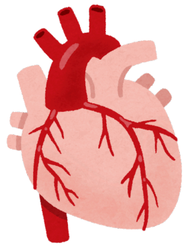
Maharshi Karve Stree Shikshan Samstha's
Cummins College of Engineering for Women, Pune
(An autonomous Institute affiliated to Savitribai Phule Pune University)
Department of Computer Engineering

Course Code : 23PCCE501L

Course Name : Artificial Intelligence and Machine Learning Laboratory

End Semester Examination , A.Y. 2025-26, Semester I
DIV : A BATCH : A3



HEART DISEASE RISK PREDICTION USING MACHINE LEARNING ALGORITHMS

A Performance Comparison of LR, DT, SVM, and XGBoost for Clinical Risk Detection

The Problem : A Global Health Challenge

- Heart diseases are a major public health burden and a leading cause of death worldwide.
- A significant portion of these deaths are preventable through early diagnosis and timely intervention.
- Traditional diagnostic procedures can be time-consuming, involve advanced equipment, and are not easily accessible to all populations.
- Machine learning (ML) offers a transformative tool to analyze complex clinical data, identify hidden patterns, and predict disease risk early.

Project Objective

Primary Goal : To design and implement a system that predicts the probability of heart disease, classifying patients as **High Risk or Low Risk**.

Model Comparison : To apply and evaluate two popular ML algorithms->Logistic Regression and Decision Tree Classifier.

Interactive System : To create a user-friendly, interactive system (using Streamlit) for data upload, model training, and real-time prediction.

Clinical Impact : To assist and augment healthcare professionals in making informed decisions and identifying high-risk patients earlier.

The Data: Fueling the Models

The system is designed to accept any user-uploaded CSV dataset. Common features in such datasets include:

- **Demographics :** Age, Sex
- **Symptoms :** Chest Pain Type, Exertional Angina
- **Vitals :** Resting Blood Pressure, Max Heart Rate
- **Lab Results :** Serum Cholesterol, Fasting Blood Sugar
- **ECG Data :** Resting ECG, Oldpeak (ST depression)

Target Variable : The presence (1) or absence (0) of heart disease.

Preparing the Data: The Preprocessing Pipeline

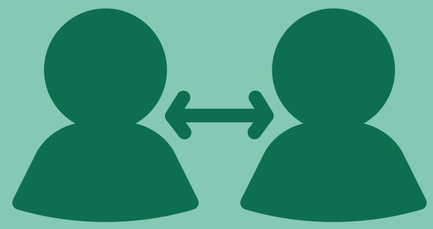
Train–Test split->80% data for **training** the model and **20%** data for **testing** the model.

1. Manage Missing Data

To ensure data integrity and prevent biased training, any entry (row) with missing values is excluded from the dataset.

2. Identify Target

The system automatically examines column names to find the target variable (e.g., "target", "output", "diagnosis").



3. Label Encoding

Categorical labels (e.g., "presence", "absence") are transformed into numerical binary values (1 and 0) for the models.

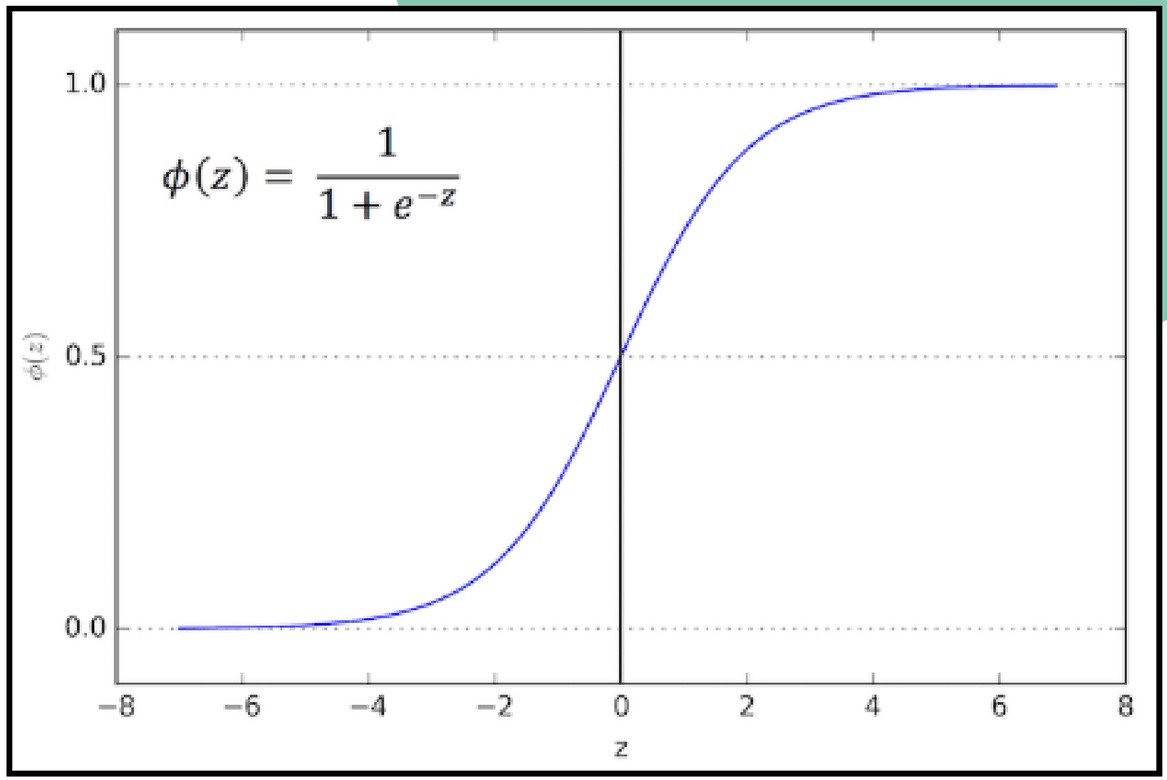
4. Feature Scaling

Using StandardScaler, all numerical features are standardized (mean=0, std=1) to prevent features with large ranges from dominating the model.

Model 1 : Logistic Regression

A widely-used statistical algorithm for binary prediction tasks.

