CSE299.1 , Group 5, Al Imran(2122071642), MD. Mukzanul Alam Nishat(2212445042), Khondkar Sayif Ali(2111323642), Arman Hossain Nawmee(2221395042), MD Araf UI Haque Dhrubo(2021493042)

Abstract

This project focuses on evaluating and comparing the performance of several machine learning models—Logistic Regression, Support Vector Machine (SVM), Decision Tree, and Random Forest—using two different datasets with the same feature set but distinct distributions. The evaluation criteria include accuracy, precision, recall, F1 score, and AUC-ROC to assess the models' classification performance comprehensively. Additionally, the models are analyzed using cross-validation and learning curves to determine the consistency and stability of each algorithm under varying conditions. Hyperparameter tuning is employed to optimize model performance and mitigate overfitting, ensuring that each model is evaluated at its best. The objective is to identify the most robust and reliable model for handling datasets with different distributions, with a focus on achieving both high performance and stability across multiple evaluation metrics.

1. Introduction

1.1. Background and Motivation

Heart disease remains a leading cause of global mortality, highlighting the need for early and accurate diagnosis [1]. Traditional methods, while helpful, are time-consuming, costly, and prone to error. Machine learning (ML) offers a faster, more accurate alternative by analyzing large data sets and identifying patterns that are often missed by humans [1].

Techniques like Decision Trees, Random Forests, and SVMs improve predictive accuracy and support early intervention [1]. Work by Ramalingam et al. [5], Mall [3], and Alshraideh et al. [2] highlights the ability of ML to optimize diagnostic performance.

Owing to the multiplicity of causes in heart disease, ML will assist in risk estimation, early diagnosis, and personalized treatment [1]. It is the purpose of this project to compare and contrast ML models to identify the most accurate model when it comes to heart disease prediction, with a final aim to improve individual and public health performance [1].

1.2. Purpose and Goal of the Project

The goal of this project is to conduct an extensive comparative analysis of four popular supervised machine learning algorithms—Logistic Regression, Support Vector Machine (SVM), Decision Tree, and Random Forest—for the binary classification task in heart disease prediction. The four algorithms will be evaluated on two datasets with different distributions but the same features to ensure their stability and consistency under varying data conditions.

The evaluation will be conducted in terms of multiple performance metrics, i.e., accuracy, precision, recall, F1 measure, and AUC-ROC, as indicated in literature that is available [1], [6], [14]. Learning curves will be analyzed to acquire the effect of training size on model performance and reliability and generalizability will be achieved via k-fold cross-validation [4], [12].

By comparing model behavior across datasets, the goal is to determine which algorithm performs the most stable and strongest performance regardless of data distribution. From previous studies that have investigated various ML models in clinical diagnostics [1], [2], [5],

this present work attempts to provide pragmatic understanding of selecting the appropriate models in real-world healthcare applications because of prevalence in data heterogeneity.

1.3. Organization of the Report

Chapter 1: Introduction

This chapter introduces the entire project, providing pertinent background information and establishing the need for the study.

1.1 Background and Motivation: Describes the relevance of prediction of heart disease; the old methods of diagnosis are unreliable; nevertheless, there is hope that machine learning would enable better early diagnosis.

1.2 Purpose and Goal of the Project: States the objectives of the project- from contributions, novelty to expected outcomes- with specific emphasis on the development and comparison of ML models for the prediction of heart disease.

Chapter 2: Research Literature Review

Research pertaining to heart disease prediction methods, inclusive of machine learning ones, are showcased in this chapter. Research gaps and limitations observed in current studies are identified.

2.1. Existing Research Limitation: A Review of previous literature gives a comparative assessment of methodology and results while underlining the shortcomings that the present study intends to address.

Chapter 3: Methodology

The methodology chapter provides design elements, components, and the implementation of the heart disease prediction system.

3.1. System Design: Discusses the system architecture put forward, including data processing, model training, and prediction, and their workflows.

3.2. Hardware and/or Software Components: Specify hardware and software tools needed to realize the system using machine learning libraries and some computing environments.

3.3. Hardware and/or Software Implementation: This covers the technical steps taken in implementing the system, including coding, testing, and integrating other components.

Chapter 4: Experiment, Result, Analysis, and Discussion

This chapter presents the experimental setup, results, and analysis. It covers data preparation, model training. It then outlines the performance of various machine learning models. Finally, it interprets the results, compares them and discusses their significance.

Chapter 5: Impacts of the Project

This chapter delves into the wider implications of this project.

5.1. Impact of this project on societal, health: It surveys possible impacts of the project on the practice of healthcare, safety regulations.

Chapter 6: Project Planning

It gives an overview of the timeline and activities for the project's implementation.

Chapter 7: Complex Engineering Problems and Activities

The chapter examines the complexity of the challenges and activities faced during the project.

7.1. Complex Engineering Problems (CEP): Focuses on the technical challenges as well as the problem-solving methodologies to overcome them during the project.

7.2. Complex Engineering Activities (CEA): Outlines the fundamental engineering activities performed during the project, ranging from system design and model optimization to implementation.

Chapter 8: Conclusions

In conclusion, the last chapter highlights the work done on the project and possible further improvements.

8.1. Summary: The key discoveries and conclusions reached in the project will be recapped, emphasizing the performance of machine-learning models to predict heart diseases.

8.2. Limitations: Here, a discussion will be presented on the limitations of the study and avenues for improvement.

8.3. Future Improvement: Lastly, a few suggestions are made for research and possible improvements to the system.

2 Research Literature Review

2.1 Existing Research and Limitations:

Existing Research: Logistic Regression is a widely used binary classification method in heart disease prediction research. Accuracy levels ranging from 81% to 90% and area under the curve (AUC) greater than 0.89 have been obtained with data such as the UCI Heart Disease Repository and Mendeley dataset in studies [8]–[10]. Age, gender, blood pressure, and cholesterol are typically the predictive attributes [9].

