|  |
| --- |
| TECHNICAL REPORT |

|  |
| --- |
| **DATA SCIENCE** |

|  |  |
| --- | --- |
| FALL 23 |  |



Contents

Heart Attack Risk 🩺 Predictive Analysis [2](#_Toc96341549)

[Executive Summary 2](#_Toc96341550)

[Highlights of Project 3](file:///C:\Users\ardia\Desktop\TEACHING\SPRING%2022\DSCI6002%20Data%20Science\Projects\Technical%20report%20Template.docx#_Toc96341552)

Abstract…………………………………………………..4

[Cover Page 4](#_Toc96341554)

Introduction……………………………………………5

Review…………………………………………………….6

Methodology…………………………………………..7

[Results…………………………………………………….](#_Toc96341556)9

[Discussion…………………………………………….](#_Toc96341556)19

[Conclusions………………………………………….](#_Toc96341556)21

Contributions/References[…………………….](#_Toc96341556)22

|  |
| --- |
| Heart Attack Risk 🩺Predictive Analysis |

|  |  |  |
| --- | --- | --- |
| Executive Summary Heart attack prediction is one of the serious causes of morbidity in the world’s population. The clinical data analysis includes a very crucial disease i.e., cardiovascular. Data Science and machine learning (ML) can be very helpful in the prediction of heart attacks in which different risk factors like high blood pressure, high cholesterol, abnormal pulse rate, diabetes, etc... can be considered. | | |
| person at a table writing in a notebook with people around | | |
| **Team Members:**  **Mohammad Abdul Imran**  **​(Data Engineer)**  **Venkata Raja Varaprasad Sanayila**  **(ML Engineer)**  **Rachana Chinthanippu**  **​(Data Analyst)** | **Questions?**  Contact :vsana4@unh.newhaven.edu |  |

|  |
| --- |
| **Heart Attack Risk 🩺 Predictive Analysis** |

|  |  |
| --- | --- |
| Submitted on 12/06/2023 Highlights of Project - Introduction of an AI-driven heart disease detection system using advanced machine learning algorithms.  - Core objective: Demonstrating the efficacy of machine learning in forecasting the likelihood of developing heart disease.  - Development involves creating a Python-based application for healthcare research with reliability and versatility.  - Phases of application development: Database collection, logistic regression implementation, and comprehensive evaluation of dataset attributes.  - Construction of a robust random forest classifier algorithm for identifying heart diseases with 83% accuracy over training data.  - In-depth exploration of the random forest classifier algorithm, detailing experiments and results for superior accuracies.  - Focus on predicting presence or absence of heart diseases, including arrhythmia and locomotor disorders, using machine learning.  - Methodology involves implementing the random forest algorithm in Python for personalized patient diagnosis and care.  - ML model demonstrates an 83% accuracy rate over training data and approximately 70% accuracy rate on test data.  - Random forest classifier highlighted for its flexibility, accuracy, and applicability to regression and classification tasks.  - Ensemble Health Data Prediction System (EHDPS) utilizes medical parameters (age, sex, blood pressure, cholesterol, obesity) for predicting heart diseases.  - EHDPS provides substantial information, health correlations, and trends related to heart diseases based on predictive analysis. Submitted on:12/06/2023 |  |

## 

## Abstract

This project introduces an artificial intelligence-driven heart disease detection system employing advanced machine learning algorithms. The core objective is to demonstrate the efficacy of machine learning in forecasting the likelihood of an individual developing heart disease. The development involves creating a Python-based application tailored for healthcare research, ensuring reliability and versatility for diverse health monitoring applications. The project encompasses crucial data processing steps, including the handling of categorical variables and the conversion of categorical columns.

The application's development unfolds in distinct phases: database collection, logistic regression implementation, and a comprehensive evaluation of dataset attributes. A pivotal aspect involves the construction of a robust random forest classifier algorithm, designed to identify heart diseases with heightened accuracy. Notably, data analysis showcases the application's significance, achieving an impressive 83% accuracy rate over training data.

The paper delves into the intricacies of the random forest classifier algorithm, detailing the experiments conducted and the corresponding results, which demonstrate superior accuracies for research diagnoses. The study concludes by outlining the project's objectives, acknowledging its limitations, and highlighting valuable research contributions in the domain of AI-driven healthcare applications.

