Group Members

ITBNM-2110-0053

ITBNM-2110-0082

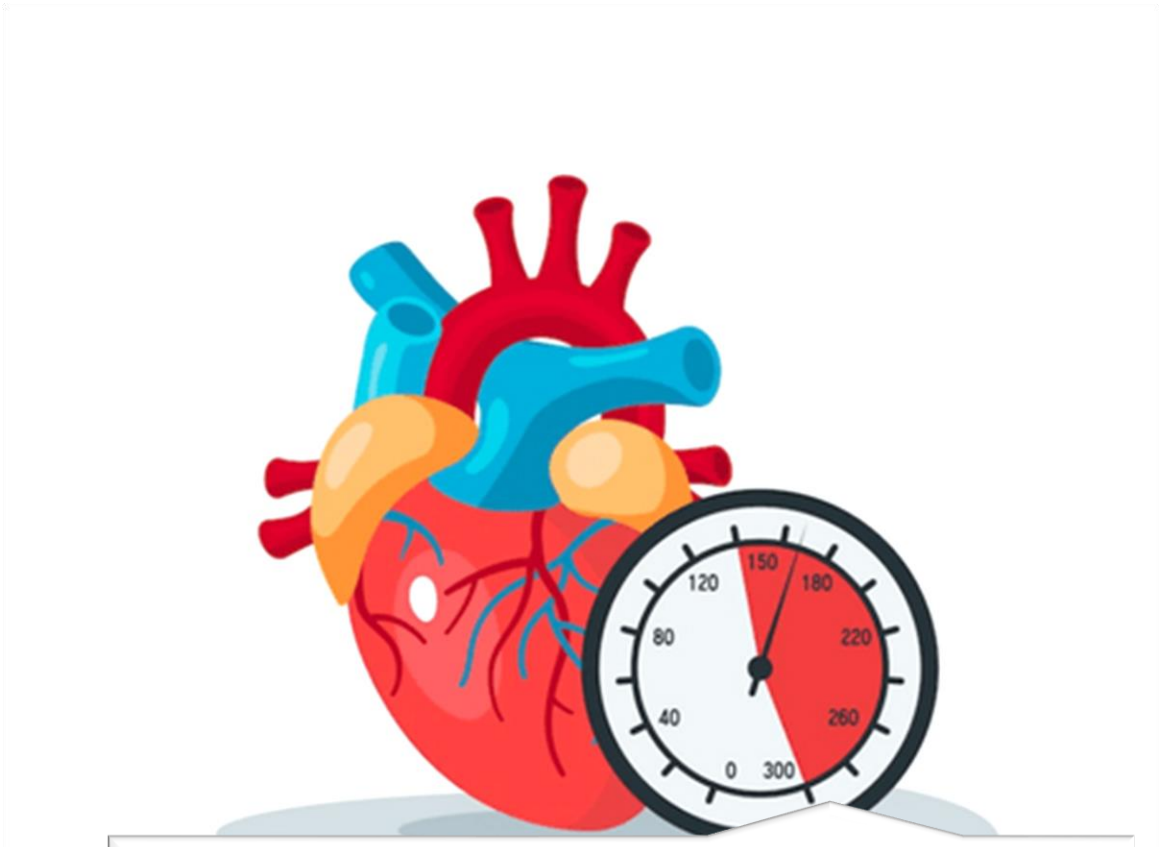
ITBNM-2110-0089

ITBNM-2110-0093

ITBNM-2110-0100

NATURE INSPIRED ALGORITHMS

MINI-PROJECT



Predicting the Probability of Cardiovascular Disease in Hypertension Patients Using Machine Learning

Problem

Cardiovascular disease (CVD) is a leading cause of mortality worldwide, and hypertension is a major risk factor. Accurately predicting the likelihood of a hypertension patient developing CVD can lead to better preventive measures and personalized treatment plans. The problem addressed in this project is the development of a machine learning model that predicts the probability of a hypertension patient developing CVD based on demographic, clinical, and lifestyle data.

