University of North Texas

Software Development for Artificial Intelligence – CSCE 5214

Group Number - 2

Project Report

Title

Heart Disease Risk Prediction

Team Members-

Laasya Priya Jyesta
Mohan Kalyan Guntupalli
Nagalakshmi Reddy
Vinay Kumar Parvathini

Table of Contents

1 Abstract:

The study delves into how meteorological conditions impact transportation traffic, evaluating various analytical methods. Specifically, it juxtaposes the methodologies outlined in 'A Big Data Science Solution for Transportation Analytics with Meteorological Data' against a simulated approach. It affirms the effectiveness of the proposed data analytics method, noting its simplicity, interpretability, and adeptness in pinpointing crucial weather factors affecting traffic flow and crafting predictive models. The simulated results align closely with those presented in the paper. Moreover, it explores potential applications of this data analytics method in transportation planning, management, traffic advisories, public transportation scheduling, and optimizing ride-hailing services.

2 Introduction:

Heart disease remains one of the most prevalent and life-threatening conditions worldwide, accounting for millions of deaths each year. According to the World Health Organization, cardiovascular diseases are responsible for approximately 31% of all global deaths. Among these, Coronary Heart Disease (CHD) is particularly concerning, as it often progresses silently until a major cardiac event occurs.

Early identification of individuals at risk is crucial because timely interventions — such as medication, dietary changes, and exercise — can drastically reduce mortality rates.

However, accurately predicting CHD is challenging due to the multifactorial nature of the disease, where several variables such as age, cholesterol, blood pressure, glucose, smoking, and diabetes interact in complex and often nonlinear ways.

Machine Learning (ML) provides a powerful and innovative approach to this problem. By analyzing large datasets and detecting subtle patterns that traditional methods may overlook, ML enables early risk assessment with higher precision. This project utilizes ML techniques to analyze data from the Framingham Heart Study and develop predictive models capable of estimating a person's ten-year risk of developing CHD.

The study compares multiple ML models, evaluates their accuracy, and identifies the most effective approach. The findings are expected to aid healthcare professionals in data-driven decision-making and support early preventive measures.

3 Problem Statement:

Heart disease continues to pose a major global health challenge, claiming millions of lives annually. Despite advances in medical diagnostics, early prediction of CHD risk remains limited due to the complex interdependence of numerous clinical and behavioral factors.

Conventional scoring systems like the Framingham Risk Score often consider only a fewparameters and assume linear relationships, which may oversimplify real-world interactions.

For example, two patients with identical cholesterol levels might have vastly different risk levels depending on their blood pressure, glucose, or smoking habits. Such complexity demands computational methods capable of understanding nonlinear patterns and interactions among multiple features simultaneously.

The objective of this project is to build and evaluate ML models that can accurately predict whether an individual will develop CHD within ten years, using data from the Framingham

Heart Study. By leveraging machine learning, we aim to overcome the limitations of traditional methods and create a predictive framework that supports personalized, data-driven healthcare.

Existing Methods/Algorithms for CHD Prediction

A wide range of computational and analytical approaches have been developed to predict Coronary Heart Disease (CHD), each differing in complexity, data requirements, and interpretability. Some of the most commonly utilized methods include:

3.1. Statistical / Traditional Risk Scoring Models

These approaches rely on well-established clinical relationships between risk factors such as age, cholesterol, blood pressure, and smoking behavior. The Framingham Risk Score (FRS) is one of the most widely used tools, estimating the 10-year probability of developing CHD based on aggregate point values assigned to risk factors.

- Examples:
  - Framingham Risk Score (FRS)
  - SCORE (Systematic COronary Risk Evaluation)
  - Pooled Cohort Equations (ACC/AHA)

While easy to interpret and clinically accepted, these models assume linear relationships among variables and often overlook complex interactions, reducing accuracy for diverse populations and individual-level risk variations.

3.2. Machine Learning Models

Machine learning techniques learn complex, nonlinear relationships directly from data. They can analyze multiple clinical and behavioral features simultaneously to identify individuals at high CHD risk.

