Classification and Comparison of Supervised Machine Learning Algorithms Based on UCI Heart Disease Dataset

Team: **Ghosh Bros**
Niladdri Ghosh **Arnab Ghosh**
Supervisor: **Br. Bhaswarachaitanya** (Tamal Maharaj)

MSc. Computer Science, RKMVERI

30 November, 2024

Contents

Introduction

Introduction

- **Objective:** Predict heart disease risk using the UCI Heart Disease dataset.
- **Techniques:** Evaluated multiple supervised machine learning algorithms like KNN, SVM, Logistic Regression, Random Forest, and XGBoost.
- **Approach:** Preprocessed data with feature engineering, imputation, and scaling for optimal model performance.
- **Metrics:** Compared models using accuracy, precision, recall and F1-score.

Objective

Objective

- **Objective:** To evaluate and compare ML algorithms for heart disease prediction.
- **Scope:**
  - Predict heart disease on UCI Heart Disease Dataset.
  - Focus on accuracy, precision, recall, and F1-score for performance evaluation.

Dataset Description

Dataset Description

Dataset Overview:

- Few rows of the UCI Heart Disease Dataset:
- Dataset has **920 instances** & **16 atrributes**

	id	age	sex	dataset	cp	trestbps	chol	fbs	restecg	thalch	exang	oldpeak	slope	ca	thal	num
0	1	63	Male	Cleveland	typical angina	145.0	233.0	True	lv hypertrophy	150.0	False	2.3	downsloping	0.0	fixed defect	0
1	2	67	Male	Cleveland	asymptomatic	160.0	286.0	False	lv hypertrophy	108.0	True	1.5	flat	3.0	normal	2
2	3	67	Male	Cleveland	asymptomatic	120.0	229.0	False	lv hypertrophy	129.0	True	2.6	flat	2.0	reversable defect	1
3	4	37	Male	Cleveland	non-anginal	130.0	250.0	False	normal	187.0	False	3.5	downsloping	0.0	normal	0
4	5	41	Female	Cleveland	atypical angina	130.0	204.0	False	lv hypertrophy	172.0	False	1.4	upsloping	0.0	normal	0

Dataset Description

Continuous Features:

- Age: Continuous (in years).
- Resting Blood Pressure (trestbps): Continuous (in mmHg).
- Serum Cholesterol (chol): Continuous (in mg/dL).
- Maximum Heart Rate Achieved (thalach): Continuous (in bpm).
- ST Depression (oldpeak): Continuous (difference during exercise).

Categorical Features:

- Sex: Binary (Male/Female).
- Fasting Blood Sugar (fbs): Binary ($>$120 mg/dL: Yes/No).
- Exercise-Induced Angina (exang): Binary (Yes/No).
- Resting ECG Results (restecg): Discrete categories (0, 1, 2).
- Number of Major Vessels (ca): Discrete categories (0–3).

Dataset Description (continued)