Cover Page

A white circle with text on it

Description automatically generated

Introduction

Heart disease stands as the foremost global cause of mortality, claiming a staggering 17.9 million lives annually, according to data from the World Health Organization (WHO) [1]. Lifestyle choices, encompassing unhealthy dietary habits, sedentary lifestyles, smoking, and alcohol consumption, contribute significantly to the behavioral risk factors associated with cardiovascular disease and stroke [1]. The intricate nature of heart-related afflictions, such as heart attacks triggered by arterial plaque buildup and strokes induced by thrombus formation, further underscores the urgency of precise prediction and timely identification [2]. Compounded by symptoms that often overlap with those of other ailments and the potential for misinterpretation as indicators of aging, the diagnosis of heart conditions remains a formidable challenge for practitioners.

Recognizing the pivotal role of data-driven healthcare advancements, this project aims to leverage machine learning (ML) techniques for the precise prediction of heart disease, thereby enhancing patient survival rates. The surge in medical data collection provides practitioners with unprecedented opportunities to revolutionize healthcare diagnostics. ML, with its versatile applications in diverse domains like text detection, early prediction, power quality disturbance detection, and agriculture, has emerged as a cornerstone in the healthcare sector [3–7].

Currently, conventional methods for predicting and diagnosing cardiac diseases rely heavily on practitioners' subjective evaluations, encompassing patient medical histories, observable signs, and physical assessment reports. However, with the exponential growth of accessible patient data stored in healthcare databases, there exists a burgeoning need for advanced tools to harness this wealth of information effectively. In this context, the article employs the Cleveland Heart Disease dataset from the UCI Machine Learning repository to develop a robust predictive model. Through supervised machine learning, the model is trained to discern patterns inherent in the dataset features, with a specific focus on classification—a potent approach for disease prediction when appropriately trained with substantial data [8].

The primary objective of this work is to employ contemporary ML techniques to construct an efficient predictive model for heart disease in the realm of healthcare. The Cleveland Heart Disease dataset undergoes scrutiny with various ML algorithms, including Support Vector Machines with Radial Basis Function (RBF) kernel, Gaussian Naive Bayes, logistic regression, LightGBM, XGBoost, and the random forest algorithm. The outcome aims to identify the most effective predictive model for the early diagnosis of heart disease.

Review

The endeavor to detect heart diseases using Python underscores a pivotal intersection of technology and healthcare. The dataset at the core of this project encapsulates crucial factors such as Chol, treetops, sex, age, and more, each serving as a unique input for the detection model. The utilization of key Python libraries, including but not limited to matplotlib, Numpy, Pandas, and warnings, showcases the comprehensive technological arsenal harnessed for this initiative. The amalgamation of diverse tools, such as the correlation matrix, histogram analysis, and classification algorithms like Support Vector Classifier, K Neighbors Classifier, Random Forest Classifier, and Decision Tree Classifier, underscores the multidimensional approach employed to extract meaningful outcomes from the dataset, all within the Python programming language.

Python's role in healthcare extends beyond its prowess as a programming language; it symbolizes an open-source gateway to innovation in healthcare solutions. Its versatility facilitates the development of cutting-edge applications that not only enhance patient outcomes but also streamline care delivery processes. Importantly, Python aligns with the Health Insurance Portability and Accountability Act (HIPAA) checklist, ensuring the security and confidentiality of medical information, a critical aspect in the healthcare data landscape.

Delving into the crux of heart disease, a multifaceted spectrum of causes emerges, ranging from diabetes and obesity to an unhealthy diet, excessive alcohol use, and physical inactivity. The interconnectedness of these factors forms the backdrop against which the detection model operates, seeking patterns indicative of potential heart diseases. Atherosclerosis, characterized by the hardening of arteries, is a significant component, often associated with arrhythmia. The symptoms manifest during a heart attack, encompassing pain radiating to the arm, dizziness, light-headedness, throat constriction, snoring, and profuse sweating. Notably, these symptoms serve as crucial inputs for the detection model, aligning with the diverse manifestations of heart-related ailments.

As the prevalence of heart-related issues tends to increase with age, the project assumes added significance in addressing the healthcare needs of the aging population. Heart attacks, strokes, coronary heart disease, heart failure, and coronary artery disease collectively highlight the pervasive impact of cardiovascular issues, with a notable emphasis on their prevalence in individuals over the age of 65. This literature review underscores the project's contextual relevance, integrating technological innovation with a nuanced understanding of the complex landscape of heart diseases and their manifestations.

On the other hand, if the arguments are more nuanced with caveats aplenty, then you must cite the relevant research to offer adequate context before you embark on your analysis. You might use the literature review to highlight gaps in the existing knowledge, which your analysis will try to fill. This is where you formally introduce your research questions and hypothesis.

