**Project - Heart Disease Prediction**

**Author: Venkatalakshmi Kottapalli**

**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.

As a data scientist collaborating with healthcare organizations, my goal is to reduce the risk of heart attacks by developing models that can predict which individuals are at high risk. By analyzing health indicators such as blood pressure, cholesterol levels, lifestyle habits, and demographics, we can provide earlier interventions and allocate healthcare resources more effectively.

**Stakeholder:**

My primary stakeholder is a healthcare organization aiming to improve the early detection of heart disease among patients. They seek to leverage predictive modeling to enhance their diagnostic capabilities, allowing for timely interventions and better patient outcomes.

**Problem Statement:**

The organization is trying to solve the problem of accurately predicting heart disease based on patient data. Early detection is critical for effective treatment and reducing the risk of severe health complications. The challenge lies in achieving high accuracy and reliability in predictions to support clinical decision-making.

**Dataset:**

**Source**: [Heart Disease Health Indicators Dataset (kaggle.com)](https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset)

**Insights from EDA:**

1. **Hear Disease and High Cholesterol**

**A graph of heart disease

Description automatically generated**

Comment:

More than half of the individuals with heart disease also have high cholesterol, indicating that high cholesterol has a significant impact on heart disease.

1. **Heart Disease and High BP**

**A graph of blood pressure

Description automatically generated**

Comment:

More than half of the individuals with heart disease also have high BP, indicating that high BP has a significant impact on heart disease.

1. **Heart Disease and Smoking status**

**A graph of a graph with text

Description automatically generated with medium confidence**

**Comment:**

More than half of the individuals with heart disease also Smoking habbit, indicating that Smoking has a significant impact on heart disease.

**4. Heart Disease and Age**

**A diagram of a heart disease

Description automatically generated**

**Comment:**

Most of the individuals with heart disease are elders, indicating that age has a significant impact on heart disease.

**5. Heart Disease and Sex**

**A graph of a heart disease

Description automatically generated**

**Comment:**

More than half of people with heart disease are men,indicating that high sex has a significant impact on heart disease**.**

**6. Heart Disease and Chol Check**

**A graph of a heart disease

Description automatically generated**

**Comment:**

Most of the individuals with heart disease are checking cholesterol and most of the people with no heart disease are not checking cholestrom indicating that cholesterol check has no impact on heart disease.

**7. Heart Disease and No doctor because of cost**

**A graph with blue and orange squares

Description automatically generated**

**Comment:**

Most of the individuals with heart disease are visiting doctors, indicating that no doctor because of cost has no impact on heart disease.

**8. Heart Disease and Any Health care**

**A graph of heart disease

Description automatically generated**

**Comment:**

Most of the individuals with heart disease are taking health care, most of the people with no heart disease are taking health care, indicating that any health care has no impact on heart disease.

**9. Heart Disease and Alcohol Consumption**

**A graph of heart disease

Description automatically generated**

**Comment:**

Most of the individuals with heart disease do not consume alcohol and most of the people with no heart disease do not consume alcohol, indicating that cholesterol check has no impact on heart disease.

**Selected Features:**

For predicting heart disease, I selected the following features based on exploratory data analysis insights:

* **Health Indicators:** Blood Pressure (BP), Cholesterol Levels, Diabetes Status, Body Mass Index (BMI)
* **Lifestyle Factors:** Smoking Status, Physical Activity, Fruit and Vegetable Consumption
* **Demographics:** Gender, Age, Education Level, Income

**Removed Features**

The following features were excluded based on exploratory data analysis insights, which showed they do not significantly impact heart disease:

* Alcohol Consumption
* No Doctor Due to Cost
* CholCheck
* Any Healthcare

**Models Tried:**

1. **Naive Bayes:** Chosen for its simplicity in implementation and understanding.
2. **Decision Trees:** Selected for their ability to model decision-making processes using a tree-like structure.
3. **Random Forest:** Used for its ensemble learning approach, leveraging multiple decision trees for improved accuracy.
4. **Gradient Boosting:** Chosen for its ensemble learning technique and utilization of gradient descent to optimize performance.
5. **Logistic Regression:** Selected for its straightforward interpretation of probabilities in classification tasks.
6. **XGBoost:** Chosen for its speed and performance, especially with large datasets.

**Models Results:**

**Naive Bayes**

* **Accuracy**: 71.5% - overall correctness in predictions.
* **ROC-AUC**: 0.80 - good at distinguishing between heart disease and no heart disease.
* **Precision**: No Heart Disease: 69%, Heart Disease: 75% - reflects accuracy in positive and negative predictions.
* **Recall**: No Heart Disease: 78%, Heart Disease: 65% - captures sensitivity to actual cases.
* **F1-Score**: No Heart Disease: 73%, Heart Disease: 70% - balances precision and recall.

**Decision Tree**

* **Accuracy**: 90.1% - strong overall accuracy.
* **ROC-AUC**: 0.90 - good distinction between classes.
* **Precision**: No Heart Disease: 89%, Heart Disease: 91% - reflects high accuracy in both predictions.
* **Recall**: No Heart Disease: 91%, Heart Disease: 89% - high sensitivity for true cases.
* **F1-Score**: No Heart Disease: 90%, Heart Disease: 90% - balances precision and recall effectively.

