Stage 1

Machine Learning Course

Major: Cloud Computing and Cybersecurity (Group CCC1)

Supervised By:

Nédra MELOULLI
Christophe RODRIGUES

Written By:

Alex QUILIS VILA
Carolina ROMERO DELGADO
Livia GAMERO HAMMERER

Date Last Edited: November 20, 2024

Contents

List of Tables

Chapter 1

Business Scope

The goal of this project is to develop a machine learning model capable of predicting the presence and severity of heart disease based on patient data. By accurately identifying at-risk individuals, the model can assist healthcare professionals in providing timely and personalized interventions. Early detection and accurate risk assessment can significantly improve patient outcomes, reduce healthcare costs, and enhance the overall efficiency of healthcare systems.

Chapter 2

Problem Formalization and Methods

Problem: Multi-class classification problem, where the target variable indicates the presence and severity of heart disease (values 0–4). Methods: Data preprocessing, exploratory analysis, and application of machine learning models such as Logistic Regression, Decision Tree, and Random Forest.

2.1 Algorithm Description

- **Logistic Regression**: A linear model used for probabilistic predictions.
- **Decision Tree**: A non-linear model that splits data into branches based on feature thresholds.
- **Random Forest**: An ensemble method that builds multiple trees and aggregates their results for better accuracy and robustness.

2.2 Limitations

- **Imbalanced target variable classes**: Could bias predictions toward dominant classes.
- **Potential overfitting**: Decision Trees are prone to overfitting without proper tuning.

Chapter 3

Methodology

3.1 Data Description and Exploration

The dataset consists of 17 features, with the target variable being `HeartDisease` (Yes/No).

- **Exploratory Analysis**:
  - **Correlation Matrix**: Used to assess relationships between features. No strong correlations were found between variables.
  - **Statistical Summary**: Basic statistics (mean, median, standard deviation, min, and max) for numerical features were generated using `.describe()`.
  - **Variable Types**: Analyzed using `.info()` to identify data types for each feature.
- **Data Cleaning**:
  - **Duplicated Rows**: Duplicates were removed.
  - **Imbalanced Data**: SMOTE was applied to address class imbalance in the `HeartDisease` target variable, generating synthetic samples for underrepresented classes.
  - **Normalization**: Numerical features were normalized to enhance model performance.

3.2 Data Splitting for Train/Test

The data was split into an 80% training set and a 20% testing set.

3.3 Model Implementation and Hyperparameters

The following models were implemented with their respective hyperparameters:

- **Logistic Regression**:
  - **Hyperparameters**: Maximum iterations (max_iter = 1000) and regularization parameter ($C = 0.1$).
- **Random Forest**:
  - **Hyperparameters**: Number of estimators ($n\_estimators = 100$), maximum depth (max_depth = 10), and minimum samples required to split (min_samples_split = 10).
- **Decision Tree**:
  - **Hyperparameters**: Maximum depth (max_depth = 5).

Chapter 4

Results

4.1 Metrics

- **Accuracy**: Measures the overall correctness of the model.
- **Precision**: Focuses on the relevance of positive predictions.
- **Recall**: Highlights the ability to capture all true positives.
- **F1-Score**: The harmonic mean of precision and recall.
- **ROC AUC**: Measures the area under the receiver operating characteristic curve, assessing the model's ability to discriminate between classes.

Models	Accuracy	Precision	Recall	F1-Score	ROC AUC
Logistic Regression	0.6995	0.1820	0.6661	0.2859	0.6844
Random Forest	0.6968	0.1952	0.7547	0.3101	0.7229
Decision Tree	0.6526	0.1797	0.7987	0.2935	0.7184

Table 4.1: Model Results

4.2 Cross-Validation Results

- **Logistic Regression**: Cross-validation scores show moderate performance with a mean score of 0.7415.
- **Random Forest**: Cross-validation yielded consistent performance, with a mean score of 0.7593.
- **Decision Tree**: Cross-validation scores indicate stable performance, with a mean score of 0.7468.

