Machine Learning Project.

Group members: Ramiz Allahverdiyev, Jalil Verdiyev, Parviz Mehrali, Rashad Qarayev, Xanahmad Mammadov.

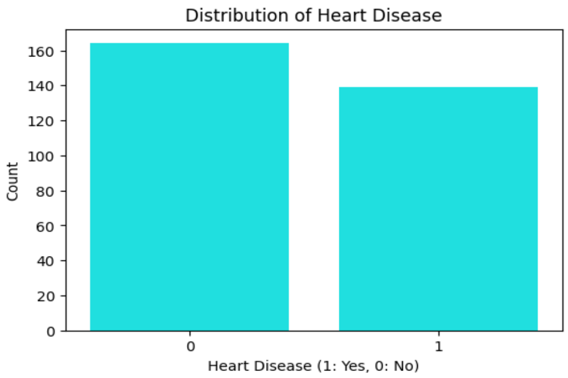
Predictive Modeling for Heart Disease Diagnosis Using Machine Learning.

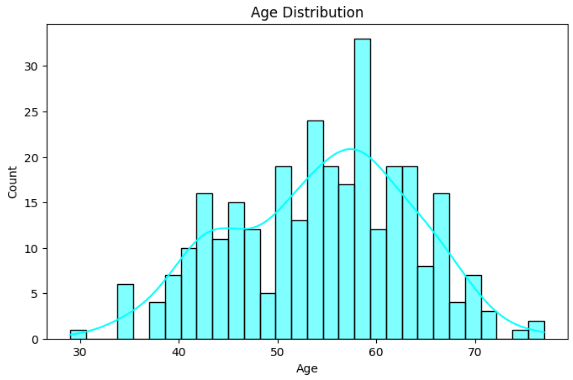
1. Introduction

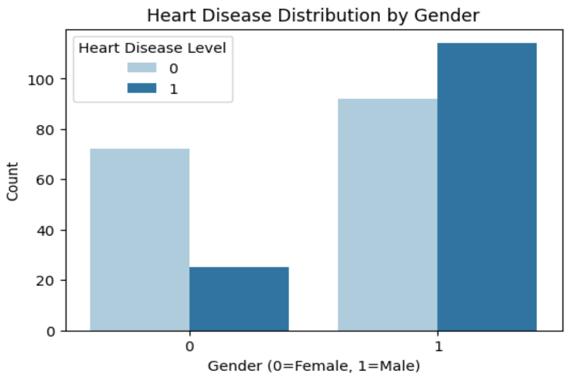
Cardiovascular diseases are still the leading cause of mortality worldwide. Early diagnosis is essential to reducing the risk of severe outcomes such as heart attacks or strokes. In this project, we tackle the binary classification problem of predicting the presence of heart disease using clinical features. Our solution uses various machine learning algorithms to compare performance and figure out the most reliable model. The primary goal is not only to optimize predictive accuracy but also to analyze model behavior through real-world metrics like ROC AUC, precision, and recall.

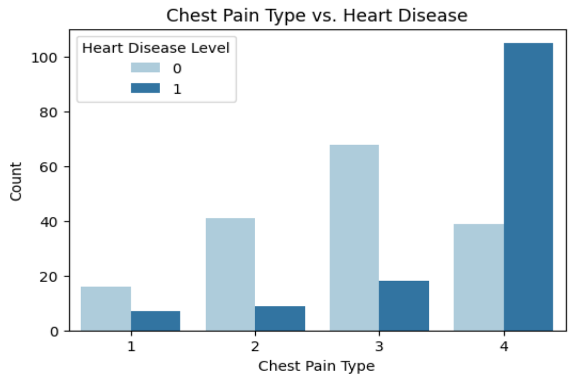
2. Data Processing and Visualizations

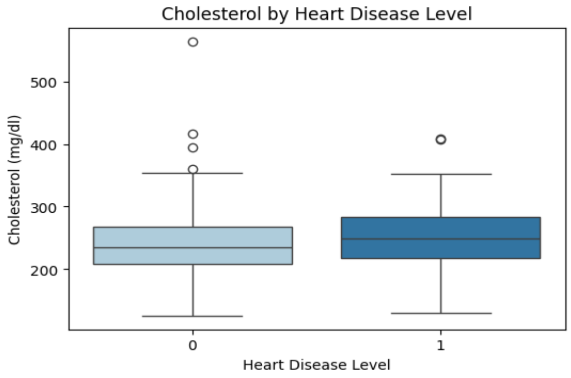
The dataset, sourced from the UCI Machine Learning Repository, includes 13 original features and one target label (num) showing heart disease status (0: no disease, 1: present disease). Initial data cleaning revealed missing values in the ‘ca’ and ‘thal’ columns, which were imputed using the mode due to their categorical nature.