- Common Methods Used:
  - Logistic Regression: Simple, interpretable baseline model.
  - Decision Trees: Rule-based predictions, but prone to overfitting.
  - Random Forest: Ensemble of decision trees providing better stability.
  - K-Nearest Neighbors (KNN): Instance-based learning sensitive to feature scaling.
  - Support Vector Machine (SVM): Effective in modeling nonlinear boundaries.
  - Gradient Boosting & XGBoost: Iterative boosting methods achieving high predictive accuracy.

Machine learning models offer high predictive power, but often require extensive preprocessing, balanced datasets, and careful tuning to avoid bias and overfitting.

3.3. Deep Learning Approaches

Deep learning models, particularly neural networks, process large and complex datasets to uncover hidden patterns that traditional ML may miss.

- Approaches Include:
  - Multi-Layer Perceptrons (MLPs): Fully connected neural networks.
  - Convolutional Neural Networks (CNNs): Used when incorporating ECG or imaging data.
  - Recurrent Neural Networks (RNNs): Useful for temporal health data and longitudinal patient monitoring.

Although they are powerful, deep learning models may lack transparency, and their performance strongly depends on the availability of large, well-labeled datasets.

3.4. Expert Systems

Expert systems simulate human clinical reasoning by incorporating medical knowledge and heuristic rules defined by cardiologists and healthcare professionals.

- Techniques Used:

  • Rule-Based Reasoning: If-then clinical decision rules.

  • Fuzzy Logic: Handles uncertainty in clinical measurements.

  • Genetic Algorithms: Optimizes classification rules and thresholds.

Expert systems provide interpretable and clinically intuitive predictions, but may depend on incomplete or subjective expert knowledge, making them less adaptable to varied patient populations.

4. Method/Algorithm Presented in the Paper:

The research presented in this project introduces a data-driven machine learning framework designed to predict the ten-year risk of Coronary Heart Disease (CHD).

This approach combines statistical preprocessing, feature selection, and predictive modeling to uncover relationships among clinical and behavioral health indicators. The methodology involves several systematically executed stages:

4.1 Data Collection and Preprocessing

The dataset utilized originates from the **Framingham Heart Study**, a well-recognized longitudinal cardiovascular dataset. This dataset includes demographic attributes (e.g., age, gender), lifestyle indicators (e.g., smoking and alcohol consumption), and clinical metrics (e.g., cholesterol, blood pressure, BMI).

To ensure reliability and consistency, the data undergoes several preprocessing steps:

- Removal of records with missing or invalid entries, particularly in cholesterol and glucose variables.
- Normalization of continuous clinical variables using **Z-score standardization** to maintain uniform scale across features.
- Correction of class imbalance between CHD and non-CHD cases using **SMOTE (Synthetic Minority Oversampling Technique)**, enabling the model to learn effectively from minority class patterns.
- Partitioning of the dataset into training and testing groups (80% / 20%) to evaluate model generalizability.

This structured preprocessing ensures high-quality input for subsequent modeling and enhances the predictive reliability of machine learning algorithms.

4.2 Feature Analysis and Selection

Statistical and algorithmic feature selection methods are employed to identify the most influential predictors of CHD.

- **Correlation analysis** evaluates the strength and direction of relationships among clinical variables.
- The **Boruta feature selection algorithm** is then used to isolate the most significant

predictors while eliminating redundant or noisy features.

Key features identified include:

- **Age**
- **Systolic and Diastolic Blood Pressure**
- **Glucose Level**
- **Total Cholesterol**
- **BMI**
- **Cigarettes Per Day**
- **Heart Rate**

These variables are subsequently used as the primary input features for prediction models.

4.3 Model Development and Evaluation

Multiple machine learning algorithms are implemented to model CHD risk:

Model	**Description**
Logistic Regression	Baseline linear model, highly interpretable
Decision Tree	Rule-based model capable of capturing variable interactions
Random Forest	Ensemble method improving stability and reducing overfitting
K-Nearest Neighbors (KNN)	Distance-based classifier sensitive to scaling and data density
Support Vector Machine (SVM)	Efficient at modeling nonlinear relationships through kernel functions
Gradient Boosting & XGBoost	Boosted ensembles optimizing error reduction sequentially

Model performance is evaluated using metrics such as:

- **Accuracy**

- **Precision**
- **Recall (Sensitivity)**
- **F1-Score**
- **ROC-AUC**

Among all models, **SVM** demonstrated the strongest performance, achieving a **ROC-AUC score of 0.93**, indicating exceptional discriminative capability in predicting CHD.

