytical techniques and everyday clinical practice. By converting sophisticated data-driven insights into an accessible and practical tool, we aim to enhance the accuracy of CVD predictions, thereby aiding healthcare providers in early diagnosis and targeted intervention. This strategic approach not only improves patient outcomes but also contributes to the broader field of medical informatics by demonstrating the practical benefits of integrating machine learning into healthcare.

V. PROPOSED METHOD

Our study employs a structured methodology to utilize advanced machine learning technologies with the aim of improving predictive models for cardiovascular disease (CVD). This methodology encompasses several stages, from data collection to the development of a practical predictive tool, each designed to optimize the quality and applicability of our findings.

1) Data Collection: Our research started by gathering a large dataset from the Huggingface platform. This dataset included 70,000 records filled with important information about people's health. It features basic demographic details like age, gender, and height, as well as more detailed clinical information such as blood pressure, cholesterol levels, and glucose levels. We also included lifestyle information like smoking habits, alcohol consumption, and physical activity levels to see how these might influence heart health. The choice of this dataset was critical. We made sure it was detailed and accurate, representing a broad spectrum of individuals to ensure our findings would be applicable to the general population at risk for or suffering from heart diseases. [15]

2) Data Cleaning & Preparation: Following the collection of the dataset, we undertook a rigorous data cleaning and preparation process to ensure the data's quality and usability. First, we converted age from days into years to enhance interpretability and relevance to clinical assessments. We also scrutinized the dataset for outliers, particularly in blood pressure readings, removing any values that were clinically implausible, such as systolic pressures exceeding 1000 mmHg. This step was critical to maintaining the integrity of our analyses. Additionally, we checked for duplicates and inconsistencies, confirming the dataset's completeness and accuracy. Lastly, we standardized the variables to ensure uniformity in measurement scales across the dataset, setting a solid groundwork for the subsequent exploratory data analysis and modeling. This thorough preparation phase was pivotal in ensuring that the dataset was primed for detailed analysis, allowing us to rely on its accuracy for developing predictive models.

3) Exploratory Data Analysis (EDA): Following the meticulous preparation of our dataset, we embarked on an exploratory data analysis (EDA) phase. This crucial step involved a deep dive into the dataset to uncover underlying patterns, identify key trends, and detect any potential anomalies that could influence our subsequent modeling efforts. The EDA was conducted using a variety of statistical tools and visualization techniques to provide a comprehensive understanding of the data's characteristics. We began by assessing the distribution of each variable, such as age, weight, cholesterol levels, and blood pressure, to gain insights into the general health status represented within the dataset. This initial analysis helped us to understand the typical ranges and variances in our data, which are essential for setting realistic parameters in our predictive models.

Next, we examined correlations between the various features to identify relationships that could be predictive of cardiovascular disease. For instance, we explored how factors like age and cholesterol levels interact with blood pressure readings, and whether lifestyle choices such as smoking and physical activity correlate with increased or decreased cardiovascular risk. These correlations were visualized using heatmaps and scatter plots, providing a clear visual representation of the relationships.

Additionally, we conducted subgroup analyses to compare the characteristics of individuals with and without cardiovascular disease. This approach allowed us to pinpoint specific patterns and risk factors that are more prevalent in the affected group. Visual tools like box plots and histograms were employed to illustrate these differences, making it easier to identify which features might play a more significant role in the development of cardiovascular conditions. The insights gained from the EDA were instrumental in guiding our feature selection and engineering phases. By understanding which variables had the most significant associations with cardiovascular disease, we could focus our modeling efforts on the most impactful predictors. This not only enhanced the efficiency of our modeling process but also improved the potential accuracy and clinical relevance of our predictive models. Overall, the exploratory data analysis provided a solid foundation for the advanced analytical techniques that followed. It ensured that our approach to modeling was informed by a thorough understanding of the data, setting the stage for more targeted and effective predictive analysis.

4) Feature Selection and Engineering: After completing the exploratory data analysis, we focused on feature selection and engineering to refine our predictive model for cardiovascular disease. This stage was crucial for enhancing the model's accuracy by identifying the most impactful predictors and minimizing redundancy. During the feature selection process, we have decided to scale the data. We decided to scale the data to make the model's training process more efficient and maximize the model's performance. Standardized scaling method was used. We then, carefully analyzed the relationships between various features and the target variable 'Cardio'. This helped us isolate key predictors such as Age, Weight, Ap_hi, Ap_lo, Cholesterol levels. Our goal was to streamline the model to focus on these variables, reducing the complexity and improving the model's generalizability. Feature engineering was conducted to maximize the predictive value of the data we collected. This involved adjusting existing variables to better suit our analytical models. For instance, age was converted from days to years to make it more interpretable. We also addressed discrepancies and outliers in blood pressure measurements, ensuring that our dataset reflected realistic and clinically relevant values. By meticulously selecting and refining features, we ensured that our predictive model was not only accurate but also efficient. This step was instrumental in improving the model's performance, enabling it to provide reliable predictions that could potentially aid in early detection and personalized treatment of cardiovascular disease. [16]