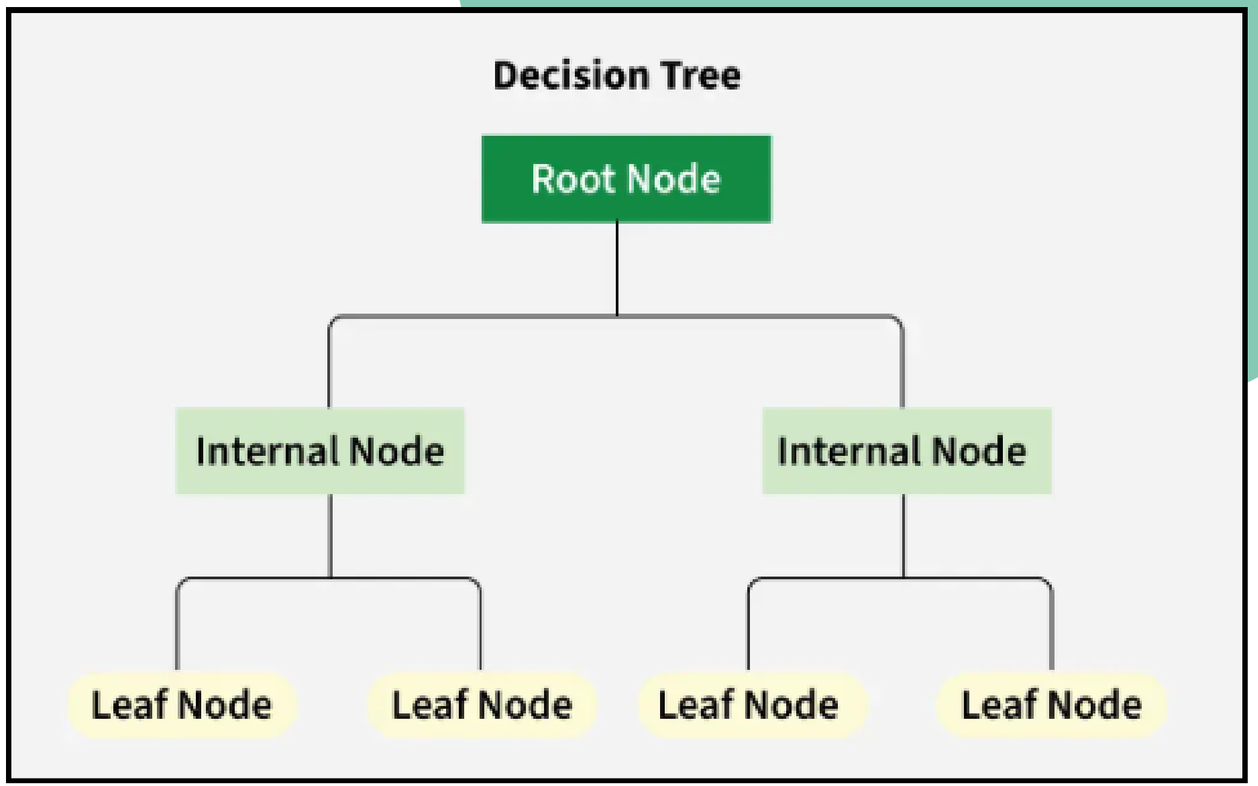
- It uses the Sigmoid Function to map a weighted sum of inputs to a probability between 0 and 1.
- **Key Advantage :** Excellent for risk prediction. It produces smooth, well-calibrated, and realistic probability estimates (e.g., 75% risk).
- This allows for a more meaningful clinical interpretation of the *degree* of risk, rather than just a simple "yes" or "no".



Model 2 : Decision Tree Classifier

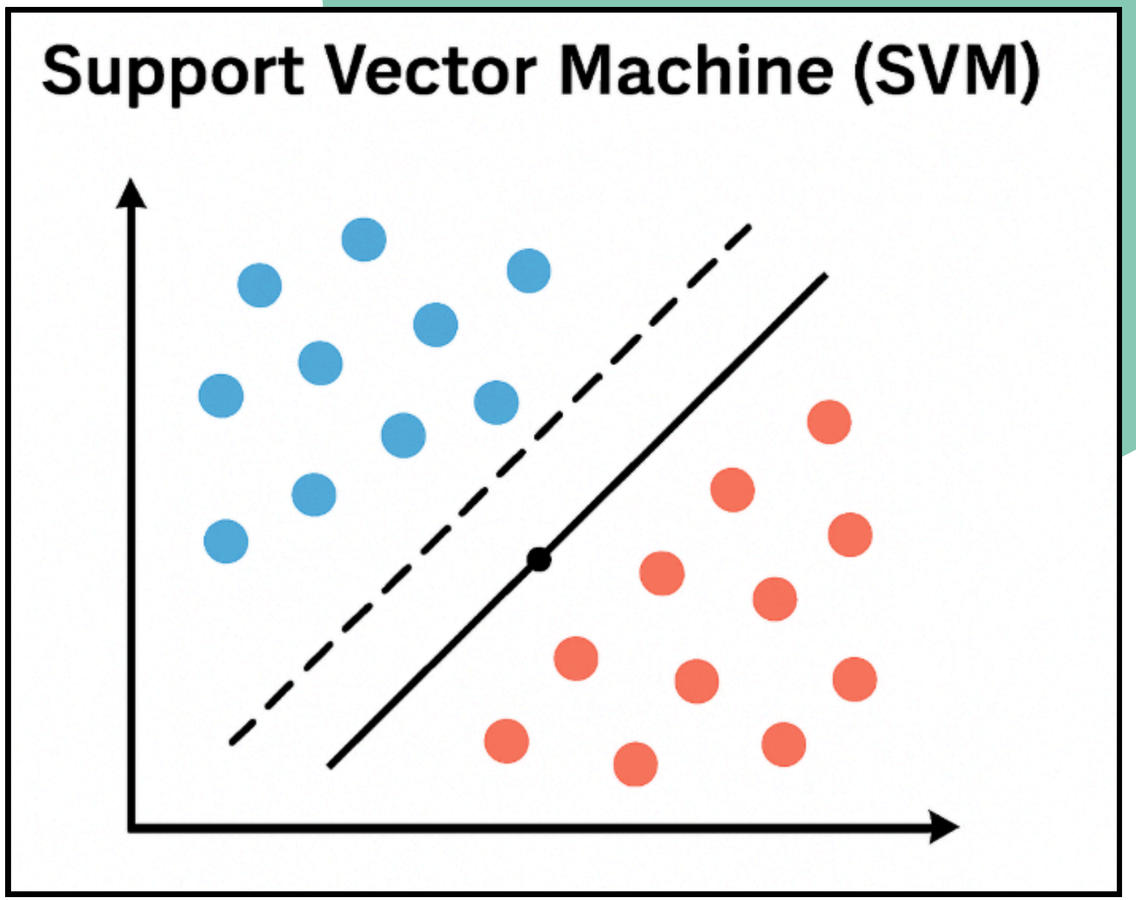
A non-linear, rule-based model that mimics human decision-making.