Background

Hypertension, or high blood pressure, is a significant public health concern, affecting an estimated 1.13 billion people worldwide and contributing to approximately 10 million deaths annually (WHO, 2019). It is a major risk factor for cardiovascular disease (CVD), which remains the leading cause of mortality globally.[1]

Relationship Between Hypertension and Cardiovascular Disease

Hypertension's role in the development of CVD is well-documented. Studies have shown that elevated blood pressure leads to endothelial damage, arterial stiffness, and atherosclerosis, all of which are critical factors in the pathogenesis of CVD[2] [3]. The Framingham Heart Study, a landmark longitudinal study, highlighted that individuals with hypertension have a significantly higher risk of coronary artery disease, stroke, and heart failure compared to those with normal blood pressure.[4]

Predictive Analytics in Healthcare

The advent of predictive analytics and machine learning has revolutionized the healthcare industry, offering new avenues for early disease detection and personalized treatment plans. Machine learning models can analyze vast amounts of patient data, uncovering patterns and relationships that traditional statistical methods might miss.[5]

Several studies have successfully utilized machine learning techniques to predict cardiovascular outcomes. For instance, Khera [6] developed a predictive model using a combination of genetic and clinical data, significantly improving the accuracy of CVD risk prediction. Similarly, Weng [7] demonstrated that machine learning algorithms outperformed traditional risk prediction models in identifying patients at high risk of CVD.

Importance of Early Prediction

Early prediction of CVD in hypertension patients is crucial for implementing preventive measures and improving patient outcomes. Current guidelines emphasize the need for early intervention in high-risk individuals to prevent the progression of hypertension to more severe cardiovascular conditions. [8]

Data Sources and Model Development

Utilizing large-scale, high-quality datasets is essential for developing accurate and generalizable predictive models. The Framingham Heart Study dataset, with its comprehensive longitudinal data on cardiovascular risk factors, provides an excellent resource for model development. [9] Additionally, the MIMIC-III Clinical Database offers extensive clinical data from critical care patients, enhancing the model's robustness.[10]

In this project, we aim to develop a machine learning model to predict the probability of CVD in hypertension patients using data from these rich sources. By leveraging demographic, clinical, and lifestyle data, we aim to create a tool that can aid healthcare professionals in early identification and intervention for high-risk patients.

Availability of Data

Several publicly available datasets contain the necessary information to develop and validate the predictive model. The following datasets are particularly relevant:

1. **Framingham Heart Study Dataset**
    - This dataset includes a range of demographic, medical history, and clinical measurement data collected from participants in the Framingham Heart Study, a long-term cardiovascular cohort study.
2. **MIMIC-III Clinical Database**
    - The MIMIC-III database contains de-identified health-related data associated with over forty thousand critical care patients, including demographics, vital signs, laboratory tests, medications, and clinical outcomes.
3. **National Health and Nutrition Examination Survey (NHANES)**
    - NHANES provides comprehensive health and nutritional data, including variables related to hypertension and cardiovascular health, collected through interviews, physical examinations, and laboratory tests.
4. **Kaggle - Cardiovascular Disease Dataset**
    - This dataset contains information about patients with and without cardiovascular disease, including clinical measurements and lifestyle factors.
5. **Kaggle - Hypertension Dataset**
    - This dataset focuses on hypertension and includes various clinical and lifestyle factors. These datasets provide a robust foundation for developing and validating a machine learning model to predict the probability of CVD in hypertension patients.

References

[1] K. T. Mills, A. Stefanescu, and J. He, "The global epidemiology of hypertension," *Nat. Rev. Nephrol. 2020 164*, vol. 16, no. 4, pp. 223–237, Feb. 2020, doi: 10.1038/s41581-019-0244-2.

[2] P. K. Whelton, R. M. Carey, G. Mancia, R. Kreutz, J. D. Bundy, and B. Williams, "Harmonization of the American College of Cardiology/American Heart Association and European Society of Cardiology/European Society of Hypertension Blood Pressure/ Hypertension Guidelines Comparisons, Reflections, and Recommendations Keywords antihypertensive agents • blood pressure • cardiovascular diseases • hypertension • life style • practice guideline • public health," *Eur. Heart J.*, vol. 43, pp. 3302–3311, 2022, doi: 10.1161/CIRCULATIONAHA.121.054602.

[3] A. V. Chobanian *et al.*, "The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure: The JNC 7 Report," *JAMA*, vol. 289, no. 19, pp. 2560–2571, May 2003, doi: 10.1001/JAMA.289.19.2560.