## Methodology

**Overview of Methodology:**

The primary focus of this project is the application of Machine Learning (ML) to predict the presence or absence of heart diseases, including arrhythmia, locomotor disorders, and other related conditions. The chosen methodology centers around the implementation of the random forest algorithm using Python, aiming to provide valuable insights to physicians for personalized patient diagnosis and care. The algorithm stands out for its flexibility, accuracy, and applicability to both regression and classification tasks.

The ML model employed in this project has demonstrated a noteworthy 83% accuracy rate over the training data, a crucial metric that validates its effectiveness in capturing underlying patterns [25]. This accuracy is rigorously assessed using a confusion matrix, offering a comprehensive view of the model's performance. The model exhibits an approximate 70% accuracy rate when evaluated against test data, showcasing its reliability in real-world scenarios.

**Research Method:**

The initial steps involve importing essential libraries such as NumPy for array operations, Pandas for data manipulation, Matplotlib for data visualization, and scikit-learn for machine learning functionalities. The dataset is then loaded and explored using Pandas dataframe.info() and describe() functions, offering a concise statistical overview. Correlation matrices aid in understanding data relationships, and visualizations are created using pyplot to enhance interpretability.

**Use of Algorithm with Justification:**

The project extensively relies on the random forest algorithm due to its inherent flexibility, ease of use, and proven excellence in various machine learning scenarios. Even without hyperparameter tuning, the algorithm consistently delivers outstanding results. As a supervised learning algorithm, random forests are adept at tasks such as categorizing loan applicants, detecting fraudulent behavior, and predicting diseases. The ensemble learning principle forms the algorithm's foundation, utilizing multiple decision trees to enhance predictive accuracy.

The working process involves selecting random subsets of the training data, building decision trees for each subset, and combining the results to make predictions. This approach leverages the diversity of decision trees to create a robust and accurate model.

**Data Analysis:**

Data analysis encompasses the treatment of categorical variables, involving the creation of dummy columns to represent binary values. The gender column, for instance, is transformed into two columns (1 for males, 0 for females). The random forest classifier utilizes decision trees, each constructed from a random selection of features, contributing to the model's resilience against overfitting.

**Data Output:**

The random forest classifier emerges as the preferred choice among classifiers, offering superior accuracy in predicting heart diseases. Medical parameters such as age, sex, blood pressure, cholesterol, and obesity play pivotal roles in the prediction model. The Ensemble Health Data Prediction System (EHDPS) provides predictions based on these parameters, unraveling substantial information, health correlations, and trends related to heart diseases.

**Use of Required Software with Justification:**

The project necessitates the use of Python, specifically leveraging the Jupyter Notebook editor. Jupyter Notebook seamlessly integrates code, numerical performance, explanatory text, and multimedia tools into a single document, fostering collaboration and knowledge sharing. The choice of Python aligns with its status as a versatile programming language, supported by a rich ecosystem of libraries including NumPy, Pandas, and scikit-learn.

While Jupyter Notebook stands as a preferred choice for its interactivity and real-time result display, the project foregoes Anaconda in favor of Python's package manager, pip. This decision is motivated by the project's focus on utilizing specific libraries and maintaining a streamlined environment.

In summary, the methodology employs a systematic approach, utilizing the random forest algorithm, comprehensive data analysis, and Python-based tools to predict heart diseases effectively. The choice of software aligns with the project's objectives, ensuring a seamless integration of code, analysis, and visualization.

## Results

### Introduction

Python stands out as a highly effective and versatile programming language, particularly renowned for its utility. Within the realm of this machine learning endeavor, Python harnesses a plethora of libraries integral to the project's success. This methodology encapsulates a facet of the broader Artificial Intelligence model, specifically within the domain of Machine Learning. The predictive capabilities are realized through the utilization of Python libraries, with SKLEARN emerging as a prominent tool in the arsenal of machine learning predictions.

### Critical analysis regarding the description of heart disease detection

The healthcare sector continually generates extensive datasets, commonly referred to as big data, with heart disease ranking as a leading global cause of mortality, presenting a substantial public health challenge. Early detection of heart disease poses a significant dilemma in medical science. Researchers, such as Ramalingam et al. [35], focus on analyzing characteristics like RR interval, QT interval, and QRS interval for effective heart disease detection, aiming to ascertain the patient's overall health status.

Contrastingly, Subhadra and Vikas [36] highlight heart disease as a pervasive and fatal affliction, prevalent in both developed and developing nations, progressively leading to fatalities. Their proposed system addresses the challenge by classifying large and intricate medical datasets, utilizing a map-reduce algorithm to both identify heart disease and streamline dataset size.

