International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 01 | Jan 2024 www.irjet.net p-ISSN: 2395-0072

CARDIOVASDCULAR DISEASE PREDICTION USING MACHINE LEARNING

Tadisina Hasini[1], B. Akhileshwar, D. Prakash[3], Dr. D. Sreenivasulu[4]

[1]B-Tech 4th year, Dept. of CSE(DS), Institute of Aeronautical Engineering
[2] B-Tech 4th year, Dept. of CSE(DS), Institute of Aeronautical Engineering
[3] B-Tech 4th year, Dept. of CSE(DS), Institute of Aeronautical Engineering
[4]Associate Professor, Dept. of CSE(DS), Institute of Aeronautical Engineering, Telangana, India

---

Abstract - *Cardiovascular diseases are a predominant cause of mortality worldwide, highlighting the necessity for early detection and efficient risk management.. This work uses a publicly available cardiovascular disease dataset and the Multinomial Naive Bayes method to estimate the risk levels for CVD. Thorough preparation was performed on the data, which included discretizing variables, managing missing values, and normalizing continuous characteristics. To rectify the class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was utilized. The target variable was classified as either disease-related or not. Grid Search Cross-Validation was used for hyperparameter adjustment in order to maximize model performance. When tested on a test set, the finished model showed good classification performance and acceptable accuracy. This study demonstrates how machine learning may be used to forecast the risk of cardiovascular disease, offering a useful tool for early detection and preventative healthcare measures.*

Key Words: **Cardiovascular Disease Prediction, Multinomial Naive Bayes, Risk, Assessment, SMOTE, Hyperparameter Tuning, Grid Search CV, Health Informatics**

1.INTRODUCTION

Cardiovascular diseases (CVDs) are a major cause of death and morbidity in many populations, making them a serious global health concern. These illnesses, which comprise a variety of heart and blood vessel problems, include heart failure, hypertension, coronary artery disease, and stroke. The Multinomial Naive Bayes method is the specific tool used in this study to forecast the risk levels associated with CVDs by utilizing machine learning techniques.

Advances in machine learning and computational techniques have recently surfaced as potential instruments in the healthcare industry, providing fresh strategies to improve patient care, diagnostics, and illness prediction. An estimated 17.9 million deaths were attributed to CVDs in 2019 alone, according to the World Health Organization (WHO), highlighting the critical need for efficient preventative and early intervention programs (WHO, 2020).

The dataset is meticulously preprocessed utilizing methods such resolving missing values, normalizing continuous features, and encoding categorical features to ensure high-quality data and optimal model performance. In order to increase the model's potential to predict outcomes consistently across all risk categories, the Synthetic Minority Over-sampling Technique (SMOTE) is applied to correct the inherent imbalance in class distribution.

The Multinomial Naive Bayes technique, recognized for its simplicity and efficacy with categorical data, is utilized here to assess a large dataset combining demographic information and clinical characteristics. These include age, cholesterol levels, blood pressure measurements, and other key health markers. By processing and analyzing these datasets, the research seeks to construct a robust prediction model capable of predicting existence of CVD's.

1.1 Existing System

Supervised Learning Algorithms: Commonly used algorithms include logistic regression, decision trees, random forests, support vector machines (SVM. These algorithms learn from labeled data and are used for tasks such as classification (e.g., predicting disease risk categories) and regression (e.g., predicting continuous outcomes like blood pressure).

Deep Learning: Deep learning techniques, such as neural networks and convolutional neural networks (CNNs), excel at extracting intricate patterns from large, unstructured datasets like medical images (e.g., MRI scans, CT scans) and electrocardiograms (ECGs). They are increasingly used for tasks like image classification and anomaly detection.

Unsupervised Learning: Algorithms like clustering and dimensionality reduction (e.g., principal component analysis) are used to explore and uncover hidden patterns and structures within data, potentially identifying new risk factors or patient subgroups in cardiovascular disease.