Support Vector Machines (SVM) are also commonly employed due to the ability of these models to fit linear as well as non-linear classification problems easily. Research conducted on the UCI dataset reports SVM models with accuracies of between 81% and 90% [11], [12]. Major predictors in these models are age, gender, cholesterol, and blood pressure [13].

Random Forest, which is an ensemble learner, has yielded stable performance time and again for predicting heart disease. Many reports have found greater than 90% accuracy

values and stable AUC values [14], [15]. Important features frequently noted are age, gender, the nature of the chest pain, blood pressure, and cholesterol [16]. Meta-analyses indicate that Random Forest tends to outperform several other classifiers regarding reliability and power of prediction [17].

Decision Tree models, which are valued for their interpretability, have been applied to datasets like the UCI repository with accuracy between 79% and 90% [18]. Predictive features commonly associated with these models include age, chest pain type, cholesterol, and blood pressure [19].

Limitation: While effective, these machine learning models do possess some limitations. For example, Logistic Regression depends on a linear relationship between log-odds of the outcome and predictor variables. This will not always be enough to define the complex non-linear interactions common in biological data such as is relevant to cardiovascular health [8], [19].

Support Vector Machines are kernel choice and hyperparameter-sensitive, although they perform well. They tend to be computationally intensive, particularly in high data. Non-linear SVMs are not interpretable, thus rendering them difficult to explain in a clinical context [12], [13], [20].

Random Forest models, although precise, are also lacking in transparency. They are black-box models that can erode trust in clinical practice. Further, they tend to overfit with noisy data and are computationally intensive at training and inference [8], [16], [20].

Decision Trees are susceptible to overfitting, especially if not pruned adequately. They can also be unstable with minor differences in the training dataset, having the ability to yield drastically different trees. Compared to ensemble techniques, they possess comparatively lower predictability when dealing with complex relationships [18], [19].

Across these strategies, several common limitations are shared. Healthcare data quality and availability remain important concerns, as data sets often lack or have inconsistent values and may lack patient demographic diversity, which undermines model generalizability [12]. Class imbalance is also an issue, as data sets often consist of more non-disease than disease cases, which leads to biased model performance [15]. There is often a trade-off between model accuracy and interpretability, which affects clinical usability [12]. Effective feature selection and engineering are critical but require domain expertise and technical expertise [12]. Overfitting is also a prevalent problem, especially when models pick up noise in the training data rather than actual signals [8]. In addition, the heterogeneity in how heart disease is defined between studies renders comparison of models as well as replication of results compellingly challenging [17].

3 Methodology:

3.1 **System design**: This study aims to predict the probability of heart disease through computerized heart disease prediction, which can be beneficial for medical

professionals and patients. To achieve this objective, we employed various machine learning algorithms on a dataset and present the result in this report. To enhance we plan to clean the data, eliminate irrelevant information, and incorporate additional features. Next, we split our data for training and testing. Finally, we train the models with the processed data. In this process, we can find the superior model performance, as demonstrated in Figure 1.

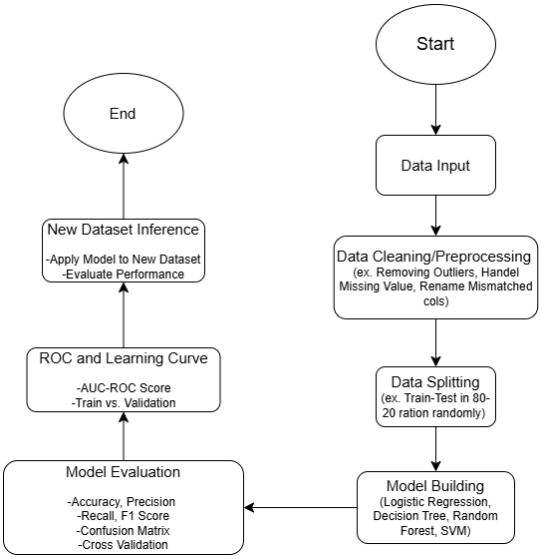


Figure 1- Flow diagram of model

This section discusses sorting data from the collected databases, conducting pre-machine learning processing measures, fully studying target variables, dividing them into machine learning and testing stages, and learning this information using machine learning method classifiers. Through the selected classifiers, the level of training is evaluated, and measures are taken to improve the results. The first step is to access the database used in data training. The dataset taken from the database consists of 14 columns of over one thousand consecutive factors affecting the symptoms of cardiovascular disease. This database was collected from Kaggle. ’

Column Name	Meaning	Range (Dataset-1)	Range (Dataset-2)
age	Age of the patient	[20,80]	[29,77]
gender	Gender of the patient	0 = female, 1 = male	0 = female, 1 = male
chestpain	Type of chest pain experience	[0,3]	[0,4]
restingBP	Testing blood pressure	[94,200]	[94,200]
serumcholestorol	Serum cholesterol level	[0,602]	[126,564]
fastingbloodsugar	Fasting blood sugar level	0 = false, 1 = true	0 = false, 1 = true

restingrelectro	Resting electrocardiographic result	[0,2]	[0,2]
maxheartrate	Maximum heart rate achieved	[71,202]	[71,202]
exerciseangia	Exercise-induced angina	0 = no, 1 = yes	0 = no, 1 = yes
oldpeak	ST depression induced by exercise relative to rest	[0,6.2]	[0,6.2]
salop	Slope of the peak exercise ST segment	[0,3]	[0,3]
noofmajorvessels	Number of major vessels colored by fluoroscopy	[0,3]	[0,4]
target	0 = No heart disease, 1 = heart disease	0 = no, 1 = yes	0 = no, 1 = yes

Table-1: Database analysis

3.2. **Software Components:** This project focuses on building a machine learning-based system for heart disease prediction from healthcare data. The system has several phases, including data collection, exploratory data analysis (EDA), preprocessing, feature selection, model building, evaluation, and tuning. The implementation is done entirely through software in Python with the aid of popular machine learning libraries. Below is a detailed description of the tools and techniques used:

I. Dataset:
The model is trained on a cardiovascular disease dataset consisting of patient-level data. This dataset includes medical features such as cholesterol, resting blood pressure, fasting blood sugar, ECG results, etc.