[4] W. B. Kannel, T. R. Dawber, A. Kagan, and N. Revotskie, "Factors of Risk in the Development of Coronary Heart Disease-Six-Year Follow-up Experience The Framingham Study", Accessed: Jul. 29, 2024. [Online]. Available: http://annals.org/

[5] Z. Obermeyer and E. J. Emanuel, "Predicting the Future — Big Data, Machine Learning, and Clinical Medicine," *N. Engl. J. Med.*, vol. 375, no. 13, pp. 1216–1219, Sep. 2016, doi: 10.1056/NEJMP1606181/SUPPL_FILE/NEJMP1606181_DISCLOSURES.PDF.

[6] "From the Center for Human Genetic Re-search and Cardiology Division, Massa-chusetts General Hospital (A," *N Engl J Med*, vol. 375, pp. 2349–58, 2016, doi: 10.1056/NEJMoa1605086.

[7] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?," 2017, doi: 10.1371/journal.pone.0174944.

[8] W. P.K., C. R.M., A. W.S., and E. Al., "Acc/aha/aapa/abc/acpm/ags/APhA/ASH/ASPC/nma/pcna guideline for the prevention, Detection, evaluation, and management of high blood pressure in adults: a Report of the American College of Cardiology/American heart Association. Task force on clinical practice guidelines // J. Am. Coll. Cardiol. - 2017. - Nov 13," *Почки*, vol. 7, no. 1, pp. 68–74, Sep. 2018, doi: 10.22141/2307-1257.7.1.2018.122220.

[9] S. S. Mahmood, D. Levy, R. S. Vasan, and T. J. Wang, "The Framingham Heart Study and the epidemiology of cardiovascular disease: A historical perspective," *Lancet*, vol. 383, no. 9921, pp. 999–1008, Mar. 2014, doi: 10.1016/S0140-6736(13)61752-3.

[10] A. E. W. Johnson *et al.*, "Data Descriptor: MIMIC-III, a freely accessible critical care database," 2016, doi: 10.1038/sdata.2016.35.

Research paper outline

- **Introduction:**
  - Briefly introduce hypertension and its connection to cardiovascular disease (CVD).
  - State the problem: the need for an accurate predictive model for identifying CVD risk in hypertensive patients.
  - Introduce the importance of machine learning in healthcare.
- **Background and Literature Review:**
  - Discuss the relationship between hypertension and CVD.
  - Review relevant studies that have used machine learning models for CVD risk prediction.
  - Highlight the gaps in the literature that your study addresses.
- **Data and Methodology:**
  - **Dataset Description:** Provide details about the dataset used (Link to the dataset: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset)
  - **Preprocessing:** Describe the data cleaning and preprocessing steps (e.g., handling missing values, feature engineering).
  - **Feature Selection:** Explain why certain features were selected for the model (e.g., clinical relevance, correlation with the target variable).
  - **Machine Learning Models:** Describe the models used and the rationale behind choosing them.
  - **Evaluation Metrics:** Define the metrics used to evaluate the models (accuracy, precision, recall, AUC).
- **Results:**
  - Present the results of your model, including the performance of different machine learning algorithms.
  - Highlight the most important features contributing to CVD prediction.
  - Provide visualizations (e.g., ROC curve, confusion matrix) to support your results.
- **Discussion:**
  - Interpret the results in the context of clinical importance.
  - Compare your findings with previous studies.
  - Discuss the potential clinical applications of your model and its role in preventing cardiovascular events in hypertensive patients.
- **Conclusion and Future Work:**
  - Summarize your key findings.
  - Discuss the limitations of your study (e.g., dataset size, generalizability).
  - Propose future research directions, such as incorporating more diverse datasets or improving model interpretability.

