Predicting Cardiovascular Disease Risk using Supervised Machine Learning Models

Atul Kumar Tiwary
Khoury College of Computer Sciences
Northeastern University
Boston, MA, USA
tiwary.a@northeastern.edu

***Abstract*—This project builds and evaluates a supervised machine learning pipeline to predict cardiovascular disease (CVD) using the Kaggle Cardiovascular Dataset. I handled the full workflow for this project, including data preprocessing, exploratory analysis, feature engineering, model training, and finally, model calibration and interpretation. Three models : Logistic Regression, Decision Tree, and Random Forest : were implemented with hyperparameter tuning via GridSearchCV. Model performance was evaluated using accuracy, precision, recall, F1-score, and AUC. Random Forest achieved the best discrimination (AUC = 0.7958), while Logistic Regression provided the most reliable probability estimates (Brier = 0.188). The feature-importance results highlighted systolic and diastolic blood pressure, age, cholesterol, and BMI as key predictors, which also matched expectations from clinical guidance discussed in course. This project shows that classical supervised learning models can be effective for early CVD risk prediction, and I specifically focused on the trade-offs between a model's predictive power, its interpretability, and its calibration.**

***Index Terms*—cardiovascular disease prediction, supervised learning, logistic regression, decision tree, random forest, model calibration, healthcare analytics**

I. Introduction

Cardiovascular disease (CVD) is a leading cause of death globally, which makes early risk detection a critical part of preventive healthcare. Machine learning can capture relationships among health parameters that may not be obvious from standard medical rules, which is why I chose to evaluate it for automated risk prediction here. My paper explores multiple supervised ML algorithms to predict CVD using structured health indicators such as age, blood pressure, cholesterol, and body mass index (BMI).

This project applies concepts from my CS6140 Machine Learning course, specifically modules covering supervised classification, model tuning, and calibration [1]. The objective is to evaluate traditional classifiers for medical prediction tasks and analyze their balance between accuracy, interpretability, and calibration reliability.

II. Dataset and Exploratory Analysis

A. Dataset Description

The dataset used is the publicly available *Cardiovascular Disease Dataset* from Kaggle, containing 70,000 records with 13 attributes such as age (in days), gender, height, weight, systolic/diastolic blood pressure, cholesterol, glucose, and physical activity indicators. The target variable (`cardio`) denotes presence (1) or absence (0) of cardiovascular disease.

B. Data Cleaning and Preparation

Exploratory Data Analysis (EDA) revealed no missing values or duplicates. Age was converted from days to years, BMI was derived from height and weight, and categorical variables (`cholesterol`, `gluc`) were encoded. I scaled numeric variables using RobustScaler after noticing outliers in blood pressure and weight distributions, which affected some early model runs. The data was split into 80% training and 20% testing sets with stratified sampling to maintain class balance.

C. Statistical Summary

After cleaning, the dataset had the following properties:

- Mean age: 52 years; mean BMI: 26.9
- Balanced classes: 49.9% negative, 50.1% positive
- Outliers capped using IQR-based filtering for blood pressure and weight

The overall preprocessing ensured numerical stability and prevented model bias toward extreme measurements.

III. Model Training and Evaluation

A. Methodology

Three supervised models were trained:

1) Logistic Regression – baseline linear classifier
2) Decision Tree – interpretable non-linear classifier
3) Random Forest – ensemble model for robust predictions

I tuned each model's hyperparameters using GridSearchCV with five-fold cross-validation. Evaluation metrics included accuracy, precision, recall, F1-score, and Area Under the ROC Curve (AUC).

B. Feature Selection

I used two methods to find the most important features: SelectKBest (with an ANOVA F-test) and Recursive Feature Elimination (RFE). The most significant predictors were: `ap_hi`, `ap_lo`, `age_years`, `cholesterol`, `BMI`, and `gluc`. These features align with medical literature highlighting blood pressure and lipid levels as core CVD indicators.