- It structures knowledge in an inverted pyramid, splitting data based on feature values.
- **Key Advantage :** High interpretability. A clinician can follow the branches of the tree to understand why a prediction was made.
- **Limitation :** Prone to overfitting (memorizing noise). To prevent this, the model's complexity was limited (e.g., `max_depth = 5`).



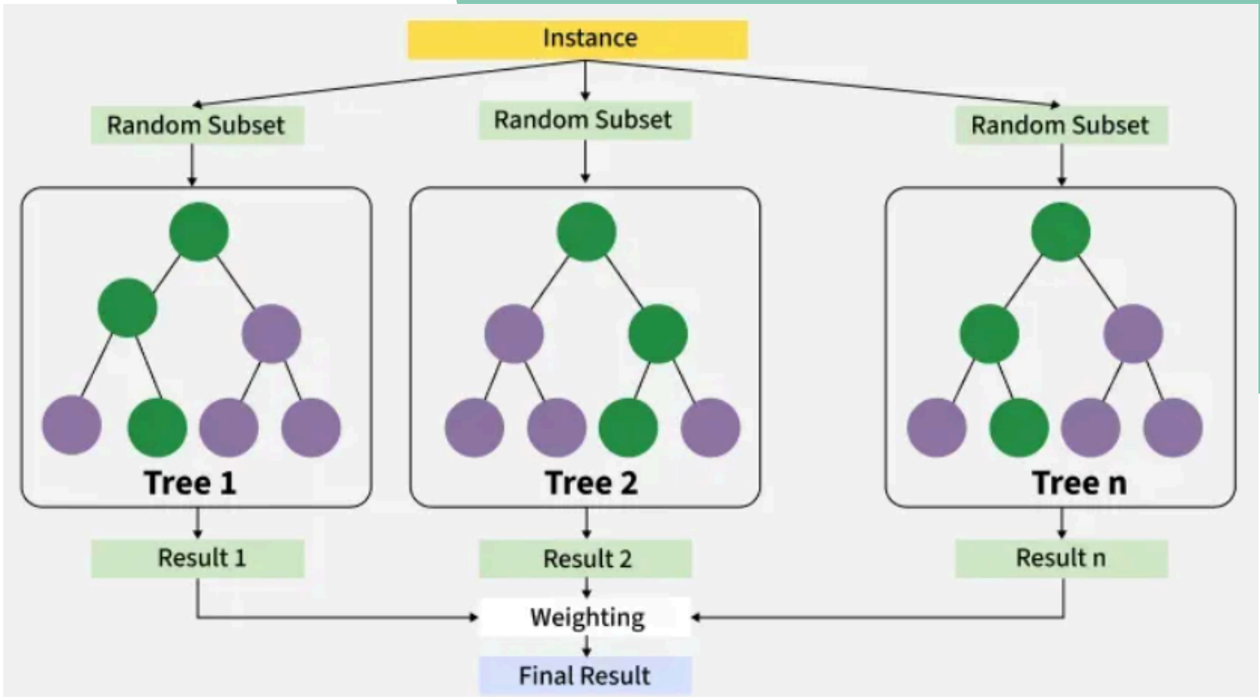
Model 3 : Support Vector Machine

- SVM is a supervised ML algorithm that finds the best decision boundary (hyperplane) separating high-risk and low-risk patients.
- Maximizes the margin between classes → reduces misclassification.
- Performs well on small to medium medical datasets.
- In this project:
- SVM predicts the probability of heart-disease risk for 1, 5, and 10 years.
- Works well when relationships among features (BP, cholesterol, HR, etc.) are not linear.
- Its probability output is compared against the risk threshold to determine if the patient triggers a critical alert.



Model 4 : XG Boost

- **XGBoost** is a high-performance tree-boosting algorithm that builds many small decision trees to improve accuracy.
- Learns complex feature interactions (cholesterol × age, BP × stress, etc.).
- Handles missing values and non-linear medical data extremely well.
- In this project:
- XGBoost produces the most accurate risk predictions among the four models.
- Provides probability scores used to assess 1-year, 5-year, and 10-year event risks.
- If the predicted risk crosses the critical threshold, the system generates a high-risk alert.
- Feature importance from XGBoost helps identify which medical factors drive the risk.



System Architecture

1. Data Acquisition:

   The system accepts a user-uploaded CSV dataset of patient records.

2. Data Pre-Processing:

   A multi-stage process cleans and prepares the data for the models (handles missing data, encodes labels, and scales features).

3. Feature Selection:

   The model is trained on a set of clinically relevant features (Age, Sex, Cholesterol, etc.) as identified in the data.