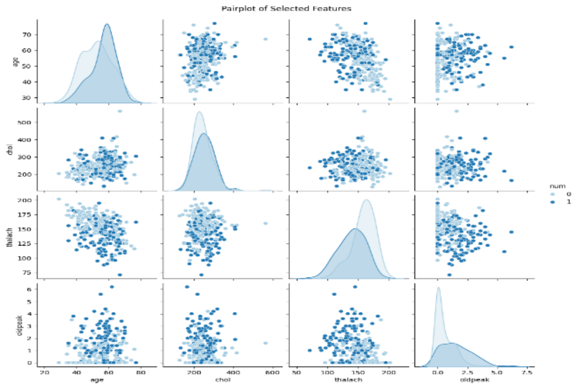


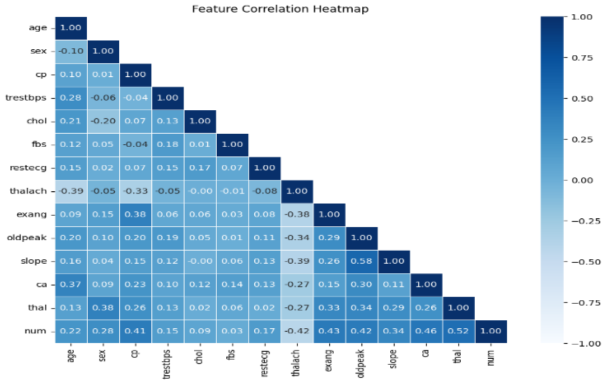












3. Methodology

Four machine learning models were developed and evaluated. These models were Logistic Regression, KNN, SVC and Random Forest.

All models were implemented using scikit-learn libraries. Feature scaling was applied where necessary using StandardScaler. The dataset was split into training and testing sets with an 80:20 ratio. To enhance the reliability of evaluation metrics, we used StratifiedKFold cross-validation (k=5), which supports the target class distribution across folds.

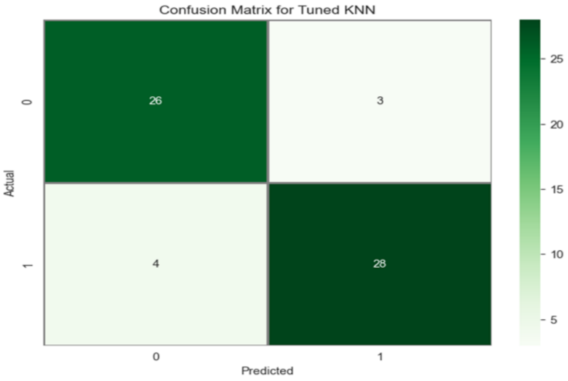
Hyperparameter tuning was conducted via GridSearchCV within each cross-validation fold for models like Random Forest, SVC, and KNN. Evaluation metrics included ROC AUC Score, Precision, Recall, F1-Score, Accuracy and Confusion matrix.

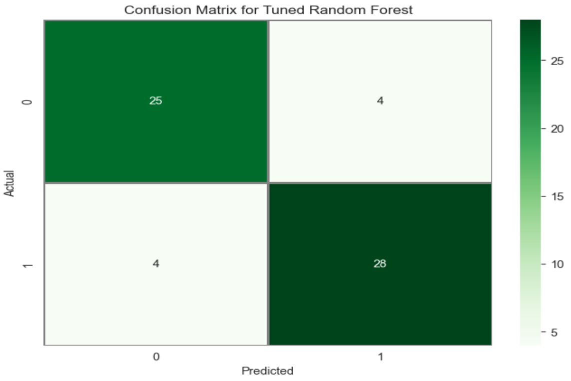
The primary metric for model comparison was ROC AUC, which provides a balanced evaluation of performance across both classes, particularly important in medical diagnostics where false positives and false negatives have different real-world costs.