The ensuing output manifests as a histogram generated through the code df.hist(figsize=(10,10)), a visual representation designed to interpret discrete data and provide a succinct summary. This histogram aids in visually assessing data points within specific value ranges, facilitating a focused understanding of crucial numerical data points. This aligns with Exploratory Data Analysis (EDA), a process that uncovers relevant data, detects errors, validates assumptions, and discerns correlations among explanatory variables. Ambale-Venkatesh et al. [37] assert the significance of considering risk factors using the random forest algorithm for heart disease prediction, utilizing publicly accessible data encompassing 304 records with 14 attributes such as age and gender.

Dogan et al. [38] delve into classification performances, pre-processing methods, and evaluation metrics, demonstrating that visualized data results in accurate predictions, with an 83 percent precision rate. They introduce an Artificial Neural Network (ANN)-based three-stage method for heart disease prediction, emphasizing the efficacy of deep neural networks in achieving favorable outcomes, as depicted in Figure.

A graph of different types of land

Description automatically generated

Fig. Bar graph comparison of Hemispheres

A graph of different colored lines

Description automatically generated

Fig. Bar graph comparison over the Countries

A graph of different colored bars

Description automatically generatedA graph of different colored bars

Description automatically generatedA graph of a number of people

Description automatically generated

Fig. Bar graph comparison over the sex genders

A diagram of a network

Description automatically generated

Fig. Heart disease detection using a [deep learning](https://www.sciencedirect.com/topics/mathematics/deep-learning) approach.

### Data interpretation of the selected dataset

Data interpretation involves the meticulous review of data through predefined processes, serving the purpose of assigning significance to the data and drawing inferences on relationships. According to Mahdavinejad et al. [39], data interpretation is integral to data analysis, collection, and presentation, presenting outcomes in a structured format. In the realm of machine learning, interpretation refers to the process of comprehending the predictions made by a machine learning model. Python libraries play a crucial role in constructing the data interpretation aspect of machine learning, aiding in understanding the classifier's choices and interpreting models based on predictions and parameters.

For heart disease detection in machine learning, this interpretative process involves two pivotal phases: rigorous monitoring of evaluation metrics and experimentation with algorithmic selections to enhance and fortify the overall approach. It becomes imperative to interpret models by scrutinizing parameters and predictions, unraveling the rationale behind the classification of a specific class. Tauzin et al. [40] emphasize that data interpretation has evolved as an integral part of modern data analysis, driven by cost-effective technologies, particularly within the healthcare sector dealing with emergencies and disease outbreaks.

In the context of "machine learning," data interpretation encompasses a sophisticated form of data processing, automating the construction of analytical models. This data analysis, rooted in artificial intelligence, operates under the premise that computers can learn, recognize patterns in data, and make decisions with minimal human intervention.A diagram of a model

Description automatically generated

Fig. Data interpretation.

### Data interpretation strategies using python language in detecting heart problem

### Stats models in Python serve as a valuable tool for users to delve into data, conduct statistical tests, and estimate statistical models. It has become the preferred language for data analysts, relying on crucial data analysis libraries, data pre-processing, and exploratory data analysis, as highlighted by Peters et al. [41]. These data analysis libraries, comprising open-source packages, are widely embraced for efficient data manipulation.

### In the realm of fundamental scientific computing, Numpy emerges as a numeric library in Python, proficient in tasks such as linear algebra, Fourier transforms, and random number generation. Complementing this, SciPy stands out as a scientific library with a high-level science module, contributing to advanced scientific computations.

### The interpretation of machine learning approaches involves a set of criteria and exploration techniques, contingent on the specific context. This study delves deep into understanding existing model interpretation approaches, scrutinizing their drawbacks and challenges. It also navigates the longstanding trade-off between model accuracy and interpretability. Lastly, it examines some of the most widely embraced model interpretation strategies.

### Data manipulation and data visualization

### Pandas plays a pivotal role in both data analysis and machine learning, primarily functioning through data frames. It facilitates data handling by offering support for various file formats like CSV, Excel, plain text, JSON, SQL, and more, as elucidated by Peters et al. [41]. Data manipulation, involving transformation, formatting, and structuring, is a crucial aspect addressed by the Pandas library.

### For plotting and visualizing data, the Matplotlib library proves instrumental. Capable of creating diverse visualizations, including graphs, heatmaps, line plots, histograms, and more, Matplotlib is seamlessly integrated into GUI toolkits, as depicted in Figure

### The bar graph in the counterplot signifies the dataset's target variable, indicating the gender of patients with a higher incidence of heart disease. Here, the code snippet sns.set\_style('whitegrid') is utilized.