4.4 Strengths and Weaknesses of the Methodology

Strengths

- The use of multiple machine learning models allows comprehensive comparison and ensures robust performance evaluation.
- SMOTE effectively addresses class imbalance, improving the detection of high-risk individuals.
- Ensemble models such as Random Forest and XGBoost provide **feature importance insights**, enhancing clinical interpretability.
- The methodology captures **nonlinear and multifactorial interactions**, which traditional medical scoring systems typically overlook.

Weaknesses

- Model performance depends heavily on the **quality and completeness** of clinical data.
- Deep interpretability is reduced in models like SVM and XGBoost, which may limit direct clinical explanation without additional explainable AI methods.
- External factors such as lifestyle behavior changes, family history, or genetic predisposition may not be fully accounted for in the dataset, potentially influencing prediction accuracy.

4.5 Possible Applications of the Method/Algorithm

The predictive framework developed in this project offers numerous real-world medical applications:

- **Early Risk Screening:** Healthcare providers can identify high-risk individuals and initiate preventive interventions.
- **Clinical Decision Support Systems:** Hospitals can integrate the model into electronic health records to generate automated risk alerts.
- **Personalized Treatment Planning:** Risk score outputs guide patient-specific lifestyle and medication strategies.
- **Public Health Policy:** Population-level risk prediction supports cardiovascular disease prevention programs.

