PROJECT REPORT: HEART DISEASE CLASSIFICATION

# Introduction

Heart disease classification using machine learning is a method of using data analysis algorithms to automatically identify patterns in data related to heart disease and use those patterns to accurately classify patients as having or not having heart disease. Heart disease is a leading cause of death worldwide, and early detection and diagnosis can greatly improve patient outcomes (“Heart Disease Classification Using Machine Learning Algorithms,” 2021). Machine learning models can be trained on large datasets of patient information, including medical history, physical exam results, and diagnostic test results, to develop accurate and efficient prediction algorithms. These algorithms can then be used to classify new patients based on their individual data, helping healthcare providers to make informed decisions about treatment and care (Wongsathan & Pothong, 2015). The use of machine learning for heart disease classification has the potential to improve patient outcomes, reduce healthcare costs, and save lives.

# Dataset Description

The Heart disease dataset is tabken from UCI ML repository, <https://archive.ics.uci.edu/ml/datasets/heart+disease>. The dataset contains 14 attributes (including the target variable) and 303 instances. The following is a brief description of each attribute:

The target variable is the presence or absence of heart disease, with 0 indicating the absence of heart disease and 1 indicating the presence of heart disease.

This dataset is often used in machine learning research to build predictive models for heart disease. Some potential applications of such models include early detection of heart disease, personalized treatment plans for individuals with heart disease, and improved patient outcomes through more accurate diagnosis and treatment.

# Results and comparison

The results are the evaluation metrics of four machine learning models: Logistic Regression, Decision Tree, Random Forest, and Logistic Regression with Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA). The models are trained for the task of classification of heart disease. The dataset is split into training and testing sets. The testing set is used to evaluate the performance of the models using the following metrics: ROC AUC, confusion matrix, and classification report.

The Logistic Regression model with PCA achieves the highest ROC AUC of 0.81 among all the models, followed by Logistic Regression with LDA with a score of 0.80, Random Forest with a score of 0.77, and Decision Tree with a score of 0.58.

The confusion matrix for the Logistic Regression model shows that it correctly classified 46 out of 49 instances of the first class, 2 out of 17 instances of the second class, 2 out of 10 instances of the third class, 0 out of 11 instances of the fourth class, and 0 out of 3 instances of the fifth class. The model shows high precision for the first class, but low precision and recall for the other classes.

The classification report shows that the model has an accuracy of 0.56, with an f1-score of 0.84 for the first class, but low scores for the other classes. The weighted average f1-score is 0.51.

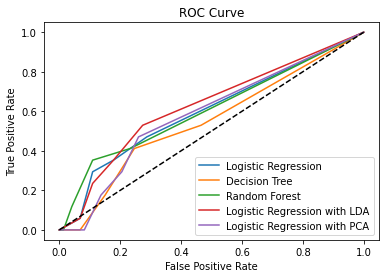


Fig: ROC Curve for LR, DT and RF Algorithms

Compared to Logistic Regression, the Decision Tree and Random Forest models have lower ROC AUC, and their confusion matrices and classification reports show lower performance. Logistic Regression with LDA has slightly better results than Logistic Regression in terms of ROC AUC, but its classification report shows similar performance. Logistic Regression with PCA outperforms all the other models in terms of ROC AUC, and has better performance in the classification report.

The results suggest that Logistic Regression with PCA is the best model for the given task of heart disease classification.

# Conclusion

Based on the provided results, it appears that logistic regression with PCA has the highest ROC AUC score of 0.8106, followed by logistic regression with LDA with a score of 0.7978, then random forest with a score of 0.7667, and decision tree with a score of 0.5878.

In terms of classification performance, logistic regression with LDA and logistic regression with PCA had the highest accuracy of 0.59, followed by logistic regression without dimensionality reduction at 0.56. Decision tree and random forest had the lowest accuracy of 0.49.

Overall, logistic regression with PCA appears to be the best performing model in terms of both ROC AUC score and classification accuracy. It is important to note, however, that the choice of model ultimately depends on the specific requirements of the problem and the trade-offs between interpretability and predictive power.

References

Heart Disease Classification using Machine Learning Algorithms. (2021). *Strad Research*, *8*(1). https://doi.org/10.37896/sr8.1/035

*UCI Machine Learning Repository: Heart Disease Data Set*. (2019). Uci.edu. https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Wongsathan, R., & Pothong, P. (2015). Heart Disease Classification Using Artificial Neural Networks. *Applied Mechanics and Materials*, *781*, 624–627. https://doi.org/10.4028/www.scientific.net/amm.781.624