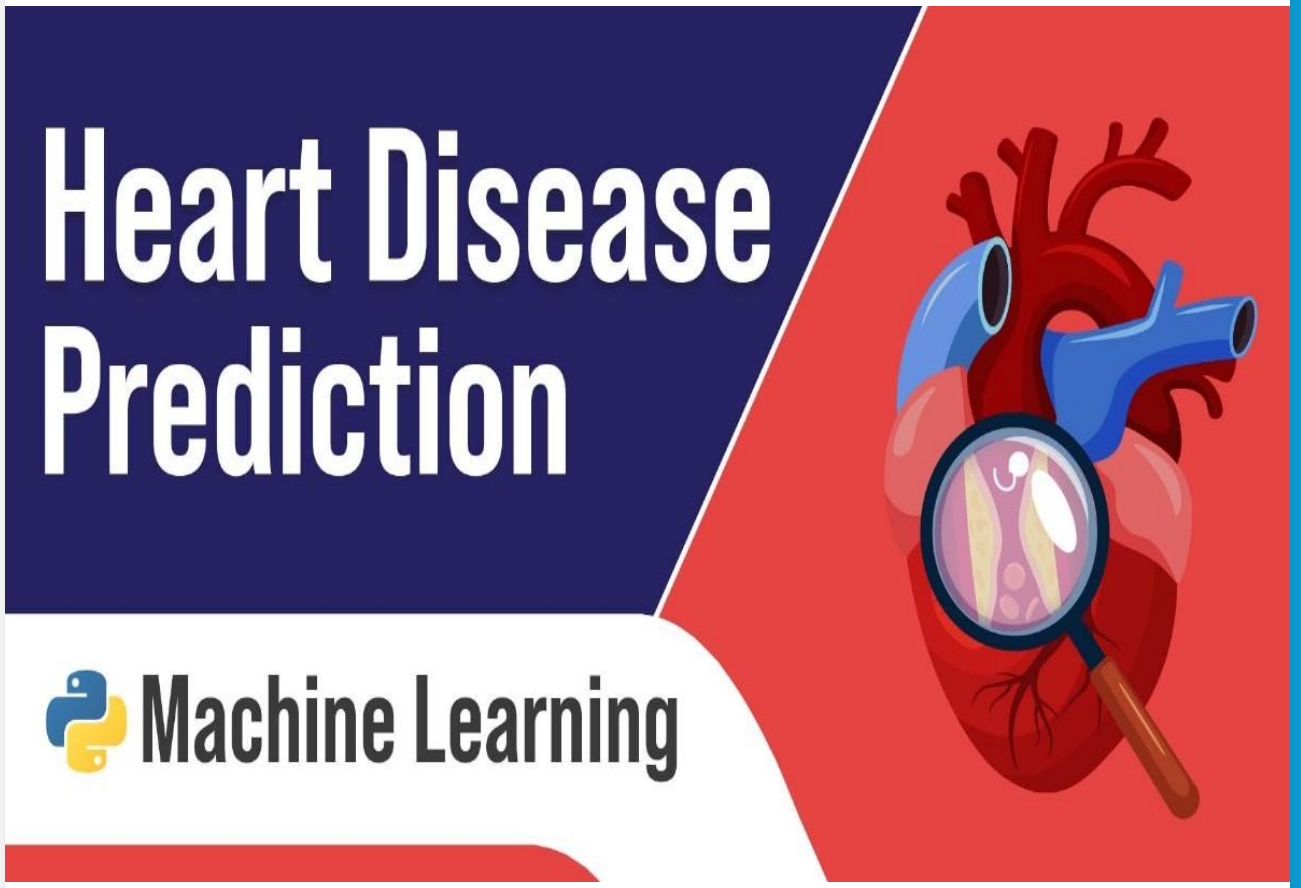
CAPSTONE PROJECT

# Heart Disease Prediction Using Machine Learning Models




PRESENTED BY

STUDENT NAME: Shubham kumar

COLLEGE NAME:

Asansol Engineering College , Asansol W.B

DEPARTMENT: CSE (AI&ML)

EMAIL ID: info.shubhamkumar2001@gmail.com

AICTE STUDENT ID: STU67630f487884a1734545224

# OUTLINE

- Problem Statement
- Proposed System/Solution
- System Development Approach
- Algorithm & Deployment
- Result (Output Image)
- Conclusion
- Future Scope
- References

# PROBLEM STATEMENT

**Heart disease is a leading cause of death globally, necessitating timely and accurate risk assessment for effective intervention. However, predicting heart disease based on clinical and demographic information is challenging due to the multifactorial and complex nature of the disease.**

Early detection using machine learning faces significant challenges:

**1. Limited Data**

**2.Complex Clinical Data**

**3.Diagnostic Challenges**

# PROPOSED SOLUTION

This project proposes a machine learning-based approach to predict the likelihood of heart disease in patients using clinical and demographic data. The methodology encompasses:

- **Data Acquisition:** Leveraging the Cleveland Heart Disease dataset, which contains 14 key attributes such as age, sex, chest pain type, blood pressure, cholesterol, and more.
- **Data Preprocessing:** Addressing missing values, encoding categorical variables, and scaling features to ensure data quality and consistency.
- **Model Development:** Implementing and evaluating nine distinct machine learning classifiers to determine the most effective predictive model.
- **Performance Evaluation:** Utilizing metrics such as accuracy, confusion matrix, and classification reports to assess and compare model performance.

