A blue and white logo

Description automatically generated

**Machine Learning**

*Michael Zelenetz*

**Final Project: Heart Disease Prediction**

Report

Submitted by

*Venkatalakshmi Kottapalli*

Artificial Intelligence

Katz School of Science and Health

Yeshiva University

New York

December 2024

**Introduction:**

The Heart disease is one of the leading causes of death globally. Early identification of individuals at risk, particularly for heart attacks, can significantly improve survival rates by enabling timely interventions and preventive care. With the advancements in machine learning, predictive models can analyze health data and accurately predict heart disease risk, helping medical professionals identify and treat high-risk patients more effectively.

**Stakeholder and Problem statement:**

I am a Data Scientist collaborating with a healthcare organization that is focused on improving early detection and reducing the risk of heart disease. The organization aims to identify individuals at high risk of developing heart disease before they experience severe health complications such as heart attacks, strokes, or other cardiovascular events.

My role involves developing predictive models to assess heart disease risk in patients and provide actionable insights, enabling timely interventions and personalized care plans.

**Solution:**

The link to the code: https://github.com/Venkatalakshmikottapalli/Heart-Risk-prediction

**Dataset:**

The dataset used for this project is [Heart Disease Health Indicators Dataset (kaggle.com)](https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset). This dataset containing diverse features such as HeartDiseaseorAttack, HighBP, HighChol, CholCheck, BMI, Smoker, Stroke, Diabetes, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, GenHlth, MentHlth, PhysHlth, DiffWalk, Sex, Age, Education, and Income. These features provide a comprehensive view of patients' health, lifestyle, and socio-economic factors, making the dataset well-suited for developing robust heart disease prediction models.

**Feature Engineering and Preprocessing:**

* **Data Cleaning:** Removed less important columns like 'NoDocbcCost,' ‘AnyHealthcare,’ 'HvyAlcoholConsump,' and 'CholCheck' from dataset based on feature importance from the Random Forest model and EDA insights.
* **Feature Creation**: Introduced two new features 'Red Flag' (combining HighBP and HighChol) and combined 'PreDiabetes' and 'Diabetes' into a single 'Diabetes' feature. However, these features did not improve model performance and instead slightly reduced its performance. Therefore, it seems that the original features were more effective for the model than the newly engineered ones.
* **Encoding:** Applied one-hot encoding for categorical variables and ordinal encoding for ordered data.
* **Scaling:** Standard scaling was applied to ensure consistent data range across features.
* **Class Balancing:** Used Synthetic Minority Over-Sampling Technique (SMOTE) to address class imbalance in the dataset.

**Models Employed:**

I employed the following machine learning models, evaluating each to determine the most effective solution for predicting heart disease risk.

**Naive Bayes (NB):**

Naive Bayes acts like a quick triage system in a clinic. It assesses each health factor independently, giving a probabilistic score of heart disease risk based on individual features like smoking or BMI.

**Why NB Works for Heart Disease Prediction:**

* It provides fast and interpretable predictions, making it useful for initial screenings.
* Despite assuming feature independence, it efficiently handles categorical and continuous data, providing a baseline for comparison.

**Logistic Regression (LR):**

Logistic Regression is like a simplified risk calculator that estimates the likelihood of heart disease based on weighted factors such as age, cholesterol, and smoking status.

**Why LR Works for Heart Disease Prediction:**

* It is transparent and easy to interpret, helping healthcare professionals understand the relationship between features and outcomes.
* Effective for identifying linear relationships and serves as a baseline model for more complex approaches.

**Decision Trees (DT):**

A Decision Tree works like a decision-making flowchart a doctor might use. It asks a series of "yes/no" questions about patient features (for example: "Is the BMI high?") to determine heart disease risk.

**Why DT Works for Heart Disease Prediction:**

* Provides clear and interpretable rules, useful for explaining predictions to medical staff.
* Handles both numerical and categorical data, modeling non-linear relationships effectively.

**Random Forest (RF):**

Random Forest acts like a panel of medical experts. Each "expert" (decision tree) reviews different aspects of the patient’s health and makes a prediction. The final decision is based on a majority vote.