5) Model Development: In the model development phase of our study, we tested a broad array of machine learning algorithms to establish the most effective model for predicting cardiovascular disease. This phase was crucial for identifying a robust model that could handle the complexities of our dataset, which includes demographic, clinical, and lifestyle variables. [19] [20] [22]

We evaluated fourteen different models, each known for its strengths in binary classification tasks:

- Logistic Regression
- K-Nearest Neighbors
- Support Vector Machine (both linear and kernel versions)
- Decision Trees
- Gradient Boosting Algorithm
- Random Forest
- Naive Bayes
- Bagging Aggregating Classifier
- Voting Classifier (hard and soft vote)
- ADA Boost Classifier
- XGB Classifier
- Ridge Classifier

Each model was trained using a standardized portion of our dataset and then validated to assess its performance on unseen data. The models' effectiveness was measured primarily by their accuracy, which is crucial for ensuring reliable predictions in a clinical setting. The careful selection of machine learning models and subsequent evaluation based on accuracy and other performance metrics ensured that we chose a model not only based on theoretical suitability but proven practical effectiveness. This approach significantly contributed to our study's goal of developing a reliable predictive tool for cardiovascular health assessment.

6) Model Evaluation and Validation: Throughout the project, a comprehensive evaluation and validation process was critical in identifying the most effective predictive model for cardiovascular disease. This process involved rigorously testing fourteen different models to assess their performance in accurately predicting cardiovascular conditions. The models were evaluated primarily based on their accuracy and the area under the receiver operating characteristic (ROC) curve (AUC). The Gradient Boosting Algorithm emerged as the most effective model, achieving the highest accuracy of 74% and an AUC of 81%. These metrics indicate the model's excellent capability in distinguishing between patients with and without cardiovascular disease, making it particularly valuable for clinical applications.

Validation involved applying these models to a test set of data, which was crucial for assessing how well the models could generalize to new, unseen data. This step ensured that our findings were robust and applicable beyond the controlled conditions of our study. The evaluation and validation process confirmed that the Gradient Boosting Algorithm, with its superior accuracy and AUC, was the most reliable and effective tool among the tested options for predicting cardiovascular disease. This model's success underscores the potential of advanced predictive analytics in improving diagnostic accuracy and enhancing preventive healthcare strategies. [19] [20] [21] [22]

7) Development of Prediction Tool: In our project, the culmination of our predictive analytics efforts was the creation of the Cardiovascular Disease Prediction Tool, developed using the DASH App Framework. This interactive web application integrates the Gradient Boosting Algorithm, which our evaluations identified as the most accurate model with an accuracy of 74% on the test set. The tool features a user-friendly interface designed to allow healthcare providers to input patient data across several critical health indicators, including age, gender, weight, height, blood pressure (both systolic and diastolic), cholesterol levels, glucose levels, smoking status, alcohol consumption, and physical activity levels. The simplicity and intuitiveness of the interface ensure that it can be seamlessly integrated into clinical workflows, enabling healthcare professionals to quickly assess an individual's risk of cardiovascular disease. [17]

Upon entering the patient data and clicking the "Predict" button, the tool instantly evaluates the risk based on the inputted data and displays whether the patient has been diagnosed with cardiovascular disease along with the model's accuracy percentage. This immediate feedback is crucial for facilitating timely medical advice and interventions. The tool's backend, built on Python's DASH framework, not only supports these real-time predictions but also ensures that the data handling is secure and efficient, which is vital for maintaining patient confidentiality and trust in medical IT systems. This prediction tool represents a significant advancement in our ability to combat cardiovascular disease by providing a means for early detection and proactive management based on personalized health data. It exemplifies how data-driven insights can be effectively translated into practical tools that enhance patient care and support healthcare decision-making.

VI. PRELIMINARY RESULTS

Our research into developing predictive models for cardiovascular disease (CVD) has produced promising preliminary results, demonstrating the efficacy of machine learning in medical diagnostics. Below, we detail the outcomes beginning with data cleaning and spanning through model testing to the initial use of our predictive tool.

1) Data Cleaning and Preparation: The initial dataset comprised 70,000 records with key health indicators but required thorough cleaning to ensure accuracy and reliability for analysis. We found no null values, which simplified the cleaning process, but addressed several discrepancies. The age data, originally recorded in days, was converted to years to align with clinical standards and enhance interpretability.

Notably, we corrected outliers in blood pressure measurements (ap_hi and ap_lo) where values such as -150 or 11,000 were deemed clinically implausible and were removed to maintain the integrity of our dataset. We also eliminated the ID column from the dataset as it was unnecessary for our analysis, focusing instead on relevant health indicators. Fig. 3 shows the Summary Statistics of the Dataset before Data Cleaning and Preparation process.

Index	age	ap_hi	ap_lo
count	70000	70000	70000
mean	19468.9	128.818	96.6304
std	2467.25	154.011	188.473
min	10798	-150	-70
25%	17664	120	80
50%	19703	120	80
75%	21327	140	90
max	23713	16020	11000

Fig. 3. Summary Statistics of the Dataset before Data Cleaning and Preparation process

2) Final Dataset: After cleaning, the dataset contained 68,692 records with 12 essential features, including age, gender, height, weight, blood pressure, cholesterol, glucose levels, smoking status, alcohol consumption, physical activity, and the presence of cardiovascular disease. Statistical summaries of the cleaned data show realistic ranges and distributions, for instance, blood pressure readings are now within a plausible range (systolic 60-240 mmHg and diastolic 20-182 mmHg), reflecting actual physiological conditions. Fig. 4 shows the Summary Statistics of the Final Dataset that is, after Data Cleaning and Preparation process.