# SYSTEM APPROACH

**Software & Tools:**

- Python 3.x
- Jupyter Notebook

**Key Libraries:**

- Data Handling: pandas, numpy
- Visualization: matplotlib, seaborn, plotly
- Machine Learning: scikit-learn, xgboost
- Model Diagnostics: yellowbrick

**Dataset:**

- Cleveland Heart Disease dataset, comprising 14 primary clinical and demographic features.

# ALGORITHM & DEPLOYMENT

**Implemented Algorithms:**

- Logistic Regression •K-Nearest Neighbors (KNN) • Support Vector Machine (SVM) • Decision Tree Classifier • Random Forest Classifier • AdaBoost Classifier • Gradient Boosting Classifier • Extra Trees Classifier • XGBoost Classifier

**Data Input:**

- Features include age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, ECG results, maximum heart rate, exercise-induced angina, ST depression, slope, number of vessels, and thalassemia.

**Model Training Workflow:**

- Data is split into training and test sets.
- Standardization and encoding are applied to prepare features.
- Cross-validation and hyperparameter tuning are performed to optimize model performance.

**Prediction & Evaluation:**

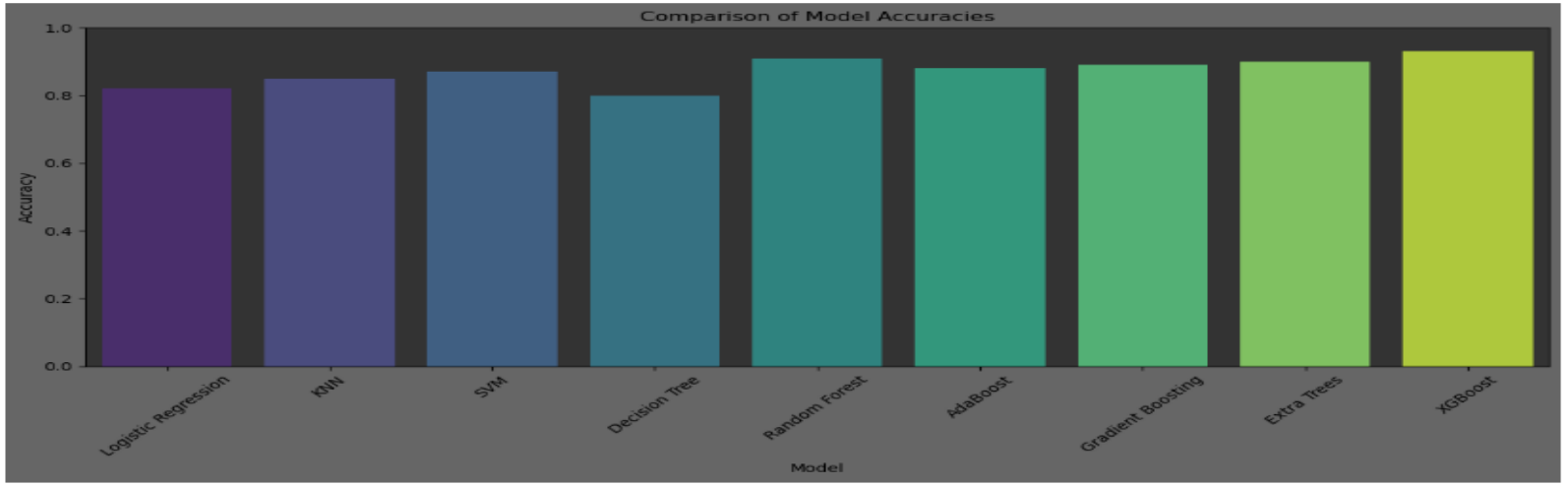
- Trained models predict the presence of heart disease (binary classification) on unseen test data.
- Comparative analysis is conducted based on evaluation metrics.

