

Assignment Stage 2: Report

Data Science for Innovation

Submitted By-

Kushal Ahuja (14191922)

Table of Contents

[Section 1: Literature review: 3](#_bookmark0)

[Section 2: Setup 4](#_bookmark1)

[Section 3: Approach 5](#_bookmark2)

[Section 4- Result 6](#_bookmark3)

[Section 5: Conclusion 12](#_bookmark4)

[References 13](#_bookmark5)

# Section 1: Literature review:

**Paper 1:** View of Heart Disease Prediction Using Logistic Regression. (n.d.). https://penerbit.uthm.edu.my/periodicals/index.php/ekst/article/view/2163/1026 The paper discusses applying binary models of logistic regression utilizing various

techniques, identifying important factors that influence the existence or nonexistence of heart disease. The dataset contains data from 270 individuals, including age, resting BP, cholesterol levels, the max cardiovascular rate reached, sex, symptoms of chest pain, its slope, number of main arteries, and thalassemia. BLR models, BLR using LQD, & BLR using MAD are the three approaches used. According to the results, the BLR using the MAD model has the greatest success percentile. The research results present information regarding the important variables that contribute to cardiovascular disease and illustrate the necessity of spreading knowledge and measures to prevent it.

**Paper 2:** Heart Disease Prediction Using Machine Learning Algorithms. (2020b, February 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/9122958 In the paper, authors examined the accuracy of knn, the decision tree, and linear regression, along with support vector machine learning approaches for estimating the risk of cardiovascular disease. The UCI repository site provides access to all of the resources for prediction. Python programming is used for algorithm implementation. All algorithms are run in a Jupyter Notebook. According to the experimental results, the authors attained the greatest accuracy of 87% by utilizing the k-nearest neighbor method, then utilized a support vector machine at 83 percent, a decision tree at 79 percent, and linear regression at 78 percent for anticipating cardiovascular disease.

**Paper 3:** Pal, M., & Parija, S. R. (2021). Prediction of Heart Diseases using Random Forest. Journal of Physics, 1817(1), 012009. https://doi.org/10.1088/1742-6596/1817/1/012009 Data extraction technology, which entails collecting significant data from massive databases, is widely used in a variety of sectors, particularly healthcare. The dataset was obtained from Kaggle and consists of 303 observations with 14 characteristics. The random forest approach was used to analyze and categorize the dataset using Python code, in Jupyter Notebook. In this heart disease prediction, the research attained a success rate of 86.9%, sensibility of 90.6%, and precision of 82.7%. The analysis of receiver parameters of operation produced a diagnostic rate of 93.3%. The random forest algorithm's efficiency in heart disease categorization was emphasized.

**Paper 4:** Saraswathi, R. V., Gajavelly, K., Nikath, A. K., Vasavi, R., & Anumasula, R. (2022). Heart Disease Prediction Using Decision Tree and SVM. In Algorithms for intelligent systems (pp. 69–78). Springer Nature. https://doi.org/10.1007/978-981-16-7389-4\_7

This article presents a technique for predicting cardiac disease by combining DT & SVM. The researchers gathered a dataset, preprocessed it, and then used machine learning methods. Random forest obtained 88.1% accuracy, whereas decision trees obtained 89.6%, SVM had an accuracy of 87.6%, whereas Nave Bayes had an accuracy of 85.6%. Decision trees partition data recursively based on feature entropy, whereas SVM transforms data with low dimensions to a viable high-level space. The suggested system provides rapid analysis and accuracy similar to complicated algorithms. It makes a contribution to data analysis and medical research by highlighting the significance of correct categorization in cardiovascular disease prediction.

# Section 2: Setup

Research Question – Can we predict the chances of the occurrence of heart disease using the given features in the dataset?

(The target variable is a binary variable indicating the presence or absence of a heart disease.) Null hypothesis – There is no relationship between the given features and the occurrence of a heart attack. The model’s prediction of the occurrence of heart attack is not statistically significant, and any observed association between the attributes and the target variable is due to random chance alone.

