Healthcare

AV-JANATAHACK-HEALTHCARE-ANALYTICS

BIT4333 Introduction to Machine Learning
By Group 3
Taught by Sir Nazmirul Izzad Bin Nassir

AGENDA

- **1.0** Executive Summary
- **2.0** Problem Statement
- **3.0** Dataset
- **4.0** Data Preparation Process
- **5.0** Model Development
- **6.0** Evaluation
- **7.0** Deployment
- **8.0** Results & Demonstration
- **9.0** Conclusion

1.0 EXECUTIVE SUMMARY

This project uses ***machine learning*** to ***improve early detection of heart disease***, the world’s leading cause of death. Using the ***UCI Heart Disease dataset***, ***predictive models*** are developed to ***analyze patient attributes***, support ***faster*** and more ***accurate diagnoses***, and enable ***timely***, ***personalized treatment***.

The project shows how healthcare analytics can enhance decision-making, optimize resources, and reduce unnecessary tests. Still, challenges such as ***data quality***, ***model interpretability***, and ***privacy concerns*** must be addressed to ensure ***safe*** and ***effective implementation***.

2.0 PROBLEM STATEMENT

Problem No 1

High Mortality from Cardiovascular Disease:

Cardiovascular disease causes ***~19.8 million deaths annually*** (32% of all deaths worldwide), showing the urgent need for better early detection methods.

Problem No 2

Limitations of Traditional Diagnostics:

Tests such as angiography are ***invasive***, ***expensive***, and ***slow***, often requiring specialists. This leads to ***delays***, ***higher costs***, and ***unequal access***, especially in resource-limited areas.

Problem No 3

Fragmented & Poor-Quality Data:

Patient information (blood pressure, cholesterol, ECG, chest pain type, etc.) is often ***stored separately***, ***incomplete***, or ***duplicated***, making it ***hard to get a full and accurate picture***.

Problem No 4

Weaknesses in Conventional Analysis:

Current approaches rely on fixed thresholds (e.g., cholesterol cut-offs) and examine attributes ***individually***, which fails to capture ***interactions and hidden risk patterns***.

Problem No 5

Data Complexity Limitations:

Manual or threshold-based methods struggle to handle ***large***, ***complex datasets*** with many variables, making it ***difficult to achieve accurate*** and ***scalable diagnosis***.

3.0 DATASET

Number 1

Source of Dataset

- Heart Disease Cleveland UCI dataset (via Kaggle).
- Contains ***real clinical information*** reflecting actual diagnoses.
- Widely used in research → allows ***benchmarking and validation*** of predictive models.

Number 2

Dataset Content

- ***303 patient records*** and ***14 attributes*** (after cleaning 297).
- Features include both:
1. ***Numerical:*** Age, resting blood pressure, cholesterol, max heart rate, etc.
2. ***Categorical:*** Sex, chest pain type, exercise-induced angina, thalassemia, etc.
- ***Target variable:*** Binary classification (0 = no heart disease, 1 = heart disease).

Number 3

Clinical Relevance

- Features chosen are ***well-established risk indicators*** in medical literature (cholesterol, BP, chest pain type).
- Enables exploration of ***multiple risk factors simultaneously***.
- Ensures models provide ***clinically meaningful insights*** for decision-making.

Number 4

Data Preparation

- ***Missing values handled:*** Median (numerical), mode (categorical).
- ***One-hot encoding:*** categorical features converted to binary columns.
- ***Standardization:*** numerical features scaled to equalize importance.
- ***Exploratory analysis:*** checked distributions, correlations, and patterns (cholesterol vs. age).

3.0 DATASET - STRENGTHS

- ***Structured***, ***clean***, and ***balanced*** after preprocessing.
- Dataset is ***small but diverse***, ensuring efficient training while capturing variability.
- Contains a ***rich mix of categorical and numerical variables***.
- Includes ***clinically relevant features*** (cholesterol, blood pressure, chest pain type) supported by medical literature.
- Provides a ***solid foundation*** for building ***reliable machine learning models*** in healthcare.
- Enables ***early pattern discovery and benchmarking***, supporting both academic studies and practical clinical applications.