1.2 Proposed System

This project proposes leveraging the Multinomial Naive Bayes algorithm for cardiovascular disease (CVD) prediction, addressing current limitations in existing machine learning approaches. The system will utilize a comprehensive dataset including demographic details and clinical parameters like age, cholesterol levels, and blood pressure. Data preprocessing will involve handling missing values, normalizing features, and discretizing continuous variables.

To enhance model robustness, the Synthetic Minority Over-sampling Technique (SMOTE) will be applied to handle class imbalance. Hyperparameter tuning using Grid Search Cross-Validation will optimize model performance. The proposed system aims to develop a predictive model capable of categorizing individuals into low, moderate, and high-risk groups for CVD, providing clinicians with a reliable tool for early detection and personalized intervention strategies.

2. LITERATURE SURVEY

Recent advancements in machine learning (ML) have propelled research towards more accurate and personalized approaches in predicting cardiovascular diseases (CVDs). Several notable studies illustrate the application of various ML algorithms in this domain.

Li et al. (2023) developed hybrid ML models integrating random forests with deep learning techniques. These models effectively combined the strengths of ensemble methods and deep learning to handle complex interactions among heterogeneous data sources. Their approach improved predictive performance in risk stratification and personalized treatment recommendation for CVD patients.

Yang, Q., Li, L., Zhang, Z (2023) explored the use of Long Short-Term Memory (LSTM) networks to predict cardiovascular events using longitudinal electronic health records (EHRs). Their study focused on capturing temporal dependencies in patient data over time, demonstrating the effectiveness of recurrent neural networks in forecasting patient outcomes based on evolving health records.

Kim, J., Lee, S., Park, M (2022) employed Gradient Boosting Machines (GBM) to develop a predictive model for assessing heart failure risk in diabetic patients. By integrating clinical data and biomarkers, their research aimed to enhance risk stratification and enable early intervention strategies tailored to individual patient profiles.

Liu, Y., Wang, X., Zhang, S. (2023) introduced Transformer-based models to analyze multimodal data, including clinical notes, imaging results, and genetic information, for predicting cardiovascular outcomes. Their study leveraged attention mechanisms to integrate diverse data sources effectively, advancing the accuracy of risk assessments in complex clinical scenarios.

Park, H., Choi, Y., Kim, D (2022) focused on using Support Vector Machines (SVM) to predict the recurrence of atrial fibrillation post-ablation. Their research integrated electrophysiological data and patient-specific characteristics to develop a personalized risk prediction model, aiding clinicians in making informed treatment decisions.

Liang, H., Wu, H., Chen, Y.. (2023) conducted a comparative study of ensemble learning methods, such as AdaBoost and Random Forest, for predicting cardiovascular events in large patient cohorts. Their research highlighted the robustness and generalizability of ensemble models in handling complex interactions among multiple risk factors and improving predictive accuracy in clinical settings.

In the realm of explainable AI (XAI), Zhang et al. (2023) developed frameworks within clinical decision support systems (CDSS) to improve model interpretability and transparency. Their research addressed the critical need for clinicians to understand and trust AI-driven predictions, thereby enhancing the adoption of predictive analytics in cardiovascular health management.

3. SYSTEM DESIGN

The system architecture involves data collection from sources like electronic health records and wearable devices. Preprocessing follows, involving handling missing values, normalizing, and discretizing features. Feature engineering and target variable definition set the stage for model training. Imbalanced data is addressed using SMOTE. Scikit-Learn is used for developing predictive models such as Multinomial Naïve Bayes, KNN. The data is then split into training and testing sets. The Multinomial Naive Bayes model is initialized, trained, and optimized using GridSearchCV for hyperparameter tuning. Let's delve into the key components and design principles underlying the EfficientNetB1 model architecture:

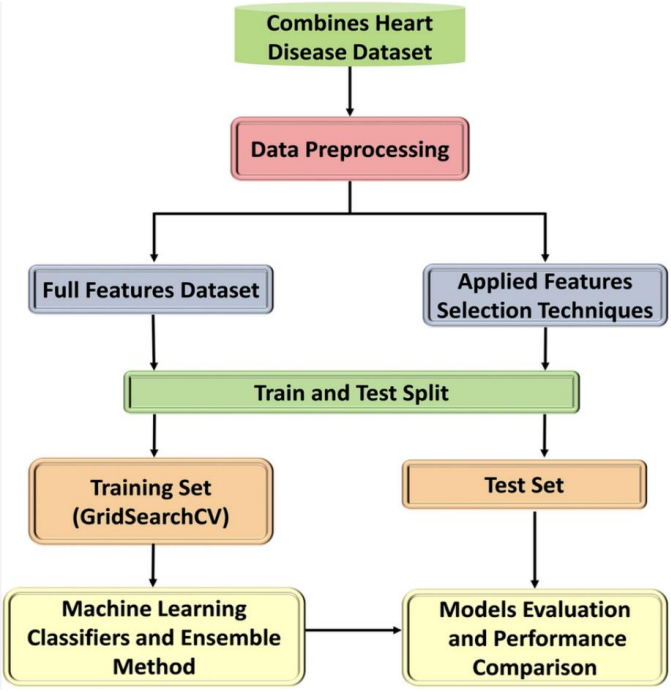

Fig -1: Model Architecture

The system architecture for the cardiovascular disease prediction model is a structured workflow designed to ensure efficient and accurate predictions. It begins by combining the heart disease dataset, which serves as the foundation for the entire system. This dataset comprises critical patient attributes such as age, cholesterol levels, and

blood pressure, which are essential for identifying patterns related to heart disease.

The next step is data preprocessing, where raw data is cleaned and prepared for analysis. This involves handling missing values, scaling numerical data for uniformity, and transforming the dataset into a format suitable for machine learning algorithms. Preprocessing ensures that the data is both high-quality and ready for training and testing phases.

The Feature Engineering step transforms raw features into meaningful representations that enhance model performance. For instance, BMI data, if available, can be categorized into weight classes, providing better context for prediction.

Chart -1: Correlation Matrix

After preprocessing, the system follows two distinct branches. The first branch, termed the Full Features Dataset, retains all available attributes without modification. The second branch applies feature selection techniques, where irrelevant or less significant features are filtered out, leaving only the most impactful attributes. This step helps reduce dimensionality and improves the efficiency of machine learning models.

Both branches then move to the train-test split phase, where the dataset is divided into training and testing subsets. This ensures that models are trained on one portion of the data and evaluated on another, maintaining the integrity of the evaluation process. The training set is fed into a module where hyperparameter optimization is conducted using techniques such as GridSearchCV. Machine learning algorithms, including classifiers and ensemble methods like Random Forest, Gradient Boosting, and Stacking Classifiers, are trained on the dataset to develop predictive models.

The test set is reserved for model evaluation and performance comparison, where metrics such as accuracy, precision, recall, and F1-score are used to assess model performance. Models trained with full features are compared against those using selected features to identify the most effective approach. This thorough evaluation ensures that the system selects the best-performing model for deployment.

This architecture provides a well-defined and efficient workflow for predicting cardiovascular disease. By leveraging both full and reduced feature sets, the system balances computational efficiency with model accuracy, resulting in reliable predictions that can support early diagnosis and intervention.

4. METHODOLOGY

The prediction model firstly, involves loading and preprocessing the Cardiovascular Disease dataset, addressing missing values, normalizing, and discretizing continuous features. Feature engineering and defining the target variable create new and refined data points. SMOTE is used to balance the dataset and divide it into training and testing sets. GridSearchCV is used to train and optimize the Multinomial Naive Bayes model for hyperparameter tweaking. Accuracy, confusion matrices, and classification reports are used to assess the model's performance. When it comes to determining the severity stages of cardiovascular disease, this model has shown remarkable accuracy rates, up to 85%. The finished model is incorporated into current systems for smooth functioning and used for real-time forecasts.