Index	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
count	68692	68692	68692	68692	68692	68692	68692	68692	68692	68692	68692	68692
mean	52.8286	1.34863	164.362	74.1178	126.674	81.2904	1.3646	1.22569	0.0879433	0.0533395	0.803369	0.49473
std	6.76901	0.47654	8.18313	14.3309	16.6951	9.50595	0.678831	0.571536	0.283214	0.224711	0.397454	0.499976
min	29	1	55	11	60	20	1	1	0	0	0	0
25%	48	1	159	65	120	80	1	1	0	0	1	0
50%	53	1	165	72	120	80	1	1	0	0	1	0
75%	58	2	170	82	140	90	2	1	0	0	1	1
max	64	2	250	200	240	182	3	3	1	1	1	1

Fig. 4. Final Dataset (after Data Cleaning and Preparation process)

3) Statistical Analysis Data Distribution: Exploratory data analysis (EDA) followed the data cleaning process, where we delved into the dataset to uncover underlying patterns and relationships. Our analysis provided insightful observations on how variables such as age, weight, and cholesterol levels correlate with the presence of CVD. The distribution plots and histograms in Fig. 5 and Fig. 6, crafted from our data illustrated these relationships vividly, showing distinct patterns that could potentially inform our feature selection and predictive modeling.

The cleaned and analyzed data have proven invaluable in laying a solid foundation for our subsequent modeling efforts. Our detailed statistical understanding of various risk factors has informed the selection of features and the construction of our predictive models.

These efforts are crucial as we aim to develop not only accurate predictive models but also a user-friendly predictive tool that can seamlessly integrate into clinical workflows, thereby enhancing the capability of healthcare providers to assess and manage cardiovascular disease risks effectively.

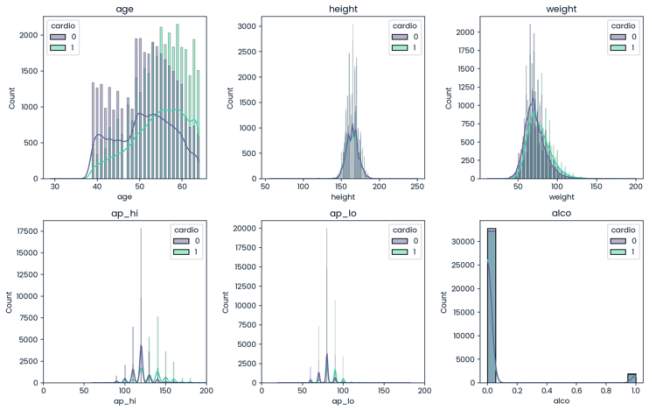

Fig. 5. Statistical Analysis Data Distribution (1)

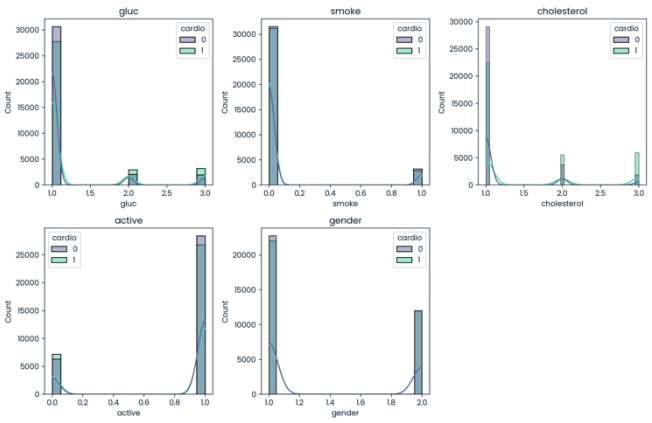

Fig. 6. Statistical Analysis Data Distribution (2)

4) Correlation Analysis Findings: The correlation matrix we developed as part of our statistical analysis highlights strong positive correlations among several critical attributes: Age, Weight, systolic blood pressure (Ap_hi), diastolic blood pressure (Ap_lo), and Cholesterol.

These findings are crucial as they directly inform our predictive modeling. There is a notable increase in CVD risk as age increases, which aligns with existing medical understanding. Both systolic and diastolic pressures show a strong correlation with CVD occurrence, underlining their importance in cardiovascular health assessments.

These factors are well-documented risk factors for cardiovascular disease and their strong correlations in our analysis confirm their predictive value. Based on these correlations, our subsequent analyses and model development have focused on these attributes, enhancing our ability to predict CVD with greater accuracy and reliability. Fig. 7 shows the Correlation analysis to visualize the relationship with all attributes of the dataset.

