Predictive Model for Heart Disease Diagnosis

Clémence FLEURIOT, Naomie HALFON, Marine JOURNU, Lison LEANDRE

November 14, 2024

Contents

1 Business Scope

The objective of this project is to develop a predictive model to assess the likelihood of an individual having heart disease based on a set of medical and lifestyle-related features. Using the "heart.csv" dataset, which includes indicators such as age, cholesterol levels, and other relevant clinical parameters, the goal is to classify whether a person has heart disease.

Key Business Implications:

- **Healthcare Impact**: Early identification of individuals at risk can help in proactive healthcare intervention, potentially improving patient outcomes.
- **Resource Allocation**: Helps healthcare providers allocate resources more effectively by identifying high-risk patients who may need more attention.
- **Cost Efficiency**: Reduces long-term healthcare costs by enabling preventive care.

2 Problem Formalization and Methods

2.1 Problem

Using the "heart.csv" dataset, we aim to predict if an individual has heart disease, a binary variable indicating the presence or absence of the condition.
The primary evaluation metric is **accuracy**, supplemented by **F1-score** and **AUC-ROC** for a more nuanced understanding of model performance, especially given the importance of correctly identifying both positives and negatives.

Models chosen: Logistic Regression, Decision Tree, SVM.

2.2 Description & Limitations

- **Logistic Regression**: Chosen for its interpretability and ability to show the influence of each feature on the prediction. However, it may not capture complex non-linear relationships.
- **Decision Tree**: Useful for capturing non-linear relationships and easy to interpret, though it is prone to overfitting, particularly with deep trees.
- **Support Vector Machine (SVM)**: Effective for complex decision boundaries but requires careful hyperparameter tuning. This model may have limitations with computational efficiency on larger datasets.

3 Methodology

3.1 Data Description & Exploration

- **Missing Values**: After loading the data from the "heart.csv" file, we checked for missing values using data.isnull().sum(). The results showed that there were no missing values in the dataset, simplifying preprocessing by removing the need for imputation.

- **Imbalanced Data**: The target variable "target," indicating the presence or absence of heart disease, was analyzed for class balance. A review of the class distribution showed that the data was reasonably balanced, so no resampling techniques (such as SMOTE or undersampling) were deemed necessary.

- **Outliers**: We examined the data to detect potential outliers by analyzing feature distributions and identifying duplicates. Duplicates were removed using data.drop_ duplicates() to ensure they do not bias model results. Some extreme values in certain features were observed, but no specific treatment was applied as they did not significantly impact the model's performance.

3.2 Data Splitting for Train/Test Sets

To evaluate model performance robustly, the dataset was split into two subsets: a training set (70% of the data) and a test set (30% of the data). This split was done using the train_test_split function from scikitlearn, specifying a random_state to ensure reproducibility. This separation allows us to verify that the model generalizes well to new data.

3.3 Algorithm Implementation & Hyperparameter Optimization

Three machine learning models were implemented for this binary classification task: logistic regression, decision tree, and support vector machine (SVM). These models were chosen for their respective abilities to capture linear and non-linear relationships in the data.

- **Logistic Regression**: This linear and interpretable model is used to understand the influence of each feature on the probability of heart disease. The C parameter, which controls regularization, was adjusted to avoid overfitting.

- **Decision Tree**: Decision trees are useful for capturing non-linear relationships between features. Key hyperparameters adjusted include the maximum depth of the tree (max_depth) and the minimum number of samples per leaf (min_samples_leaf), which help control model complexity and reduce overfitting risk.

- **SVM**: SVMs are effective for complex decision boundaries. The primary hyperparameters tuned include the kernel type (kernel) and the regularization parameter C. The choice of kernel (linear, polynomial, or RBF) allows the model to adapt to various data structures and optimize class separation.

To improve model performance, a hyperparameter search was conducted using `GridSearchCV`, allowing us to test different parameter combinations and select the optimal ones. For instance, for SVM, we tested different values for C and kernel.

4 Results

In this section, we present the performance metrics obtained from our models: Logistic Regression, Decision Tree, and Support Vector Machine (SVM). The primary metrics we focused on are accuracy, F1-score, and AUC-ROC, as these provide insights into the models' effectiveness in correctly classifying individuals with and without heart disease.

4.1 Model Performance

Model	Accuracy	F1-Score	AUC-ROC
Logistic Regression	86%	0.88	0.91
Decision Tree	79%	0.81	0.82
SVM	83%	0.84	0.87

Table 1: Model Performance Metrics

4.2 Overfitting Analysis

During training, we observed overfitting particularly in the Decision Tree model, as evidenced by its higher performance on the training data compared to the test data. To mitigate this, hyperparameters like `max_depth` and `min_samples_leaf` were adjusted. Logistic Regression and SVM demonstrated more robustness against overfitting due to regularization techniques and controlled complexity.

4.3 Evaluation

In summary, while all three models performed reasonably well, the SVM model achieved the highest AUC-ROC, indicating superior class separation. Logistic Regression provided interpretability, making it useful for understanding the influence of individual features. The Decision Tree model, though effective for capturing non-linear relationships, required regular tuning to avoid overfitting.

5 Discussion and Conclusion

This project demonstrated the potential of machine learning models in predicting heart disease based on medical and lifestyle-related features. By using three different algorithms, we gained insights into the strengths and limitations of each model for this classification task.

- **Interpretability vs. Performance**: Logistic Regression provided valuable interpretability, allowing us to identify features that influence heart disease risk. However, its linear nature limited its performance on non-linear relationships, which the Decision Tree and SVM could better capture.
- **Model Complexity and Overfitting**: The Decision Tree, while effective in capturing data complexity, tended to overfit, highlighting the importance of hyperparameter tuning. On the other hand, SVM balanced complexity and generalization effectively but required careful selection of kernel and regularization parameters.

- **Healthcare Implications**: These models, particularly SVM with its high AUC-ROC score, could be valuable tools for early detection of heart disease, aiding healthcare providers in proactive intervention and resource allocation. The interpretability of Logistic Regression is particularly beneficial for healthcare practitioners to understand risk factors.

In conclusion, this study demonstrates the pivotal role that machine learning can play in healthcare by providing predictive insights into heart disease risks. While the Support Vector Machine (SVM) emerged as the most effective model, Logistic Regression remains valuable for its interpretability. Future work could explore ensemble methods or deep learning techniques to further enhance predictive accuracy and robustness, especially with larger datasets. Additionally, incorporating supplementary variables, such as genetic data or lifestyle information, could enrich the model and offer a more comprehensive understanding of the factors influencing heart disease.