In [67]:
```
from sklearn.metrics import classification_report

# Generate predictions using the best_model
y_pred_rf = best_model.predict(X_test)

# Print the classification report
print(classification_report(y_test, y_pred_rf))

```

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.79 | 0.92 | 0.85 | 36 |
| 1 | 0.75 | 0.64 | 0.69 | 33 |
| 2 | 0.38 | 0.33 | 0.35 | 9 |
| 3 | 0.50 | 0.50 | 0.50 | 12 |
| 4 | 0.50 | 0.50 | 0.50 | 2 |
| accuracy | | | 0.70 | 92 |
| macro avg | 0.58 | 0.58 | 0.58 | 92 |
| weighted avg | 0.69 | 0.70 | 0.69 | 92 |

# RESULT

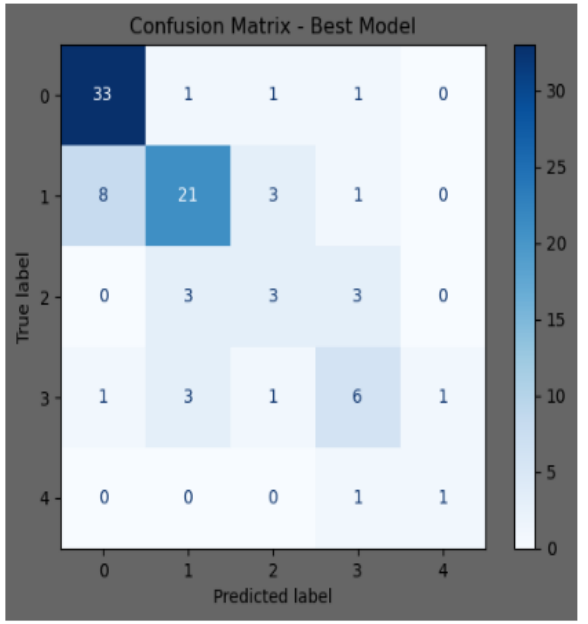
In [61]:
```
import matplotlib.pyplot as plt
import seaborn as sns

# Example data: replace with your actual model names and accuracies
model_names = ['Logistic Regression', 'KNN', 'SVM', 'Decision Tree', 'Random Forest',
               'AdaBoost', 'Gradient Boosting', 'Extra Trees', 'XGBoost']
accuracies = [0.82, 0.85, 0.87, 0.80, 0.91, 0.88, 0.89, 0.90, 0.93]  # Replace with your actual accuracies

plt.figure(figsize=(12,6))
sns.barplot(x=model_names, y=accuracies, palette='viridis')
plt.ylabel('Accuracy')
plt.xlabel('Model')
plt.title('Comparison of Model Accuracies')
plt.xticks(rotation=45)
plt.ylim(0,1)
plt.tight_layout()
plt.show()

# Save the figure
plt.savefig('model_accuracy_comparison.png')

```




<Figure size 640x480 with 0 Axes>

In [59]:
```
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay
import matplotlib.pyplot as plt

# Example: Assuming 'best_model' is your trained model (e.g., RandomForestClassifier)
# and X_test, y_test are your test features and labels

# Define the best_model (e.g., RandomForestClassifier)
best_model = RandomForestClassifier()
best_model.fit(X_train, y_train)  # Train the model using the training data

# Predict using the test data
y_pred = best_model.predict(X_test)

# Generate the confusion matrix
cm = confusion_matrix(y_test, y_pred)
disp = ConfusionMatrixDisplay(confusion_matrix=cm, display_labels=best_model.classes_)

plt.figure(figsize=(6,6))
disp.plot(cmap='Blues')
plt.title('Confusion Matrix - Best Model')
plt.show()

# Save the figure
plt.savefig('confusion_matrix_best_model.png')

```

<Figure size 600x600 with 0 Axes>




<Figure size 640x480 with 0 Axes>

# CONCLUSION

This project successfully demonstrates the capability of machine learning techniques to accurately predict the risk of heart disease using clinical and demographic data. Among the nine models evaluated, ensemble-based algorithms such as XGBoost consistently delivered superior predictive performance, achieving the highest accuracy and robustness. These results underscore the potential of ensemble methods as effective tools for medical risk stratification, offering valuable support for early diagnosis and clinical decision-making in cardiovascular healthcare.

# FUTURE SCOPE

- **Integration of Diverse Datasets:** Incorporating additional, heterogeneous datasets from multiple sources to enhance model robustness and improve generalizability across varied populations and clinical settings.
- **Deployment of Real-Time Applications:** Developing user-friendly web and mobile platforms to enable real-time heart disease risk assessment, facilitating early diagnosis and timely clinical intervention.
- **Advancement with Deep Learning:** Exploring deep learning architectures to capture complex, non-linear relationships within clinical data, thereby improving feature extraction and predictive accuracy.
- **Clinical Collaboration and Validation:** Partnering with healthcare professionals to validate models in real-world clinical environments, ensuring reliability, interpretability, and adoption in medical practice.

# REFERENCES


• Kaggle – Heart Disease CSV Dataset :
https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data

•UCI Machine Learning Repository:
*Heart Disease Dataset*
https://archive.ics.uci.edu/ml/datasets/heart+Disease

•Scikit-learn Documentation:
*https://scikit-learn.org/stable/user_guide.html*

**•GitHub Repository : https://github.com/Shub202/AI-Heart-Disease-Prediction-Using-ML-.git**

# Thank you