Results -

Figure 4.1

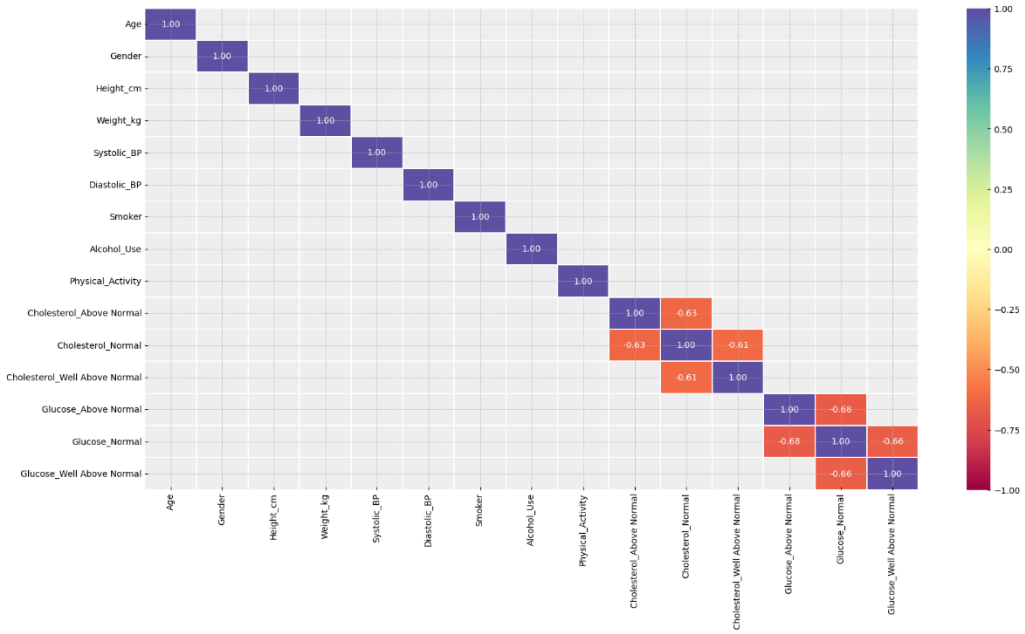


Figure 4.2

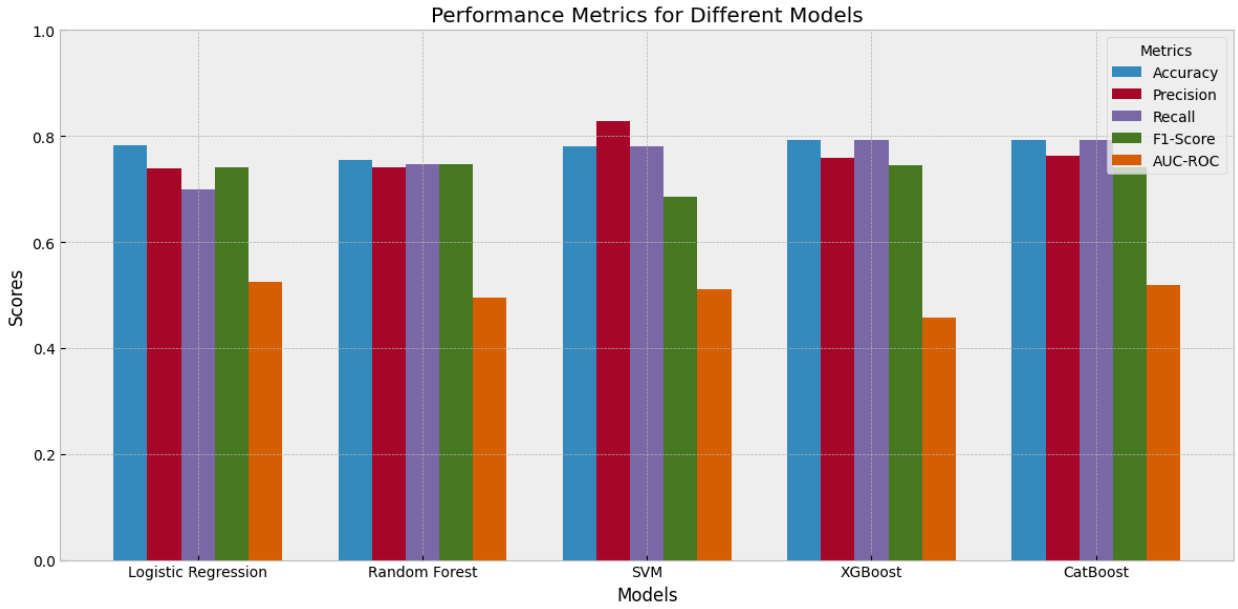


Figure 4.3

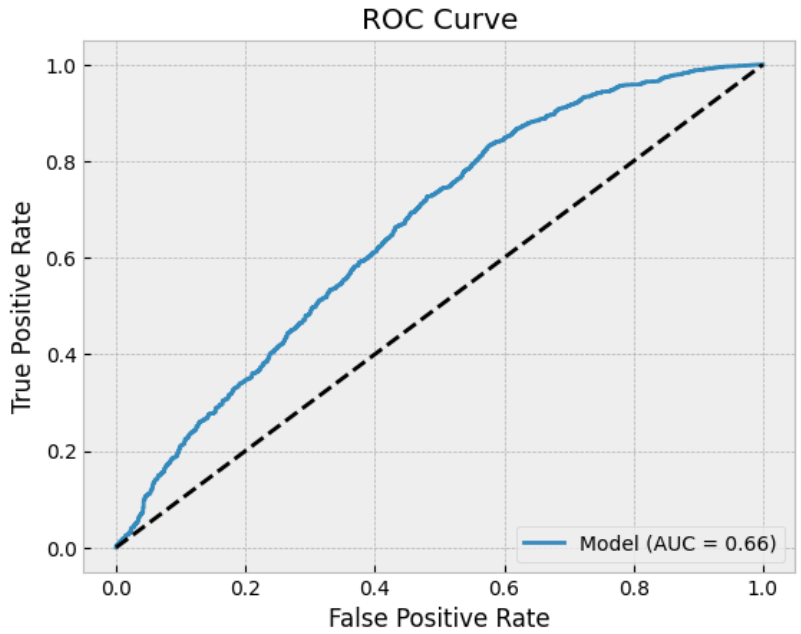


Figure 4.4

