ML Project Report [Team 52]

1. Contribution and Roles table

Name	Role	Key responsibility
Achal Nanjundamurthy	Model Building, Cross-validation Research	Implemented and trained models, ran cross-validation & tuning, produced final metrics/plots.
Khushi Bijkal	Pre-processing, Visualisation, Reporting	Handled preprocessing + EDA visuals, interpreted results, wrote and finalized the report.
Suman Neupane	Model Building, Error Analysis, Research	Supported model building, did error/metric analysis, gathered related research papers and references.
Durga Prasad Narsing	Analysis, Ethical and Legal considerations	Contributed ethical/legal write-up and brief analysis support.

2. Title: *Predicting Cardiovascular Disease Using Machine Learning Techniques*

Link to code :
https://gitlab.computing.dcu.ie/suman.neupane2/predicting-cardiovascular-disease-ml-project/-/blob/1d3e15f70022fa83b61f19a8dcd14fbd612dafae/STATFINAL.ipynb

Abstract:

Cardiovascular disease is one of the leading causes of death, and being able to predict risk early can help guide timely interventions. This project focuses on publicly available cardio train dataset, which includes 70,000 anonymised patient records and build machine learning models for predicting cardiovascular disease.

The dataset was cleaned by removing unrealistic values, and created features such as age in years and BMI. Trained models include logistic regression (baseline), random forest, gradient boosting, and XGBoost, evaluating them with accuracy, F1 score, ROC AUC, confusion matrices, and cross-validation. XGBoost (GPU) performed best, offering best accuracy and efficiency. The results show that carefully designed machine learning pipelines can support early CVD risk detection while emphasising responsible use of health data.

3. Problem Definition

The project aims to build a supervised model to predict cardiovascular disease, where **0 indicates healthy** and **1 indicates the condition**.

Aim and Success Metrics

Aim was to compare simple and ensemble models on health data, identify key predictive factors, and evaluate performance using accuracy, precision, recall, F1 score, ROC AUC, and confusion matrices to understand errors.

Dataset Description

- **Source**: Cardiovascular Disease dataset from Kaggle [1]
- **Original size**: 70,000 rows × 13 columns
- **Dataset link** : cardio_train.csv from Kaggle
- **Initial features:** id, age, gender, height, weight, ap_hi, ap_lo, cholesterol, gluc, smoke, alco, active, and cardio (target variable).

Ethical Notes

The dataset contains no personal identifiers, includes common clinical or lifestyle indicators, suitable for academic research, and should be used responsibly with appropriate medical oversight.

4. Methods

4.1 Preprocessing

- Dataset was loaded using **read_csv()** with a semicolon separator.
- Selected features were checked for **missing values and duplicates**, and **none were found.**
- Physiologically unrealistic values for blood pressure, height, and weight were removed, reducing the dataset to **68,591 rows.**
- Age in days was converted to age_years, and BMI was calculated from height and weight.
- The data was split into **80% training** and **20% test** sets using **stratified sampling**.
- Continuous variables such as height, weight, and blood pressure were normalized using StandardScaler.

4.2 Models Used

Logistic Regression (baseline model): Simple and provides probability estimates to understand feature influence. Assumes linear relationship between features and target variable [8].

Random Forest Classifier: Ensemble of decision trees that can capture non linear patterns, handle larger datasets, robust to noisy data. Also produces feature importance scores, helping identify influential variables [10].

Gradient Boosting (GB): Builds trees sequentially, each tree corrects the errors of the previous ones, improving predictive performance iteratively [9].

Extreme Gradient Boosting (XGBoost): Optimized version of Gradient Boosting, offers high performance, especially on large datasets. GPU acceleration and hyperparameter tuning were used to further enhance accuracy and efficiency [5].

4.3 Validation Protocol

Five-fold stratified cross-validation method was applied to XGBoost to ensure that performance was consistent throughout. After tuning, the model retrained on a full training set and evaluated on the held out 20% test set to provide a fair estimate of real-world performance.

5. Results

5.1 Accuracy and F1 Score

Models	Accuracy	F1 Score
Baseline-Logistic Regression	0.7205	0.7001
Random Forest	0.7077	0.6992
Gradient Boosting (GB)	0.7281	0.7124
Extreme Gradient Boosting (XGBoost)	0.7271	0.7109
Extreme Gradient Boosting (XGBoost+GPU)	0.7299	0.7135

XGBoost with GPU acceleration and hyper-tuning delivered **best overall results** [2] [6]. Gradient Boosting was close, Logistic Regression and Random Forest lagged, Boosting models outperformed simpler approaches.

5.2 Confusion Matrix

Models	True Negative	False Negative	True Positive	False Positive
Logistic Regression	5408	2311	4476	1524
Random Forest	5049	2127	4660	1883
Gradient Boosting	5369	2167	4620	1563
XGBoost (GPU)	5370	2152	4635	1562

XGBoost with GPU performed best, striking a good balance with fewer misclassifications while correctly identifying most positive cases.

5.3 Cross Validation

5-Fold Stratified Cross-Validation was performed using XGBoost **with GPU** to evaluate stability and generalisability of the training data. Stratification ensures that the proportion of target classes (cardio 0/1) is maintained across all folds.

