Comparative Analysis of Logistic Regression and Decision Tree for Predicting the Risk of Heart Attacks

Prepared by 2441638

Abstract

This project evaluates the performance of two machine learning models, Logistic Regression and Decision Tree, in predicting heart attack risks based on a medical dataset. The dataset includes various health-related features such as cholesterol levels, blood pressure, and heart rate. Logistic Regression was chosen for its ability to model linear relationships, while Decision Tree was selected for its flexibility in capturing complex patterns. The models were trained and tested using the dataset, and their accuracies were compared. Logistic Regression outperformed Decision Tree, achieving higher average accuracy. This was likely due to the linear nature of the dataset, which aligned well with Logistic Regression's assumptions. In contrast, the Decision Tree struggled with overfitting and high variance, which reduced its ability to generalise to unseen data. Methods in which the models could be improved were explored, including feature engineering, data augmentation, and hyperparameter tuning. Additionally, ensemble methods like Random Forest and Gradient Boosting are suggested to enhance Decision Tree performance. Overall, the study highlights the importance of aligning model selection with dataset characteristics to achieve the best results.

Keywords: Logistic Regression, Decision Tree, Machine Learning, Classification Models

Contents

1. Introduction

When working with machine learning (ML), selecting the most suitable classification model is important to be able to achieve optimal performance on any given dataset. This project explores two widely used classification algorithms (Logistic Regression and Decision Tree) by applying them to a dataset focused on predicting the risk of a heart attack. The object of this project is to evaluate the performance of the two algorithms, identify the strengths and weakness of each and suggest strategies to improve model accuracy. This project provides a systematic approach to building, evaluating, and refining these models. Logistic Regression (LR), being a linear model, is compared against Decision Tree (DT), a non-linear model, to assess their performance in handling the dataset's characteristics. The analysis involves training the models, measuring their accuracy through repeated experiments with varying random states, and calculating the average accuracy. Additionally, the report delves into explaining the underlying mechanisms of each algorithm and their suitability for this specific classification task.

Through this comparative study, insights into the practical applications of these algorithms will be provided. Their limitations, and recommend data preprocessing and feature engineering techniques to enhance their predictive power will be discussed. This comprehensive approach ensures a robust understanding of how these models function and how they can be optimised for real-world problems.

2. Logistic Regression

LR, a statistical model, is commonly used for binary classification problems. It predicts the probability of a given outcome based on input features by applying the logistic (sigmoid) function, mapping any real-valued number into a range between 0 and 1. The model assumes a linear relationship between the input features and the log-odds of the target variable, making it a linear classifier. LR calculates the probability of a binary outcome using the sigmoid function (Boateng and Abaye, 2019). The probability $P(y = 1|x)$ is given by:

$$P(y = 1|x) = \frac{1}{1 + e^{-z}}, \quad wherez = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$$

Here:

- z represents the linear combination of input features x_i weighted by their respective coefficients β_i and an intercept term β_0.
- The sigmoid function ensures the output probabilities lie in the range $[0, 1]$.

The model parameters (β) are estimated by minimising the log-loss function:

$$Log - Loss = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

Overfitting in LR can be addressed using technqiues such as regularisation L1(Lasso) and L2 (Ridge) can be applied and cross-validation for hyperparamter tuning

2.1. Advantages of Logistic Regression

LR is straightforward to implement and interpret, offering clear insights into how each feature contributes to the prediction through its coefficients. The model is computationally efficient and performs well with small to medium-sized datasets. Its coefficients (β_i, beta_iβi) **what is beta_iβi here? Is the coeff not just β_i here** provide direct interpretability by indicating the magnitude and direction of each feature's impact on the target variable. To prevent overfitting, regularisation techniques such as L1 (Lasso) and L2 (Ridge) can be applied, penalising large coefficients and improving the model's generalisation to unseen data (Zou and Hastie, 2005).

2.2. Limitations of Logistic Regression

LR assumes a linear relationship between input features and the log-odds of the target variable. This makes it less effective for datasets with non-linear patterns unless feature transformations are applied (Hosmer et al., 2013). The model is sensitive to the scale of input features, requiring standardisation or normalisation for optimal performance, particularly when features have different ranges. Additionally, it is sensitive to outliers, which can significantly affect the estimated coefficients and lead to suboptimal predictions. Identifying and addressing outliers is crucial before applying the model.

3. Decision Tree

DT are a non-linear ML model used for both classification and regression tasks. The model works by recursively splitting the dataset into subsets based on feature thresholds, aiming to create homogeneous groups at the leaf nodes.

A Decision Tree represents a flowchart-like structure where each node signifies a decision based on a feature, and each leaf node represents the output (class label or regression value).