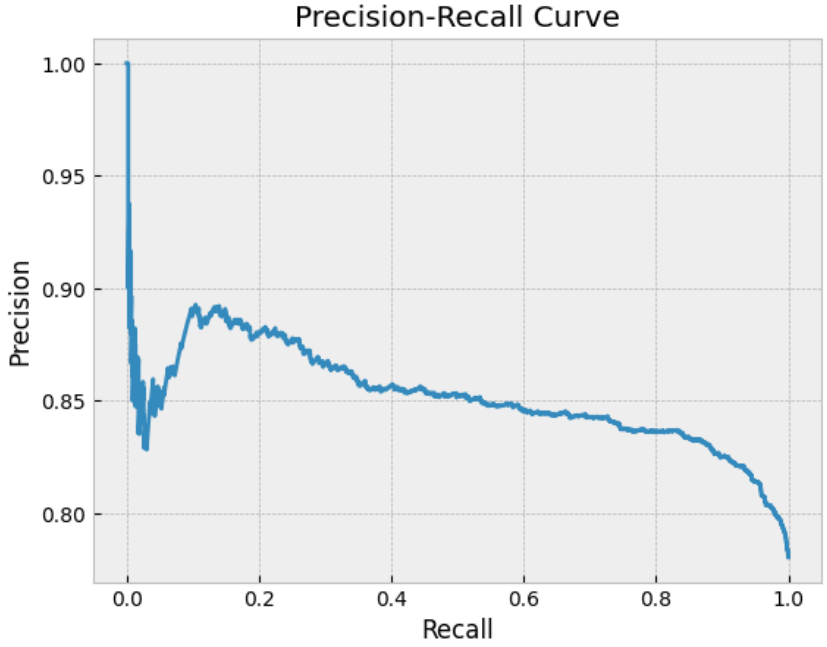


Figure 4.5

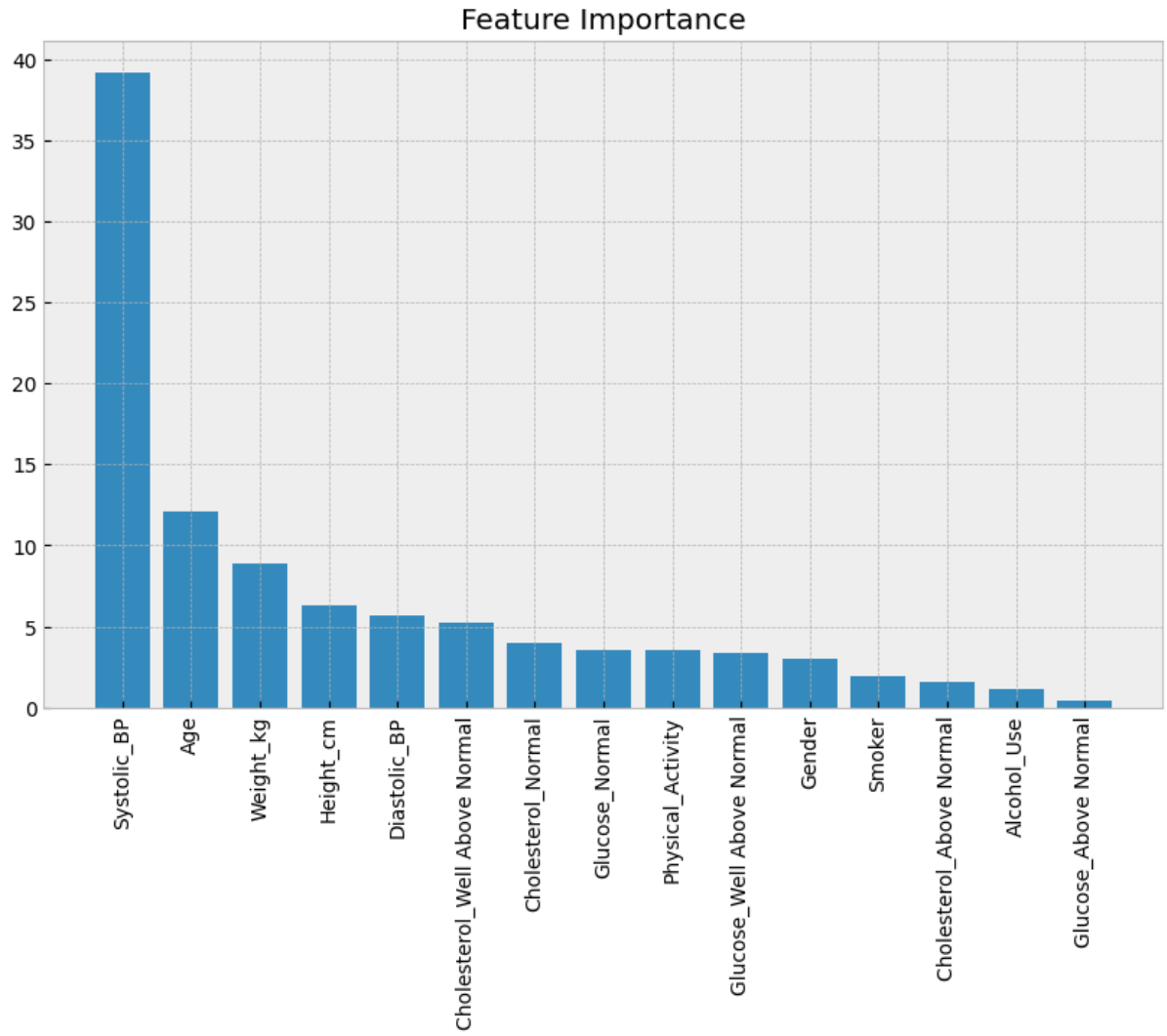

Feature Importance
40
35
30
25
20
15
10
5
0
Systolic_BP
Age
Weight_kg
Height_cm
Diastolic_BP
Cholesterol_Well Above Normal
Cholesterol_Normal
Glucose_Normal
Physical_Activity
Glucose_Well Above Normal
Gender
Smoker
Cholesterol_Above Normal
Alcohol_Use
Glucose_Above Normal