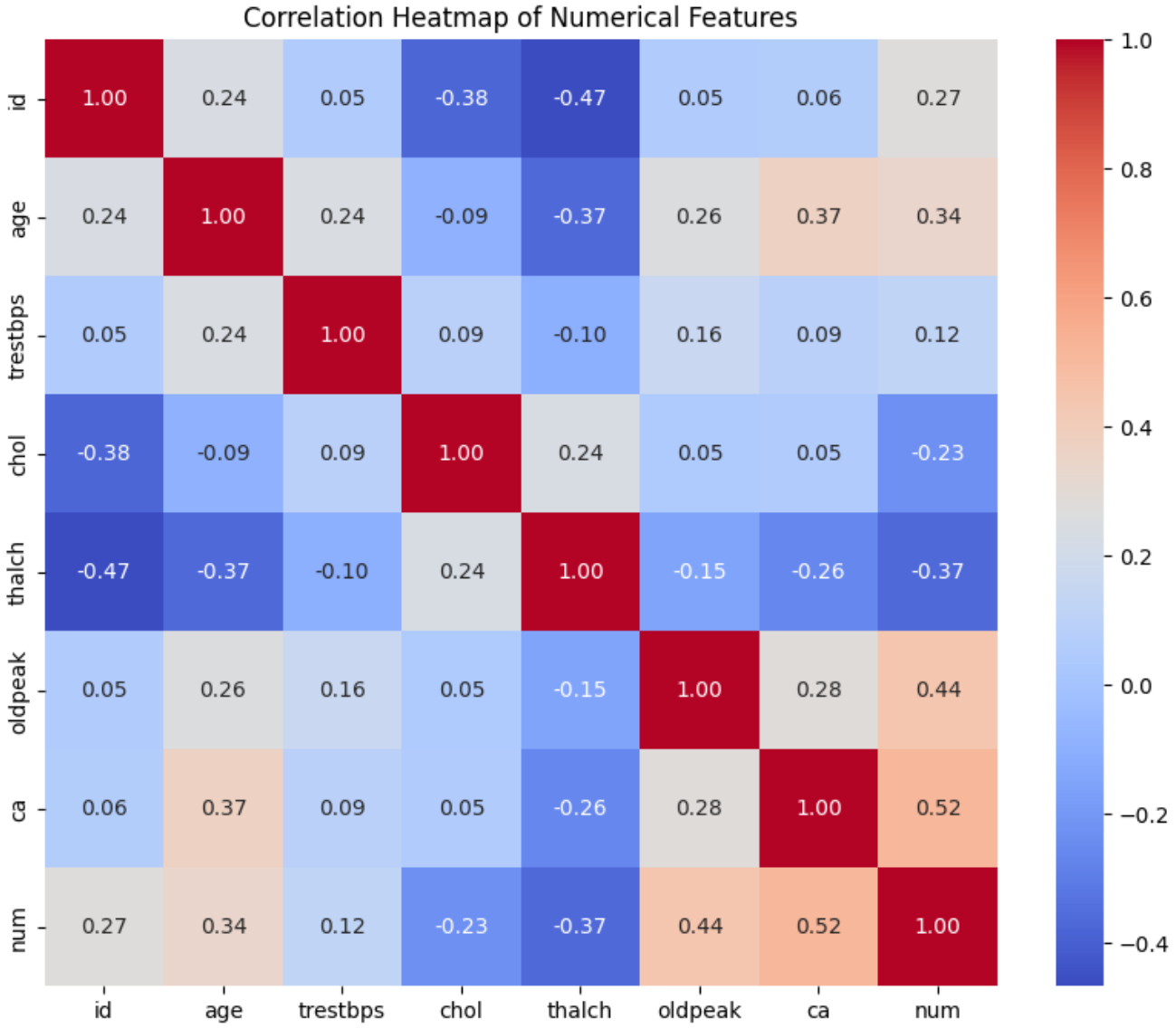
Nominal Features:

- Chest Pain Type (cp): Nominal (Typical Angina, Atypical Angina, Non-Anginal, Asymptomatic).
- Dataset ID: Nominal (indicating data source/location).
- Thalassemia (thal): Nominal (Normal, Fixed Defect, Reversible Defect).

Ordinal Features:

- Slope of the ST Segment (slope): Ordinal (Upsloping, Flat, Downsloping).

Correlation Heatmap



Dataset Insights and Preprocessing Overview

- **Target Variable:** Predicts the presence and severity of heart disease (multi-class classification).
- **Data Imbalance:** Males dominate the dataset (79%), and asymptomatic chest pain is the most common feature (54%).
- **Preprocessing:** Addressed missing values, outliers, and performed feature engineering to enhance the dataset's utility for machine learning models.