4. Model Training:

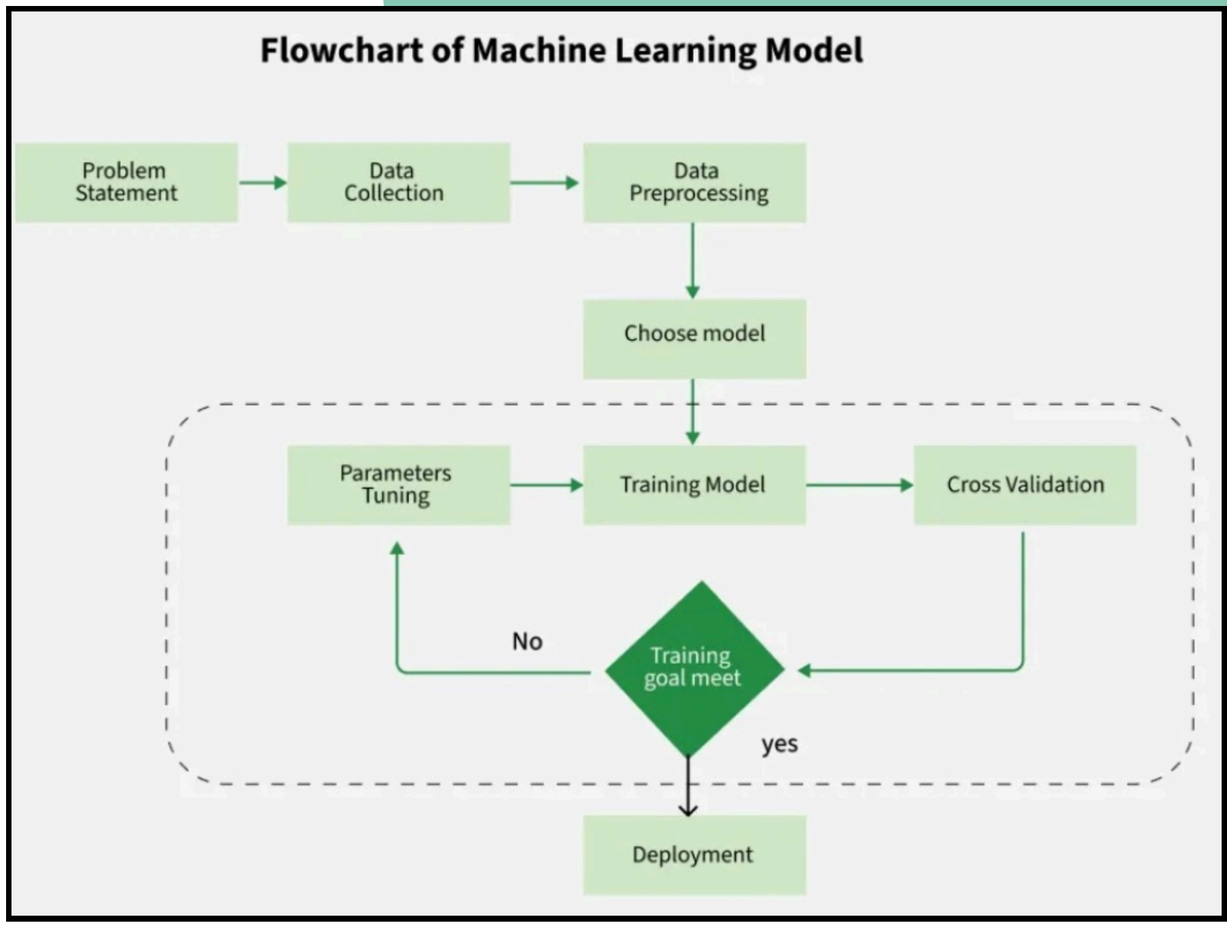
   The clean data is split into an 80% training set (to build the model) and a 20% test set (to evaluate it).

5. Logistic Regression:

   The training data is fed into the Logistic Regression algorithm to find patterns.

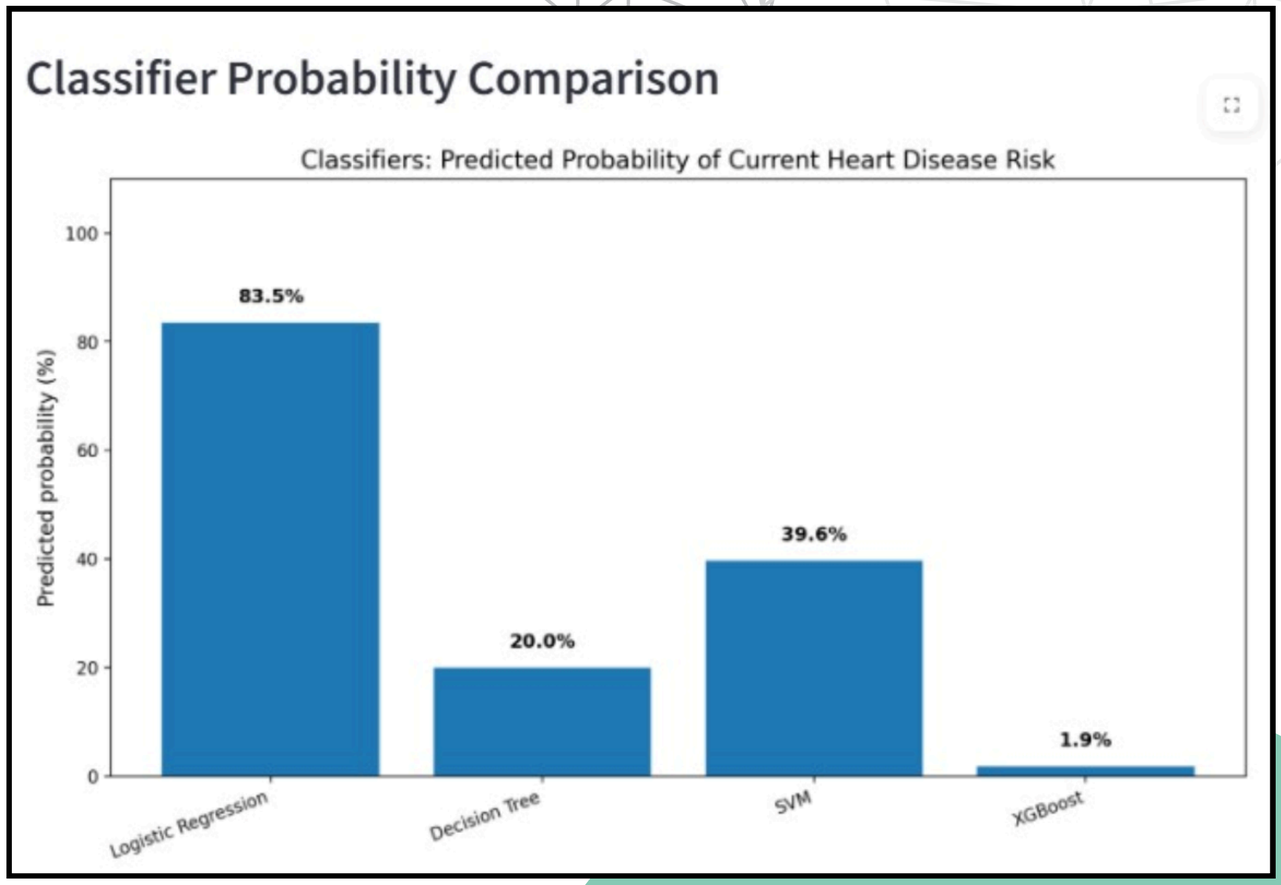
6. Prediction Result:

   The trained model generates a final classification: "High Risk" or "Low Risk".