### In the realm of machine learning, Scikit Learn stands out as a comprehensive and free library built upon NumPy, SciPy, and Matplotlib, offering efficient components for statistical model development. Carleo et al. [42] highlight its capabilities in executing various classification, regression, and clustering algorithms, seamlessly integrating with Pandas through data frame operations.

### The process initiates by importing essential libraries like NumPy and Pandas, followed by dataset loading. The commonly used CSV format is employed for presenting machine learning data. The CSV file, when headerless, automatically assigns names, or else, manual labeling of attributes is carried out for each dataset column.

### Exploratory data analysis

### Exploratory Data Analysis (EDA) serves as a valuable tool for gaining deeper insights into data and unraveling its patterns. Resembling a form of storytelling for statisticians, EDA employs visual methods to uncover trends and observations within the data. Often positioned as the initial phase in the data modeling journey, EDA diligently explores the dataset.

### This comprehensive process involves various steps, including the handling of missing values, treating outliers, encoding categorical variables, normalization, and, ultimately, scaling and eliminating duplicates. EDA plays a crucial role in laying the groundwork for understanding data intricacies and informing subsequent analytical endeavors.

### Evaluation of data

### Data evaluation introduces the utilization of Python, a mathematical data processing language that leverages Pandas Data Frame objects for data storage. The tasks involved in data analysis encompass importing, cleansing, and transforming data to ensure its readiness for review.

### Data cleaning

Data cleansing involves the preparation of data for analysis by cleaning raw data, enabling data visualization, and facilitating data prediction. It is the process of eliminating inaccurate, corrupted, improperly formatted, and redundant data from a dataset to ensure accuracy, integrity, and completeness.

**Correlation Matrix Plot**

According to Harper et al. [[43]](https://www.sciencedirect.com/science/article/pii/S2772442522000016#b43), a correlation Matrix Plot is a [covariance matrix](https://www.sciencedirect.com/topics/mathematics/covariance-matrix) that is a metric called the correlation that defines the strength of the linear association. The Correlation matrix sums up the power and direction of a [linear relationship](https://www.sciencedirect.com/topics/mathematics/linear-relationship) between two variables, and it allows values between −1 as well as +1. The feature of the correlation matrix displays the correlation between the coefficients. A particular random variable is considered correlated with each of its other values. This represents an excellent way to check correlations among features by visualizing the correlation matrix as a heat map.

The relationship between all the attribute columns is expressed as a graph. The linear relationship between two continuous variables is defined using the dataset’s correlation.

A screenshot of a graph

Description automatically generated

Fig. Correlation of each feature in the dataset using the heatmap.

## If there's a requirement to employ statistical models, you may consider utilizing regression models or categorical analysis. Additionally, results from alternative empirical techniques falling within the broader category of data mining can be reported. It's worth mentioning that numerous reports in the business sector often opt for a more accessible presentation style by withholding intricate statistical details and relying on illustrative graphics for summarizing the results.

## Discussion

### Random forest classifier

Random forest classification is considered an ensemble learning approach used for solving machine learning challenges such as classification and regression. This random forest classification of heart disease detection algorithm works by constructing multiple decision trees. Apart from this, this classifier uses a technique called “bootstrap aggregation”.

The likelihood of a reduction in node impurity is weighted, and it can be improved by hitting the node to determine feature significance. With only two child nodes, Scikit-learn calculates a node’s importance using Gini Importance. Between Eqs. [(2)](https://www.sciencedirect.com/science/article/pii/S2772442522000016#fd2), [(6)](https://www.sciencedirect.com/science/article/pii/S2772442522000016#fd6), we show key formulas for random forest classifiers.(2) 

ni sub (j) = node j significance

W sub (j) = weighted number of samples hitting node j

C sub (j) = node j impurity value j

left (j) = child node from left split on node j

right (j) = child node from right split on node j

sub () is used because subscript is not available in medium.

After that, the value of each function on a decision tree is determined as follows:



normfi sub i = the feature’s normalized importance i

fi sub i = the significance of i

By dividing by the number of all function importance values, these can be normalized to a value between 0 and 1.



At the Random Forest stage, the final feature importance is the average of all the trees. The total number of trees is divided by the amount of the feature’s significance value on each tree.