Figure 1: Class Distribution Before and After SMOTE

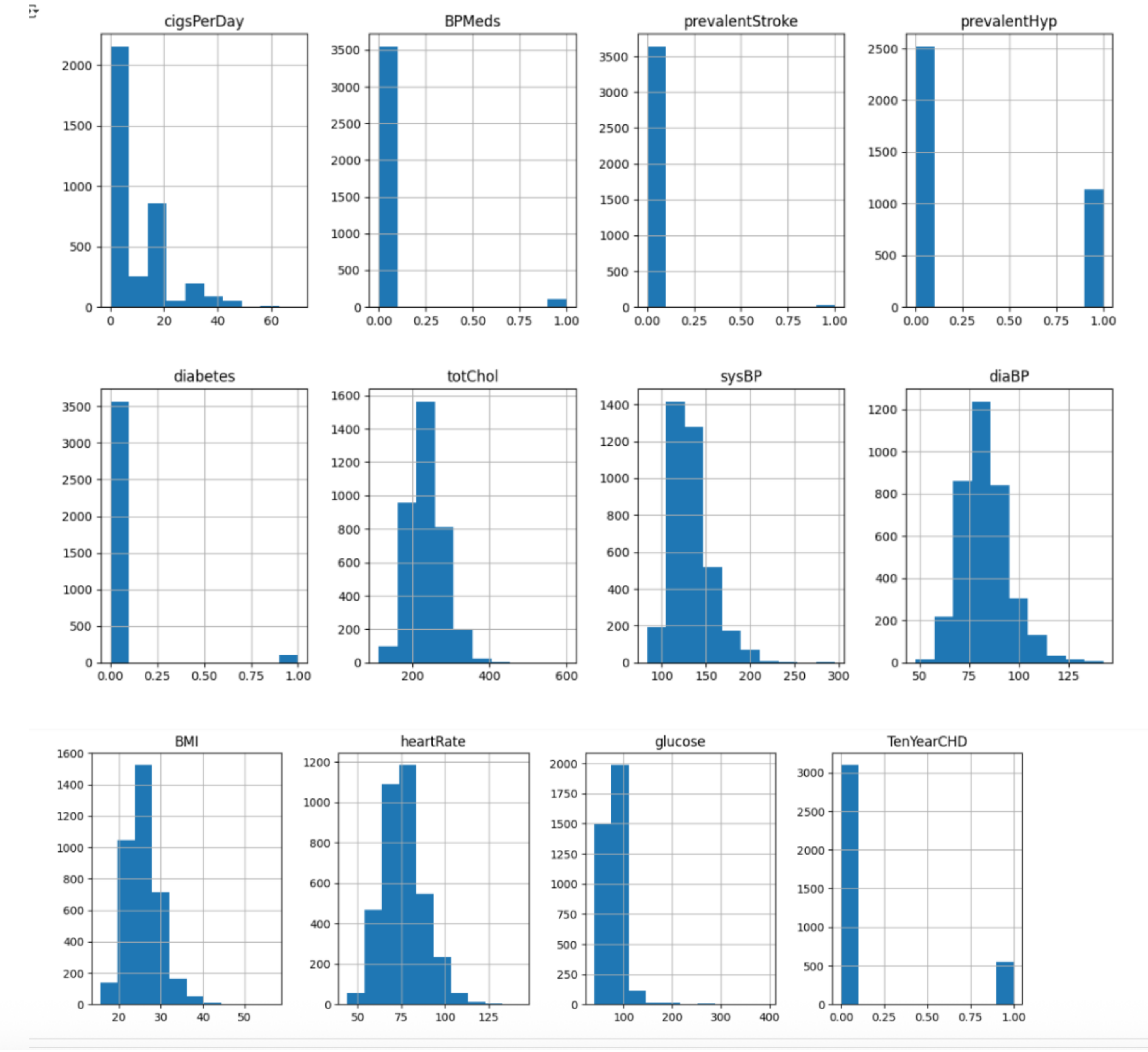
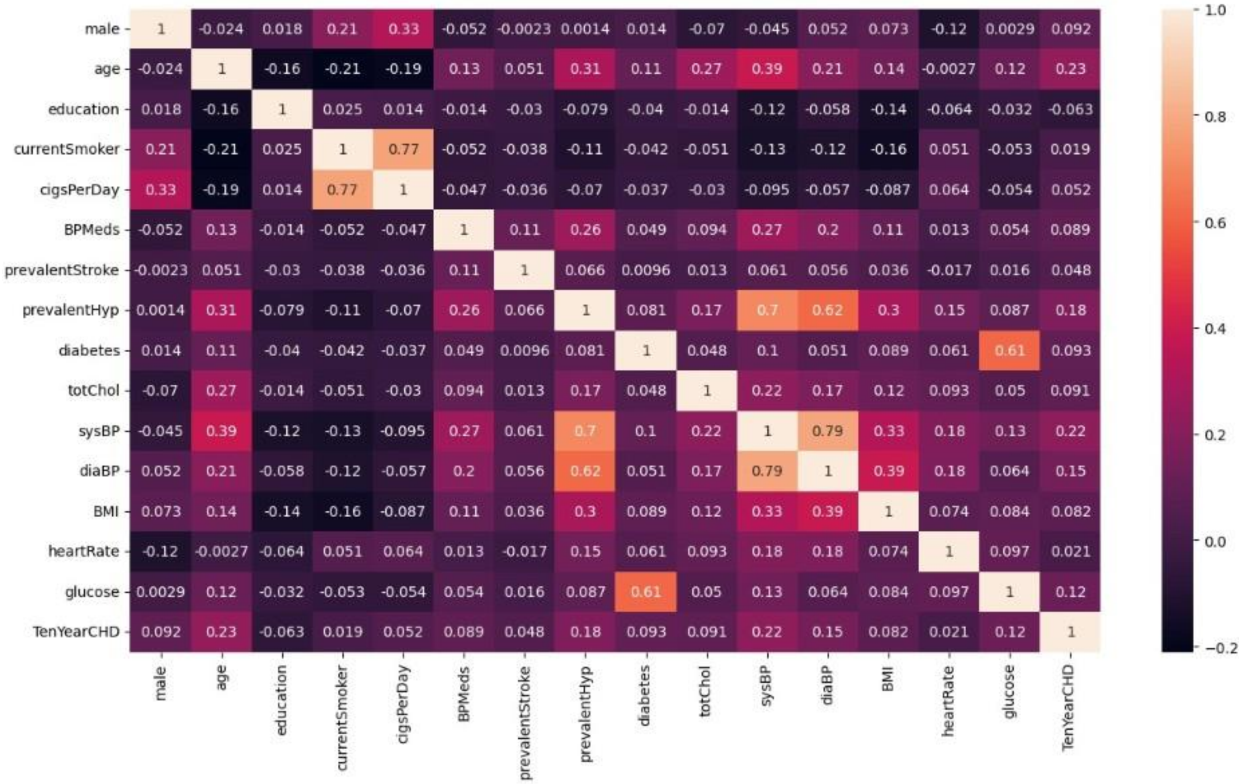
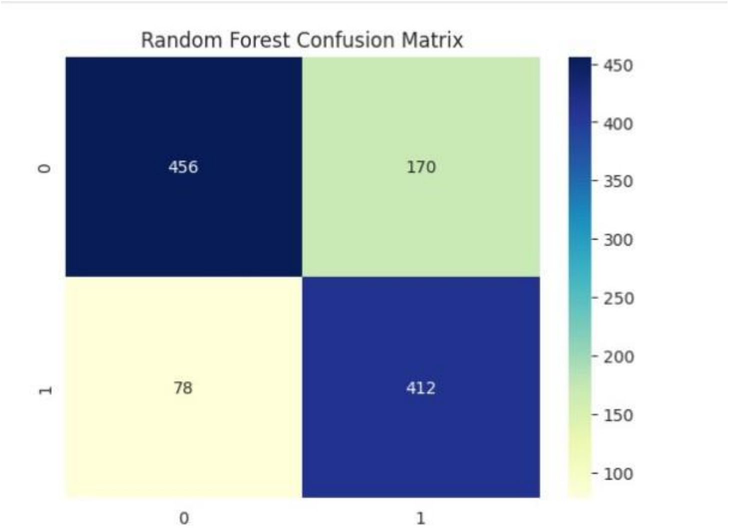