**Why RF Works for Heart Disease Prediction:**

* Captures complex interactions between features (e.g., how age and cholesterol interact).
* Robust against overfitting by averaging predictions from multiple trees.
* Provides feature importance rankings, highlighting critical risk factors for targeted interventions.

**Gradient Boosting (GB):**

Gradient Boosting is like a series of specialists correcting each other’s mistakes. Each model improves on the errors of the previous one, refining the diagnosis.

**Why GB Works for Heart Disease Prediction:**

* Builds highly accurate models by focusing on difficult-to-predict cases.
* Effective for handling imbalanced datasets, ensuring that high-risk patients are detected.

**XGBoost (XGB):**

XGBoost is like an advanced diagnostic tool optimized for speed and accuracy. It quickly processes large amounts of patient data to make precise risk assessments.

**Why XGB Works for Heart Disease Prediction:**

* Efficient and scalable, ideal for large datasets.
* Handles complex patterns and imbalances, ensuring high-risk individuals are accurately identified.

**AdaBoost:**

AdaBoost acts like a team of doctors who focus on difficult cases. Each iteration emphasizes cases that were previously misclassified, improving overall accuracy.

**Why AdaBoost Works for Heart Disease Prediction:**

* Enhances weak learners, ensuring subtle patterns are not overlooked.
* Effective in boosting performance on complex datasets.

**Artificial Neural Networks (ANN):**

ANN functions like an advanced brain scan, detecting complex, non-linear patterns in patient data that might not be obvious to traditional methods.

**Why ANN Works for Heart Disease Prediction:**

* Captures intricate relationships between features, such as how lifestyle and medical history combine to impact risk.
* Flexible and powerful, suitable for handling complex healthcare data.

**Convolutional Neural Networks (CNN):**

CNN is like analyzing X-rays in detail, though applied here to tabular data by reshaping inputs to find hidden patterns.

**Why CNN Was Tested:**

* Investigates deep hierarchical relationships between features.
* Explores unconventional approaches to enhance prediction accuracy.

**Recurrent Neural Networks (RNN):**

RNN is like tracking a patient’s health history over time. Although not sequential data, it was tested to uncover any temporal dependencies.

**Why RNN Was Tested:**

* Evaluates potential sequences or patterns within health data.
* Ensures no time-based relationships are missed, providing a comprehensive evaluation.

**Models Results:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **ROC-AUC** | **Precision (No/Yes)** | **Recall (No/Yes)** | **F1-Score (No/Yes)** | **Comment** |
| **Naive Bayes** | 0.712 | 0.799 | 0.69 / 0.74 | 0.77 / 0.66 | 0.73 / 0.70 | Basic model; limited performance on complex data. |
| **Logistic Regression** | 0.763 | 0.837 | 0.78 / 0.75 | 0.73 / 0.80 | 0.75 / 0.77 | Solid, interpretable; struggles with non-linear patterns. |
| **Random Forest** | 0.94 | 0.984 | 0.92 / 0.96 | 0.96 / 0.92 | 0.94 / 0.94 | Excellent accuracy; handles complex data well. |
| **AdaBoost** | 0.85 | 0.931 | 0.86 / 0.84 | 0.84 / 0.86 | 0.85 / 0.85 | Strong performance; sensitive to noisy data. |
| **Gradient Boosting** | 0.893 | 0.964 | 0.90 / 0.89 | 0.89 / 0.90 | 0.89 / 0.90 | Very accurate; captures complex patterns. |
| **XGBoost** | 0.938 | 0.98 | 0.90 / 0.98 | 0.98 / 0.89 | 0.94 / 0.94 | High accuracy; optimized for speed and performance. |
| **ANN** | 0.869 | 0.949 | 0.85 / 0.89 | 0.89 / 0.85 | 0.87 / 0.87 | Good performance; requires more tuning. |
| **RNN** | 0.859 | 0.946 | 0.81 / 0.92 | 0.93 / 0.79 | 0.87 / 0.85 | Performs well; better suited for sequential data. |
| **CNN** | 0.88 | 0.96 | 0.87 / 0.89 | 0.89 / 0.87 | 0.88 / 0.88 | Strong performance; computationally intensive. |

Based on the results above, Random Forest and XGBoost stand out as the best performing models compared to the others.

