Heart Disease Prediction: Dataset Analysis and Model Evaluation

Ziad Ahmad
IEEE, ACM Member, Researcher, CS Student

Muayad Walid
CS Student

November 22, 2024

1. Dataset Overview

The dataset used in this analysis is sourced from Kaggle: Heart Disease Dataset. It contains valuable information regarding patients with heart disease and is highly beneficial for understanding and predicting the risk factors involved in developing heart conditions. The dataset is designed to assist healthcare professionals by providing a wide range of clinical and health-related attributes. The dataset allows for a detailed exploration of various risk factors that are crucial for determining the likelihood of heart disease, such as cholesterol levels, resting blood pressure, and exercise-induced responses. By analyzing these factors, healthcare professionals can better understand the complex interactions that contribute to cardiac health. Additionally, the dataset provides a foundation for building predictive models that can be used to evaluate patient health, improving diagnostic accuracy and supporting preventive measures.

The dataset is well-structured and provides extensive details about various medical metrics, making it an invaluable resource for predictive modeling. It serves as a representative collection of the different health conditions that can lead to heart disease. Moreover, this dataset can be used as a benchmark for comparing different machine learning and deep learning techniques aimed at predicting heart disease. The rich features allow practitioners to train sophisticated models that are capable of making accurate predictions, thereby enhancing patient care and medical interventions.

2. Key Dataset Features

The dataset consists of 1,035 entries, with 16 features per patient. Below is a summary of the most important columns included in the dataset:

- **sex**: The gender of the patient, either "male" or "female." Gender is known to be a significant factor in determining heart disease risk, as males are generally considered at higher risk compared to females.
- **age**: Age of the patient at the time of evaluation. Age is a crucial predictor, as the risk of heart disease typically increases with age.
- **cp (Chest Pain Type)**: Type of chest pain experienced (0-3), which may provide insight into the patient's cardiac status. Chest pain is one of the primary symptoms that may indicate heart disease, and the different types (typical angina, atypical angina, non-anginal pain, or asymptomatic) help in categorizing the risk.

- **resting_BP**: Resting blood pressure of the patient in mm Hg. High blood pressure is a well-known risk factor for heart disease, and monitoring this metric is essential for early diagnosis.
- **chol**: Cholesterol level of the patient (in mg/dL). Elevated cholesterol levels can lead to arterial blockages, which significantly increase the risk of cardiac problems.
- **fbs**: Fasting blood sugar (> 120 mg/dL or not). High fasting blood sugar is indicative of diabetes, which is a major risk factor for heart disease.
- **restecg**: Resting electrocardiographic results (0, 1, or 2). These values represent different levels of abnormalities in heart function detected at rest, providing insight into potential electrical activity issues.
- **thalach**: Maximum heart rate achieved. This metric is used to assess the heart's response to physical activity. Higher values typically indicate better cardiovascular health.
- **exang**: Exercise-induced angina (0: No, 1: Yes). This feature indicates whether the patient experienced chest pain during exercise, which is an important indicator of coronary artery disease.
- **oldpeak**: ST depression induced by exercise relative to rest. The value reflects changes in the ST segment of the ECG, which can be a marker of myocardial ischemia.
- **slope**: Slope of the peak exercise ST segment (0, 1, 2). This parameter is used to assess the heart's reaction during physical activity. A decreasing slope might indicate abnormal heart function.
- **ca**: Number of major vessels colored by fluoroscopy (0-3). This feature provides information on the number of major blood vessels that are obstructed, which helps in determining the severity of the condition.
- **thal**: Thalassemia type (1: normal, 2: fixed defect, 3: reversible defect). This genetic condition can have significant effects on heart health, especially when it presents as a fixed or reversible defect.
- **Max Heart Rate Reserve**: Difference between calculated maximum heart rate and recorded heart rate. This indicates the remaining heart rate reserve, which is used to evaluate cardiac efficiency.
- **Heart Disease Risk Score**: Score indicating the likelihood of developing heart disease based on other features. It helps in quantifying the patient's overall risk of cardiac conditions.
- **target**: Indicates presence of heart disease (0: No heart disease, 1: Heart disease present). This is the output variable used for model training and evaluation.

3. Data Preprocessing

To effectively train machine learning and deep learning models, the dataset required several preprocessing steps:

- **Categorical Encoding**: The categorical column 'sex' was encoded using label encoding, where "male" became 1 and "female" became 0. This transformation was necessary to convert categorical data into a numerical format that could be utilized by machine learning algorithms.