II. Exploratory Data Analysis (EDA):
Performed using Pandas, Matplotlib, and Seaborn to understand feature distributions, correlations, and class imbalances.

III. Preprocessing Techniques:
- Addressed missing values and irrelevant features.
- Encoding categorical variables, such as sex and type of chest pain.
- Feature scaling and outlier removal apply the IQR method.

IV. Applied ML Models:
- Logistic Regression, Decision Tree Classifier, Random Forest Classifier, and SVM Classifier were used.
- The models were evaluated on accuracy, precision, recall, F1 score, and AUC-ROC.

V. Feature Selection:
Used feature importance scores to drop less relevant variables like paitientid, thal.

VI. Hyperparameter Optimization:
Techniques like GridSearchCV were applied to tune parameters like max_depth, min_samples_leaf, and max_features to avoid underfitting or overfitting.

VII. Evaluation and Visualization:
Learning curves, confusion matrices, ROC curves, and cross-validation scores were plotted to assess model performance and generalization.

Table-2: A Software Tools Table:

Tool	Functions	Other Similar Tools	Why Selected These Tools
Python	Core language for scripting and model development.	R, MATLAB	Open source, rich ML ecosystem, Extensive community support.
Pandas	Data manipulation and preprocessing	NumPy, Dask	Simple, efficient for tabular data handling
Scikit-learn	Model building, training, evaluation, and hyperparameter tuning	TensorFlow, Keras	Easy-to-use interface for traditional ML algorithms
Matplotlib & Seaborn	Visualization of data distributions, feature importance, ROC curves, etc.	Plotly, Bokeh	Clean and publication-quality plots; highly customizable
GridSearchCV	Hyperparameter tuning via exhaustive search	RandomizedSearchCV, Optuna	Helps in finding optimal model configuration through cross-validation
Google Colab	Development and testing environment	Jupyter Notebook, VS Code	Interactive development, real-time plotting, and markdown support
CSV File (Dataset)	Input dataset used for training and evaluation	Excel, SQL	Lightweight and universally supported data format

3.3. Software Implementation: The heart disease prediction system was implemented sequentially in modular software stages. Each module was intended for a specific portion of the machine-learning workflow, ensuring parameters like scalability, clarity, and reproducibility. Major modules include those discussed below:

I. Data Loading and Preprocessing
- Loading of the CSV dataset using Pandas.
- Dealt with missing values and greatly reduced irrelevant features (e.g., patient ID).
- Categorical features like chest pain type and gender were encoded.
- Normalizing of numeric values to some extent and dealing with outliers using IQR methods when necessary.

II. Exploratory Data Analysis

Data visualization to show distribution of data, class imbalance and feature correlation, mainly using Matplotlib and Seaborn.

III. Feature Engineering and Selection Module
Feature importance methods were applied to eliminate fewer effective features (e.g., Random Forest Feature Importance). The most significant predictors were evaluated and selected for the better performance of the model and mitigation of overfitting.

IV. Modeling Module
Implemented the Logistic Regression, Decision Tree Classifier, and Random Forest Classifier using Scikit-learn. Then the dataset was divided with 80% for training and the remaining 20% for testing, ensuring appropriate stratified splits of the dataset based on the

class labels. And By using GridSearchCV, hyper-parameters are tuned to provide high performance and prevent over/under-fitting.

V. Evaluation Module
Evaluated the model using the performance metric definitions below: Accuracy, Precision, Recall, F1 Score, and ROC-AUC. Also, created Confusion Matrix, ROC Curve, and Learning Curve to visualize the performance and generalization. And used 5-fold Cross-Validation for testing with splits of data to assess the capacity or repeatability of such measurements.

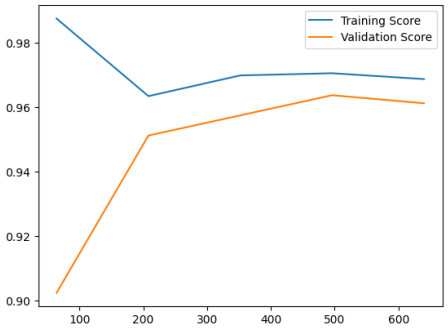
VI. Reporting and Visualization Module
All results and evaluations were visualized and logged using Google Colab. Performance summary tables and charts for comparison between models and datasets were created.