Methodology

Algorithm

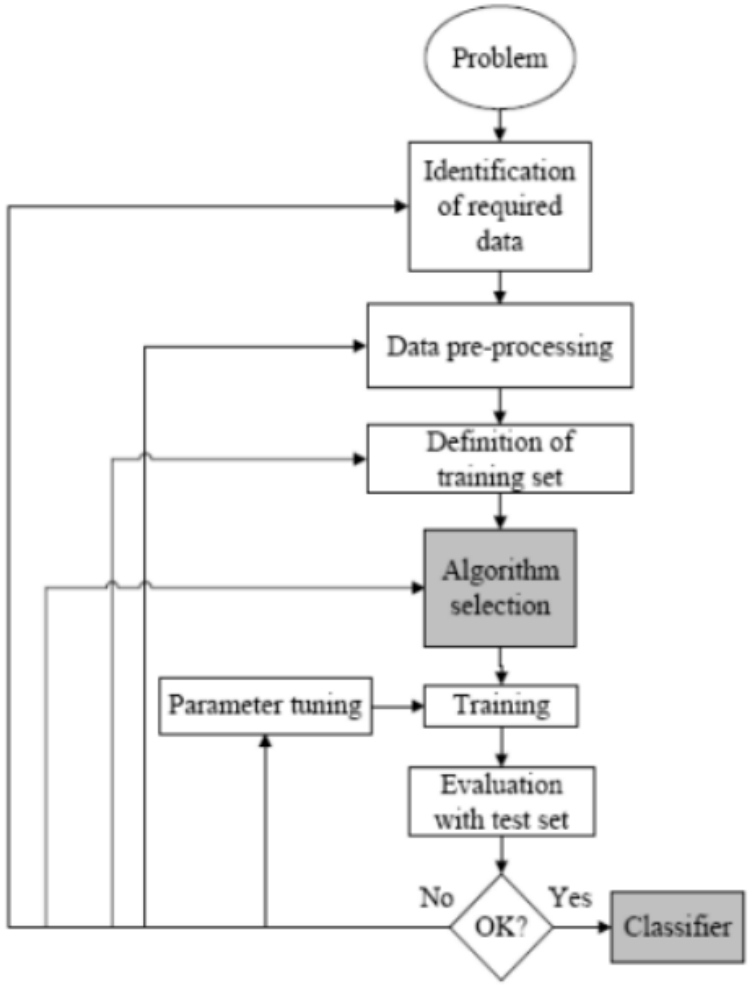


Figure: **The process of supervised Machine Learning**

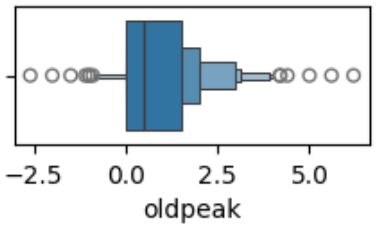
Machine Learning Workflow

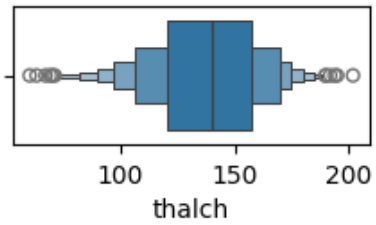
Data Preprocessing:

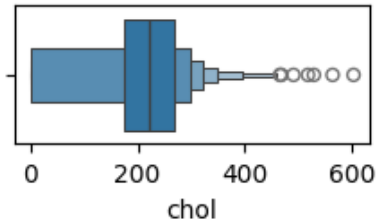
Handled missing values, outliers, and performed feature engineering. Scaled and encoded features for consistent data representation.

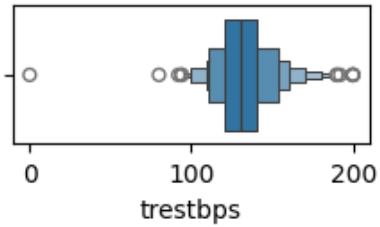
Model Training:

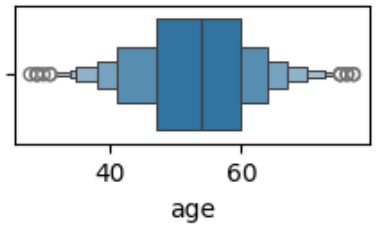
Split dataset into training (80%) and testing (20%). Used multiple algorithms for classification.











Models Implemented

Algorithms:

- K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression.
- Decision Tree, Random Forest, Gradient Boosting, XGBoost, Naive Bayes.

Reason for Selection:

- KNN and Logistic Regression: Baseline models for comparison.
- Random Forest, Gradient Boosting, XGBoost: Handle non-linear relationships and offer high accuracy.
- Naive Bayes: Simple and computationally efficient.

Technology and Purpose

- **Programming Language:** Python.
- **Libraries/Frameworks:**
  - Data Processing: Pandas, NumPy.
  - Visualization: Matplotlib, Seaborn.
  - Machine Learning: Scikit-learn, XGBoost, Joblib.
- Ensure data quality through robust preprocessing techniques.
- Use diverse models to compare performance and identify the best classifier.
- Leverage ensemble techniques for better generalization and accuracy.
- Provide reliable metrics to validate the effectiveness of machine learning in heart disease prediction.