- **F1 Scores:** [0.7219, 0.7207, 0.7101, 0.7246, 0.7280]
- **Mean CV F1 Score:** 0.7211
    - This shows model's consistent predictive power (72%).
- **Standard Deviation:** 0.0060
    - Low standard deviation shows the model's stable performance across folds

5.4 Feature Importance (Top 6)

BMI	341
ap_hi	301

height	297
age_years	291
weight	274
ap_lo	204

Feature importance analysis ranking shows that **BMI** is the strongest predictor [7].

6. Analysis & Discussion

- **Goal** was to compare machine learning models for predicting cardiovascular disease, starting with Logistic Regression as a baseline and then evaluating using ensemble models. Performance was measured using accuracy and F1-score.

- **Error analysis** showed XGBoost (GPU) produced 2152 false negatives and 1562 false positives, slightly more missed cases but fewer overall misclassifications than Logistic Regression and Random Forest, making it more balanced.

- XGBoost with GPU was selected over Gradient Boosting because it trained faster, scaled better, and delivered slightly higher accuracy and F1 score, making it the more reliable option for cardiovascular risk prediction.

- **Key features** such as BMI, blood pressure, age, height and weight aligned with known clinical risk factors, showing the model learned meaningful patterns.

- We **illustrated** feature correlations, showing BMI and blood pressure as key predictors. XGBoost's ROC curve (AUC 0.795) confirmed strong discrimination, and model comparison showed XGBoost with GPU outperformed simpler models while remaining efficient and interpretable.

7. Ethical & Legal Considerations

Although the cardio train dataset is anonymised, using health features like blood pressure and glucose still requires responsible handling. This model is only for academic use and cannot replace clinical judgement, as errors could affect real patient decisions. Any real deployment would require GDPR compliance, strong data security, and clear communication about model limits to ensure safe and ethical use [11].

8. Conclusion:

This project built a **CVD prediction pipeline** on the cardio_train dataset. EDA showed no missing or duplicate records; unrealistic blood pressure and height/weight values were removed, and age_years plus BMI were created. Logistic Regression was compared with Random Forest, Gradient Boosting, and XGBoost, with boosting models performing best. Unlike studies reporting >85% accuracy on only a few hundred samples [3][4], results stayed strong on this larger cohort. XGBoost was selected for tuning because it achieved the top accuracy/F1 while training efficiently on GPU [2]. Feature importance ranked BMI as the strongest predictors. The final tuned model generalized stably **(CV F1 ≈ 0.72)** and reached **~0.73 test accuracy**, supporting early CVD risk screening.

Limitations & future work:

Although the dataset is large, it may still include measurement noise or bias, which can affect performance, and findings may not fully generalise to other populations or clinical settings. Future work should validate on external cohorts, add richer clinical/lifestyle and longitudinal features, test stronger or hybrid ensemble models, and use SHAP/LIME to keep predictions transparent and clinically interpretable.

9. Bibliography

[1] Ulianova, S. (2018). Cardiovascular Disease Dataset. Kaggle. Retrieved from https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

[2] Mitchell, R., & Frank, E. (2017). Accelerating the XGBoost algorithm using GPU computing. PeerJ Computer Science, 3, e127. Retrieved from https://peerj.com/articles/cs-127

[3] Tang, J. (2025). Comparison of Prediction Models for Heart Disease Data: Logistic Regression, Random Forest, and Extreme Gradient Boosting. HSETDATA Journal. Retrieved from https://hsetdata.com/index.php/ojs/article/view/1052/980

[4] Meti, M., & Lingraj, D. (2025). HeartBoost: Clinical Data-Driven Heart Disease Prediction Using XGBoost. International Research Journal of Advanced Engineering and Health. Retrieved from https://irjaeh.com/index.php/journal/article/view/1046/954

[5] XGBoost Documentation. (2025). XGBoost Python Package. Retrieved from https://xgboost.readthedocs.io/en/stable/python

[6] XGBoost GPU Guide. (2025). XGBoost documentation: GPU Support. Retrieved from https://xgboost.readthedocs.io/en/stable/gpu/index.html

[7] Antigone Oreopoulos, & Raj Padwal. (2016).Body mass index and mortality in heart failure: a meta-analysis. NLM. Retrieved from https://pubmed.ncbi.nlm.nih.gov/18585492/

[8] Scikit-learn. (2025). *sklearn.linear_model.LogisticRegression*. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[9] Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics, 29*(5), 1189-1232. Retrieved from https://www.cmi.ac.in/~madhavan/courses/dmml2025/literature/Friedman-Gradient-Boosting-Machine-2001.pdf

[10] Scikit-learn. (2025). sklearn.ensemble.RandomForestClassifier. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[11] NC Medical Journal. (2020). Machine Learning in Health Care: Ethical Considerations Tied to Privacy, Interpretability, and Bias. Retrieved from https://ncmedicaljournal.com/article/120562-machine-learning-in-health-care-ethical-considerations-tied-to-privacy-interpretability-and-bias