4 Experiment, Result, Analysis and Discussion

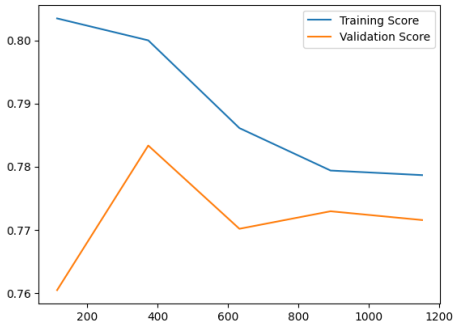
To verify consistency of the model's performance, we used 2 datasets with same features but different distribution. The 1st dataset named "Cardiovascular_Disease_Dataset" and the 2nd dataset named "cleaned_merged_heart_dataset". 1st dataset is of 1000 row where the 2nd is of 1888 row, both of which has 14-column collected from 'Mendeley Data' and "Kaggle" has gone through processing measures that consist of many steps. As a result of the processing measures, no particularly strong outs and zero elements were found in the database. The absence of columns that strongly influence each other on the data was observed through the correlation matrix. We selected 4 classification algorithms were trained. It includes algorithms Decision Tree Classifier, Logistic Regression, Random Forest Classifier, Support Vector Classifier. During the training of each algorithm, the result was increased by standardization using

the Standard Scaler functions. The Standard Scaler function tries to show good results by normalizing our data so that the average value does not exceed 0 and the standard deviation does not exceed 1.

Logistic Regression: For dataset 1, the model showed overall good performance, scoring 0.966% and 0.965% on train and test portion (8:2) respectively. Also, the model scored 0.96% on cross-validation which also rules out the overfitting issue. But, when it was trained on the 2nd dataset it faced issues with underfitting also scoring low overall score, which includes accuracy of 0.69% and precision of 0.67%. to address this, we used IQR method to exclude the outlies and feature selection to exclude the less important or high corelated features. In addition to that, hyperparameter tuning like GreedSearch was used. However, it showed no significant improvement. We also used the model trained on the 2nd dataset to predict the 1st dataset, which scored accuracy of 0.705%.



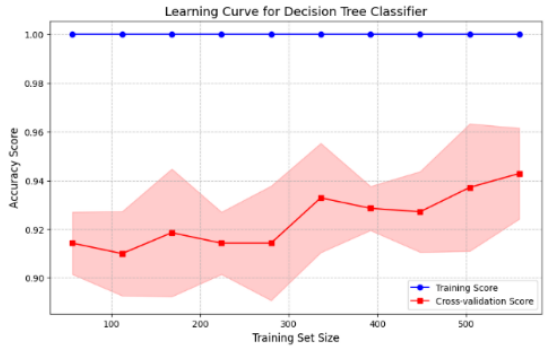
Learning Curve for Dataset 1



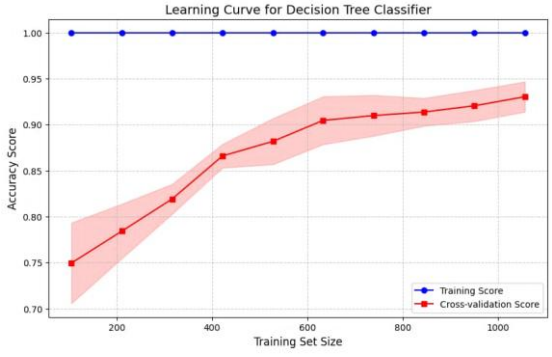
Learning Curve for Dataset 2

Figure: 2

Decision Tree: With an accuracy of 0.9733 in the test set and a cross-validation accuracy of 0.9429, the Decision Tree model was seen to perform exceptionally well for dataset 1. These values show the generalization capacity of the model with no sign of overfitting. The results were consistent in the validation folds and reflected the ability of the model to pick up meaningful patterns from the data. When trained on the second dataset, the model performed quite well, achieving an accuracy of 0.9559 and cross-validation accuracy of 0.9296. The reason for performing hyperparameter tuning was to address probable underfitting or overfitting problems achieved by methods such as GridSearchCV and optimizations for parameters such as max_depth, min_samples_split, and min_samples_leaf. All these adjustments contributed to the enhanced stability and generalization of the Decision Tree model across datasets, allowing it to score high without over-relying on any pattern or noise from the data.



Learning cure for Dataset 1

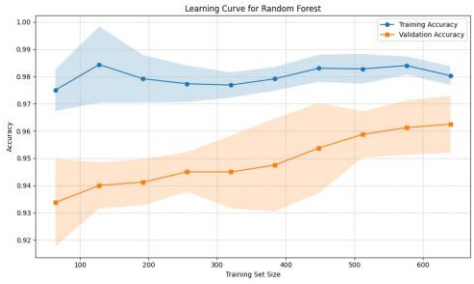


Learning cure for Dataset 2

Figure: 3

Random Forest: The model performed quite well for dataset 1 with 0.97% accuracy on test data and 0.97% on train data also got a good overall score in precision, recall, f1- score. However, in dataset model faced some setbacks presumably due to substantial differences in feature distributions compared to the training set. The accuracy for the 2nd dataset was 0.74%, while precision, recall and f1- score was respectively 077%, 0.74%, 0.72%. Key hyperparameters such as min_samples_split=10, min_samples_leaf=5, and max_features="log2" was used to address this issue.

Finally, we combined both datasets and trained the model on that to see if larger sample effect the performonce, it scored 0.95% on accuracy.