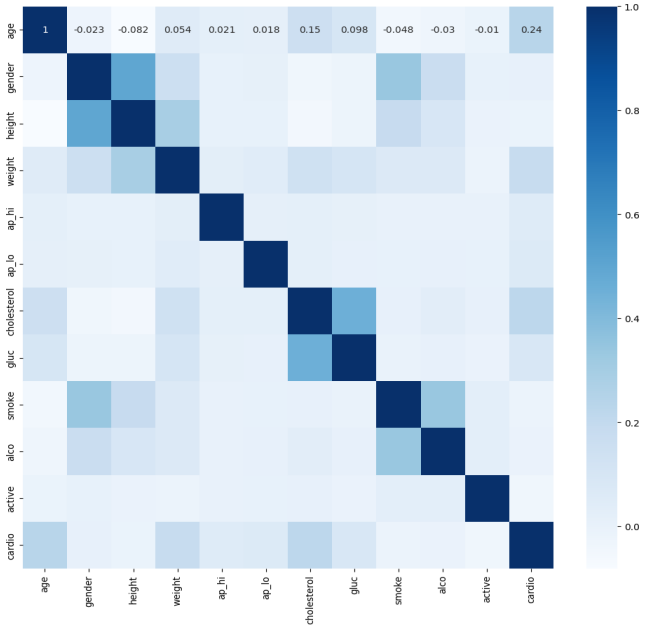

Fig. 7. Correlation analysis to visualize the relationship with all attributes of the dataset

5) Scatter Plot Analysis: Furthering our analysis, we utilized a scatter plot to visualize the average risk of cardiac disease by age and cholesterol level. This visualization, in Fig. 8 provides a clear, intuitive understanding of how age and cholesterol levels interact to influence the risk of cardiac disease. The plot uses a color gradient from blue to red where blue represents a minimal risk (0%) and red denotes higher risk percentages, reaching up to 71%.

For instance, in Fig. 8 the data points show that younger individuals around the age of 30 with a cholesterol level of 1 have a very low risk of developing cardiac disease. In contrast, individuals aged 58 with a cholesterol level of 3 face a significantly higher average risk, approximately 71%. This gradient not only highlights the increasing risk with age and higher cholesterol but also visually emphasizes the gradient of risk across the age spectrum.

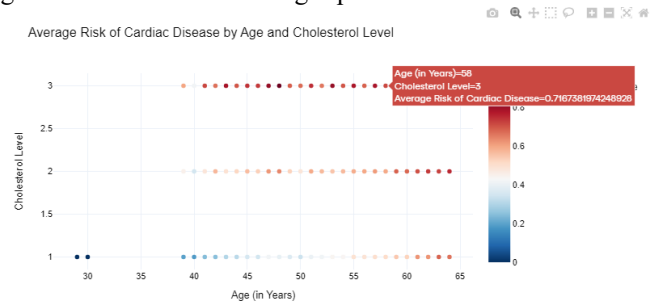

Fig. 8. Scatter plot analysis to visualize the average risk of CVD by age and cholestrol levels

6) Age-Based Risk Analysis: Expanding on our visual analyses, another graph in Fig. 9 presents a compelling view of how the probability of cardiovascular disease escalates with age. This bar chart, through a gradient of blue to darker blue, shows a progressive increase in risk as age advances, a trend that is consistent with both medical literature and our dataset's patterns. For instance, the graph clearly delineates that individuals in their early thirties have a significantly lower risk of cardiac diseases, which gradually increases with each decade. By the time individuals reach the age of 63, the average risk surges to approximately 72%, underscoring age as a pivotal factor in cardiac risk assessments. This visualization not only corroborates the data derived from our correlation analysis but also serves as a stark illustration of the critical need for targeted preventive measures among older populations.

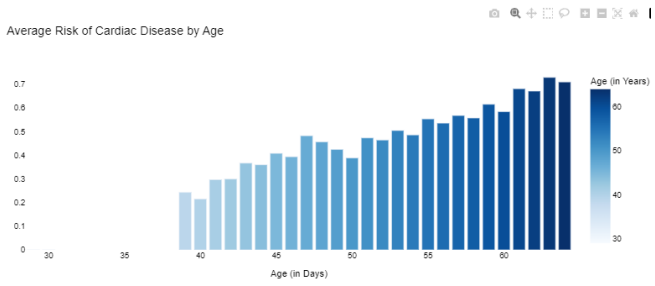

Fig. 9. Bar plots to visualize average risk of Cardiac disease by Age

7) Cholesterol Level Impact on Cardiac Disease Risk: In addition to the previous analyses, we also explored the influence of cholesterol levels on cardiac disease through a series of pie charts in Fig. 10.

These charts effectively segment the risk percentages based on cholesterol categories—normal, above normal, and well above normal.

Normal Cholesterol Levels: Individuals with cholesterol levels within the normal range showed a 44% risk of developing cardiac diseases. This baseline provides a comparative standard for assessing higher cholesterol levels.

Above Normal Cholesterol Levels: As cholesterol levels rise above the normal threshold, the risk of cardiac diseases significantly increases to 60.2%. This substantial rise highlights the critical impact of moderately elevated cholesterol on cardiovascular health.

Well Above Normal Cholesterol Levels: The risk escalates further for individuals with cholesterol levels well above normal, reaching a stark 76.5%. This dramatic increase underscores the severe threat posed by high cholesterol levels to cardiac health.