Figure 2: Correlation Heatmap of Risk Factors and CHD



	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
male	1	-0.024	0.018	0.21	0.33	-0.052	-0.0023	0.0014	0.014	-0.07	-0.045	0.052	0.073	-0.12	0.0029	0.092
age	-0.024	1	-0.16	-0.21	-0.19	0.13	0.051	0.31	0.11	0.27	0.39	0.21	0.14	-0.0027	0.12	0.23
education	0.018	-0.16	1	0.025	0.014	-0.014	-0.03	-0.079	-0.04	-0.014	-0.12	-0.058	-0.14	-0.064	-0.032	-0.063
currentSmoker	0.21	-0.21	0.025	1	0.77	-0.052	-0.038	-0.11	-0.042	-0.051	-0.13	-0.12	-0.16	0.051	-0.053	0.019
cigsPerDay	0.33	-0.19	0.014	0.77	1	-0.047	-0.036	-0.07	-0.037	-0.03	-0.095	-0.057	-0.087	0.064	-0.054	0.052
BPMeds	-0.052	0.13	-0.014	-0.052	-0.047	1	0.11	0.26	0.049	0.094	0.27	0.2	0.11	0.013	0.054	0.089
prevalentStroke	-0.0023	0.051	-0.03	-0.038	-0.036	0.11	1	0.066	0.0096	0.013	0.061	0.056	0.036	-0.017	0.016	0.048
prevalentHyp	0.0014	0.31	-0.079	-0.11	-0.07	0.26	0.066	1	0.081	0.17	0.7	0.62	0.3	0.15	0.087	0.18
diabetes	0.014	0.11	-0.04	-0.042	-0.037	0.049	0.0096	0.081	1	0.048	0.1	0.051	0.089	0.061	0.61	0.093
totChol	-0.07	0.27	-0.014	-0.051	-0.03	0.094	0.013	0.17	0.048	1	0.22	0.17	0.12	0.093	0.05	0.091
sysBP	-0.045	0.39	-0.12	-0.13	-0.095	0.27	0.061	0.7	0.1	0.22	1	0.79	0.33	0.18	0.13	0.22
diaBP	0.052	0.21	-0.058	-0.12	-0.057	0.2	0.056	0.62	0.051	0.17	0.79	1	0.39	0.18	0.064	0.15
BMI	0.073	0.14	-0.14	-0.16	-0.087	0.11	0.036	0.3	0.089	0.12	0.33	0.39	1	0.074	0.084	0.082
heartRate	-0.12	-0.0027	-0.064	0.051	0.064	0.013	-0.017	0.15	0.061	0.093	0.18	0.18	0.074	1	0.097	0.021
glucose	0.0029	0.12	-0.032	-0.053	-0.054	0.054	0.016	0.087	0.61	0.05	0.13	0.064	0.084	0.097	1	0.12
TenYearCHD	0.092	0.23	-0.063	0.019	0.052	0.089	0.048	0.18	0.093	0.091	0.22	0.15	0.082	0.021	0.12	1