Model Evaluation

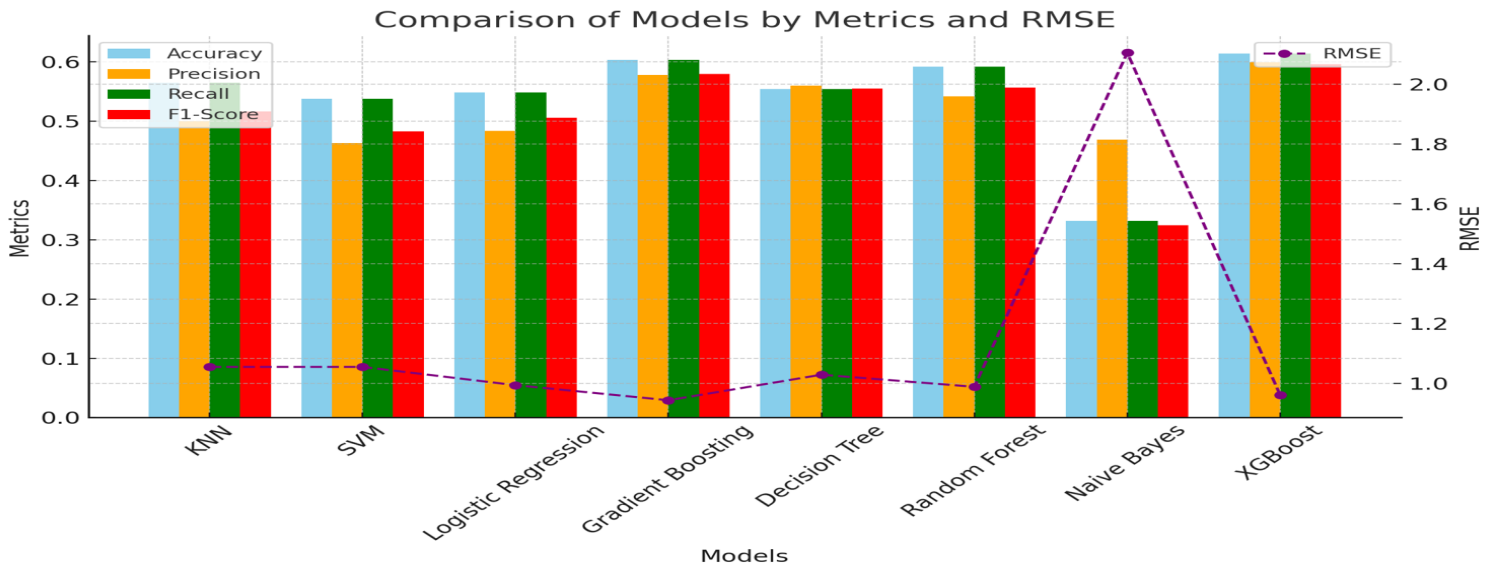
Metrics Used:

- **Accuracy:** Percentage of correctly classified instances.
- **Precision:** Focuses on reducing false positives.
- **Recall:** Measures the ability to identify true positives.
- **F1-Score:** Balances precision and recall.