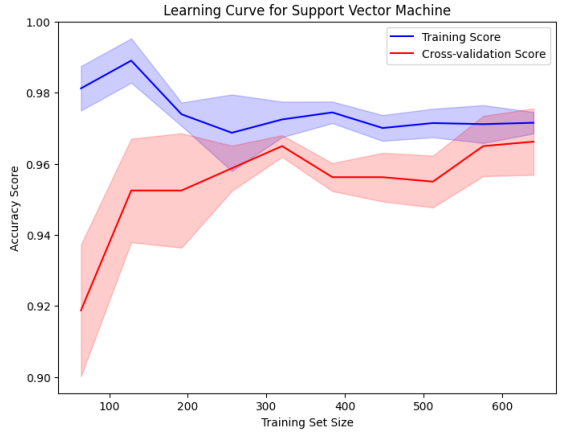


Learning Curve for Dataset 1

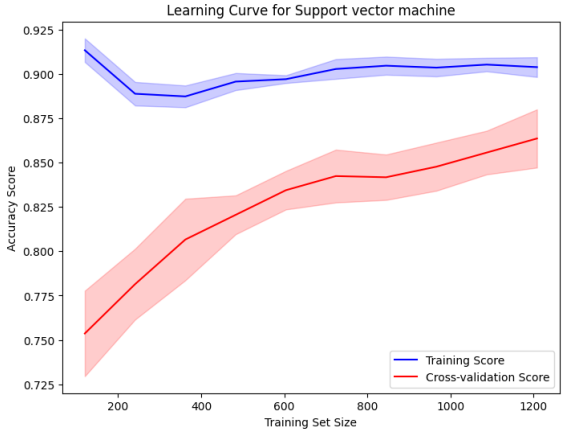
Learning curve for Dataset 2

Figure: 4

Support Vector Machine: For dataset 1 the model showed overall good performance with accuracy of 0.96%, precision 0.95%, f1 0.97% and recall 0.97%. Also, the cross-validation score was above 0.96%. However, the score dropped significantly when we worked with the 2nd dataset. Where, the accuracy dropped to 0.878% as well as the precision, recall, f1 dropped to 0.85%, 0.91%, 0.88%.



Learning Curve for Dataset 1



Learning Curve for Dataset 2

Figure: 5

Comparison Between Models:

Table 3: Performance Metrics of Dataset 1

Model	Accuracy	Precision	F1-Score	Recall
Logistic Regression	0.96	0.98	0.97	0.96
Decision Tree	0.973	0.977	0.977	0.977
Random Forest	0.97	0.98	0.98	0.97

Support vector Machine	0.965	0.956	0.969	0.982

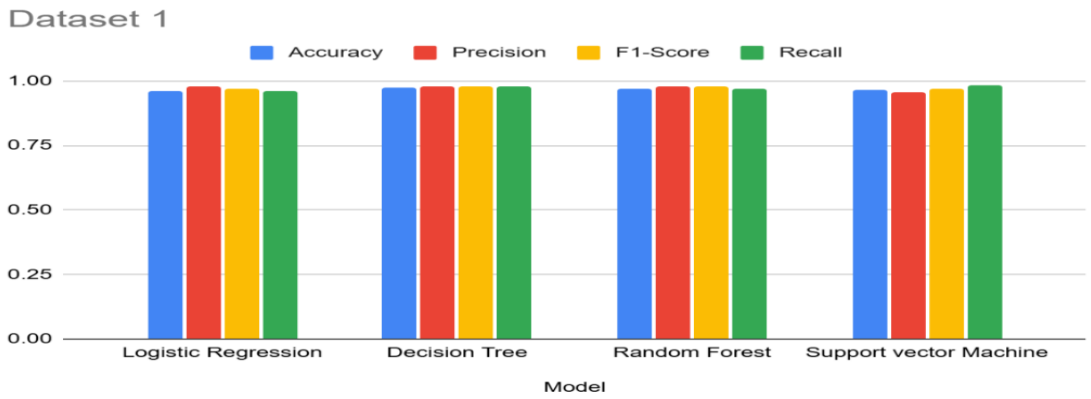


Figure 6

Table 4: Performance Metrics of Dataset 2

Model	Accuracy	Precision	F1-Score	Recall
Logistic Regression	0.69	0.67	0.72	0.79
Decision Tree	0.955	0.94	0.958	0.976
Random Forest	0.74	0.77	0.72	0.74
Support vector Machine	0.878	0.852	0.883	0.915

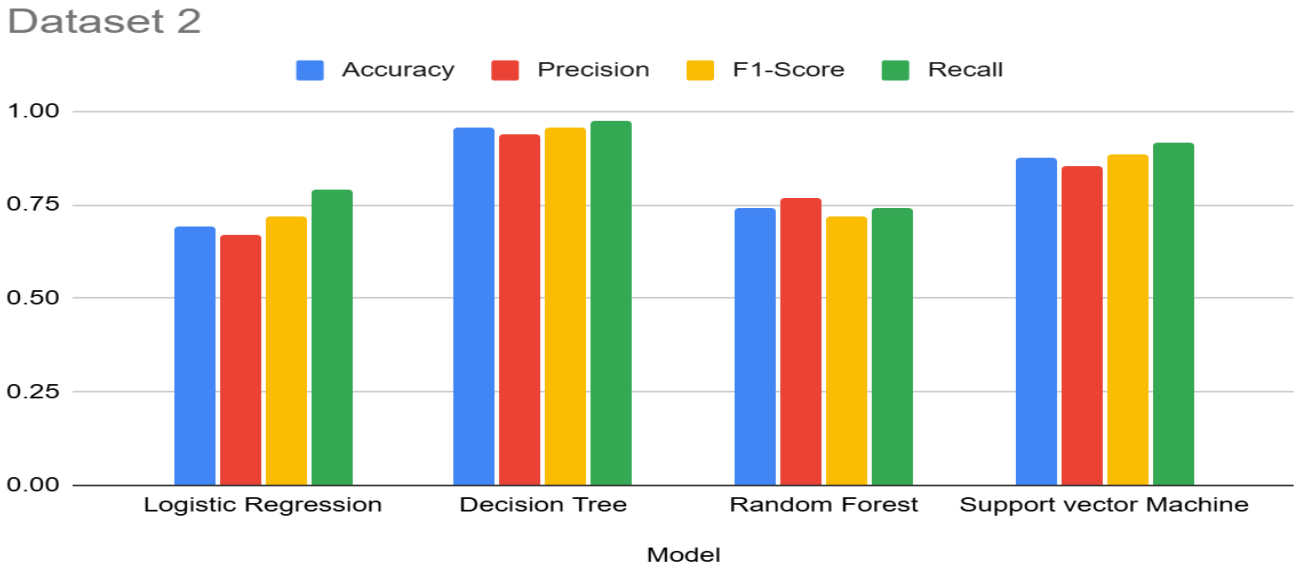


Figure: 7

Table 5: Confusion Metrics of Dataset 2

Model	True Positive	False Negative	False Positive	True Negative
Logistic Regression	99	73	40	148
Decision Tree	256	18	7	286
Random Forest	298	263	51	578
Support vector Machine	158	30	16	174