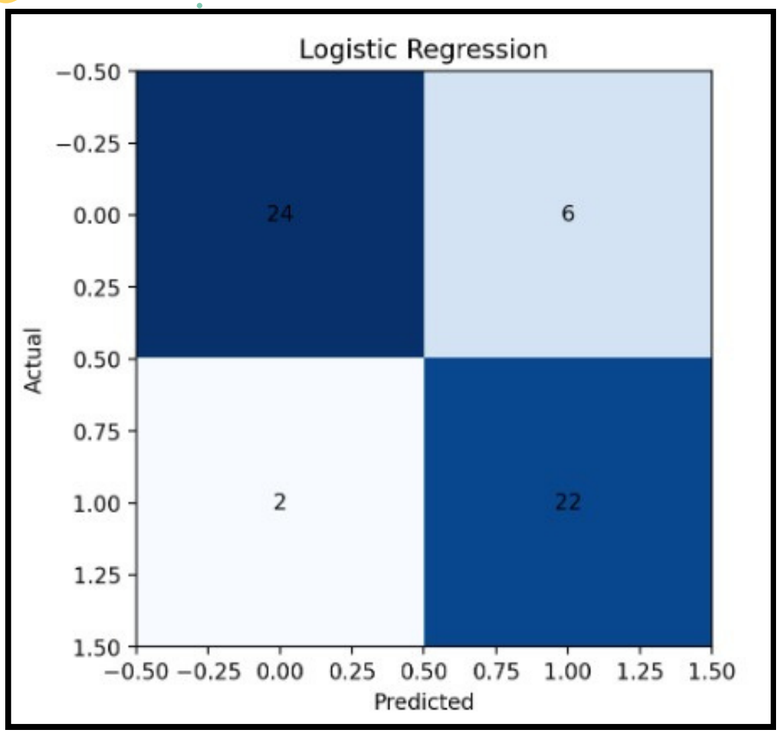


Results

- Logistic Regression: 83.5%
  - Highest predicted risk among all models
  - Exceeds typical clinical alert thresholds
  - Suggests strong likelihood of current heart disease presence
- Decision Tree: 20.0%
  - Low predicted risk
  - More conservative model output
- SVM: 39.6%
  - Moderate risk indication
  - Reflects a non-linear boundary detection
- XGBoost: 1.9%
  - Very low predicted risk
  - Produces the lowest probability among models



Results : Confusion Matrix (Test Set)

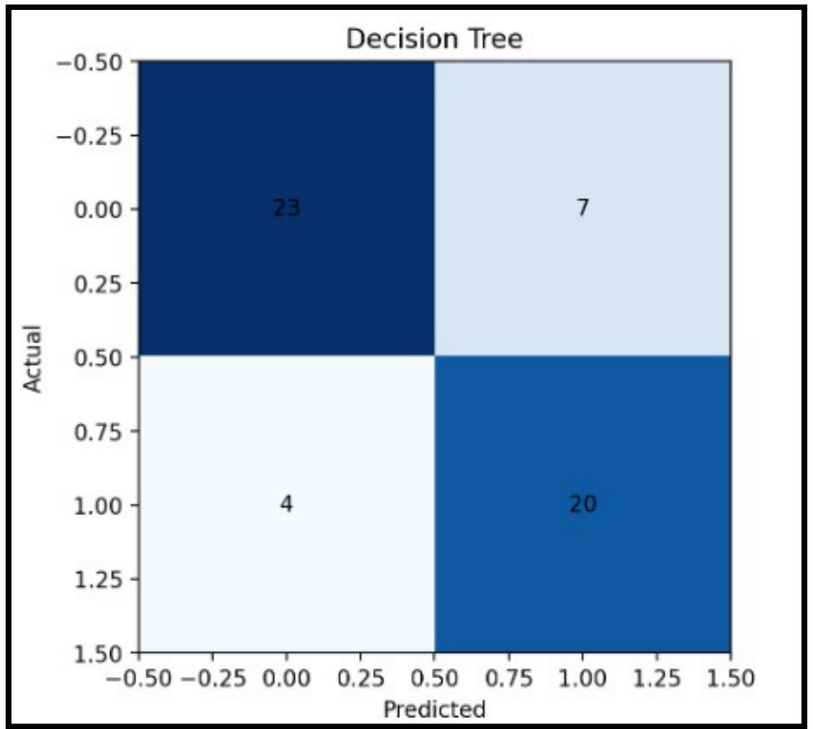


Key Insights : Logistic Regression(LR)

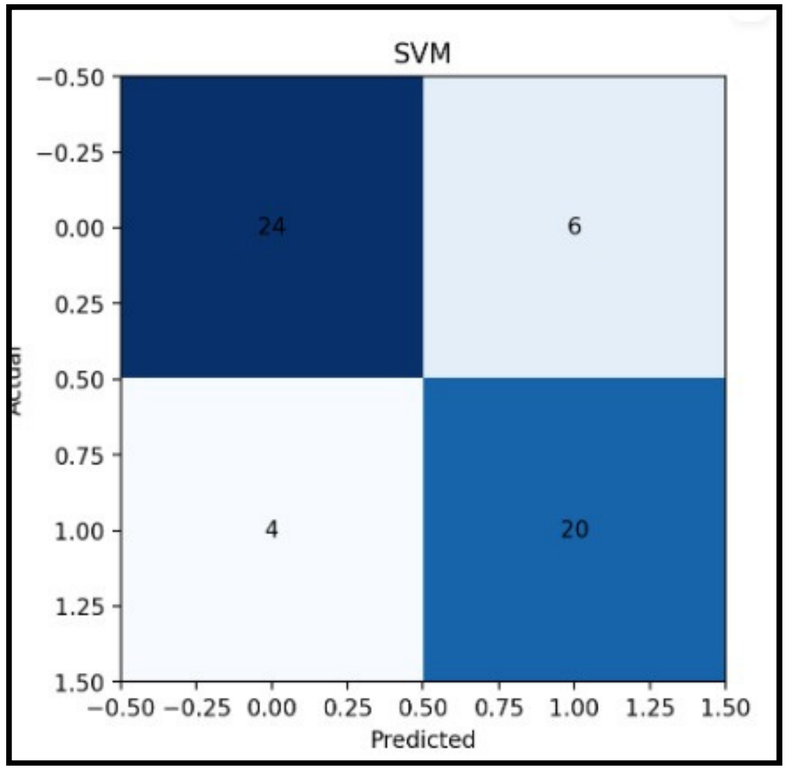
- **High Specificity :** The model correctly identified 24 out of 30 patients without heart disease.
- **Good Sensitivity :** It correctly identified 22 out of 24 patients with heart disease.
- **Low Error Rate :** Only 8 total misclassifications on the test set, showing a balanced and reliable model.