Results & Analysis

Results & Analysis

- **Classification score for models:**

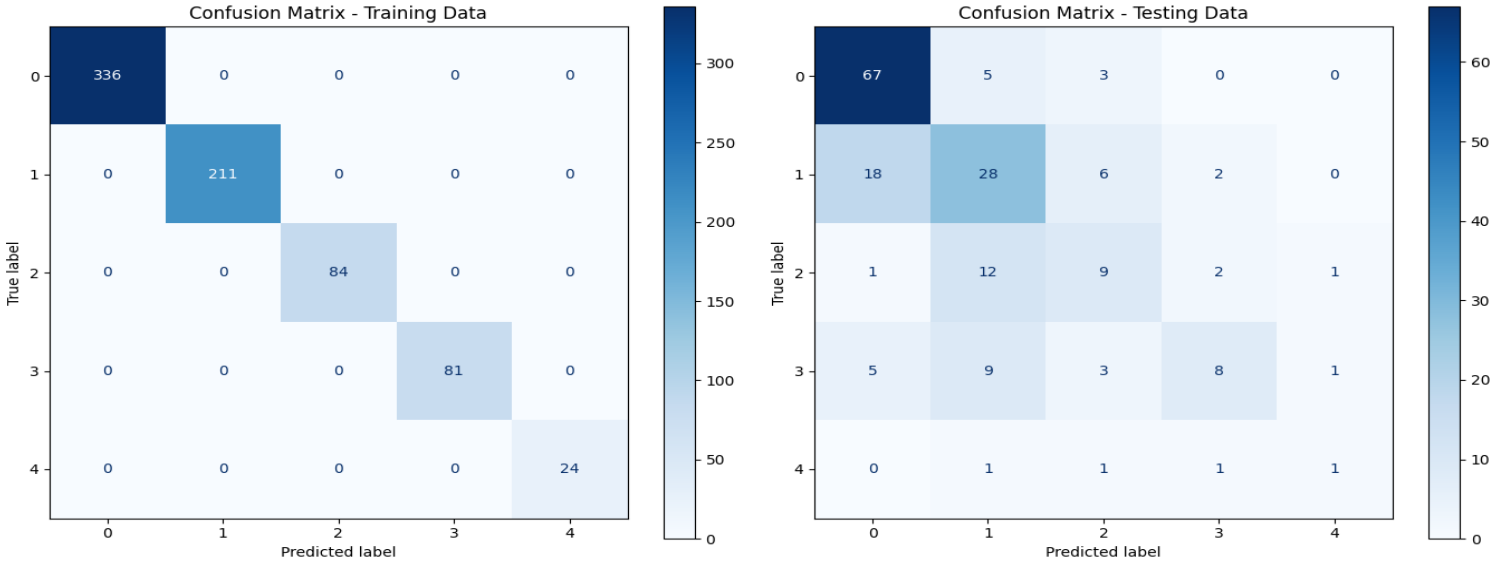


Key Results

- **Top-Performing Models:** Gradient Boosting, Random Forest, and XGBoost showed the best accuracy and robustness.
- **Best Model:** XGBoost performed the best on the test data with:
  - Accuracy: 61%
  - Precision: 60%
  - F1-Score: 0.60
- **Model Insights:**
  - Simple models like Logistic Regression performed reasonably but lagged behind ensemble techniques.
  - Naive Bayes struggled due to its assumptions of feature independence.

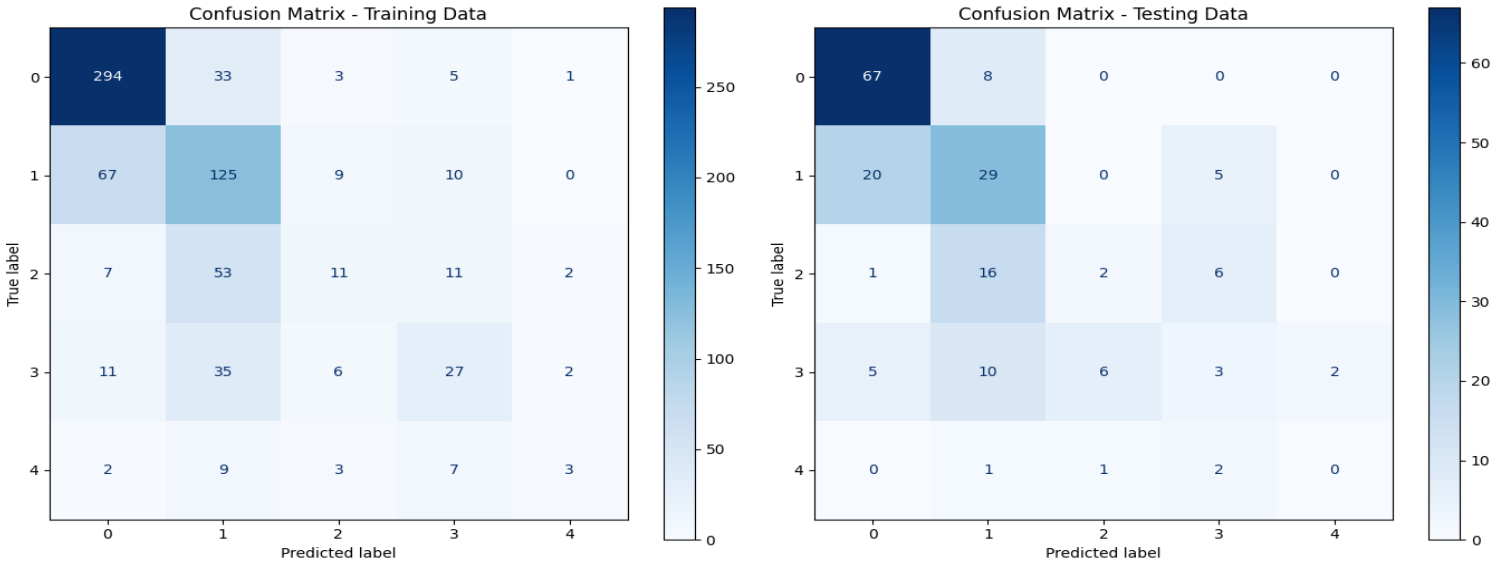
Confusion Matrices

- **XGBoost:** Fewer false positives and false negatives compared to other models.