The model uses criteria such as Gini Index or Entropy to determine the best splits. Gini measures impurity by calculating the probability of misclassification:

$$Gini = 1 - \sum_{i=1}^{n} P_i^2$$

Random State	Logistic Regression (%)	Decision Tree (%)
Random_state=0	86.89	73.77
Random_state=3	78.69	75.41
Random_state=7	80.33	68.85
Random_state=10	86.89	70.49
Random_state=25	85.25	80.33
Random_state=42	86.89	70.49
Random_state=123	85.25	80.33
Random_state=2023	88.52	75.41
Random_state=88	81.97	77.05
Random_state=55	77.05	73.77
Average Score	**83.77**	**74.59**

Table 1: Average Accuracy of each Algorithm

Entropy measures information gain, calculated as:

$$Entropy = -\sum_{i=1}^{n} P_i \log_2(P_i)$$

The tree-building process continues until a stopping criterion is met, such as maximum depth, minimum samples per leaf, or no further information gain.

3.1. Advantages of Decision Tree

DT offer several advantages that make them versatile and useful in various applications. They can effectively handle complex interactions and non-linear patterns, making them suitable for diverse datasets (Breiman et al., 1984). Unlike many other models, they do not require feature scaling, as they operate based on thresholds, which simplifies preprocessing. The tree structure provides a clear and visual representation of decision paths, making it easy to interpret and explain to stakeholders. Additionally, DT can handle both numerical and categorical features, making them adaptable to datasets with mixed data types.

3.2. Limitations of Decision Tree

Despite their advantages, DT have notable limitations. They are prone to overfitting the training data, especially when allowed to grow too deep without restrictions (**?**) **Dont forget ref here**. Their high variance means that even small changes in the training data can result in significantly different trees, which can impact generalisation. Furthermore, they tend to favour features with more unique values, such as continuous variables, which can lead to biased splits. Lastly, DT struggle with extrapolation and do not perform well on unseen values that fall outside the range of the training data.

4. Evaluation of the Algorithms

LR and DT models were compared on this dataset to evaluate their strengths, limitations, and overall performance 2.LR emerged as the better performing model, largely because it suited the dataset's linear characteristics 2.2, 2. LR assumes a direct relationship between features and the target variable, which aligns well with features like thalachh (maximum heart rate)and oldpeak that exhibit linear trends. The model's simplicity and use of global decision boundaries allowed it to generalise effectively without being overly influenced by noise in the data. Standardising features using StandardScaler() ensured all inputs contributed proportionally making the model robust on unseen data. These factors, combined with the dataset's small-to-medium size, played a pivotal role in LR's superior performance. Although DT performed well, its performance accuracy was less compared to LR. While DT excels at capturing complex and non-linear relationships, their flexibility can become a drawback with small datasets. The limited data led to overfitting, with the tree creating specific splits that failed to generalise to new observations. The absence of regularisation techniques, such as pruning to limit tree depth, further exacerbated overfitting. DT also showed a bias towards features with many unique values, which skewed their decision making process. Their high variance, where slight changes in the data produce significantly different trees, compounded their instability. Overall, DT was less suited to the linear tendencies and size of this dataset, highlighting the importance of aligning model selection with data characteristics.

To improve the performance of DT, the classifier was pre-pruned using max_depth.This parameter controls the maximum depth of the tree. After setting a max_depth and running the algorithm again the model's performance improved by 4.83% bringing its accuracy score close to that of LR

LR could benefit from feature engineering techniques like adding interaction terms or polynomial features to capture non-linear relationships. For

Table 2: Comparison of Logistic Regression and Decision Tree

Aspect	**Logistic Regression**	**Decision Tree**
Assumptions	Linear relationship between features and log-odds.	No assumptions; works with non-linear data.
Feature Scaling	Requires scaling for optimal performance.	Not required.
Overfitting	Regularisation helps mitigate overfitting.	Prone to overfitting; mitigated by pruning.
Interpretability	Coefficients indicate feature importance.	Tree structure is visually interpretable.
Sensitivity to Outliers	Highly sensitive to outliers.	Less sensitive; splits are robust to extreme values.
Performance on Small Datasets	Generally reliable.	Can overfit, depending on depth and splits.

instance, combining features such as age and cholesterol could reveal hidden patterns. Additionally, addressing class imbalance through synthetic data generation, such as SMOTE, would create a richer training environment, leading to better generalisation. DT, on the other hand, could be improved by applying pruning methods to limit their depth and reduce overfitting. Exploring alternative splitting criteria like entropy instead of Gini index might refine their decisionmaking process. Feature selection, informed by feature importance scores, could also help focus the model on the most relevant predictors, reducing noise and improving accuracy. Both models could also benefit from ensemble techniques like Random Forest or Gradient Boosting. These approaches combine the strengths of multiple models to improve stability and accuracy. For instance, Random Forest reduces variance by aggregating multiple DT, while Gradient Boosting iteratively corrects errors to refine predictions. Thorough data exploration, including visualisation of feature distributions, can help identify and address outliers or skewed data, ensuring cleaner inputs for both models. Addressing class imbalances by oversampling minority classes or adjusting class weights would further enhance their robustness. With these improvements, LR and DT models could achieve better generalisation and reliability for real-world applications.