Figure 4.6

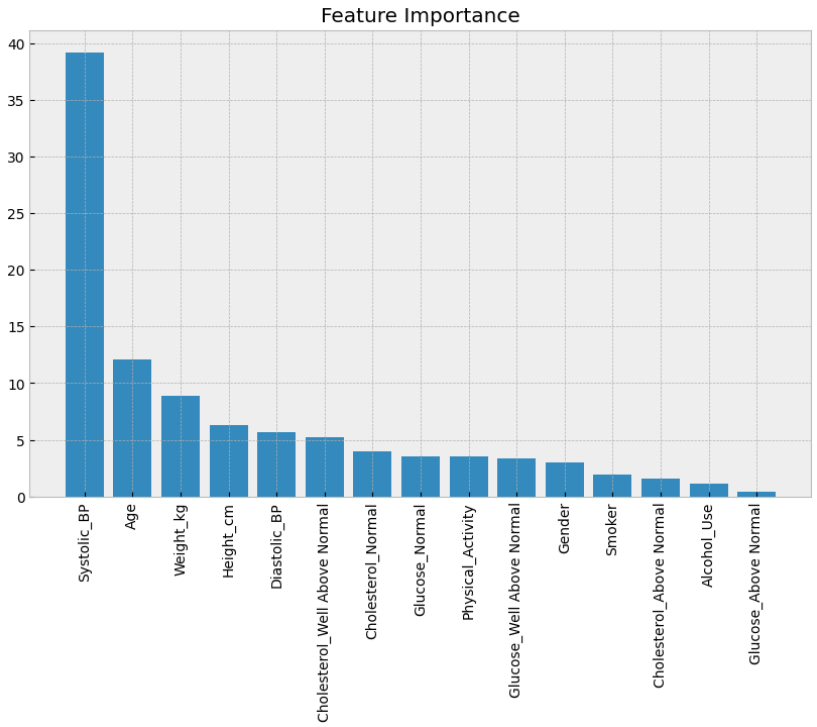
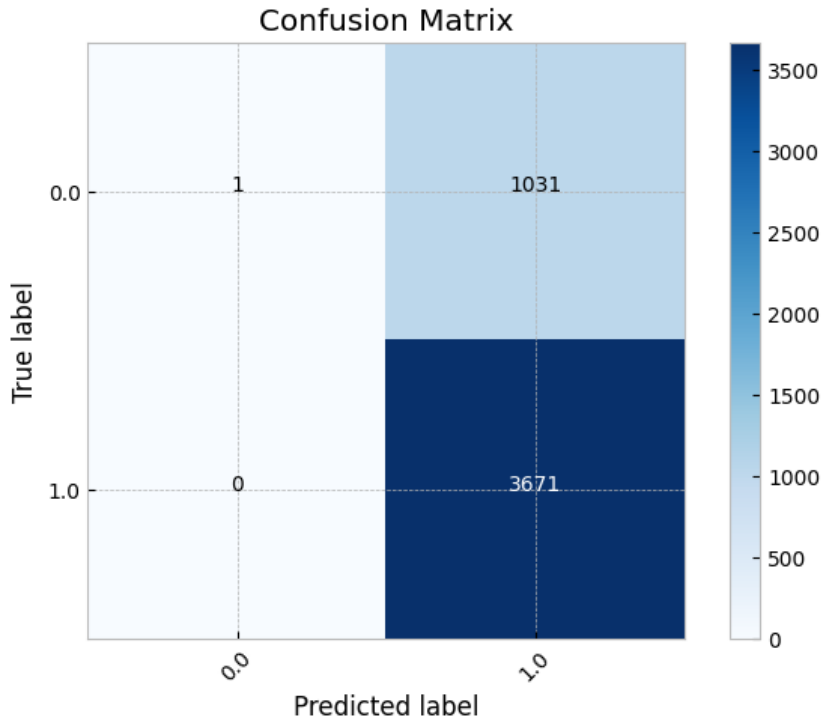


Figure 4.7



Here is a structured table based on the given performance metrics of the machine learning classifiers:

Model	Accuracy	Time Taken (seconds)	Precision (Class 0)	Precision (Class 1)	Recall (Class 0)	Recall (Class 1)	F1-Score (Class 0)	F1-Score (Class 1)	AUC-ROC
SVM Classifier	0.7808	60.6663	1.00000	0.78073	0.00097	1.00000	0.00194	0.87687	0.5111
Random Forest Classifier	0.7538	5.1240	0.42873	0.82901	0.36725	0.86244	0.39562	0.84539	0.4949
XGBoost Classifier	0.7920	0.1647	0.58654	0.80665	0.17733	0.96486	0.27232	0.87869	0.4571
CatBoost Classifier	0.7933	1.7968	0.61905	0.80319	0.15116	0.97385	0.24299	0.88033	0.5183
Logistic Regression	0.7827	0.1117	0.57576	0.78564	0.03682	0.99237	0.06922	0.87699	0.5237

Notes:

- **AUC-ROC** values represent the Area Under the Curve for the ROC curves of each model.
- Class 0 refers to the absence of CVD, and Class 1 refers to the presence of CVD.