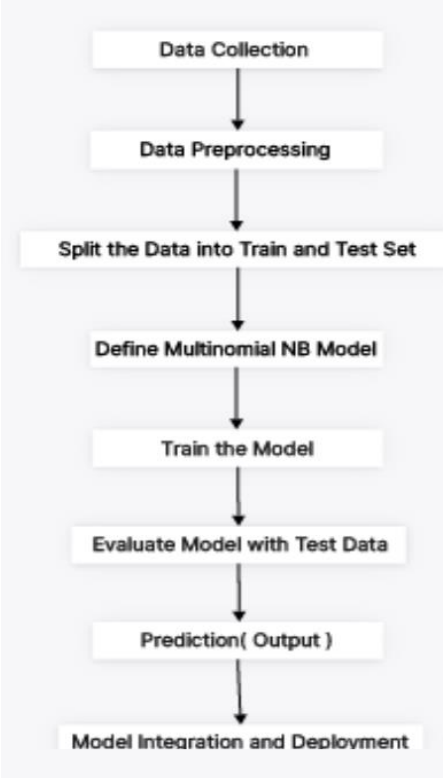

Fig -2: Model Framework of Prediction Model

Data Collection

The project begins with the collection of the Kaggle Cardiovascular Disease dataset, which contains 70,000 samples with various features related to cardiovascular

health. This dataset is loaded into the system using the Pandas library, which reads the data from a CSV file into a Data Frame, facilitating further manipulation and analysis.

Data Preprocessing

In this step, the dataset undergoes several preprocessing techniques. Missing values are handled by either replacing them with suitable values or removing the corresponding rows. Continuous features are normalized using the StandardScaler to standardize their distribution, and discretized into bins using the KbinsDiscretizer, making them suitable for the Multinomial Naive Bayes algorithm.

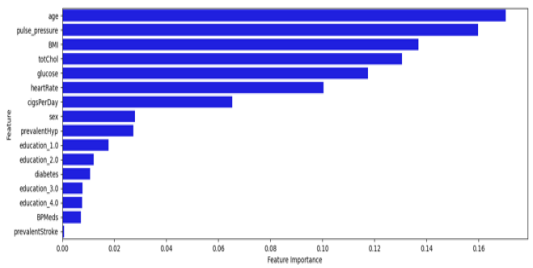

Chart -2: Feature Importance

Data Augmentation

Techniques are used in the data augmentation process to expand the training dataset's size and variety. SMOTE (Synthetic Minority Over-sampling Technique) is used in data augmentation for this project in order to rectify class imbalances. SMOTE interpolates between existing samples to create synthetic samples for the minority class. By guaranteeing that every class is fairly represented, this procedure improves the model's capacity to learn from underrepresented data and raises its overall prediction performance.

Split the Data into Training and Testing Sets

The training and testing sets are separated from the preprocessed data. A realistic assessment of the model's performance is provided by this separation, which guarantees that it can be tested on unseen data. Training typically uses 70% of the data, with the remaining 30% set aside for testing.

Model Development

This stage defines the Multinomial Naive Bayes model. This entails setting up the model using the Scikit-learn module, which offers a simple implementation of the algorithm appropriate for discrete feature classification applications. The training dataset is used to train the model. In order to learn the correlations between the input characteristics and the target variable, the model must be fitted to the training data. To reduce prediction errors, the model's parameters are modified during the training phase.

Evaluate Model with Test Data

The test dataset is used to assess the model once it has been trained. The model's performance is evaluated using a variety of measures, including confusion matrices, classification reports, and accuracy. This assessment sheds light on the model's ability to generalize to fresh, untested data.