Figure 3: Feature Importance (Random Forest , XGBoost)



Random Forest Confusion Matrix

	0	1
0	456	170
1	78	412

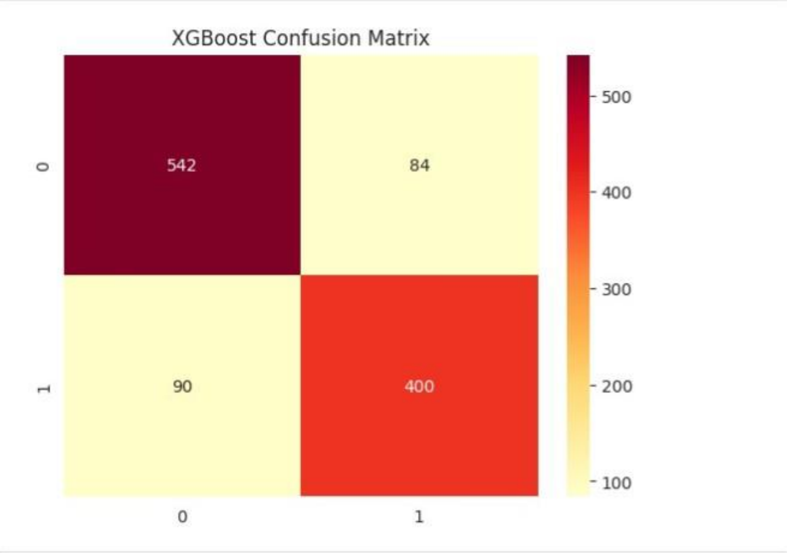
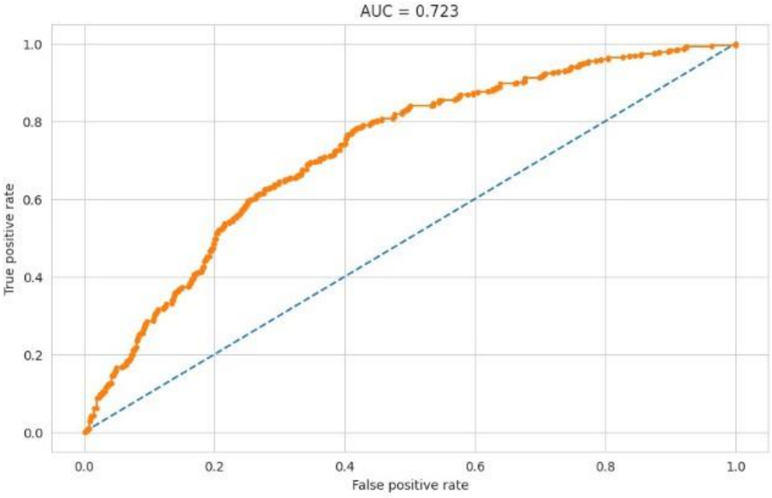
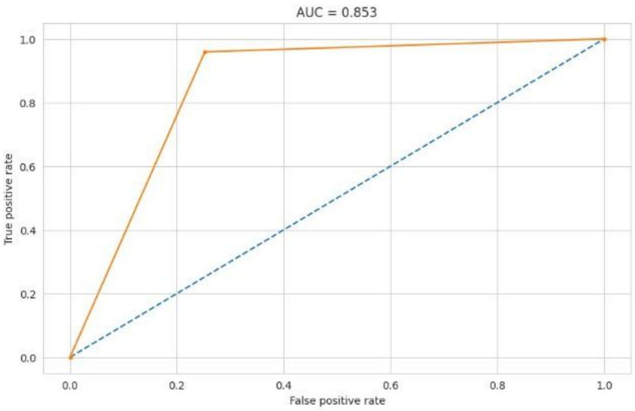