RFfi sub (i) = the importance of feature i calculated from all trees in the Random Forest model

normfi sub(ij) = the normalized feature importance for i in tree j

T = total number of trees

As Huljanah et al. [[56]](https://www.sciencedirect.com/science/article/pii/S2772442522000016#b56) stated, the classification has produced the class that the majority of the decision trees predicted in the forest. On the other hand, For regression outputs, the class with mean predictions of the individual trees. Based on the given scenario, trees and entropy criteria are used for developing the classifier. The entropy criterion is represented by



A “Random forest classifier”, according to Mehrang et al. [[57]](https://www.sciencedirect.com/science/article/pii/S2772442522000016#b57), shows a collection of decision trees built from randomly selected subsets of the training set. The [sepsis](https://www.sciencedirect.com/topics/nursing-and-health-professions/sepsis) data is then aggregated from various decision trees, determining the final class of the test object. The Random Forest Prediction depicts the model’s pictorial representation as well as its precision for the dataset’s data collection. Using the Random Forest, the accuracy of the heart disease detection system is 0.821875.

Furthermore, by inputting the XTrain and YTrain parameters required for the method of fit, the accuracy value of the “Random Forest Score model” is obtained. Then, to determine the score of this model, use the parameters XTest and YTest in the method score() to determine the Random Forest Score model’s score.

The results section is followed by the **discussion section**, where you craft your main arguments by building on the results you have presented earlier.

The "discussion section" is where you rely on the power of narrative to enable numbers to communicate your thesis to your readers. You refer the reader to the research question and the knowledge gaps you identified earlier. You highlight how your findings provide the ultimate missing piece to the puzzle.

Of course, not all analytics return a smoking gun. At times, more frequently than I would like to acknowledge, the results provide only a partial answer to the question and that, too, with a long list of caveats.

## Conclusion

In the voyage through our project, we embarked on a journey to predict heart disease using the versatile Python language. Python, with its object-oriented structure and high-level programming prowess, emerged as our guiding compass, offering quick development cycles and dynamic building options. This language proved instrumental in accurately navigating the intricate pathways of heart diseases, fostering a transformative impact on the heart care industry.

Our exploration began with the heartbeat of healthcare—data. The heart care industry, harnessing information from diverse facilities and patient interactions, strategically employed Python to craft a superior predictive model. Doctors, wielding this innovative model, showcased its prowess in treatments, promising improvements in the entire healthcare delivery system. This predictive model found its niche in heart diseases, resonating with clinicians and institutions, promising better outcomes through scalable and dynamic applications.

Chapter two unfolded the art of detecting the presence of heart diseases, relying on a rich dataset featuring crucial patient data such as age, sex, chol, and more. The application, a symphony of individually imported libraries—Matplotlib, Numpy, Pandas, and others—leveraged Python's robust capabilities to extract valuable insights from patient information. Compliant with HIPAA regulations, the application prioritized the safety of medical information.

Chapter three delved into the intricate techniques of detecting heart disease using machine learning, where the significance of predicting threats took center stage. The project gracefully followed the rhythmic beats of the random forest algorithm, a powerful force in the development of heart disease detection methodologies using Python. With an impressive accuracy rate of approximately 83% over training data, this algorithm emerged as a maestro orchestrating the methodology to detect heart diseases with precision.

A model regarding machine learning played a pivotal role in crafting accuracy through the symphony of training data. The careful selection of the random forest classification, based on the exact dataset and decision tree nuances, added depth to the narrative. The spotlight then shifted to data analysis, deftly working with categorical variables and transforming them into dummy columns, breathing life into 1s and 0s.

This section not only unveiled predictions but illuminated the data output, a tapestry of medical parameters—age, gender, blood pressure, cholesterol, and obesity—orchestrating a symphony of heart disease prediction. Python libraries, particularly SKLEARN, stood as prerequisites for this predictive journey within the realm of Artificial Intelligence. The Random Forest algorithm emerged as the crescendo, simplifying the prediction of heart diseases with a precision that surpasses the capabilities of the Decision Tree.

In this culmination, we gracefully conclude our project, having navigated the seas of Python, machine learning, and predictive precision. Our journey leaves an indelible mark on the canvas of healthcare innovation, paving the way for a future where data science and technology harmoniously unite for the betterment of patient outcomes.

## Contributions/References

1. [[1]](https://www.sciencedirect.com/science/article/pii/S2772442522000016#bb1)

Loku L., Fetaji B., Krstev A., Fetaji M., Zdravev Z.

Using python programming for assessing and solving health management issues

South East Eur. J. Sustain. Dev., 4 (1) (2020)

[Google Scholar](https://scholar.google.com/scholar_lookup?title=Using%20python%20programming%20for%20assessing%20and%20solving%20health%20management%20issues&publication_year=2020&author=L.%20Loku&author=B.%20Fetaji&author=A.%20Krstev&author=M.%20Fetaji&author=Z.%20Zdravev)

1. [[2]](https://www.sciencedirect.com/science/article/pii/S2772442522000016#bb2)

Mathur P.