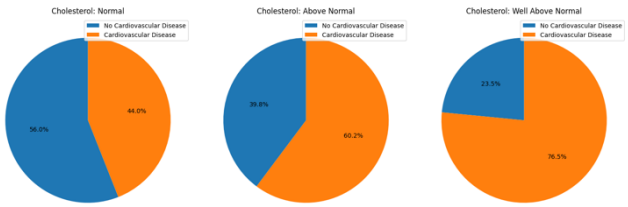

Fig. 10. Pie charts to visualize the risk of Cardiac disease by Cholesterol levels

8) Blood Pressure Variability and Cardiovascular Risk: In addition to the previous analyses, we also explored the influence of cholesterol levels on cardiac disease through a series of pie charts in Fig. 10. These charts effectively segment the risk percentages based on cholesterol categories—normal, above normal, and well above normal.

Our study's statistical analysis continues to provide critical insights into the various factors influencing cardiovascular health. A significant part of our analysis focused on blood pressure measurements—systolic (ap_hi) and diastolic (ap_lo)—and their correlation with the incidence of cardiovascular diseases.

In Fig. 11, we have been able to visually compare and analyze the variability and distribution of blood pressure between individuals diagnosed with cardiac disease and those without. These plots are instrumental in illustrating the relationship between blood pressure levels and cardiovascular health.

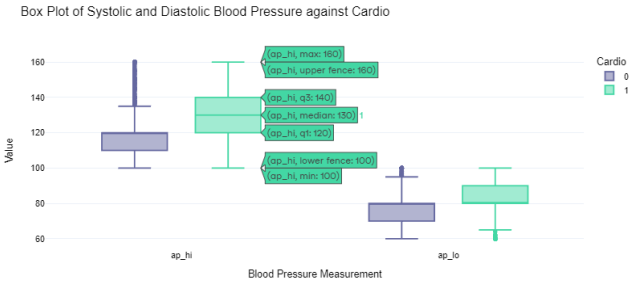

Fig. 11. Box plot to visualize the Systolic and Diastolic Blood Pressure against Cardio

9) Model Evaluation in Cardiovascular Disease Prediction: Our preliminary findings extend beyond statistical correlations and risk factor visualizations to include a comprehensive evaluation of various machine learning models aimed at predicting cardiovascular disease. This evaluation is critical in determining the most effective tools for our predictive analytics.

Comparison of Model Performances: The study tested a variety of machine learning models to identify those that provide the highest accuracy in predicting cardiovascular events.

Diverse Model Testing: We tested fourteen different models, including Logistic Regression, K-Nearest Neighbors, Support Vector Machines (both linear and kernel-based), Decision Trees, Gradient Boosting, Random Forest, and several ensemble methods like AdaBoost and XGBoost. Each model was chosen based on its potential suitability for handling binary classification tasks within the context of our dataset. Fig. 12 shows the accuracy bar graph that visually represents the performance of each model, making it easy to compare their effectiveness at a glance. [22]

Model Accuracy Insights: The Gradient Boosting Algorithm emerged as the top performer with an accuracy of 74%, closely followed by the Support Vector Machine (Kernel) at around 73.5%. XGBoost Classifier, ADA Boost Classifier and Voting Classifier. Soft Vote Classifier also showed commendable performance with accuracies around 73%. The performances of other models can be seen in the graph below. These models demonstrated their ability to effectively manage the complexities of cardiovascular disease prediction, taking into account a multitude of risk factors. [19] [22]

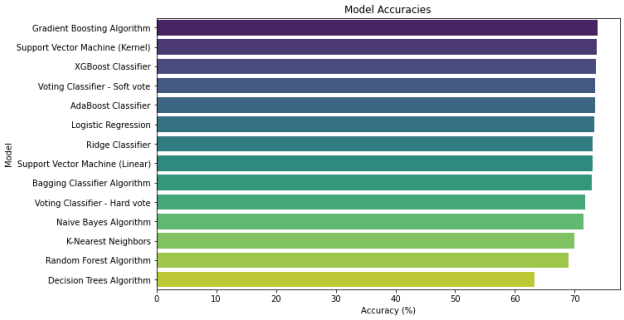

Fig. 12. Accuracy Bar Graph to visualize the performance of each model

MODEL NAME	ACCURACY
Gradient Boosting Algorithm	74%
Support Vector Machine (Kernel)	73.7%
XGBoost Classifier	73.5%
Voting Classifier – Soft Vote	73.5%
ADA Boost Classifier	73.4%
Logistic Regression	73.3%
Ridge Classifier	73%
Support Vector Machine (Linear)	73%
Bagging Classifier Algorithm	72.9%
Voting Classifier – Hard Vote	71.7%
Naïve Bayes Algorithm	71.4%
K-Nearest Neighbors	69.8%
Random Forest Algorithm	68.9%
Decision Trees Algorithm	63.2%

Table. 1: Accuracies of each model

The ROC Curve, confusion matrix in Fig. 13 and feature importances in Fig. 14 have been plotted for the model with best accuracy. For the Gradient Boosting Model, the features with the most significance seemed to be ap_hi with 0.74, age with 0.12 and cholesterol with 0.07. In Fig. 13, the confusion matrix gives more detailed picture of the error rates. [21]