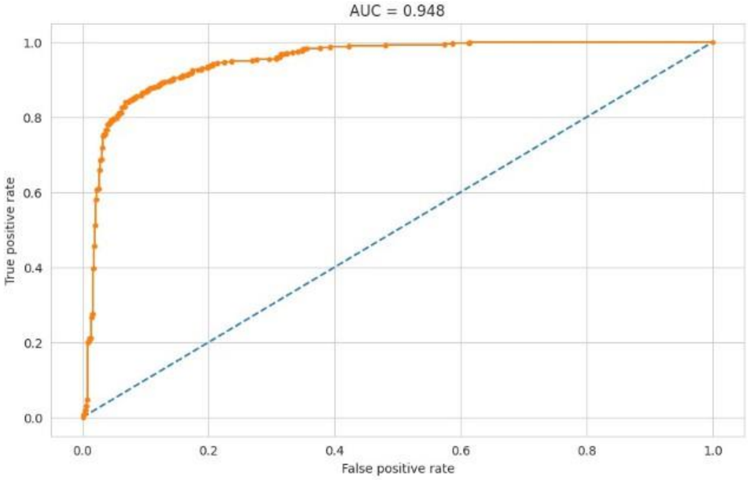
Figure 4: ROC-AUC Curve Comparing Models