Alternative hypothesis – There is a relationship between the given attributes and the target variable (predicting the chances of heart attack). The model's heart attack prediction is statistically significant and is proven using different cross-validation and model performance metrics. This signifies that the observed association between the given features and the target is not random chance alone.

For the given classification problem, the Accuracy, Precision, Recall, and F1 Scores are calculated, and the Confusion matrix is used to assess the model's reliability. The comparisons are made based on the values of these performance metrics and the confusion matrix:

1. Accuracy – Measures the overall correctness of the model by indicating the correctly predicted proportions.
2. Precision – Evaluate the ability of the model to correctly distinguish the positive prediction among the predictions labelled as positive.
3. Recall – Correctly distinguish between the positive instances out of the actual positive predicted instances
4. F1 Score – Evaluates and provides the balance score between Precision and Recall in a harmonic manner.

The model's reliability can be assessed by calculating the scores of the five different models and used to compare which model best fits the dataset. The accuracy and precision scores will provide a general idea about whether the model can accurately differentiate between the two target classes. The recall and f1 scores focus more on how the performance of the model is while predicting the Target = 1. The confusion matrix will be used to check the False Positive and False Negative values.

False negative values can significantly impact hypotheses related to predicting the chances of a heart attack in the dataset. False negatives occur when the model incorrectly predicts that an individual does not have a heart attack when, in reality, they do. In the context of hypotheses, false negatives can lead to rejecting the alternative hypothesis in favour of the null hypothesis. This means that the model may fail to identify the true association between the predictors and the occurrence of a heart attack, potentially leading to a conclusion that there is no significant relationship between the variables. As a result, critical findings regarding the predictors' influence on the chances of a heart attack might be missed, leading to potential underestimation of the model's predictive power and the significance of the predictors. It is crucial to minimise false negatives to ensure that the hypotheses accurately capture the true relationship between the predictors and the occurrence of a heart attack, enabling appropriate conclusions and informed decision-making in heart disease prevention and treatment.

Dataset Description- The dataset originally featured 76 attributes, out of which the research studies were performed on these 14 attributes specifically. They are age, sex, chest pain, blood pressure, serum cholesterol in mg/dl, fasting blood sugar, resting electrocardiographic results, maximum heart rate, Exercise-induced angina, ST depression induced by exercise

relative to rest, the slope of the peak exercise, number of significant vessels and diagnosis of heart disease status. Additionally, the dataset includes information on whether each patient has experienced any heart disease or not, which is represented by the binary target variable. These given attributes from the dataset can serve as the basis of training a binary classification model to predict the chances of heart attack from the available feature.

# Section 3: Approach

The approach to be followed as per stage 1 to achieve the goal of predicting the chances of heart attack of the patients or answering further business problems like predicting the heart attack rate for insurance companies to make data-driven decisions related to specific attributes and its insights.

**Logistic Regression** -The first model is our proposed model ,i.e., Logistic Regression. To begin with, the model is trained by defining the features and scaling them, followed by splitting the dataset, i.e. training and testing. The testing dataset is to predict the model's accuracy and evaluate the prediction rate.

In the Python notebook prototype we have made five models for linear regression for optimal results and to compare as benchmark models, after implementing the model, it is found that logistic regression with Feature Selection using SelectFromModel has the highest level of model performance, i.e., 89% by evaluating the f1 score. The confusion matrix outlines that it has more instances of true positives and true negatives.

Moreover, the accuracy of this model is also the highest when compared to other models, i.e., 89%, as this model has the highest accuracy, recall and precision, it shows that this model fits the best at correctly predicting heart disease or chances of heart attacks.

The other results of logistic regression are using model tuning, hyperparameter tuning and one-hot coding, for hyperparameter tuning, we have created a grid and classified them and then using GridsearchCV, found the best hyperparameters which is given below -

**Decision tree**

The second model which is implemented is a Decision tree; the model has performed well with having an accuracy of 84%, with the hyperparameters: 'max\_depth' of 6 and 'min\_samples\_split' of 2. These parameters imply a balance between capturing detail and preventing overfitting, as a higher max depth allows the model to make more detailed decisions, while a lower min\_samples\_split can lead to more complex trees that may not generalize well. This model is implemented because it is easy to understand and is easy to interpret

**Naive Byes**

The other model applied is naive byes, we have implemented this model because it is easy to interpret and can be scaled easily to handle the dataset,

We have applied two models, The hyperparameter tuned in Model 2 ,i.e, Gaussian Naive Bayes was 'var\_smoothing' with a best value of 0.1, which adds a portion of the most considerable variance of all features to the variances for calculating stability.