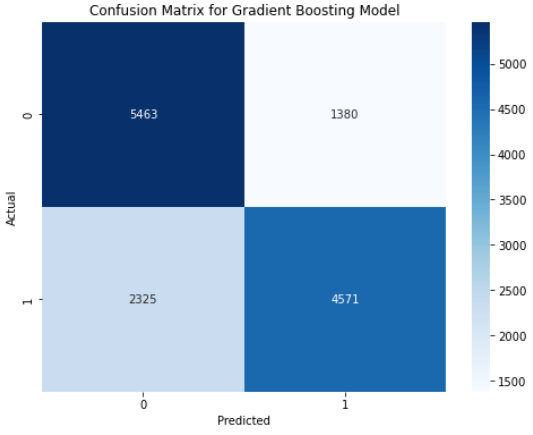

Fig. 13. Confusion Matrix for Gradient Boosting Model

Feature	Importance
ap_hi	0.746758
age	0.124769
cholestrol	0.077272
weight	0.019922
ap_lo	0.012798
active	0.005296
glu	0.004637
height	0.003528
smoke	0.002177
alco	0.001970
gender	0.000873

Fig. 14. Plotting the Feature Importances

The area under curve value (in Fig. 15) for the Grading Boosting model (best model) is 0.81 which is pretty decent. The next best area under curve value was 0.80. Three models had this value. SVM-Kernel, ADA Boost and XGBoost Classifier Models had AUC of 0.80. [21]

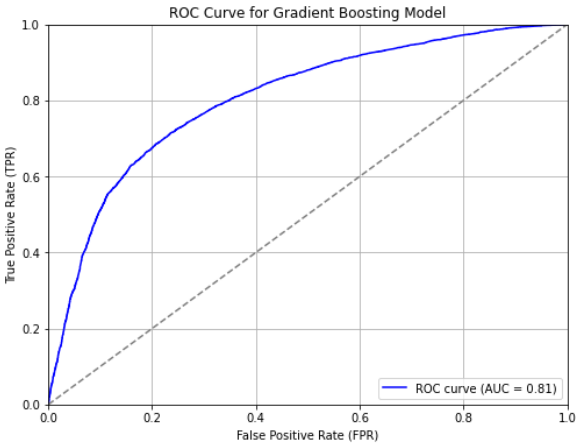

Fig. 15. ROC Curve for Gradient Boosting Model

10) Deployment of Cardiovascular Disease Prediction Tool:

Concluding our preliminary results, we highlight the practical application of our research through the development and deployment of the "Cardiovascular Disease Prediction Tool." This innovative tool embodies the synthesis of our analytical efforts and model evaluations, providing a user-friendly interface for predicting cardiovascular disease based on individual health metrics. [17]

Functionality and Interface:

The prediction tool, designed using the DASH framework, allows users to input several key health parameters, including age, gender, weight, height, blood pressure (both systolic and diastolic), cholesterol levels, glucose levels, smoking status, alcohol consumption, and physical activity levels.

Each parameter can be adjusted to reflect the user's personal health information, which is then processed by the underlying predictive model—the Gradient Boosting Algorithm, noted for its high accuracy and reliability. [17] [21]

Predictive Outcomes:

Positive Diagnosis: For instance, in Fig. 13 a user aged 55, with parameters like a cholesterol level of 3 and higher blood pressure readings (140/90), is predicted to have cardiovascular disease, reflecting an accuracy of 74% based on our model's test results. This scenario demonstrates the tool's capacity to identify higher-risk profiles effectively.

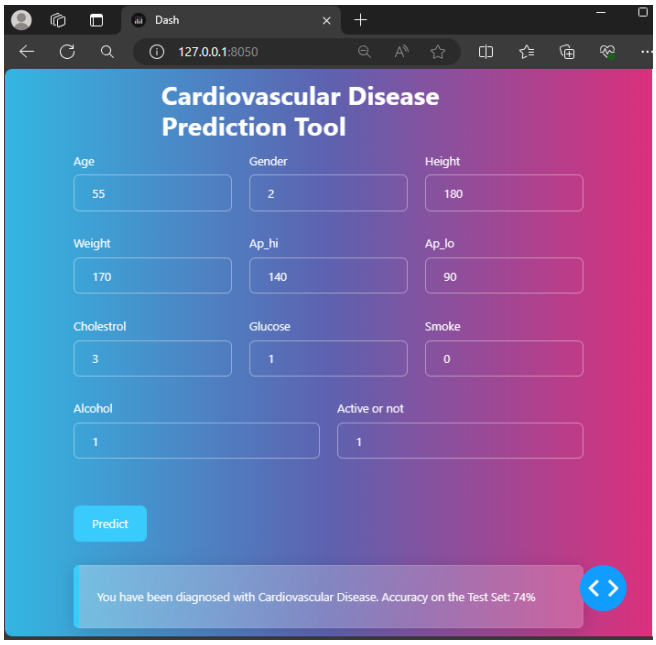

Fig. 16. Prediction Tool – Positive Diagnosis

Negative Diagnosis: Conversely, in Fig. 14 a healthier profile, such as a 50-year-old with normal blood pressure (110/70) and low cholesterol level (1), results in a negative diagnosis for cardiovascular disease. The tool reports an accuracy of 74% in this case as well, illustrating its ability to reassure individuals with lower risk factors.