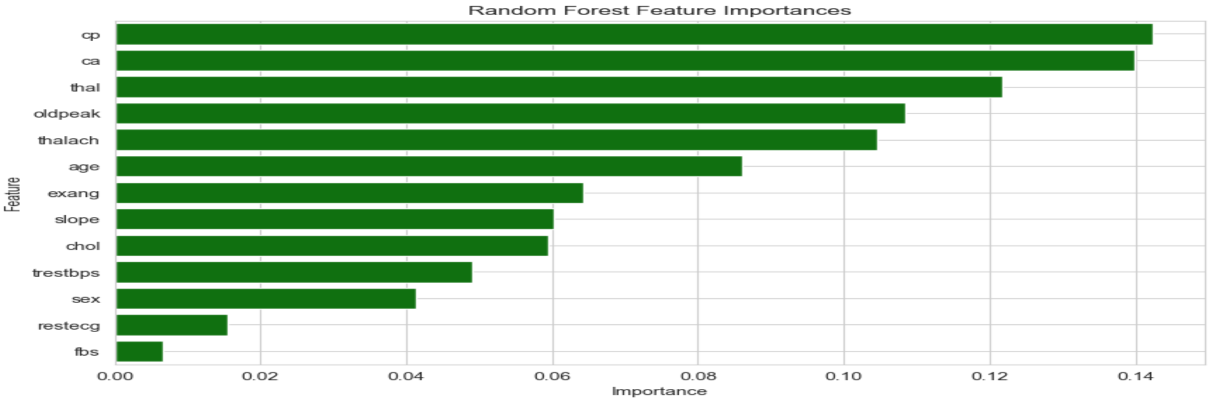
4. Results and Discussion

Comparative Metrics Summary

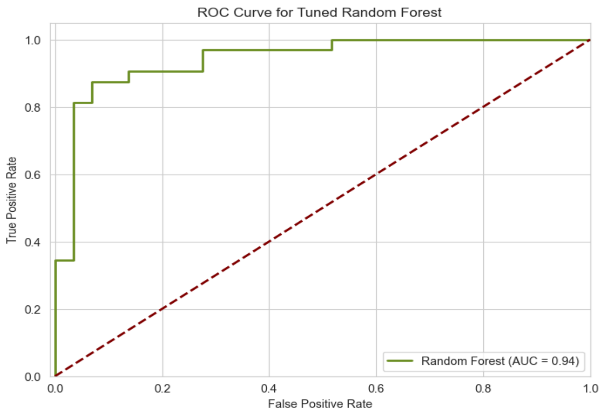
Model	ROC AUC	Precision	Recall	F1-score	Accuracy
KNN	0.94	0.94	0.91	0.92	0.918
Random Forest	0.94	0.96	0.84	0.90	0.902
SVC	0.93	0.93	0.88	0.90	0.902
Logistic Regression	0.94	0.88	0.91	0.89	0.885