Confusion Matrices

- **Logistic Regression:** Balanced but slightly more errors in both categories.



Analysis

Analysis:

- **Strengths:**
  - Ensemble models captured complex patterns effectively, providing high accuracy and balanced predictions.
  - XGBoost's regularization and robustness made it the best-performing model.
- **Weaknesses:**
  - Simpler models like Naive Bayes underperformed due to the complex relationships in the dataset.
  - Decision Tree overfitted the training data.

Overall Performance

Algorithm	Correctly Classified (%)	Incorrectly Classified (%)	Kappa	MAE	Precision (YES)	Precision (NO)
KNN	56.52	43.48	0.346	0.636	0.345	0.345
SVM	53.80	46.20	0.302	0.658	0.309	0.309
Logistic Regression	54.89	45.11	0.329	0.609	0.317	0.317
Gradient Boosting	60.33	39.67	0.415	0.538	0.447	0.447
Decision Tree	55.43	44.57	0.368	0.625	0.406	0.406
Random Forest	59.24	40.76	0.396	0.576	0.385	0.385
Naive Bayes	33.15	66.85	0.191	1.565	0.265	0.265
XGBoost	61.41	38.59	0.436	0.543	0.521	0.521

Discussions

Challenges Encountered

1. **Data Quality Issues:**
   - Missing values in key features (e.g., cholesterol, blood pressure).
   - Imbalanced dataset with male dominance (79%) and more asymptomatic cases (54%).
2. **Computational Limitations:**
   - High training time for ensemble models like Gradient Boosting and XGBoost.
   - Limited resources for hyperparameter tuning and large-scale cross-validation.
3. **Feature Engineering:**
   - Identifying meaningful derived features (e.g., BP to cholesterol ratio) was time-intensive.

Key Learnings

Key Insights About the Dataset:

- Correlation between features like cholesterol, blood pressure, and heart disease risk.
- Importance of addressing class imbalance to improve model generalization.

ML Knowledge Growth:

- Gained deeper understanding of ensemble models (e.g., XGBoost's regularization benefits).
- Importance of evaluating multiple metrics (e.g., F1-score, Precision) rather than relying solely on accuracy.

Preprocessing Value:

- Robust preprocessing significantly improved model performance and interpretability.

Conclusion

Conclusion

- **Objective:** Applied supervised ML algorithms to the UCI Heart Disease dataset for heart disease prediction.
- **Key Findings:** Gradient Boosting, Random Forest, and XGBoost were the top-performing models with strong predictive accuracy and robustness.
  - XGBoost was particularly effective during testing (Accuracy: 61.41%, F1-Score: 0.60).
- **Limitation:**
  - Naive Bayes performed poorly due to its assumption of feature independence.
  - Some models like Decision Trees showed overfitting on training data.

Future Work

- Address data imbalance through techniques like oversampling or SMOTE.
- Automate hyperparameter tuning for faster and more reliable model optimization.
- Explore advanced deep learning approaches for higher-dimensional feature interactions.
- Collaborate on better datasets to improve gender and feature representation.

References

1. Shalev-Shwartz, S. & Ben-David, S. *Understanding Machine Learning - From Theory to Algorithms*. Cambridge University Press, 2014.
2. Kotsiantis, S. B., Zaharakis, I., Pintelas, P. et al. *Supervised machine learning: A review of classification techniques*. Emerging artificial intelligence applications in computer engineering 160, 3–24 (2007).
3. Ayodele, T. O. *Types of machine learning algorithms*. New advances in machine learning 3, 5–1 (2010).
4. Pradeep, K. & Naveen, N. *A collective study of machine learning (ml) algorithms with big data analytics (bda) for healthcare analytics (hca)*. International Journal of Computer Trends and Technology 47, 149–155 (2017).
5. Zhang, S., Li, X., Zong, M., Zhu, X. & Cheng, D. *Learning k for knn classification*. ACM Transactions on Intelligent Systems and Technology (TIST) 8, 1–19 (2017).

Thank You!