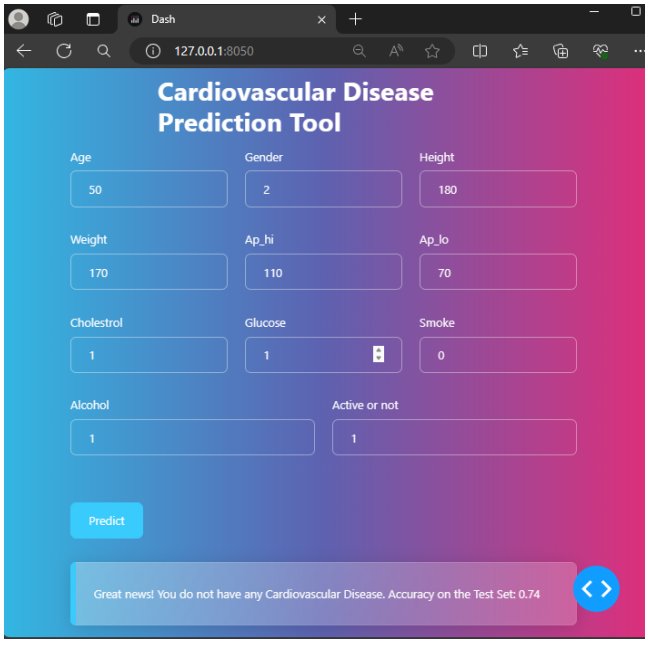

Fig. 17. Prediction Tool – Negative Diagnosis

Implications and Future Directions:

This tool not only facilitates immediate and easy-to-understand assessments of cardiovascular risk but also enhances patient engagement with their health management. By providing real-time feedback, the tool empowers individuals to understand their risk levels and seek appropriate medical advice and interventions. As we progress beyond the preliminary phase, the insights gathered from the deployment of this tool will inform further refinements to our models and strategies, aiming to improve accuracy and usability. The ongoing collection of user data and feedback will help fine-tune the tool's predictive capabilities, ensuring it remains a valuable resource in the fight against cardiovascular disease.

The preliminary results of our study have laid a solid foundation for understanding and predicting cardiovascular diseases, beginning with rigorous data cleaning and followed by detailed statistical analyses which highlighted significant correlations among key health indicators such as age, weight, cholesterol, and blood pressure. Our evaluation of various predictive models identified the Gradient Boosting Algorithm as particularly effective, demonstrated through the deployment of a user-friendly cardiovascular disease prediction tool. This tool allows individuals to input personal health data and receive real-time assessments of their cardiovascular risk, with an accuracy of 74%. This integration of analytical rigor with practical application underscores the potential of our approach to significantly enhance cardiovascular health management and prevention strategies.

VII. CONCLUSION

As we conclude our study on the predictive modeling of cardiovascular disease (CVD) using advanced machine learning techniques, we reflect on the substantial contributions and future implications of our findings. Our research demonstrates a significant stride towards enhancing the diagnostic accuracy and management of cardiovascular health. By employing a dataset of 70,000 patient records, we successfully applied and evaluated various predictive models, with the Gradient Boosting Algorithm standing out due to its superior accuracy and robustness, achieving 74% accuracy and an AUC of 81%. This research not only validates the effectiveness of machine learning in medical diagnostics but also emphasizes the critical role of data quality and detailed statistical analysis in developing reliable predictive tools. The integration of our Cardiovascular Disease Prediction Tool into clinical practice could revolutionize how healthcare providers assess and manage cardiovascular risks, offering a more personalized approach to patient care. Looking forward, the potential to expand and refine these tools is vast. Continuous advancements in machine learning and data analytics, coupled with an increasing availability of health data, may allow for even more nuanced insights into cardiovascular health. Ultimately, our work contributes to the broader goal of reducing the global burden of cardiovascular diseases, supporting early detection, and promoting preventive healthcare measures. This project not only showcases the power of predictive analytics in healthcare but also sets a foundation for future innovations that could further transform the landscape of cardiovascular disease management.

VIII. PROJECT LINKS

Dataset Link:
https://huggingface.co/datasets/AlexCambell/HeartFailureDataset
Project Website: https://mason.gmu.edu/~aanumall/

IX. PROJECT TIMELINE

We have successfully completed all phases of the project, culminating in a comprehensive study and development of a predictive tool for cardiovascular disease, which was finalized and presented as per the project schedule.