**Hyperparameter Tuning:**

Hyperparameter tuning was performed for the best-performing models Random Forest and XGBoost using RandomSearchCV to optimize their performance.

**Hyperparameters Tuned:**

**Random Forest:**

* n\_estimators: Controls the number of trees. More trees generally improve performance but increase computation time.
* max\_depth: Limits tree depth to prevent overfitting and ensure better generalization.
* min\_samples\_split: Sets the minimum samples required for a split, reducing overfitting by avoiding splits based on small samples.

**XGBoost:**

* learning\_rate: Controls the contribution of each tree to the final prediction. A lower rate helps prevent overfitting but requires more trees.
* n\_estimators: Sets the number of boosting rounds. Too few may cause underfitting; too many may cause overfitting.
* max\_depth: Controls tree complexity. Deeper trees capture more patterns but may overfit, so tuning helps balance performance.

These tuned hyperparameters helped in slight improvement in model accuracy for heart disease prediction.

**Evaluation Metrics:**

Given the class imbalance in heart disease cases, ROU-AUC, precision, recall, and F1 score were prioritized:

* **Precision:** Ensures predictions of heart disease are accurate, minimizing false positives. This is critical to avoid unnecessary alarms in a medical setting, where incorrect predictions could lead to unnecessary treatment or anxiety.
* **Recall:** Focuses on detecting actual heart disease cases, ensuring true positives are captured. This is vital to prevent missing diagnoses, which could lead to severe health complications.
* **F1 Score:** Balances precision and recall, offering a comprehensive view of model performance and ensuring that both false positives and false negatives are minimized.
* **ROC-AUC:** Measures the model’s ability to distinguish between classes, providing an overall performance indicator that is useful for comparing different models.

**Hyper Tuning Results:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Best Parameters** | **Accuracy** | **ROC-AUC** | **Precision (No HD / HD)** | **Recall (No HD / HD)** | **F1-Score (No HD / HD)** |
| **Random Forest** | n\_estimators: 200 | 94.16% | 0.984 | 0.92 / 0.96 | 0.96 / 0.92 | 0.94 / 0.94 |
| max\_depth: None |
| min\_samples\_split: 2 |
| **XGBoost** | n\_estimators: 300 | 93.85% | 0.98 | 0.90 / 0.98 | 0.98 / 0.89 | 0.94 / 0.94 |
| max\_depth: 7 |
| learning\_rate: 0.15 |

**Final Model:**

The Random Forest model emerged as the final choice due to its robust performance, high accuracy, and interpretability. It effectively captures complex relationships, making it a reliable tool for early heart disease detection.

**User Interface:**

The user interface, developed with Streamlit and Flask, provides an intuitive and efficient platform for users to interact with the heart disease prediction model. By integrating data input, processing, and output in a seamless workflow, the UI empowers individuals to gain quick insights into their heart disease risk based on their personal health information. Integration with Flask ensures smooth communication between the front-end and the model, allowing for real-time predictions and interactive user experience.

**Deployment:**

The model will be deployed through a user-friendly Streamlit interface, allowing users to input their health information and receive heart disease predictions in real-time. This interactive tool will enable healthcare providers and patients to assess heart disease risk easily, ensuring timely and informed decision-making.

**Future Work:**

* **Data Enrichment**: I would enhance the model by adding diverse data like family history, stress levels, and cholesterol results to improve predictive accuracy.
* **Marketing and Deployment**: I plan to market the model to healthcare organizations and make it accessible to patients via a web platform for early detection and personalized care.

**Recommendation**:  
Yes, I recommend the client use this model. With an accuracy of 94.16% and a ROC-AUC of 98.44%, it effectively distinguishes between heart disease and no heart disease. The high precision (0.96) minimizes false positives, while the good recall (0.92) ensures most heart disease cases are detected.

The balanced F1-score (0.94) demonstrates strong performance in both detection and accuracy, making it suitable for early heart disease detection and intervention.