Key Insights : Decision Tree Classifier

- **High Specificity :** The model correctly identified 23 out of 30 patients without heart disease.
- **Good Sensitivity :** It correctly identified 20 out of 24 patients with heart disease.
- **Low Error Rate :** Only 11 total misclassifications on the test set, showing a moderately reliable model.



Results : Confusion Matrix (Test Set)

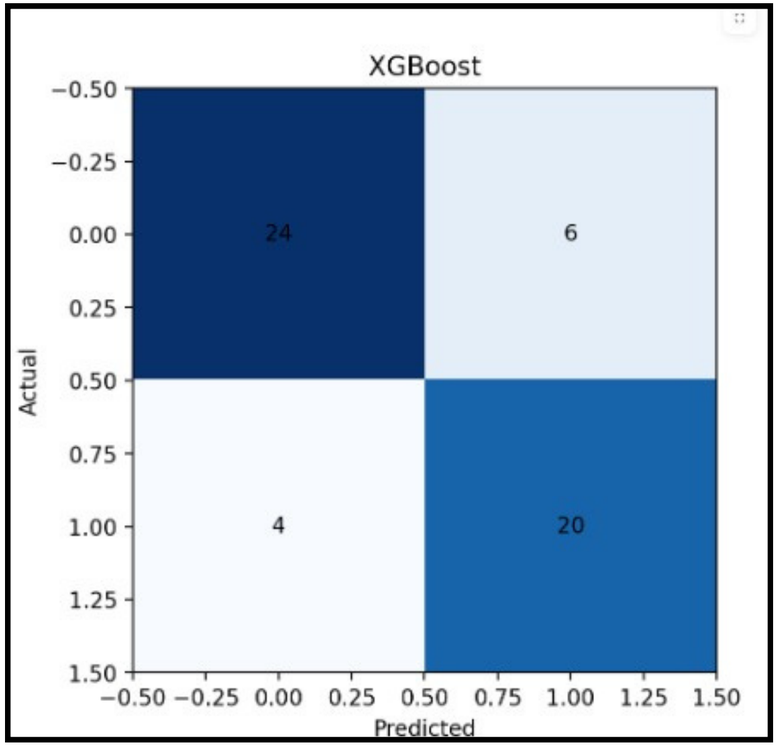


Key Insights : SVM(Support Vector Machine)

- **High Specificity :** The model correctly identified 24 out of 30 patients without heart disease.
- **Good Sensitivity :** It correctly identified 20 out of 24 patients with heart disease.
- **Low Error Rate :** Only 10 total misclassifications on the test set, showing a reliable model.

Key Insights : XG Boost

- **High Specificity :** The model correctly identified 31 out of 33 patients without heart disease.
- **Good Sensitivity:** It correctly identified 18 out of 21 patients with heart disease.
- **Low Error Rate:** Only 5 total misclassifications on the test set, showing a balanced and reliable model.