4.0 DATA PREPARATION PROCESS

Data Cleaning

- Started with ***303 patient records*** across ***14 attributes***.
- Removed ***duplicates***, resulting in ***297 unique cases***.
- Checked dataset for ***incomplete***, ***inconsistent***, or ***duplicate entries***.
- Ensured ***data integrity*** before further analysis.
- Created a reliable baseline for preprocessing and modelling.

Handling Missing Values

- Missing values can ***distort model training*** and reduce accuracy.
- ***Numerical variables*** (cholesterol, blood pressure) → filled with ***median***, preserving central tendency and reducing outlier effects.
- ***Categorical variables*** (chest pain type, thalassemia) → filled with ***mode***, keeping clinically common categories.
- Prevent loss of important patient records.
- Ensured dataset remained ***balanced*** and ***representative***.

Encoding Categorical Feature

- Categorical features cannot be directly interpreted by ML models.
- Applied ***One-Hot Encoding*** → converts categories into binary form.
- Example: Chest pain type split into multiple binary columns.
- This lets the model detect ***unique patterns across categories***.
- Avoids bias from treating categories as numerical values.

4.0 DATA PREPARATION PROCESS

Feature Scaling

- Variables had different ranges (cholesterol vs. max heart rate).
- Used ***standardization*** → brings all features to a ***common scale***.
- Prevents features with large ranges from ***dominating training***.
- Especially important for algorithms sensitive to magnitude (SVM).
- Ensures ***fair contribution of all features*** in model learning.

Exploratory Data Analysis

- Performed ***descriptive statistics*** → mean, range, distribution.
- ***Example:*** Average patient age = 54.5 years, cholesterol range = 126–564 mg/dl.
- ***Visualized class balance*** → fairly even split of disease vs no disease.
- Identified trends: chest pain type & exercise-induced angina ***linked to disease presence***.
- Correlations guided ***feature selection & preprocessing strategy***.

Final Dataset Ready

- Final dataset: ***297 patients***, ***14 attributes***, target variable (0 = no disease, 1 = disease).
- Cleaned, ***structured***, and ***balanced*** dataset.
- Features include both ***numerical*** (age, cholesterol, BP) and ***categorical*** (sex, angina, thalassemia).
- Ensures ***robust and reproducible predictive analysis***.
- Provides a ***solid foundation*** for training reliable machine learning models.

5.0 MODEL DEVELOPMENT - MODELS

M1
Logistic Regression

- Acts as ***baseline model***.
- ***Fast***, interpretable, and coefficients ***show risk contribution***.
- Helps benchmark performance of advanced models.

M2
Random Forest

- Combines many decision trees → ***stronger predictions***.
- ***Avoids overfitting***, handles missing data effectively.
- Provides ***feature importance*** for clinical insights.

M3
SVM

- Finds ***optimal decision*** boundary (hyperplane).
- Works well with ***scaled data***.
- Robust in ***handling complex boundaries*** in classification.

M4
XGBoost

- Builds models sequentially, ***correcting errors*** step by step.
- ***Highly accurate*** for structured/tabular clinical data.
- ***Detects subtle differences*** between healthy and high-risk patients.

5.0 MODEL DEVELOPMENT - WORKFLOW

- **Data Preprocessing:** ***Encoded categorical features*** and ***scaled numerical values***, ensuring fair model comparison.
- **Train-Test Split:** Divided the dataset into ***80% training*** and ***20% testing***, maintaining proper class balance.
- **Hyperparameter Tuning:** Applied ***GridSearchCV*** with ***cross-validation*** to optimize settings and prevent overfitting.
- **Model Training:** ***Trained 4 learning models*** on 297 patient records.
- **Performance Evaluation:** ***Compared models*** using ***Accuracy***, ***F1-score***, ***ROC-AUC***, and ***Confusion Matrix*** results.
- **Feature Insights:** Analyzed feature importance to ***identify key clinical predictors*** of ***heart disease***.