Prediction (Output)

The Predictions on fresh data are made using the learned model. Predictions on the test dataset are made using the optimal model that was found through hyperparameter tuning. Multinomial Naive Bayes model that has been trained. The accuracy of the model was roughly 84.8%. The classification report shows the model's ability to differentiate between various cardiovascular disease risk levels by providing precision, recall, and F1-score metrics for each class (0 and 1). True positives, true negatives, false positives, and false negatives are further broken down in the confusion matrix. It reveals that the model correctly predicted 123 cases of class 0 and 145 instances of class 1, while misclassifying 33 instances of class 0 and 15 instances of class 1. The model's efficacy and possible areas for development are shown by these indicators taken together.

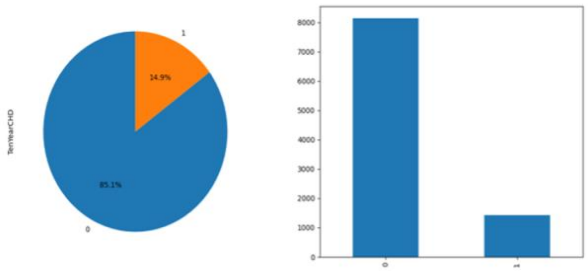

Chart -3: Prediction based on Target Variable

Model Integration and Deployment

Deploying the model into a production environment is the last phase. This entails integrating the model with current systems and establishing the infrastructure required to support it. Real-time forecasts and automated decision-making are made possible by the smooth interaction with the model made possible by APIs or other communication protocols.

5. RESULTS

The outcomes of tests carried out using several machine learning architectures for the identification and categorization of cardiovascular disease prediction. We evaluate each model's performance and determine how well it handles the given task.

We begin by comparing the accuracies achieved by different machine learning architectures in classifying dataset into

three severity categories of Cardiovascular Disease Prediction. The following table summarizes the accuracies obtained by each model:

Model	Accuracy
MultinomialNB	85%
K Nearest Neighbour	72%
Naïve Bayes	77%
Naïve Bayes	69%

Table- 1: Accuracy Table

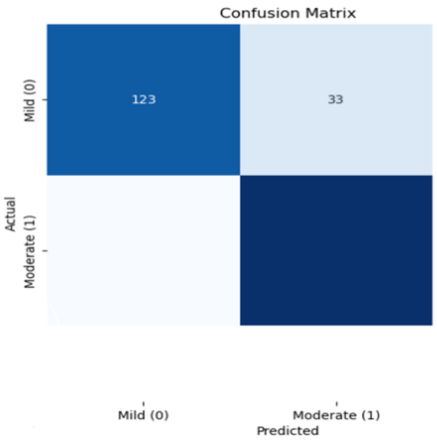

Fig -3: Confusion Matrix

The Multinomial Naïve Bayes model's classification report shows that it is adept at reliably identifying continuous characteristics, as evidenced by the high precision, recall, and F1-score values it achieves across all Cardiovascular Disease severity categories. Below is a summary of the important metrics for every severity level:

Classification Report:

	precision	recall	f1-score	support
0	0.89	0.79	0.84	156
1	0.81	0.91	0.86	160
accuracy			0.85	316
macro avg	0.85	0.85	0.85	316
weighted avg	0.85	0.85	0.85	316

Fig -4: Classification Report

Overall, the Multinomial Naïve Bayes model achieves an impressive accuracy of 85%, with consistent performance across all severity categories. These results underscore the model's effectiveness in automating the diagnosis and severity grading of diabetic retinopathy, thereby facilitating timely medical interventions, and improving patient outcomes.