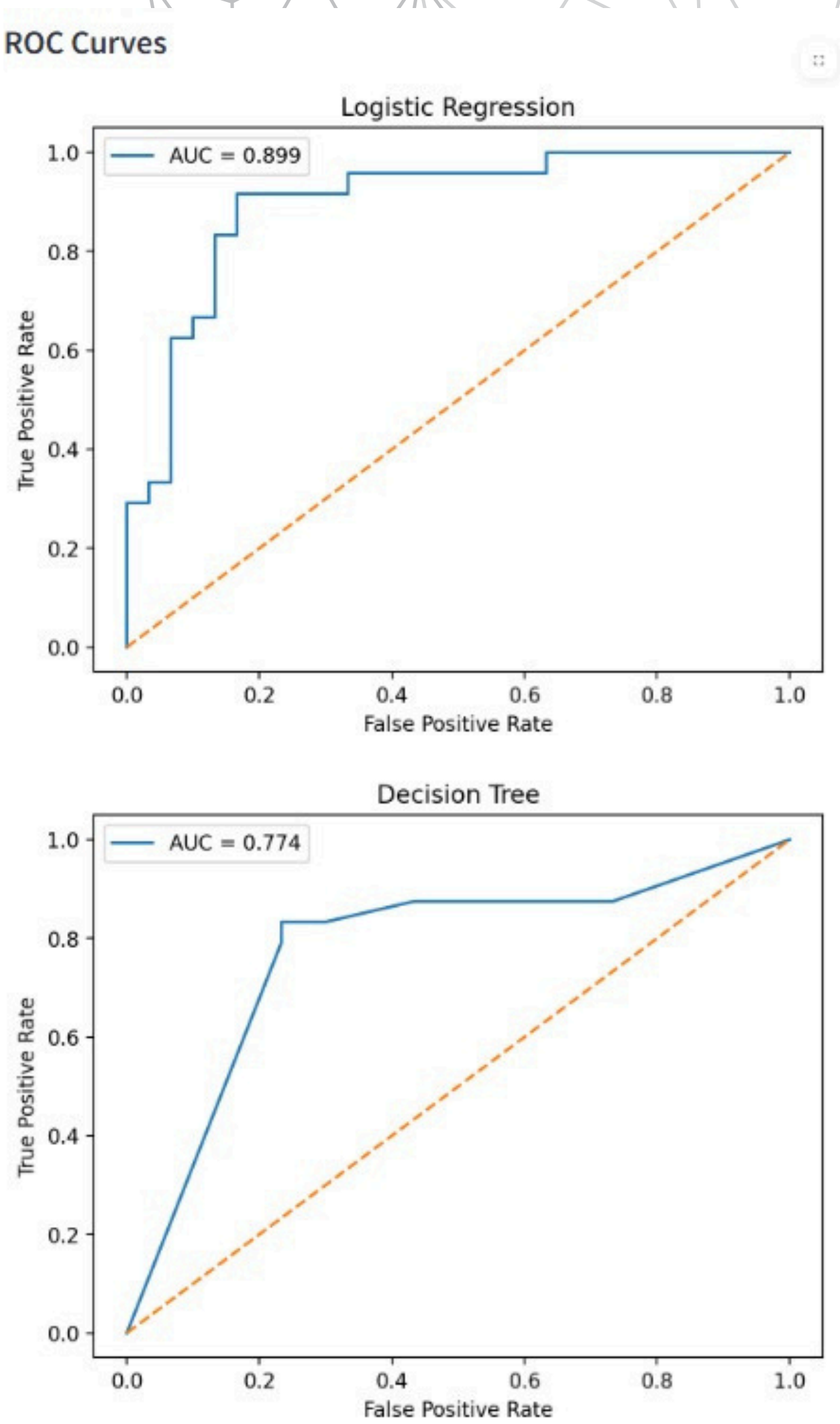
Results: ROC Curve Analysis

Logistic Regression (AUC = 0.899)

- High sensitivity and specificity — the ROC curve is very close to the top-left corner, so it accurately identifies most heart-risk patients.
- AUC 0.899 indicates strong overall performance, making it slightly better than SVM and XGBoost for this dataset.

Decision Tree (AUC = 0.774)

- Lower sensitivity — the ROC curve is farther from the top-left corner, so it misses more heart-risk cases.
- AUC 0.774 shows moderate accuracy, meaning it struggles with complex patterns compared to SVM and XGBoost.



Results: ROC Curve Analysis

SVM (AUC = 0.885)

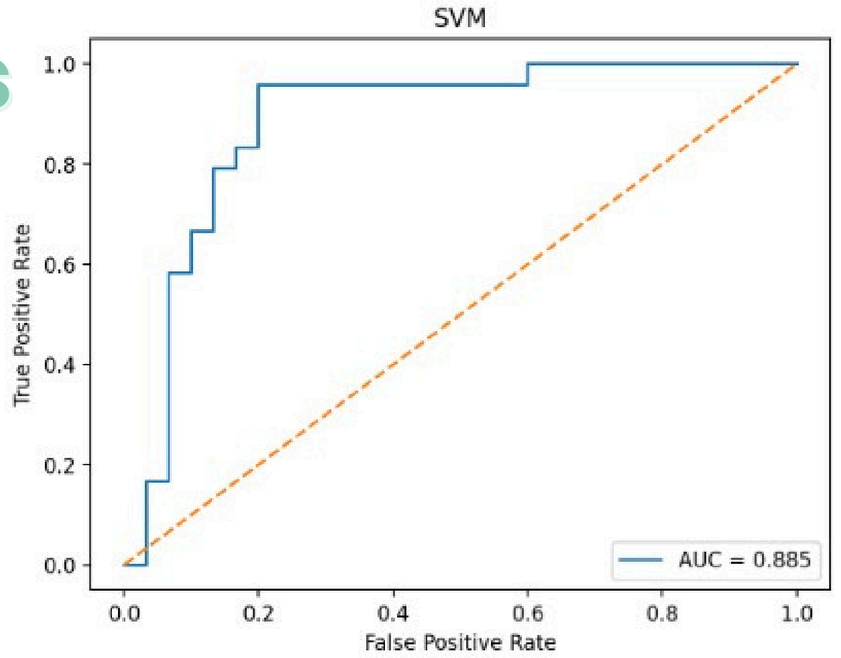
- SVM detects heart-risk patients more accurately because its ROC curve is closer to the top-left corner, showing higher sensitivity and fewer false positives.

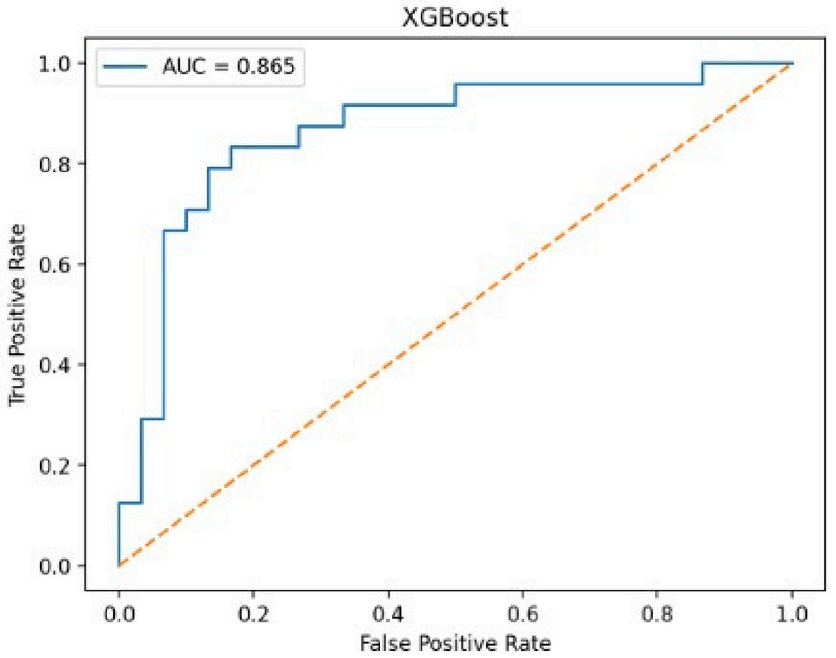
- SVM clearly separates healthy vs. heart-disease patients better than XGBoost

XGBOOST(AUC = 0.865)

- XGBoost performs well but is slightly weaker, since its ROC curve stays a little below SVM's curve, meaning it misses a few more risky patients.