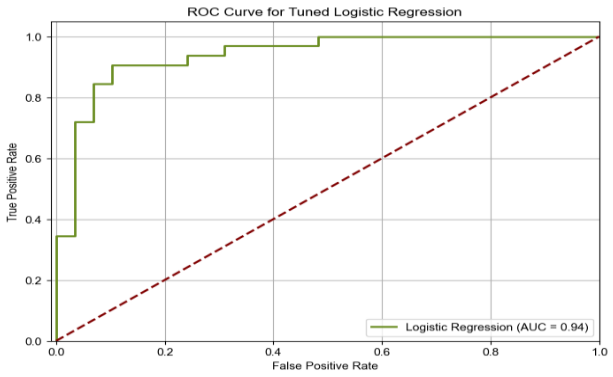

Confusion Matrix for Tuned KNN

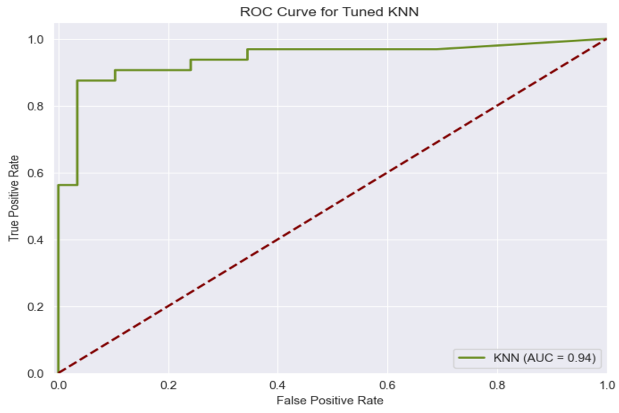

Confusion Matrix for Tuned Random Forest

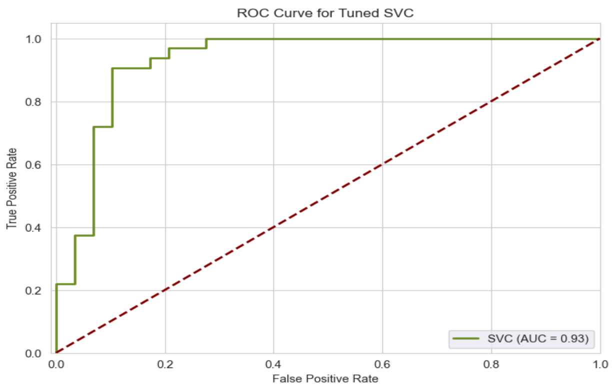

Random Forest Feature Importances

The Random Forest model's feature importance analysis reveals that chest pain type (cp) and the number of major vessels (ca) are the most influential factors in predicting heart disease. Other key indicators include thallium stress test results (thal) and exercise-induced ST depression (oldpeak). Conversely, features such as resting ECG results (restecg) and fasting blood sugar (fbs) showed minimal importance in the model's predictive capability.









All models showed high AUC values (~0.94), making AUC an insufficient differentiator alone.

Therefore, secondary metrics were critical:

1. Random Forest (AUC 0.94): Excels at finding heart disease patients while minimizing false positives (incorrectly diagnosing healthy people), especially at the critical first stages. This is highly valuable when avoiding unnecessary patient anxiety and follow-up tests is a priority.
2. Logistic Regression (AUC 0.94): Offers a strong, balanced performance, like Random Forest. Its simplicity and interpretability make it easier for clinicians to understand why a diagnosis was made, fostering trust in the model's recommendations.
3. KNN (AUC 0.94): Performs robustly, but its "stepped" curve reflects its reliance on patient similarity. It's effective in scenarios where distinct patient profiles exist and historical data can guide diagnosis.
4. SVC (AUC 0.93): Although slightly lower AUC, it is still highly effective. Its strength lies in its ability to manage complex data patterns and distinguish classes even when relationships are intricate, making it useful for challenging or novel patient presentations.

In essence, while all models are great at finding heart disease, their specific curve shapes highlight their strengths in managing the balance between catching all cases and avoiding false alarms, offering different practical advantages depending on the clinical priority:

- KNN, with the highest accuracy (0.918), also supported balanced precision and recall, suggesting reliability in both detecting disease and minimizing false alarms.
- Random Forest had superior precision (0.96), meaning it was least likely to generate false positives. In a clinical context, this is valuable when misdiagnosing healthy individuals could lead to unnecessary procedures.
- Logistic Regression had a high recall (0.91), making it preferable when not detecting an actual disease (false negatives) is costlier.

Practical Scenarios:

- High Recall (Logistic Regression): Ideal for general health screening where finding every potential heart disease case is crucial, even at the cost of some false positives.
- High Precision (Random Forest): Suitable for resource-intensive procedures (e.g., angiography), where a false positive leads to unnecessary and costly intervention.
- Balanced Performance (KNN): Best for real-time wearable health monitors where both accuracy and minimal false alerts are necessary to ensure user trust and safety.

5. Conclusion

This project applied multiple machine learning algorithms to classify heart disease presence using clinical data. After thorough preprocessing, feature engineering, and model optimization, KNN appeared as the top-performing model, though others had unique strengths.

Working with real-world medical data highlighted the importance of carefully selecting performance metrics that align with application goals. For example, high recall models help in early detection, while high precision models reduce unnecessary treatment.

If more time were available, we would:

- Expand hyperparameter grids.
- Introducing ensemble techniques (stacking/blending).
- Collect temporal data to detect disease progression.