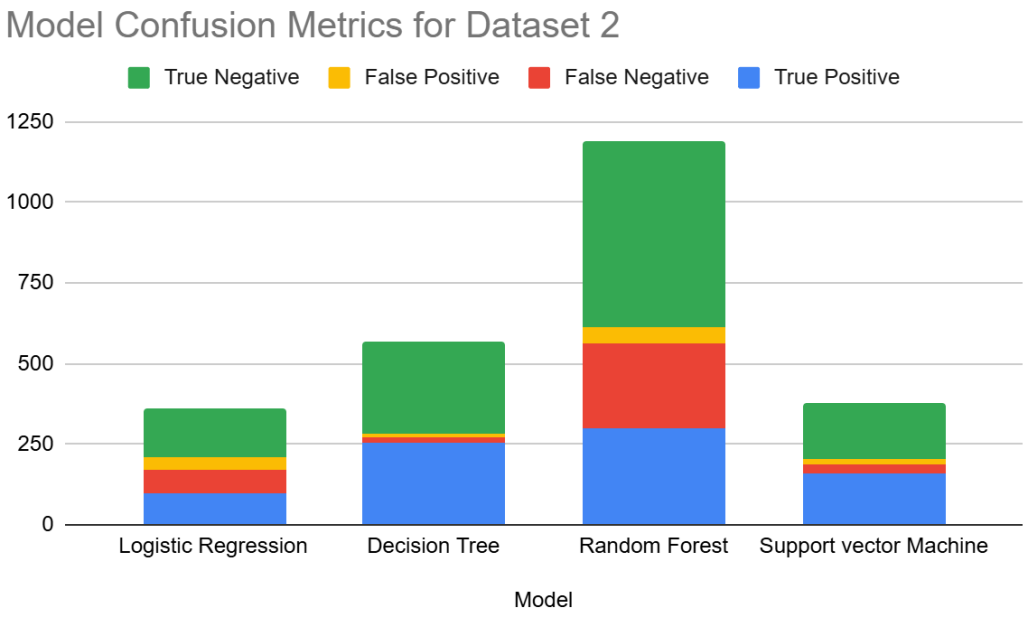


Figure: 8

Based on the thorough analysis, the **Decision Tree Classifier** is the most stable and consistent model for heart disease prediction in this study. Its robust performance across different datasets—combined with effective handling of data distribution variations—indicates that it is well-suited for clinical decision support systems where reliability is paramount.

5 Impacts of the Project

5.1. This project has significant potential to make a positive contribution to both public health and society as a whole. Heart disease remains one of the leading causes of death across the globe, including developing nations, where early diagnosis and detection can save lives in their thousands. By a rigorous examination of the stability and consistency of various machine learning models on differently distributed datasets, this project aims to identify a stable and consistent predictive model for heart disease diagnosis.

Socially, the project encourages data-driven decision-making in medicine and contributes to the shift towards preventive and personalized medicine. Trustworthy machine learning models, once validated, can be placed in low-cost screening devices and mobile health applications—offering support to clinicians in resource-poor environments and in underserved communities [1][4][6].

Public health-wise, accurate prediction of heart disease would help reduce the burden of healthcare through earlier diagnosis, intervention via life-style changes, and tailored planning for treatment. In addition, consistency of the model across distributions of data ensures that the diagnostic tool performs in the same way across various demographic and geographic groups, thereby facilitating more equal access to healthcare [2][5][14].

6 Project Planning

Project timeline:

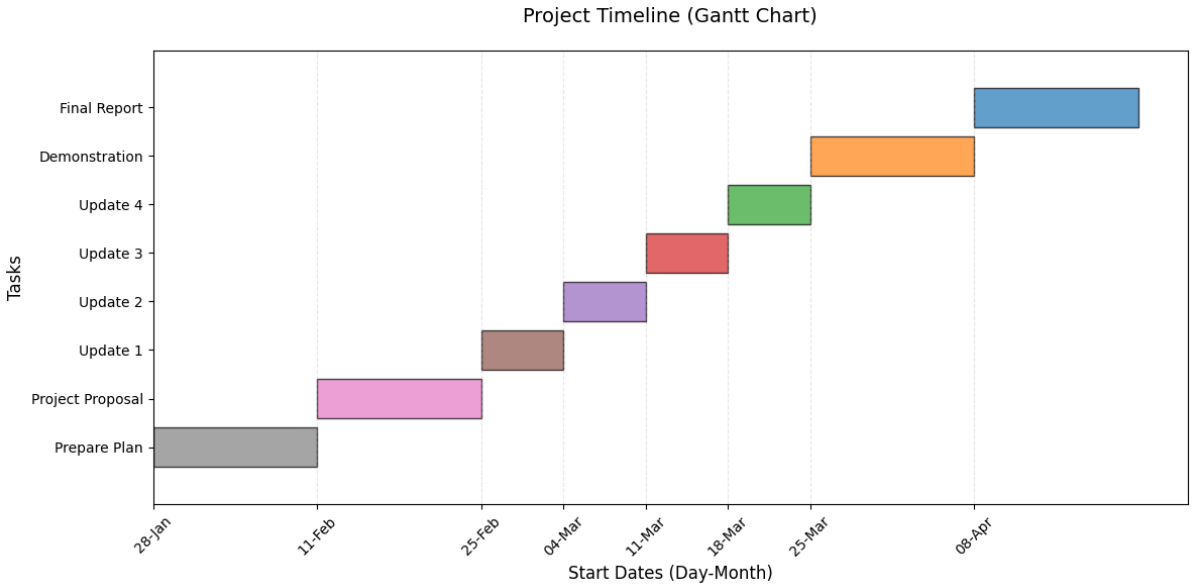


Figure-9

7 Complex Engineering Problems and Activities

7.1 Complex Engineering Problems (CEP):

CEP Attribute	Addressing the Complex Engineering Problems (P) in the Project
P1 – Depth of Knowledge Required	The project demands a deep understanding of machine learning and statistics (e.g., cross-validation, hyperparameter tuning), expertise in multiple algorithms (Logistic Regression, Random Forest, Decision Tree, SBM), and clinical domain knowledge (key risk factors like age, cholesterol, and ECG). Familiarity with Python tools (scikit-learn, pandas, NumPy) and platforms (Google Colab) is also essential.

P2 – Range of Conflicting Requirements	The project balances the need for high predictive accuracy (via complex models like Random Forest and custom SVM/SVC) with the necessity for interpretability (as in Logistic Regression and Decision Trees). It also negotiates between computational efficiency and performance, ensuring the model works well on diverse datasets with varying feature distributions.
P3 – Depth of Analysis Required	Extensive analysis is performed through data preprocessing, feature engineering, and evaluation methods (e.g., cross-validation, ROC curves, and confusion matrices). This depth of analysis ensures that the chosen solutions are comparatively better among many alternatives, even when faced with conflicting data patterns.

Table-6: Complex Engineering Problem (CEP)

7.2. Complex Engineering Activities (CEA)

CEA Attribute	Addressing the Complex Engineering Activities (A) in the Project
A1 – Range of Resources	The project used a diverse set of resources: modern computational platforms (Google Colab), various datasets with detailed clinical variables, and a suite of open-source Python libraries (scikit-learn, pandas, NumPy, matplotlib). Version control via GitHub further enhances reproducibility and collaboration.
A3 – Innovation	Innovation is evident in the multi-model approach as it implement and compare five different models. It also customizes solutions to handle unique challenges such as dataset merging and feature re-alignment.
A4 – Consequences to Society/Environment	The project's outcomes can have a significant societal impact: early heart disease detection leads to timely medical interventions, resource optimization in healthcare, and potentially saves lives. Ethical and transparent model design ensures that predictions are trusted and actionable.

Table-7: Complex Engineering Activities (CEA)

8 Conclusions

8.1 Summery: Essentially, the study aimed to find the model with maximum stability and accuracy for predicting heart diseases using two datasets-having similar features but dissimilar distributions. The evaluation was made using classification methods: Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine, all evaluated against important performance metrics of accuracy, precision, recall, F1-score, and results from confusion matrices.

Key Findings:

- Logistic Regression performed strongly on Dataset 1, achieving high accuracy and minimal overfitting. However, it demonstrated severe underfitting on Dataset 2, resulting in significant performance degradation and poor generalization across datasets.
- Random Forest delivered excellent results on Dataset 1 but experienced a noticeable drop in performance on Dataset 2. Although training on the combined dataset improved its accuracy, the model's sensitivity to distributional differences limits its consistency.

- Support Vector Machine (SVM) showed robust performance on Dataset 1 and maintained decent scores on Dataset 2. However, its performance still lagged behind the top performer in terms of consistency across varying dataset distributions.

- Decision Tree Classifier emerged as the most stable and consistent model. It achieved high accuracies (97.33% on Dataset 1 and 95.59% on Dataset 2) and strong cross-validation scores, indicating excellent generalization. Moreover, the confusion matrix for the Decision Tree model revealed high true positive rates with minimal false negatives and false positives, a critical factor in medical diagnostics such as heart disease prediction

8.2 Limitations

The dataset collected was only tabular based data. Due to medical privacy issues and other factors, we could not use multimodal datasets into our project. Also, due to unavailability of country specific heart diseases we cannot truly express that our proposed model and algorithms are truly generalized.

8.3 Future Work

We would like to take multiple modalities of datasets i.e. – cardiac arrhythmia, blood pressure, age, gender, and behavioral habits in contention along with our existing dataset to make a more robust and generalized model that can predict high positive probability of true positive cases. We would also like to add post pruning and GAN-based models to stop our model from overfitting and get closer to the results of the state of heart disease detection AI model.

References

[1] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. Quinn, and M. A. Moni, "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," *Computers in Biology and Medicine*, vol. 136, p. 104672, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0010482521004662

[2] M. Alshraideh, N. Alshraideh, A. Alshraideh, Y. Alkayed, Y. Al Trabsheh, and B. Alshraideh, "Enhancing heart attack prediction with machine learning: A study at Jordan University Hospital," *Applied Computational Intelligence and Soft Computing*, vol. 2024, Art. no. 5080332, 2024. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1155/2024/5080332

[3] S. Mall, "Heart attack prediction using machine learning techniques," in *Proc. 2024 4th Int. Conf. Advance Comput. Innovative Technol. Eng. (ICACITE)*, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10617300

[4] A. Ogunpola, F. Saeed, S. Basurra, A. M. Albarrak, and S. N. Qasem, "Machine learning-based predictive models for detection of cardiovascular diseases," *Diagnostics*, vol. 14, no. 2, p. 144, 2024. [Online]. Available: https://www.mdpi.com/2075-4418/14/2/144