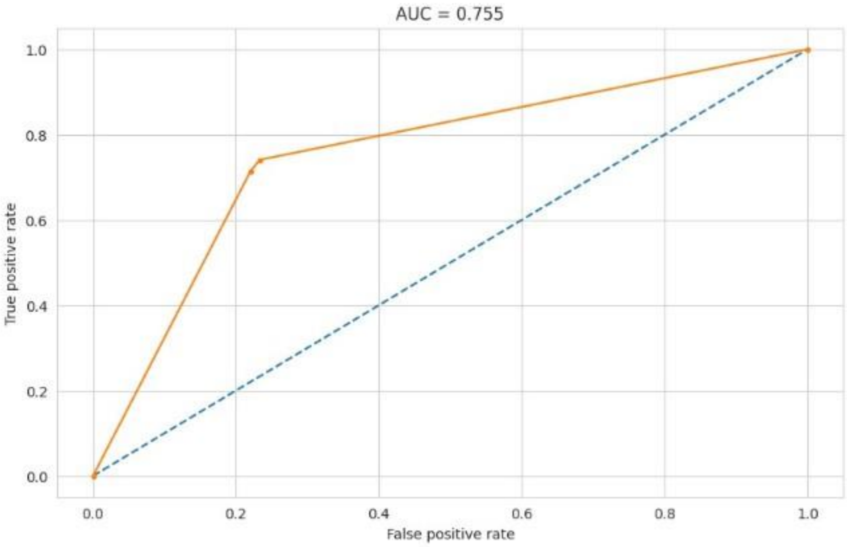
Logistic Regression



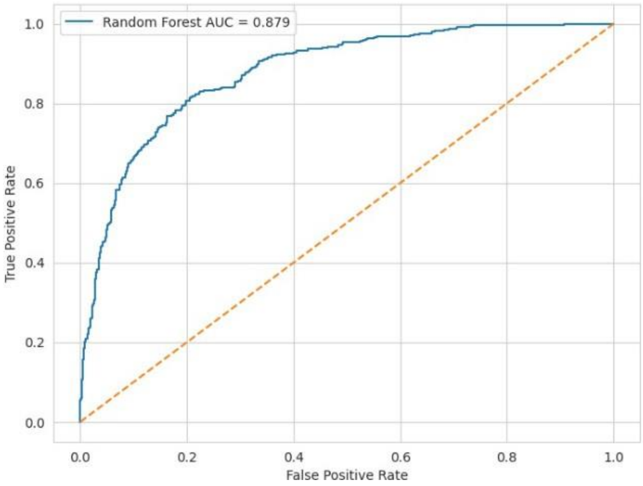
KNN



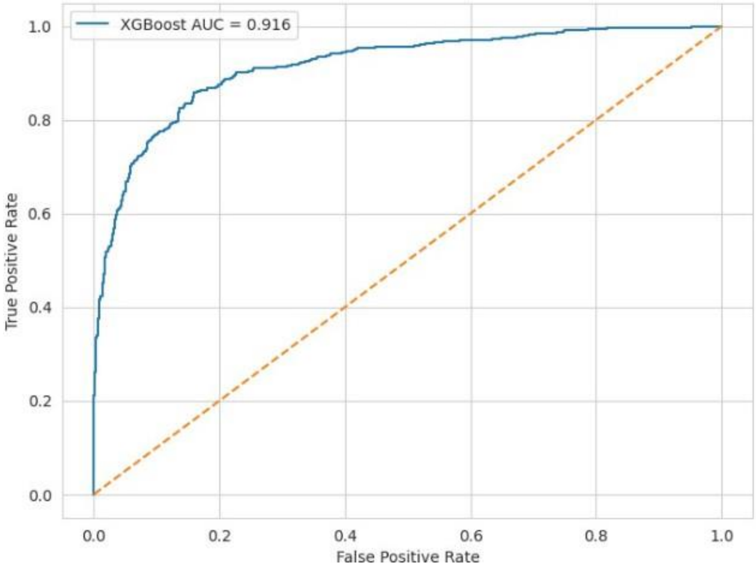
SVM



Decision tree



Random forest



XGBoost

5 Conclusion:

The comparison of the model results demonstrates that the machine learning-based approach is both consistent and effective in identifying and predicting the risk of Coronary Heart Disease (CHD). The models were able to capture the complex and nonlinear relationships among various clinical and lifestyle factors, such as age, blood pressure, cholesterol level, glucose, BMI, and smoking intensity. Through systematic preprocessing, feature selection, and model evaluation, the Support Vector Machine (SVM) emerged as the most accurate and reliable model, achieving the highest classification performance and discriminative capability.

The approach successfully highlights the most influential risk predictors while maintaining a strong balance between predictive accuracy and interpretability. Moreover, by incorporating class balancing techniques like SMOTE and feature importance analysis, the framework ensures fairness and clinical relevance in the prediction outcomes. The methodology is adaptable and can be applied to diverse patient populations and medical datasets, provided that data completeness and quality are maintained.

Overall, the findings reinforce the potential of machine learning to enhance early diagnosis and preventive healthcare strategies. By enabling timely risk assessment and individualized intervention, this approach can support healthcare professionals in reducing the burden of heart disease and improving patient outcomes on a broader scale.

6 References:

- Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., ... & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. The American Journal of Cardiology, 64(5), 304–310. https://doi.org/10.1016/0002-9149(89)90524-9

- Wilson, P. W., D’Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., & Kannel, W. B. (1998). Prediction of coronary heart disease using risk factor categories. Circulation, 97(18), 1837–1847. https://doi.org/10.1161/01.CIR.97.18.1837

- Kannel, W. B., Dawber, T. R., Kagan, A., Revotskie, N., & Stokes, J. (1961). Factors of risk in the development of coronary heart disease—six-year follow-up experience: The Framingham Study. Annals of Internal Medicine, 55(1), 33–50. https://doi.org/10.7326/0003-4819-55-1-33

- Taneja, S., & Sachdeva, A. (2020). Heart disease prediction using machine learning algorithms. International Journal of Computer Applications, 177(26), 6–11. https://doi.org/10.5120/ijca2020919938

- Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE Access, 7, 81542–81554. https://doi.org/10.1109/ACCESS.2019.2923707

- Alaa, A. M., & van der Schaar, M. (2018). Forecasting individualized disease trajectories using interpretable deep learning. Nature Communications, 9(1), 1–10. https://doi.org/10.1038/s41467-018-05581-7

- Ghosh, S., & Sinha, B. (2022). Comparative study of ML algorithms for cardiovascular disease prediction using Framingham dataset. Journal of Biomedical Informatics, 129, 104065. https://doi.org/10.1016/j.jbi.2022.104065

- Chaurasia, V., & Pal, S. (2017). A novel approach for heart disease prediction using data mining and cloud computing techniques. International Journal of Computer Science and Information Security, 15(5), 304–310.

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953