- **Normalization**: All numerical features were normalized using MinMax scaling to bring the features into the range of 0 to 1. This ensures that the models do not become biased towards features with larger numeric ranges and allows for faster convergence during training. Normalization also helps maintain numerical stability and reduces the risk of gradient explosion or vanishing issues in deep learning models.

- **Feature Selection**: All available features were retained as they each contributed useful information for predicting the target variable. Given the dataset size and the importance of each feature, no feature reduction was performed, ensuring that all potential predictive information was leveraged during model training.

- **Train-Test Split**: The dataset was divided into training (80%) and testing (20%) sets to evaluate model performance accurately. This split helps in assessing how well the model generalizes to unseen data and prevents overfitting by ensuring that the model does not simply memorize the training data.

4. Machine Learning Models Evaluation

Four machine learning models were trained on the dataset to predict the presence of heart disease. The models, along with their evaluation metrics (Accuracy, F1 Score, Precision, Recall), are as follows:

Model	**Accuracy**	**F1 Score**	**Precision**	**Recall**
Random Forest	1.000	1.000	1.000	1.000
Gradient Boosting	0.985	0.988	1.000	0.976
Support Vector Classifier	0.894	0.911	0.911	0.911
Logistic Regression	0.850	0.877	0.860	0.895

Table 1: Performance metrics for various machine learning models.

Why These Models Were Chosen

- **Random Forest**: Random Forest is an ensemble learning technique that combines multiple decision trees to improve classification accuracy and reduce overfitting. It was chosen due to its robustness in handling datasets with diverse features, providing excellent performance by averaging the output of multiple trees. This model is also less prone to overfitting compared to a single decision tree, which makes it suitable for medical datasets where generalization is crucial.

- **Gradient Boosting**: Gradient Boosting builds decision trees sequentially, where each tree is trained to correct the errors of the previous trees. This allows it to achieve high precision and recall, particularly in datasets where the relationships between features are complex and subtle. Gradient Boosting was chosen because it effectively balances bias and variance, making it a strong candidate for predictive modeling in clinical data.

- **Support Vector Classifier (SVC)**: SVC was chosen because of its capability to handle high-dimensional datasets and find the optimal hyperplane that maximizes the margin between classes. It is particularly useful for classification tasks involving complex boundaries between classes. In this scenario, SVC provides a different perspective on the data's decision boundary, giving us another baseline to compare against.

- **Logistic Regression**: Logistic Regression is one of the simplest machine learning models used for binary classification problems. It was chosen for its interpretability, allowing us to understand the direct influence of each feature on the target variable. Logistic Regression serves as a useful baseline model that offers insight into how more complex models improve upon simple linear relationships.

Key Observations

- The **Random Forest** model performed perfectly, achieving 100% accuracy, F1 score, precision, and recall on the test dataset. This indicates that it successfully classified all patients correctly, though this may also indicate potential overfitting. The model may have memorized the training data, and it is important to conduct cross-validation to determine if the perfect score generalizes well across different folds.

- **Gradient Boosting** also performed very well, with high precision and a slight drop in recall compared to Random Forest. This suggests it may be slightly less sensitive in detecting true positives, but it still maintained a strong ability to classify correctly overall. The model balances complexity and generalizability well, making it an effective choice in clinical contexts where precision is crucial.

- **Support Vector Classifier** and **Logistic Regression** performed reasonably well but had lower overall metrics compared to Random Forest and Gradient Boosting, which suggests that they may not be as effective in identifying all patterns in the dataset. Nevertheless, these models are simpler, faster to train, and may be suitable for real-time applications where interpretability and efficiency are needed.

5. Deep Learning Model Evaluation

A deep learning model was also implemented using a sequential neural network with the following architecture:

- **Input Layer**: Fully connected layer with 64 nodes and ReLU activation. This layer takes all the features and feeds them to the subsequent layers, capturing complex patterns in the data.

- **Hidden Layers**: Two layers with 32 and 16 nodes respectively, each using ReLU activation. These layers help in extracting meaningful patterns and relationships between features. The nodes progressively decrease to reduce the dimensionality, promoting more abstract and high-level features towards the output.

- **Output Layer**: A single neuron with sigmoid activation for binary classification. The sigmoid function outputs probabilities between 0 and 1, enabling the model to classify whether a patient has heart disease or not.