Overview of machine learning in healthcare

Machine Learning Applications using Python, A Press, Berkeley, CA (2019), pp. 1-11

[View article](javascript:;)

[CrossRef](https://doi.org/10.1007/978-1-4842-3787-8_1)[Google Scholar](https://scholar.google.com/scholar_lookup?title=Overview%20of%20machine%20learning%20in%20healthcare&publication_year=2019&author=P.%20Mathur)

1. [[3]](https://www.sciencedirect.com/science/article/pii/S2772442522000016#bb3)

Guleria P., Sood M.

Intelligent learning analytics in healthcare sector using machine learning

Machine Learning with Health Care Perspective, Springer, Cham (2020), pp. 39-55

[View article](javascript:;)

[CrossRef](https://doi.org/10.1007/978-3-030-40850-3_3)[Google Scholar](https://scholar.google.com/scholar_lookup?title=Intelligent%20learning%20analytics%20in%20healthcare%20sector%20using%20machine%20learning&publication_year=2020&author=P.%20Guleria&author=M.%20Sood)

1. [[4]](https://www.sciencedirect.com/science/article/pii/S2772442522000016#bb4)

Kumar V.V.

Healthcare Analytics Made Simple: Techniques in Healthcare Computing using Machine Learning and Python

Packt Publishing Ltd. (2018)

Available at: [https://books.google.com/books?hl=en&lr=&id=nwZnDwAAQBAJ&oi=fnd&pg=PP1&dq=application+of+python+programming+language+in+health+care+sectors&ots=BJz0Qe\_09q&sig=Bhv\_iaxOuZoKJu-WVEKPDr6B9y0](https://books.google.com/books?hl=en%26lr=%26id=nwZnDwAAQBAJ%26oi=fnd%26pg=PP1%26dq=application%2Bof%2Bpython%2Bprogramming%2Blanguage%2Bin%2Bhealth%2Bcare%2Bsectors%26ots=BJz0Qe_09q%26sig=Bhv_iaxOuZoKJu-WVEKPDr6B9y0), [Accessed on 5th March, 2021]

[Google Scholar](https://scholar.google.com/scholar_lookup?title=Healthcare%20Analytics%20Made%20Simple%3A%20Techniques%20in%20Healthcare%20Computing%20using%20Machine%20Learning%20and%20Python&publication_year=2018&author=V.V.%20Kumar)

1. [[5]](https://www.sciencedirect.com/science/article/pii/S2772442522000016#bb5)

BelItSoft

Python in healthcare

BelItSoft (2017)

Available at: <https://belitsoft.com/custom-application-development-services/healthcare-software-development/python-healthcare>, [Accessed on 5th March, 2021]

[Google Scholar](https://scholar.google.com/scholar_lookup?title=Python%20in%20healthcare&publication_year=2017&author=BelItSoft)

1. [[6]](https://www.sciencedirect.com/science/article/pii/S2772442522000016#bb6)

Nithya B., Ilango V.

Predictive analytics in health care using machine learning tools and techniques

2017 International Conference on Intelligent Computing and Control Systems (ICICCS), IEEE (2017), pp. 492-499

[View in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-85047615879&partnerID=10&rel=R3.0.0)[Google Scholar](https://scholar.google.com/scholar_lookup?title=Predictive%20analytics%20in%20health%20care%20using%20machine%20learning%20tools%20and%20techniques&publication_year=2017&author=B.%20Nithya&author=V.%20Ilango)

1. [[7]](https://www.sciencedirect.com/science/article/pii/S2772442522000016#bb7)

McPadden J., Durant T.J., Bunch D.R., Coppi A., Price N., Rodgerson K., Torre Jr. C.J., Byron W., Hsiao A.L., Krumholz H.M., Schulz W.L.

Health care and precision medicine research: analysis of a scalable data science platform

J. Med. Internet Res., 21 (4) (2019), pp. e13043-e13045

[View article](javascript:;)

[CrossRef](https://doi.org/10.2196/13043)[View in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-85064537463&partnerID=10&rel=R3.0.0)[Google Scholar](https://scholar.google.com/scholar_lookup?title=Health%20care%20and%20precision%20medicine%20research%3A%20analysis%20of%20a%20scalable%20data%20science%20platform&publication_year=2019&author=J.%20McPadden&author=T.J.%20Durant&author=D.R.%20Bunch&author=A.%20Coppi&author=N.%20Price&author=K.%20Rodgerson&author=C.J.%20Torre%C2%A0Jr.&author=W.%20Byron&author=A.L.%20Hsiao&author=H.M.%20Krumholz&author=W.L.%20Schulz)

1. [[8]](https://www.sciencedirect.com/science/article/pii/S2772442522000016#bb8)

Panesar A.