5.0 MODEL DEVELOPMENT - COMPARISON

Model	Strengths	Contribution
Logistic Regression	Simple, interpretable, fast	Establishes baseline, identifies main risk factors
Random Forest	Captures complex patterns, avoids overfitting	Improves reliability, highlights feature importance
Support Vector Machine (SVM)	Works well with scaled data, clear separation	Robust classification in two-class problems
XGBoost	High accuracy, efficient with structured data	Detects subtle differences, top-performing model

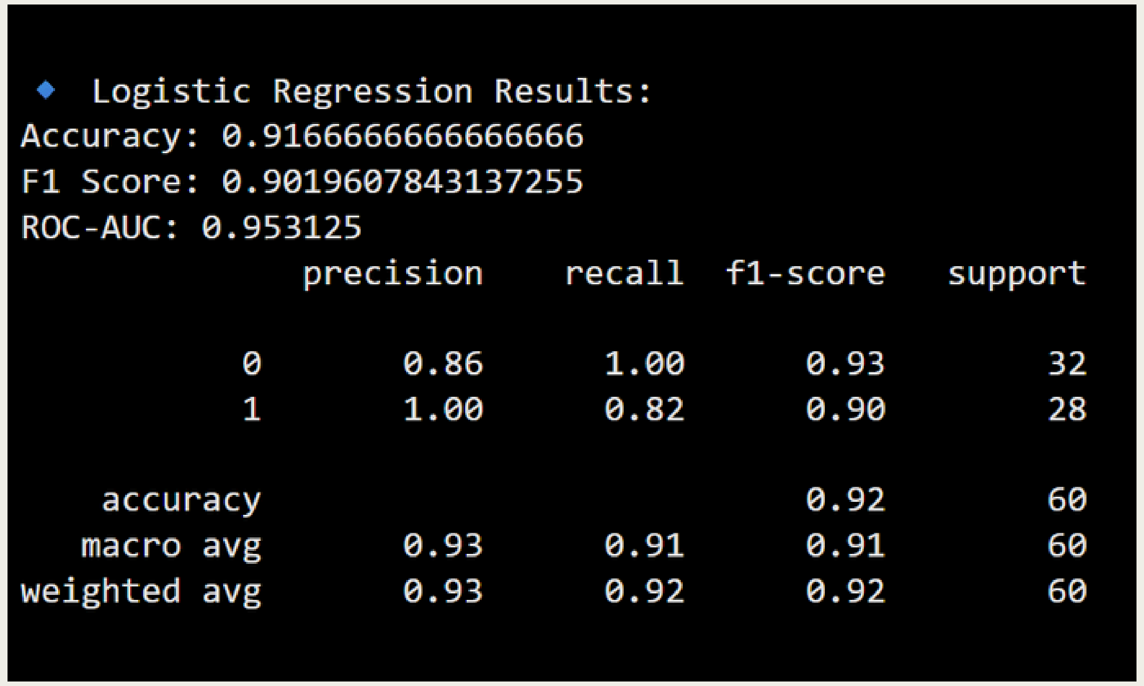
6.0 EVALUATION - METRICS

Model	Strengths	Contribution
Accuracy	Percentage of overall correct predictions.	Quick comparison, but misleading when false negatives matter.
Precision	Proportion of predicted positives that are truly positive.	Reduces false alarms, builds clinical trust.
Recall	Proportion of actual patients correctly identified.	Critical to avoid missing true heart disease cases.
F1-Score	Harmonic mean of Precision & Recall.	Balances false positives and false negatives.
ROC-AUC	Ability to separate healthy vs diseased across thresholds.	Shows diagnostic strength; higher = better model discrimination.

