Detailed Report: Heart Disease Prediction Project

1. Project Overview

This project aims to build a machine learning system to predict whether a patient has heart disease based on medical attributes. It is a **binary classification problem** where the target variable is target (0 = No Heart Disease, 1 = Heart Disease). The dataset contains 1,190 entries and 12 features, including age, cholesterol, blood pressure, chest pain type, and more.

---

2. Approach

The project follows a structured machine learning workflow:

- **Data Preprocessing**: Handling missing values, scaling, encoding categorical variables.
- **Exploratory Data Analysis**: Basic data checks and understanding.
- **Model Training**: Four models are trained and evaluated:
    1. Logistic Regression
    2. Random Forest
    3. XGBoost
    4. Support Vector Machine (SVM)
- **Hyperparameter Tuning**: Applied to Random Forest, XGBoost, and SVM (though not fully implemented in the provided code).
- **Evaluation**: Models are evaluated using accuracy, classification reports, and confusion matrices.

---

3. Data Preprocessing

Features and Target

- **Target**: target (binary)
- **Features**:
    - **Categorical**: chest pain type, resting ecg, ST slope
    - **Binary**: sex, fasting blood sugar, exercise angina
    - **Numeric**: age, resting bp s, cholesterol, max heart rate, oldpeak

Preprocessing Pipelines

- **Numeric Pipeline**: Imputation (median) + Standard Scaling
- **Categorical Pipeline**: Imputation (most frequent) + One-Hot Encoding
- **Binary Pipeline**: Imputation (most frequent) only (no encoding needed)

A ColumnTransformer combines these pipelines for full preprocessing.

---

4. Model Training and Evaluation

Models Used

1. **Logistic Regression**: Linear baseline classifier.
2. **Random Forest**: Ensemble of decision trees.
3. **XGBoost**: Gradient boosting algorithm.
4. **SVM**: Effective for complex decision boundaries.

Results Summary

Model	Accuracy	Precision (0)	Recall (0)	F1-Score (0)	Precision (1)	Recall (1)	F1-Score (1)
Logistic Regression	0.8613	0.87	0.83	0.85	0.85	0.89	0.87
Random Forest	0.9244	0.91	0.93	0.92	0.94	0.92	0.93
XGBoost	0.9160	0.91	0.91	0.91	0.92	0.92	0.92
SVM	0.8782	0.88	0.86	0.87	0.88	0.90	0.89

Key Observations

- **Random Forest** achieved the highest accuracy (92.44%), followed closely by **XGBoost** (91.60%).

- **Logistic Regression** and **SVM** performed reasonably well but lagged behind ensemble methods.
- All models show balanced performance across both classes (No Heart Disease and Heart Disease).

---

5. Confusion Matrix Insights

- **Logistic Regression**: 93 correct negatives, 19 false positives, 14 false negatives, 112 correct positives.
- **Random Forest**: 104 correct negatives, 8 false positives, 10 false negatives, 116 correct positives.
- **XGBoost**: 102 correct negatives, 10 false positives, 10 false negatives, 116 correct positives.
- **SVM**: 96 correct negatives, 16 false positives, 13 false negatives, 113 correct positives.

Random Forest and XGBoost minimized false positives and false negatives most effectively.

---

6. Model Comparison

Accuracy Ranking

1. **Random Forest**: 92.44%
2. **XGBoost**: 91.60%
3. **SVM**: 87.82%
4. **Logistic Regression**: 86.13%

Random Forest is the best-performing model for this dataset.

---

7. Limitations and Future Work

Limitations

- **Hyperparameter Tuning**: Although mentioned, hyperparameter tuning was not implemented in the provided code.
- **Feature Engineering**: No explicit feature engineering or selection was performed.

- **Class Imbalance**: The dataset appears balanced, but further analysis could confirm this.

Future Improvements

1. **Hyperparameter Tuning**: Use GridSearchCV or RandomizedSearchCV to optimize model parameters.
2. **Cross-Validation**: Implement k-fold cross-validation for more robust evaluation.
3. **Feature Importance**: Analyze feature importance (especially for tree-based models) to identify key predictors.
4. **Advanced Models**: Experiment with neural networks or other advanced algorithms.
5. **Deployment**: Develop a web app or API for real-time predictions.

---

8. Conclusion

The Heart Disease Prediction project successfully demonstrates the application of multiple machine learning models to predict heart disease. **Random Forest** emerged as the best model with **92.44% accuracy**, followed closely by **XGBoost**. The project highlights the importance of preprocessing and model selection in achieving high predictive performance. Future work should focus on hyperparameter tuning and model deployment for practical use.

---

9. Appendix

Libraries Used

- pandas, numpy: Data manipulation
- matplotlib, seaborn: Visualization
- sklearn: Preprocessing, model training, evaluation
- xgboost: XGBoost classifier

Dataset

- **Source**: dataset.csv
- **Size**: 1,190 samples, 12 features
- **Missing Values**: None (as per initial check)