Machine Learning and AI for Healthcare (1–73)

Apress, Coventry, UK (2019)

Available at <https://iedu.us/wp-content/uploads/edd/2020/01/Arjun_Panesar_Machine_Learning_and_AI_for_Health-iedu.us_.pdf>, [Accessed on 5th March, 2021]

[Google Scholar](https://scholar.google.com/scholar_lookup?title=Machine%20Learning%20and%20AI%20for%20Healthcare%20&publication_year=2019&author=A.%20Panesar)

1. [[9]](https://www.sciencedirect.com/science/article/pii/S2772442522000016#bb9)

Davenport T., Kalakota R.

The potential for artificial intelligence in healthcare

Future Healthcare J., 6 (2) (2019), p. 94

[View article](javascript:;)

[CrossRef](https://doi.org/10.7861/futurehosp.6-2-94)[View in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-85073819217&partnerID=10&rel=R3.0.0)[Google Scholar](https://scholar.google.com/scholar_lookup?title=The%20potential%20for%20artificial%20intelligence%20in%20healthcare&publication_year=2019&author=T.%20Davenport&author=R.%20Kalakota)

1. [[10]](https://www.sciencedirect.com/science/article/pii/S2772442522000016#bb10)

Vidhya Analytics

Commonly used machine learning algorithms (with python and R codes)

(2021)

Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/>, [Accessed on 5th March, 2021]

[Google Scholar](https://scholar.google.com/scholar_lookup?title=Commonly%20used%20machine%20learning%20algorithms%20&publication_year=2021&author=Analytics%20Vidhya)

1. [[11]](https://www.sciencedirect.com/science/article/pii/S2772442522000016#bb11)

Mayfield H., Smith C., Gallagher M., Hockings M.

Use of freely available datasets and machine learning methods in predicting deforestation

Environ. Model. Softw., 87 (2017), pp. 17-28

[View PDF](https://www.sciencedirect.com/science/article/pii/S1364815216308428/pdfft?md5=794302daa3d420b3ec93c0d5e47c4379&pid=1-s2.0-S1364815216308428-main.pdf)[View article](https://www.sciencedirect.com/science/article/pii/S1364815216308428)[View in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-84994000503&partnerID=10&rel=R3.0.0)[Google Scholar](https://scholar.google.com/scholar_lookup?title=Use%20of%20freely%20available%20datasets%20and%20machine%20learning%20methods%20in%20predicting%20deforestation&publication_year=2017&author=H.%20Mayfield&author=C.%20Smith&author=M.%20Gallagher&author=M.%20Hockings)

1. [[12]](https://www.sciencedirect.com/science/article/pii/S2772442522000016#bb12)

Barot P.

Why use python in healthcare applications?

BoTree Technologies. (2020)

Available at: <https://www.botreetechnologies.com/blog/python-in-healthcare-application/>, [Accessed on 5th March, 2021]

[Google Scholar](https://scholar.google.com/scholar_lookup?title=Why%20use%20python%20in%20healthcare%20applications&publication_year=2020&author=P.%20Barot)

1. [[13]](https://www.sciencedirect.com/science/article/pii/S2772442522000016#bb13)

Bhanot K.

Predicting presence of heart diseases using machine learning

Towards Data Sci. (2019)

Available at: <https://towardsdatascience.com/predicting-presence-of-heart-diseases-using-machine-learning-36f00f3edb2c>, [Accessed on 5th March, 2021]

[Google Scholar](https://scholar.google.com/scholar_lookup?title=Predicting%20presence%20of%20heart%20diseases%20using%20machine%20learning&publication_year=2019&author=K.%20Bhanot)

1. [[14]](https://www.sciencedirect.com/science/article/pii/S2772442522000016#bb14)

Zaidi J.

Project: Predicting heart disease with classification machine learning algorithms

Towards Data Sci. (2020)

Available at: <https://towardsdatascience.com/project-predicting-heart-disease-with-classification-machine-learning-algorithms-fd69e6fdc9d6>, [Accessed on 5th March, 2021]

[Google Scholar](https://scholar.google.com/scholar_lookup?title=Project%3A%20Predicting%20heart%20disease%20with%20classification%20machine%20learning%20algorithms&publication_year=2020&author=J.%20Zaidi)

1. [[15]](https://www.sciencedirect.com/science/article/pii/S2772442522000016#bb15)

Ozgur C., Colliau T., Rogers G., Hughes Z.

Matlab vs. Python vs. R

J. Data Sci., 15 (3) (2017), pp. 355-371