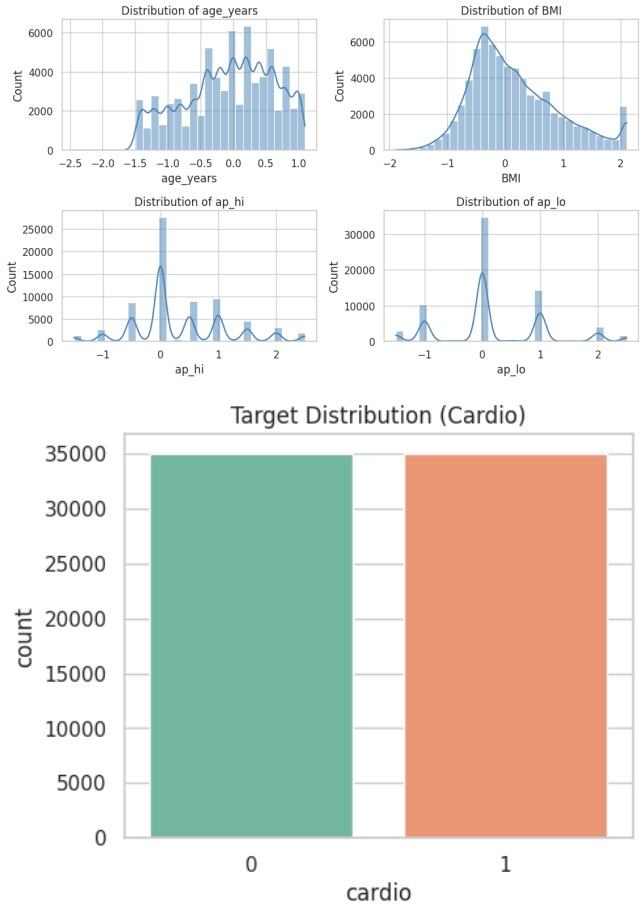

Fig. 1. Distribution of key features (*age_years*, *BMI*, *ap_hi*, *ap_lo*) and target balance after preprocessing.

TABLE I
PERFORMANCE COMPARISON OF CLASSIFIERS

Model	Accuracy	Precision	Recall	F1	AUC
Logistic Reg.	0.725	0.750	0.676	0.711	0.788
Decision Tree	0.725	0.726	0.721	0.724	0.783
Random Forest	**0.731**	**0.749**	**0.695**	**0.721**	**0.796**

C. Results

Table I summarizes the model performance metrics.

Random Forest achieved the highest accuracy and AUC, showing its advantage in capturing complex interactions, while Logistic Regression remained competitive and highly interpretable.

IV. Model Calibration and Interpretability

To assess reliability of predicted probabilities, calibration plots and Brier scores were computed. Random Forest produced the lowest Brier score (0.183), followed closely by Decision Tree (0.187) and Logistic Regression (0.188). Despite Random Forest's better calibration numerically, Logistic Regression offered smoother reliability curves in my plots, which made me more comfortable with its probability estimates even though Random Forest scored slightly better numerically.

From Fig. 3, `ap_hi` stood out, which aligns with what Prof. Tala mentioned in course lecture regarding systolic pressure acting as a high-variance predictor.

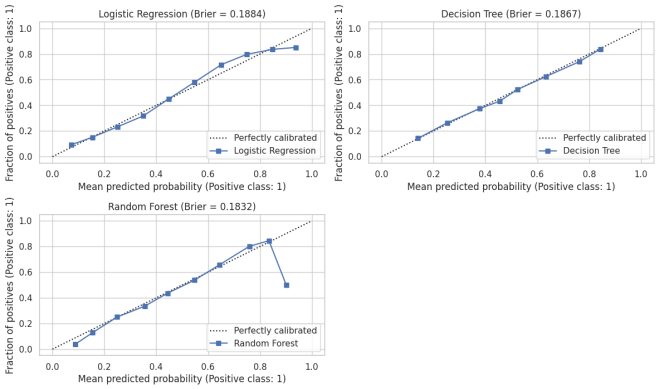

Fig. 2. Calibration curves showing probability reliability for Logistic Regression, Decision Tree, and Random Forest models. Logistic Regression demonstrates the most consistent probability alignment, while Random Forest achieves slightly better overall Brier score performance.

V. Discussion

The Random Forest model offered superior predictive capability, while Logistic Regression provided explainable and stable calibration. The Decision Tree helped me interpret which features mattered the most before moving on to the more complex ensemble model.

A. Efficiency Considerations

Random Forest incurred higher training time due to ensemble complexity, whereas Decision Tree and Logistic Regression trained faster with less memory overhead. All models achieved near real-time inference, suitable for integration into healthcare decision support systems.

B. Limitations

The dataset represents observational data and may include measurement noise in blood pressure or weight values. If time allowed, I would also try external validation datasets and gradient boosting methods like XGBoost or LightGBM to see if they generalize better.

VI. Conclusion and Future Work

This project demonstrated that classical supervised ML models can be effectively applied to predict CVD risk. The analysis balanced predictive accuracy, interpretability, and calibration quality. Key findings show that blood pressure, cholesterol, BMI, and age are strong predictors of cardiovascular disease.

For future work, the model could be improved by adding more data, like lifestyle or genetic features. It would also be valuable to try cross-institutional validation or explore stacking models. Such extensions can further improve the accuracy and clinical applicability of ML-driven risk prediction systems.

ACKNOWLEDGMENT

This project was completed as part of CS6140 (Machine Learning) coursework under the guidance of Professor Tala. I thank Northeastern University's Khoury College for academic support.

REFERENCES

[1] Course materials, "CS6140 Machine Learning – Northeastern University," 2025.

[2] Kaggle Dataset, "Cardiovascular Disease Dataset," Available: https://www.kaggle.com/datasets/sulianova/

[3] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, 2009.

[4] Scikit-learn documentation, "Model Evaluation and Calibration," https://scikit-learn.org/stable/.

[5] D. Gong, "Top 4 Linear Regression Variations in Machine Learning," Towards Data Science, 2022.