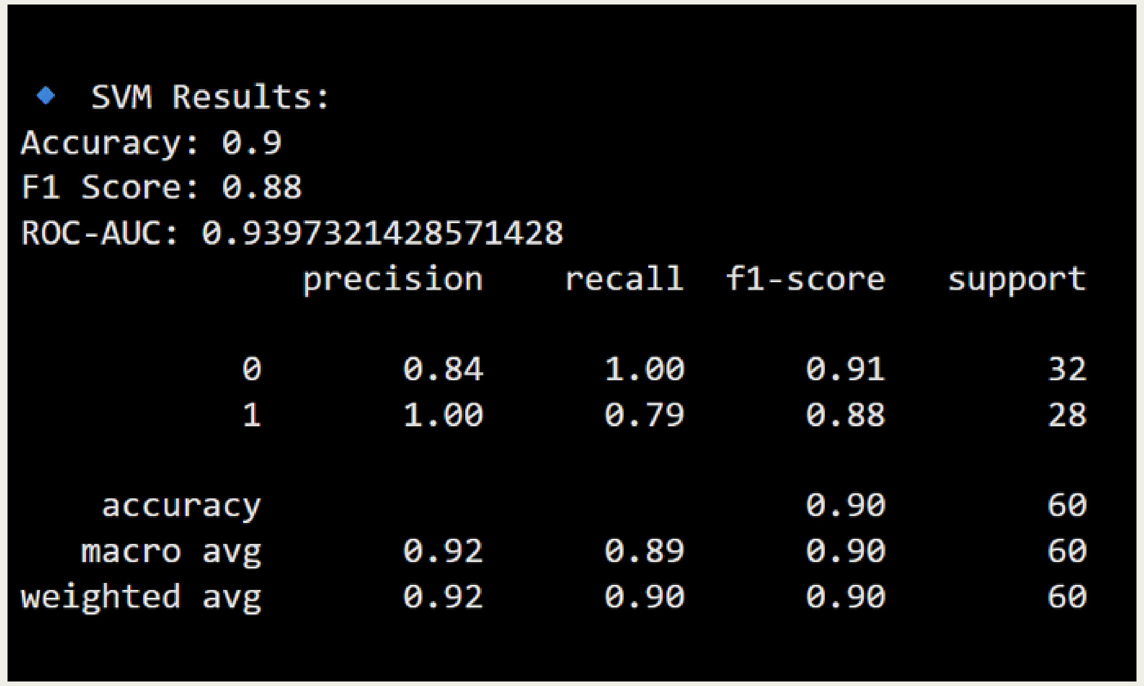
6.0 EVALUATION - COMPARISON RESULTS

Model	Performance Summary	Remarks
Logistic Regression	Moderate Accuracy, lower Recall.	Easy to interpret, weaker predictions.
Random Forest	High Accuracy, strong Recall & ROC-AUC.	Best balance of accuracy & sensitivity.
Support Vector Machine (SVM)	Good Accuracy, moderate Recall.	Finds patterns well, less transparent.
XGBoost	Highest Accuracy, very strong Recall & ROC-AUC.	Most accurate, detects subtle differences.