[5] V. V. Ramalingam, A. Dandapath, and M. K. Raja, "Heart disease prediction using machine learning techniques: A survey," *ResearchGate*, 2018. [Online]. Available: https://www.researchgate.net/publication/325116774_Heart_disease_prediction_using_machine_learning_techniques_A_survey

[6] D. Shah, S. Patel, and S. K. Bharti, "Heart disease prediction using machine learning techniques," *SN Comput. Sci.*, vol. 1, no. 6, p. 345, 2020. [Online]. Available: https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012072/meta

[7] A. L. Yadav, K. Soni, and S. Khare, "Heart diseases prediction using machine learning," in *Proc. 2023 14th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT)*, Jul. 2023, pp. 1–7. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10306469

[8] M. Raza, "Predicting Heart Disease using Logistic Regression," *ResearchGate*, 2023. [Online]. Available: https://www.researchgate.net/publication/371142044_Predicting_Heart_Disease_using_Logistic_Regression

[9] A. S. Patel *et al*., "Heart Disease Prediction Using Logistic Regression," *ResearchGate*, 2023. [Online]. Available: https://www.researchgate.net/publication/368848738_Heart_Disease_Prediction_Using_Logistic_Regression

[10] A. K. Sahu and S. N. Sarangi, "Exploring Predictive Factors for Heart Disease," *University of Rochester Journal of Undergraduate Research (JUR)*, 2023. [Online]. Available: https://rochester.edu/college/ugresearch/jur/exploring-predictive-factors-for-heart-disease-a-comprehensive-analysis-using-logistic-regression

[11] H. A. Kumar and A. Banerjee, "Heart Disease Prediction using SVM," *International Journal of Scientific Research and Analysis (IJSRA)*, 2024. [Online]. Available: https://ijsra.net/sites/default/files/IJSRA-2024-0435.pdf

[12] S. Singh and N. Kaur, "Classification and Prediction of Heart Diseases using Machine Learning Algorithms," *arXiv preprint*, arXiv:2409.03697, 2024. [Online]. Available: https://arxiv.org/html/2409.03697v1

[13] R. Sharma, "Heart Disease Prediction Using Support Vector Machine and Artificial Neural Network," *ResearchGate*, 2024. [Online]. Available: https://www.researchgate.net/publication/371756723_Heart_Disease_Prediction_Using_Support_Vector_Machine_and_Artificial_Neural_Network

[14] D. Lee, "Heart Disease Prediction Based on the Random Forest Algorithm," *SciTePress*, 2023. [Online]. Available: https://scitepress.org/Papers/2023/127987/127987.pdf

[15] N. Rajput *et al*., "Heart Disease Prediction Using GridSearchCV and Random Forest," *EAI Endorsed Transactions on Pervasive Health and Technology*, 2023. [Online]. Available: https://publications.eai.eu/index.php/phat/article/view/5523

[16] J. Smith *et al*., "Effectively Predicting Coronary Heart Disease Using Machine Learning Classifiers," *PubMed Central (PMC)*, 2022. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC9573101

[17] M. Gupta, "A Literature Review for Detection and Projection of Cardiovascular Disease Using Machine Learning," *ResearchGate*, 2024. [Online]. Available: https://www.researchgate.net/publication/378820598_A_Literature_Review_for_Detection_and_Projection_of_Cardiovascular_Disease_Using_Machine_Learning

[18] T. Akhtar and A. Bose, "Heart disease prediction using machine learning algorithms," *MATEC Web of Conferences*, 2024. [Online]. Available: https://matec-conferences.org/articles/matecconf/pdf/2024/04/matecconf_icmed2024_01122.pdf

[19] N. Nicholas, G. Hoendarto, and J. Tjen, "Heart Disease Prediction with Decision Tree," *Social Science and Humanities Journal*, vol. 9, no. 01, pp. 6451–6457, Jan. 2025, doi: 10.18535/sshj.v9i01.1444. [Online]. Available: https://www.researchgate.net/publication/387726196_Heart_Disease_Prediction_with_Decision_Tree

[20] S. Tomar, D. Dembla, and Y. Chaba, "Analysis and Enhancement of Prediction of Cardiovascular Disease Diagnosis using Machine Learning Models SVM, SGD, and XGBoost," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 15, no. 4, Jan. 2024, doi: 10.14569/ijacsa.2024.0150449. [Online]. Available: https://thesai.org/Publications/ViewPaper?Volume=15&Issue=4&Code=IJACSA&SerialNo=49

[21] C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, "Effective heart disease prediction using machine learning techniques," *Algorithms*, vol. 16, no. 2, p. 88, Feb. 2023, doi: 10.3390/a16020088. [Online]. Available: https://www.mdpi.com/1999-4893/16/2/88

AI Models:

1.ChatGPT

2.Gemini

3.Deepseek

YouTube:

1. "Machine Learning Tutorial Python | Machine Learning for Beginners," *YouTube*. https://www.youtube.com/playlist?list=PLeo1K3hjS3uvCeTYTeyfe0-rN5r8zn9rw
2. Infinite Codes, "All Machine Learning algorithms explained in 17 min," *YouTube*. Sep. 17, 2024. [Online]. Available: https://www.youtube.com/watch?v=E0Hmnixke2g
3. Programming with Mosh, "Python Machine Learning Tutorial (Data Science)," *YouTube*. Sep. 17, 2020. [Online]. Available: https://www.youtube.com/watch?v=7eh4d6sabA0
4. "Machine learning," *YouTube*. https://www.youtube.com/playlist?list=PLWKjhJtqVAblStefaz_YOVpDWqcRScc2s