**Logistic Regression**

* **Accuracy**: 76.5% - moderate classification accuracy.
* **ROC-AUC**: 0.84 - fair distinction between heart disease and no heart disease.
* **Precision**: No Heart Disease: 78%, Heart Disease: 75% - moderate positive and negative prediction accuracy.
* **Recall**: No Heart Disease: 74%, Heart Disease: 79% - fair sensitivity to true cases.
* **F1-Score**: No Heart Disease: 76%, Heart Disease: 77% - fair balance between precision and recall.

**Random Forest**

* **Accuracy**: 94.1% - high overall accuracy.
* **ROC-AUC**: 0.98 - excellent at distinguishing heart disease presence.
* **Precision**: No Heart Disease: 92%, Heart Disease: 96% - very high prediction accuracy.
* **Recall**: No Heart Disease: 96%, Heart Disease: 92% - strong sensitivity.
* **F1-Score**: No Heart Disease: 94%, Heart Disease: 94% - balanced and reliable.

**Gradient Boosting**

* **Accuracy**: 89.5% - solid performance.
* **ROC-AUC**: 0.96 - strong ability to separate classes.
* **Precision**: No Heart Disease: 90%, Heart Disease: 89% - high prediction accuracy.
* **Recall**: No Heart Disease: 89%, Heart Disease: 90% - consistent sensitivity.
* **F1-Score**: No Heart Disease: 89%, Heart Disease: 90% - effective balance.

**XGBoost**

* **Accuracy**: 93.9% - robust performance.
* **ROC-AUC**: 0.98 - excellent at class separation.
* **Precision**: No Heart Disease: 98%, Heart Disease: 89% - high precision.
* **Recall**: No Heart Disease: 89%, Heart Disease: 98% - strong sensitivity for heart disease.
* **F1-Score**: No Heart Disease: 94%, Heart Disease: 94% - balanced, highly effective performance.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **ROC-AUC** | **Precision (No Heart Disease)** | **Precision (Heart Disease)** | **Recall (No Heart Disease)** | **Recall (Heart Disease)** | **F1-Score (No Heart Disease)** | **F1-Score (Heart Disease)** |
| Naive Bayes | 71.50% | 0.8 | 69% | 75% | 78% | 65% | 73% | 70% |
| Decision Trees | 82.30% | 0.85 | 80% | 85% | 83% | 80% | 81% | 82% |
| Logistic Regression | 84.10% | 0.88 | 82% | 87% | 85% | 82% | 83% | 84% |
| Random Forest | 94.00% | 0.98 | 92% | 96% | 96% | 92% | 94% | 94% |
| Gradient Boosting | 93.50% | 0.97 | 91% | 95% | 95% | 90% | 93% | 92% |
| XGBoost | 93.90% | 0.98 | 93% | 96% | 96% | 93% | 94% | 94% |

**Model Chosen:**

**Features Selected/Engineered**

I selected the following features based on exploratory data analysis, which revealed their significant relationships with heart disease outcomes:

**HeartDiseaseorAttack, HighBP, HighChol, , BMI, Smoker, Stroke, Diabetes, PhysActivity, Fruits, Veggies, , GenHlth, MentHlth, PhysHlth, DiffWalk, Sex, Age, Education, Income**

These features were chosen due to their strong correlation with heart disease and their potential to provide meaningful insights. In contrast, features like **CholCheck**, **HvyAlcoholConsump**, **NoDocbcCost**, and **AnyHealthcare** were found to be less impactful during the analysis.

**Model Evaluation Metrics Used:**

I evaluated the model using the following metrics:

1. **Accuracy:**I chose accuracy to see how well my model performs overall. High accuracy means the model correctly classifies most heart disease and non-heart disease cases.
2. **ROC-AUC:**I selected ROC-AUC to measure the model's ability to differentiate between heart disease and no heart disease. A score close to 1 shows effective classification, which is important due to the imbalance in the dataset.
3. **Precision:**I chose precision to evaluate the accuracy of positive predictions (heart disease cases). High precision means fewer false positives, ensuring that the model correctly identifies patients with heart disease.
4. **Recall:**Recall helps me understand how well the model identifies actual heart disease cases. High recall means the model captures most true cases, reducing missed diagnoses.
5. **F1-Score:**The F1-Score combines precision and recall into one metric, providing a balanced view of the model's performance. It’s useful when both false positives and false negatives matter.
6. **Confusion Matrix:**I included the confusion matrix to visualize the model’s performance. It shows the counts of true positives, true negatives, false positives, and false negatives, helping me identify where the model may misclassify cases.

**Impact of Hyperparameter Tuning:**

Despite applying extensive hyperparameter tuning, the improvements in accuracy, precision, recall, and F1 score were minimal. Tuning did not significantly enhance the model's performance; therefore, the original model, without tuning, will be retained as it offers comparable results with less complexity.

**Future Work**

Given more time, I would:

**Neural Networks**: Experiment with neural networks to capture more complex patterns in the data, which could enhance predictive performance.

**User Interface Design**: Develop a user-friendly interface where users can input their health data, and the model can predict the likelihood of developing heart disease. This would make the model accessible to a broader audience, such as patients and healthcare providers.

**Recommendations**

I strongly recommend using this model for predictions. It demonstrates high accuracy and reliability, making it suitable for identifying individuals at risk for heart disease. Its performance indicates it can effectively assist in healthcare decision-making, benefiting both patients and medical professionals.