4.3 Overfitting

Overfitting was addressed by tuning the hyperparameters and applying cross-validation to ensure the models generalize well:

- **Logistic Regression**: To reduce overfitting, a regularization parameter $C = 0.1$ was chosen, which helps to avoid excessively complex models by penalizing large coefficients.
- **Random Forest**: The model was tuned with a maximum tree depth of 10 (max_depth = 10) and a minimum of 10 samples required to split an internal node (min_samples_split = 10), which helps to prevent the trees from growing too deep and memorizing the training data.
- **Decision Tree**: The tree depth was capped at 5 (max_depth = 5) to avoid fitting the model too closely to the training data, which can result in overfitting.

Additionally, cross-validation was used for all models to validate their performance on different subsets of the data, further helping to ensure that the models do not overfit and perform consistently across different data samples.

4.4 Evaluation

In this study, three machine learning models—Logistic Regression, Decision Tree, and Random Forest—were evaluated based on their performance in predicting heart disease. The models were compared using common classification metrics such as accuracy, precision, recall, and F1-score.

- **Logistic Regression**: This model demonstrated the lowest performance, with an accuracy of 69.95% and a precision of 18.20%. Despite being computationally efficient, it struggled to capture the complexity of the data, particularly for rare classes, as reflected by its low precision.
- **Decision Tree**: The Decision Tree model had an accuracy of 65.26% and showed the highest recall (79.87%), indicating that it was effective at identifying true positive cases of heart disease. However, it exhibited lower precision, showing that it also made many false positive predictions.
- **Random Forest**: The Random Forest model showed a balanced performance, achieving an accuracy of 69.68% and the highest precision (19.52%). It also had a higher F1-score compared to the Decision Tree model, showing a good trade-off between precision and recall.

Based on these results, the Random Forest model appears to be the most effective for this classification task, with a good balance of precision, recall, and accuracy. However, further fine-tuning and testing with additional data would be beneficial to improve model robustness and address any remaining overfitting concerns.

Chapter 5

Discussion and Conclusion

5.1 Discussion

- **Class Imbalance**: The issue of class imbalance was a significant challenge in training the models. However, the SMOTE technique proved effective in addressing this imbalance by generating synthetic examples for the minority class. This improved the model's ability to learn from the underrepresented class and enhanced its capacity to detect heart disease.

- **Best Model**: Among the models evaluated, the Random Forest classifier emerged as the top performer. It demonstrated strong predictive power, achieving high accuracy while maintaining a reasonable level of interpretability. This balance makes it suitable for healthcare applications, where both performance and understanding of predictions are important.

- **Model Evaluation**: Different models showed varying performance levels. While the Decision Tree classifier was simpler and more interpretable, the Random Forest outperformed it in terms of accuracy. Logistic Regression performed well but lacked the complexity and depth that Random Forest provided.

5.2 Conclusion

- **Recommendation**: Given its superior performance and ability to handle complex datasets, the Random Forest model is recommended as the primary choice for heart disease prediction. While it is less interpretable than the Decision Tree, its strong predictive power and robustness make it a reliable tool for healthcare applications.

- **Impact**: The successful application of machine learning models, particularly Random Forest, for predicting heart disease highlights the significant potential of AI in healthcare. These models can support early detection, guide treatment decisions, and ultimately improve patient outcomes.

5.3 Future Prospects

- **Exploring Advanced Models**: While Random Forest performed exceptionally well, further investigation into even more sophisticated models, such as Gradient Boosting or XGBoost, could offer even greater predictive power without compromising too much on interpretability.
- **Feature Engineering**: Incorporating additional relevant features, such as medical history, age, lifestyle factors (e.g., smoking or exercise habits), and genetic data, could further enhance model accuracy. However, acquiring high-quality data for these features in healthcare remains a challenge.
- **Model Deployment and Monitoring**: For practical use, the model will need to be thoroughly validated and deployed in a clinical environment. Continuous monitoring and periodic updates are essential to ensure its effectiveness as new patient data becomes available.
- **Real-World Validation**: To assess its real-world applicability, conducting pilot studies or clinical trials is essential. This will provide insights into how the model performs with actual patient data and identify any biases or limitations that were not captured in the training dataset.