TASK NAME	TENTATIVE TIME	STATUS
Project Initiation Phase	Week 1	Done
Data Collection	Week 2-3	Done
Project Proposal	Week 4	Done
Data Cleaning & Preparation Phase	Week 5-6	Done
Project Milestone 1	Week 7-8	Done
Visualizations	Week 9	Done
Model Development Phase	Week 10	Done
Project Milestone 2	Week 11	Done
Model Evaluation and Interpretation Phase	Week 12	Done
Heart Disease Prediction Tool Development	Week 13	Done
Reporting and Presentation Phase	Week 14	Done
Final Report	Week 15	Done
Final Project Submission	Week 16	Done

Table. 2: Project Timeline

X. REFERENCES

[1] S. S. Salma Banu, Prediction of heart disease at early stage using data mining and big data analytics: A survey, IEEE, 2017. https://ieeexplore.ieee.org/document/7955226

[2] R. G. Priyadarshini, Analysis of heart disease using statistical techniques, IOPScience, 2021. https://iopscience.iop.org/article/10.1088/1742-6596/1770/1/012105/meta

[3] A. P. J. Fabio Mendoza, Cardiovascular Disease Analysis Using Supervised and, JSW-Journal of Software, 2016. https://www.jsoftware.us/index.php?m=content&c=index&a=show&catid=178&id=2727

[4] Krittanawong, Machine learning prediction in cardiovascular diseases: a meta-analysis, Scientific Reports, 2020. https://doi.org/10.1038/s41598-020-72685-1

[5] D. Jones, Special Report on CVD by AHAAC, Science Direct, 2018. https://www.sciencedirect.com/science/article/pii/S0735109718390363?via%3Dihub

[6] Deepa, Dr. R., Sadu, V. B., C, P. G., & Sivasamy, Dr. A. (2022, March 27). *Early prediction of cardiovascular disease using machine learning: Unveiling risk factors from Health Records*. AIP Publishing. https://pubs.aip.org/aip/adv/article/14/3/035049/3279524/Early-prediction-of-cardiovascular-disease-using

[7] Ogunpola, A., Saeed, F., Basurra, S., Albarrak, A. M., & Qasem, S. N. (2024, January 8). *Machine learning-based predictive models for detection of cardiovascular diseases*. Diagnostics (Basel, Switzerland). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10813849

[8] (PDF) cardiovascular disease prediction combination using machine and deep learning model. (n.d.-b). https://www.researchgate.net/publication/378481511_Cardiovascular_Disease_Prediction_combination_Using_Machine_and_Deep_Learning_Model

[9] Pal, M., Parija, S., Panda, G., Dhama, K., & Mohapatra, R. K. (2022, June 17). *Risk prediction of cardiovascular disease using Machine Learning Classifiers*. Open medicine (Warsaw, Poland). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9206502

[10] Dixon, D., Sattar, H., Moros, N., Kesireddy, S. R., Ahsan, H., Lakkimsetti, M., Fatima, M., Doshi, D., Sadhu, K., Hassan, M. J., Dixon, D., Moros, N., Kesireddy, S. R., A., H., Lakkimsetti, M., F., M., Doshi, D., & Hassan, M. J. (2022, May 9). *Unveiling the influence of AI predictive analytics on patient outcomes: A comprehensive narrative review*. Cureus. https://www.cureus.com/articles/247197-unveiling-the-influence-of-ai-predictive-analytics-on-patient-outcomes-a-comprehensive-narrative-review#!/

[11] JACC journals on precision health analytics . (n.d.-a). https://www.jacc.org/doi/10.1016/j.jacc.2023.06.043

[12] (PDF) predictive analysis of heart diseases with machine learning approaches. (n.d.-c). https://www.researchgate.net/publication/359730073_PREDICTIVE_ANALYSIS_OF_HEART_DISEASES_WITH_MACHINE_LEARNING_APPROACHES

[13] Shrimali, S. (2021, April 11). *Unveiling insights: Exploring cardiovascular health through data analytics*. LinkedIn. https://www.linkedin.com/pulse/unveiling-insights-exploring-cardiovascular-health-through-shrimali-hg8uf

[14] (PDF) cardiovascular disease prediction combination using machine and deep learning model. (n.d.). https://www.researchgate.net/publication/378481511_Cardiovascular_Disease_Prediction_combination_Using_Machine_and_Deep_Learning_Model

[15] Alex, Heart Failure Dataset, Huggingface. https://huggingface.co/datasets/AlexCambell/HeartFailureDataset

[16] *Understanding feature importance in machine learning*. Built In. (n.d.). https://builtin.com/data-science/feature-importance#

[17] *Dash app by plotly*. Plotly. (n.d.). https://dash.plotly.com/tutorial

[18] HealthySimulation. (n.d.). *Behind the scenes with data: The Unsung Hero of Healthcare Simulation*. HealthySimulation.com LEARN. https://learn.healthysimulation.com/course/how-to-use-healthcare-simulation-data

[19] Keylabs. (2023, January 26). *Improving your AI model's accuracy: Expert tips*. https://keylabs.ai/blog/improving-your-ai-models-accuracy-expert-tips/amp/

[20] *Machine learning - scaling*. Python Machine Learning Scaling. (n.d.). https://www.w3schools.com/python/python_ml_scale.asp

[21] Narkhede, S. (2021, June 15). *Understanding AUC - roc curve*. Medium. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5?gi=089dfddab5fc

[22] Takyar, A. (2023, March 27). *All about machine learning techniques*. LeewayHertz. https://www.leewayhertz.com/machine-learning-techniques/

[23] *Confusion matrix: How to use it & interpret results [examples]*. V7 labs. (n.d.). https://www.v7labs.com/blog/confusion-matrix-guide