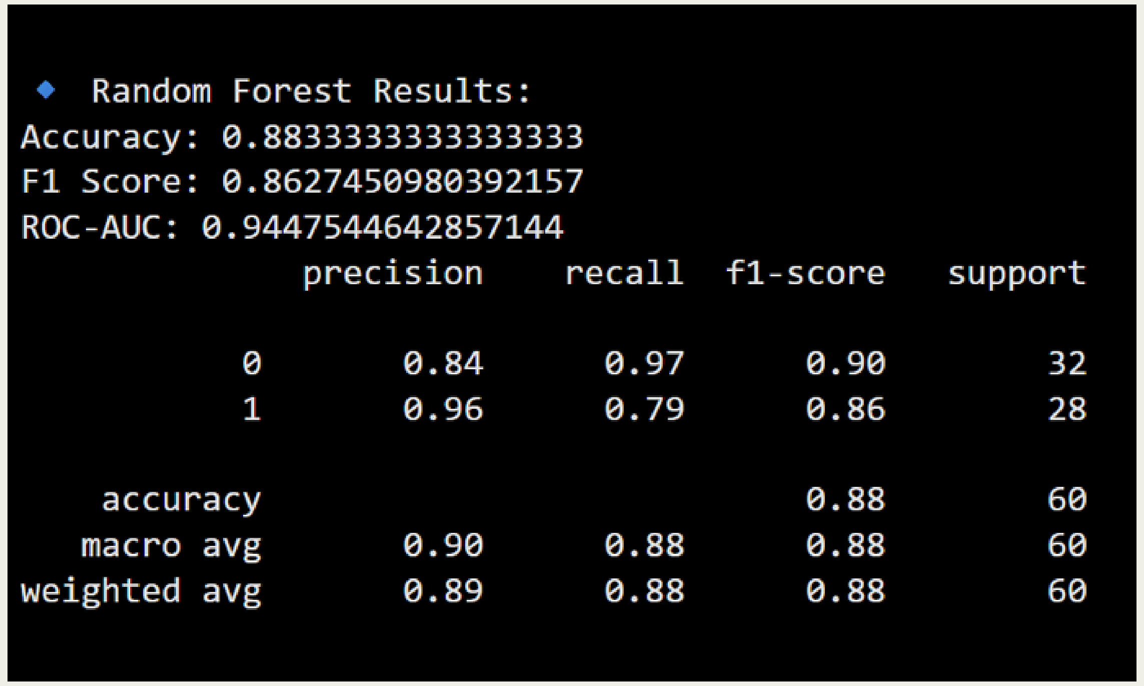
6.0 EVALUATION - RESULTS



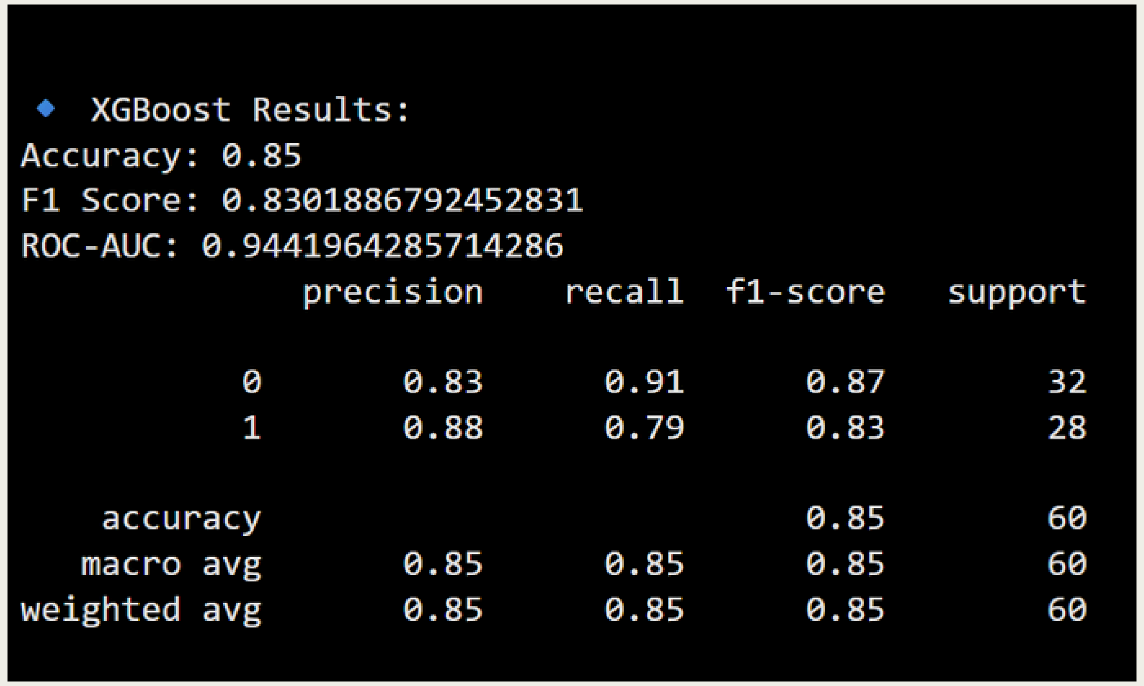
```
🔹 Logistic Regression Results:
Accuracy: 0.9166666666666666
F1 Score: 0.9019607843137255
ROC-AUC: 0.953125
              precision    recall  f1-score   support

           0       0.86      1.00      0.93        32
           1       1.00      0.82      0.90        28

    accuracy                           0.92        60
   macro avg       0.93      0.91      0.91        60
weighted avg       0.93      0.92      0.92        60
```



```
🔹 SVM Results:
Accuracy: 0.9
F1 Score: 0.88
ROC-AUC: 0.9397321428571428
              precision    recall  f1-score   support

           0       0.84      1.00      0.91        32
           1       1.00      0.79      0.88        28

    accuracy                           0.90        60
   macro avg       0.92      0.89      0.90        60
weighted avg       0.92      0.90      0.90        60
```



```
🔹 Random Forest Results:
Accuracy: 0.8833333333333333
F1 Score: 0.8627450980392157
ROC-AUC: 0.9447544642857144
              precision    recall  f1-score   support

           0       0.84      0.97      0.90        32
           1       0.96      0.79      0.86        28

    accuracy                           0.88        60
   macro avg       0.90      0.88      0.88        60
weighted avg       0.89      0.88      0.88        60
```



```
🔹 XGBoost Results:
Accuracy: 0.85
F1 Score: 0.8301886792452831
ROC-AUC: 0.9441964285714286
              precision    recall  f1-score   support

           0       0.83      0.91      0.87        32
           1       0.88      0.79      0.83        28

    accuracy                           0.85        60
   macro avg       0.85      0.85      0.85        60
weighted avg       0.85      0.85      0.85        60
```