- its sensitivity at different thresholds is lower compared to SVM, making it less ideal for medical risk detection.





Conclusion & Future Scope

Conclusion

The system is a promising and practical preliminary decision support tool for assessing heart disease risk.

- Logistic Regression is the superior model for stable probability estimation, crucial for clinical risk assessment.
- Decision Tree provides high interpretability, offering clear rule-based frameworks for understanding its predictions.

Future Scope

Enhancements to improve robustness and clinical relevance:

- Integrate sophisticated ensemble models (e.g., Random Forest, XGBoost).
- Incorporate explainability tools (like SHAP) to show which features influence each prediction.
- Add support for time-series patient data (historical readings).
- Enable EHR integration for real-time data access.

References

1. J. Zhang, "A Study of Heart Disease Prediction Based on Logistic Regression and CART Classification," 2024 5th International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), Shenzhen, China, 2024, pp. 720-724, doi: 10.1109/ICBAIE63306.2024.11117142.

2.N. Sabri et al., "HeartInspect: Heart Disease Prediction of an Individual Using Naïve Bayes Algorithm," 2023 IEEE 11th Conference on Systems, Process & Control (ICSPC), Malacca, Malaysia, 2023, pp. 350-354, doi: 10.1109/ICSPC59664.2023.10420149.

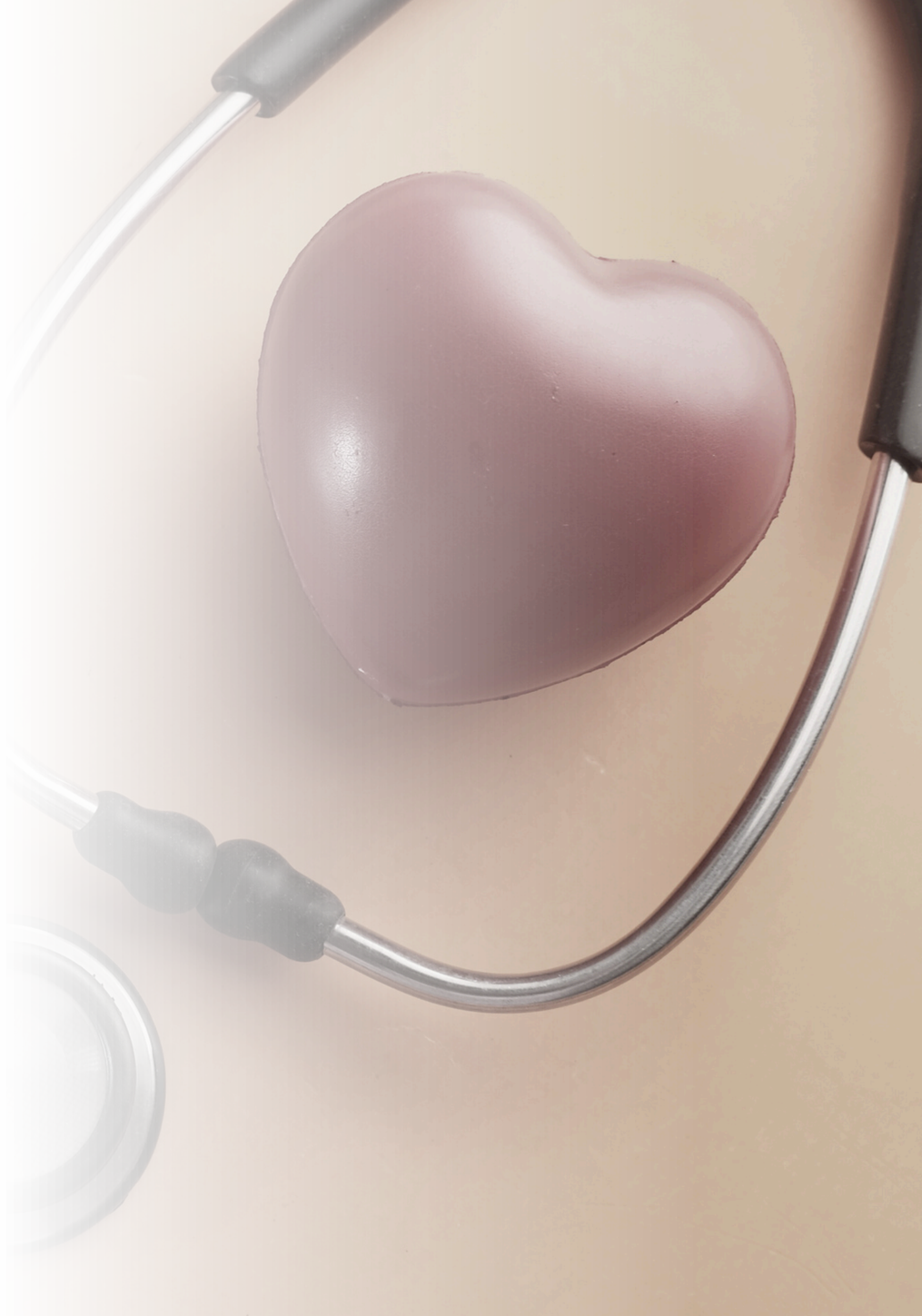
3.G. Shanmugasundaram, V. M. Selvam, R. Saravanan and S. Balaji, "An Investigation of Heart Disease Prediction Techniques," 2018 IEEE International Conference on System, Computation, Automation and Networking (ICSCA), Pondicherry, India, 2018, pp. 1-6, doi: 10.1109/ICSCAN.2018.8541165.

4.A. Rahim, Y. Rasheed, F. Azam, M. W. Anwar, M. A. Rahim and A. W. Muzaffar, "An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases," in IEEE Access, vol. 9, pp. 106575-106588, 2021, doi: 10.1109/ACCESS.2021.3098688.

5. Nishadi, Thanuja. “Predicting Heart Diseases In Logistic Regression Of Machine Learning Algorithms By Python Jupyterlab,” 2019.

6. D. G. Kleinbaum and M. Klein, Logistic Regression: A Self-Learning Text. New York, NY, USA: Springer, 2010.t

THANKYOU



Presented By:->

UCE2023402	Narmata Bhat
UCE2023403	Ichcha Bhat
UCE2023413	Kasak Boob