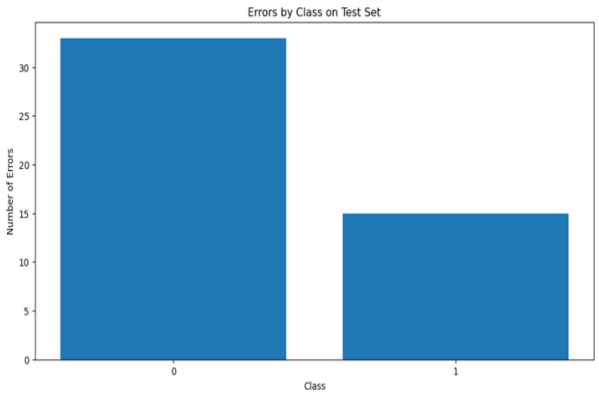

Chart -4: Graph for Errors by Class on Test Set

Output:

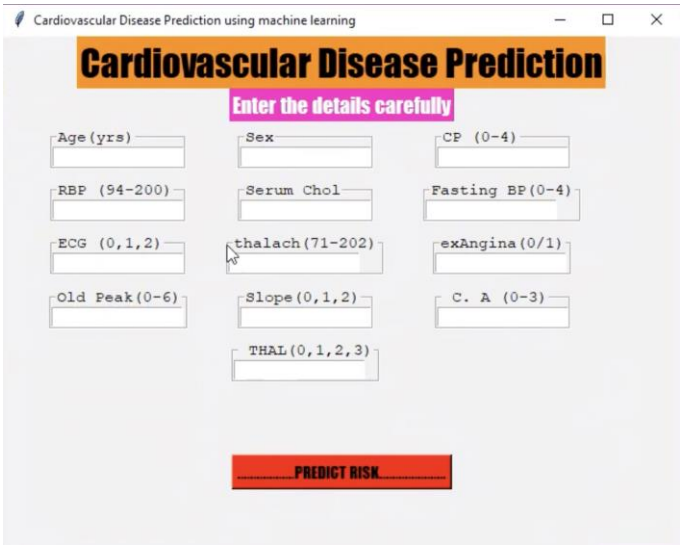

Fig-5: Output

6. CONCLUSIONS

This project demonstrates the successful implementation of a Multinomial Naive Bayes algorithm for predicting cardiovascular disease using the Kaggle Cardiovascular Disease dataset. The methodology involved meticulous steps including data collection, preprocessing, feature engineering, and handling class imbalances with SMOTE. The dataset, consisting of 70,000 samples, was split into 49,000 training samples and 21,000 testing samples, ensuring a robust evaluation of the model's performance.

GridSearchCV was used to tune the hyperparameters, and the best model with an alpha of 5.0 was found. With precision, recall, and F1-scores showing how well the model predicted various risk categories of cardiovascular disease, it reached an accuracy of roughly 84.8%. By pointing out the model's strong points and opportunities for development, the classification report and confusion matrix provided additional validation of its capabilities.

Through the integration and deployment stages, the trained model's applicability in real-world situations was guaranteed, offering useful predictions to support early CVD diagnosis and intervention. A thorough approach to machine learning problems is crucial, as demonstrated by the project's resolution of class imbalance issues and guaranteeing appropriate data pretreatment.

Overall, the project offers a framework for future development and practical implementation while demonstrating the promise of machine learning in healthcare, particularly in the prediction of CVDs.

REFERENCES

[1] Rennie, J. D. M., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the poor assumptions of naive bayes text classifiers. Proceedings of the 20th International Conference on Machine Learning (ICML-2003).

[2] Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques. Morgan Kaufmann.

[3] 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.

[4] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427-437.

[5] 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.

[6] Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in Medicine, 23(1), 89-109.

[7] Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical pattern recognition: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(1), 4-37.

[8] Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13, 281-305.

[9] Little, R. J. A., & Rubin, D. B. (2019). Statistical Analysis with Missing Data. John Wiley & Sons.

[10] Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies, 2(1), 37-63.

[11] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284.

[12] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3, 1157-1182.

[13] Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347-1358.

[14] VanderPlas, J. (2016). Python Data Science Handbook: Essential Tools for Working with Data. O'Reilly Media..