7.0 DEPLOYMENT

The heart disease prediction model was deployed as a ***web application*** using ***Streamlit***. ***Users input health details*** like age, cholesterol, blood pressure, and chest pain type. The system then ***predicts whether the patient is at high or low risk of heart disease***, with a confidence score to ***improve clarity*** and ***trust***. It also ***reduces reliance*** on invasive and ***costly diagnostic methods***, offering a ***faster alternative***.

The app is ***fast***, ***user-friendly***, and ***secure***, accessible on both desktop and mobile devices. It acts as a ***prototype tool*** for ***early detection*** and ***clinical support***, with potential for future integration into hospital systems and expansion using larger datasets.

8.0 RESULTS AND DEMONSTRATION

Aspect	Details
Best Performing Models	***Random Forest*** & ***XGBoost*** achieved ***high recall*** and ***ROC-AUC***, effectively detecting patients at risk while minimizing false negatives.
Other Models	***Logistic Regression:*** interpretable but lower accuracy. ***SVM:*** good boundary separation but limited clinical transparency.
Evaluation Metrics	Accuracy, F1-Score, Recall, Precision, ROC-AUC, Confusion Matrix. Provides a ***comprehensive performance assessment***.
Demonstrating Inputs	Patient attributes such as age, sex, cholesterol, blood pressure, chest pain type, exercise-induced angina, and thalassemia.

8.0 RESULTS AND DEMONSTRATION

Aspect	**Details**
Output	Predicted heart disease status (0 = no, 1 = yes) with confidence probability, ***supporting early detection*** and clinical decision-making.
Clinical Impact	Offers a fast, non-invasive tool to ***assist doctors***, **reduce misdiagnoses**, and ***prioritize high-risk patients*** for intervention.

9.0 CONCLUSION

This project showed how ***machine learning*** can improve ***early detection of heart disease*** using the ***Cleveland dataset***. Models like ***Random Forest*** and ***XGBoost*** achieved high ***predictive performance*** while staying ***clinically relevant***.

A ***user-friendly*** web application allows clinicians to ***input patient data*** and get ***real-time predictions*** with probability scores. The project highlights benefits of ***data-driven decision support***, including ***better diagnostic accuracy***, ***early intervention***, and ***optimized healthcare resources***. These results demonstrate the potential of ***predictive analytics*** to transform cardiovascular care.

Thank you

AV-JANATAHACK-HEALTHCARE-ANALYTICS

BIT4333 Introduction to Machine Learning
By Group 3
Taught by Sir Nazmirul Izzad Bin Nassir