Why This Deep Learning Model Architecture Was Chosen

The deep learning model was built with a relatively simple architecture to prevent overfitting on a dataset that, while informative, is not large enough for very deep or complex models. The **Input Layer** with 64 nodes was chosen to ensure that the model captures sufficient complexity from the features. The **Hidden Layers** with 32 and 16 nodes gradually reduce dimensionality, allowing the model to focus on abstract patterns without over-complicating the architecture. **ReLU** activation was chosen because it helps in avoiding vanishing gradients, a common issue with deep learning models.

The **sigmoid** activation in the output layer was selected as it is ideal for binary classification tasks, producing outputs between 0 and 1, which can be interpreted as probabilities. This architecture was designed to achieve a balance between complexity and performance, allowing the model to learn effectively without overfitting to the relatively small dataset.

The model was trained for 50 epochs with a batch size of 16. The evaluation metrics for the deep learning model are as follows:

- **Accuracy**: 0.947
- **F1 Score**: 0.954

Key Observations

- The deep learning model performed quite well, achieving high accuracy and F1 score. However, its performance was slightly below that of the Random Forest and Gradient Boosting models. This suggests that the dataset, being relatively small and structured, was perhaps more suited to classical machine learning models rather than a deep learning approach. The deep learning model's strength lies in its ability to capture complex patterns, but with smaller datasets, it may struggle to generalize effectively compared to traditional models.

Additionally, the relatively small dataset size and the presence of structured, tabular data are likely reasons why the deep learning model did not outperform simpler models. Deep learning models often require large volumes of data to fully exploit their representation capabilities, and in this scenario, the dataset may not have been extensive enough for the deep learning model to demonstrate its potential advantages.

6. Conclusion

The analysis of the heart disease dataset shows that traditional machine learning models such as **Random Forest** and **Gradient Boosting** outperformed the deep learning model in terms of accuracy, precision, recall, and F1 score. The Random Forest model, in particular, achieved perfect scores, making it the most reliable for predicting the presence of heart disease. However, the potential for overfitting should be examined carefully to ensure that the model generalizes well in real-world scenarios.

The dataset itself provides rich information regarding patient health, which can be used to effectively predict heart disease risk. The preprocessing steps, including normalization and categorical encoding, were crucial to ensure effective model training. Each feature contributes to the overall predictive power of the models, which highlights the importance of thorough feature engineering and preprocessing.

Recommendations

- Given the perfect performance of the Random Forest model, further validation should be conducted to ensure that overfitting is not an issue, possibly by applying cross-validation or increasing the dataset size. Exploring ensemble techniques that average multiple models may also reduce overfitting risks.

- Additional features or more data could improve the performance of models like Logistic Regression and SVM. Feature engineering, such as creating new interaction terms or incorporating domain-specific insights, could also enhance the models' ability to generalize.

- For real-world applications, model explainability techniques such as SHAP or feature importance plots can be employed to interpret the predictions of the Random Forest model for healthcare professionals. This helps build trust in the model's recommendations, as doctors need to understand the reasons behind each prediction.

This analysis serves as a useful baseline for heart disease prediction, and future work could involve gathering more data to enhance the model's robustness or deploying the model as part of a healthcare decision support system. By integrating the model into electronic health records, predictions can be made in real time, helping physicians make more informed decisions and potentially saving lives through early intervention.

References

1. S. N. Mahsa. *Heart Disease Dataset*. Available at: `https://www.kaggle.com/datasets/snmahsa/heart-disease/data` [Accessed: November 22, 2024].

2. A. Comprehensive Review on Heart Disease Risk Prediction using Machine Learning. Springer Link. Available at: `https://link.springer.com/article/10.1007/s11831-024-10194-4` [Accessed: November 22, 2024].

3. Early prediction of heart disease with data analysis using supervised learning models. Journal of Engineering and Applied Sciences. Available at: `https://jeas.springeropen.com/articles/10.1186/s44147-023-00280-y` [Accessed: November 22, 2024].

4. A Review of Machine Learning's Role in Cardiovascular Disease Diagnosis. MDPI Algorithms. Available at: `https://www.mdpi.com/1999-4893/17/2/78` [Accessed: November 22, 2024].

5. Enhancing Cardiovascular Disease Risk Prediction with Machine Learning Models. arXiv. Available at: `https://arxiv.org/abs/2401.17328` [Accessed: November 22, 2024].