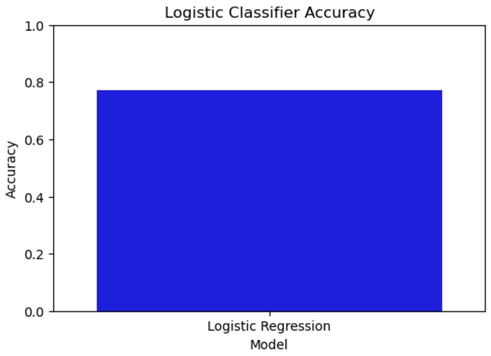

Figure 1: Logistic Regression Accuracy Score

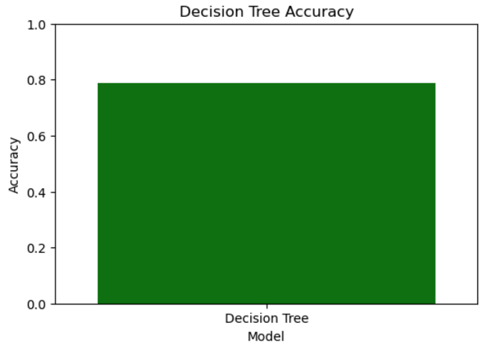

Figure 2: Decision Tree Accuracy Score

In contrast, DT struggled to perform. While they excel at capturing complex and non-linear relationships, their flexibility can become a drawback with small datasets. The limited data led to overfitting, with the tree creating specific splits that failed to generalise to new observations. The absence of regularisation techniques, such as pruning to limit tree depth, further exacerbated overfitting. DT also showed a bias towards features with many unique values, which skewed their decisionmaking process. Their high variance, where slight changes in the data produce significantly different trees, compounded their instability. Overall, DT were less suited to the linear tendencies and size of this dataset, highlighting the importance of aligning model selection with data characteristics. To improve the performance of both models, targeted strategies can be employed. LR could benefit from feature engineering techniques like adding interaction terms or polynomial features to capture non-linear relationships. For instance, combining features such as age and cholesterol could reveal hidden patterns. Additionally, addressing class imbalance through synthetic data generation, such as SMOTE, would create a richer training environment, leading to better generalisation. DT, on the other hand, could be improved by applying pruning methods to limit their depth and reduce overfitting. Exploring alternative splitting criteria like entropy instead of Gini index might refine their decision-making process. Feature selection, informed by feature importance scores, could also help focus the model on the most relevant predictors, reducing noise and improving accuracy.

Both models could also benefit from ensemble techniques like Random Forest or Gradient Boosting. These approaches combine the strengths of multiple models to improve stability and accuracy. For instance, Random Forest reduces variance by aggregating multiple DT, while Gradient Boosting iteratively corrects errors to refine predictions. Thorough data exploration, including visualisation of feature distributions, can help identify and address outliers or skewed data, ensuring cleaner inputs for both models. Addressing class imbalances by oversampling minority classes or adjusting class weights would further enhance their robustness. With these improvements, LR and DT models could achieve better generalisation and reliability for real-world applications.

5. Conclusion

In conclusion, this analysis has demonstrated the comparative strengths and limitations of LR and DT models when applied to the given dataset. LR excelled due to the dataset's predominantly linear characteristics, leveraging its simplicity, global decision boundaries, and effective feature standardisation to deliver robust and generalisable predictions. On the other hand, DT struggled due to their susceptibility to overfitting, high variance, and biases in feature selection, particularly when working with small datasets.

The findings emphasise the importance of aligning model selection with the nature of the dataset to achieve optimal performance. While LR outperformed in this instance, tailored improvements such as feature engineering, regularisation, and handling class imbalances could further enhance its capabilities. Similarly, applying regularisation techniques, pruning, and ensemble methods could significantly improve the performance and stability of DT. Ultimately, the choice of model depends on the specific requirements of the problem, and integrating domain knowledge with systematic data exploration is critical to building effective predictive models.

Bibliography

Boateng and Abaye. 2019. A review of the logistic regression model with emphasis on medical research. *Journal of Data Analysis and Information Processing*, 7:190–207.

L Breiman, J Friedman, C.J Stone, and R.A Olshen. 1984. *Classification and Regression Trees*. CRC Press.

D.W Hosmer, S Lemeshow, and R.X Sturdivant. 2013. *Applied Logistic Regression*, 3rd edition. Wiley.

H Zou and T Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.