With model 2, by use of hyperparameters tuning and feature engineering over the baseline naive byes model has shown better results and improvement.

**Random Forest**

Three models has been applied by using Random forest with having better results in model 2, in this model standard scaling was applied to the features, and the RandomForestClassifier was adjusted to handle class imbalance by setting 'class\_weight' to 'balanced’. The hyperparameters were tuned using the 'max\_depth': 2, 'min\_samples\_split': 2, 'n\_estimators': 300.

We have also implemented the feature interaction based on the feature's importance in model 3, and it is found that cp (chest pain type) and exang (exercise-induced angina) have different aspects of heart-related pain. Additionally, the interaction between oldpeak and slope attributes is related to ST segment on an EKG, which can indicate heart damage, this emphasizes more information about the ST segment.

The model tuning is done by adjusting the hyperparameter to improve the overall performance of the model, thus have used grid search and feature interaction.

**Support Vector Machine (SVM)**

Three models have been made using SVM to predict the rate, the scaling has been done using StandardScaler() and the fitted into training and testing dataset, and then the model has been defined and tuning the model using GridSearchCV to find the best hyperparameters, Different kernel functions were used in the models, including 'rbf' and 'sigmoid'. The 'C' parameter, which controls the trade-off between achieving a low error on the training data and maximizing the margin, was best set to 1 for all three models.

The analysis by models has given the insights of the rate whether the person is likely to have a risk of heart attack or not and has further solved the business problems, it is upon stakeholder to choose the result from the model and apply to combat the issue.

The analysis has also answered the research question by the results and insights given below.

# Section 4- Result

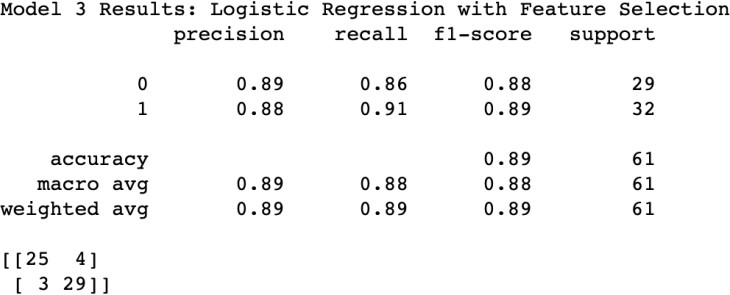
**Model Overview**

We trained several models based on our setup and approach. Here is a breakdown and analysis of each of our models:

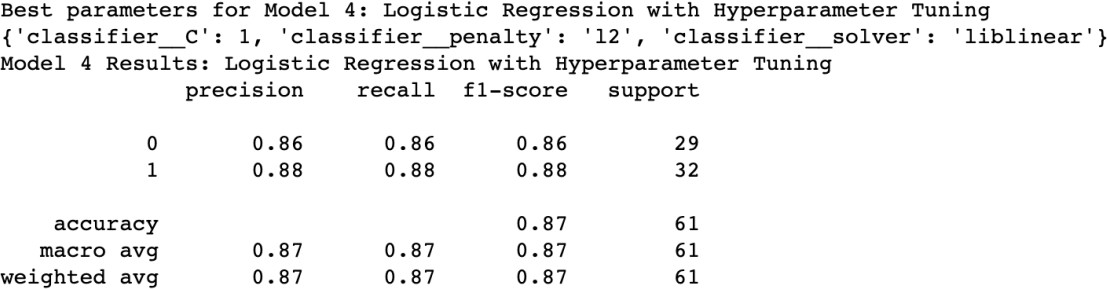
**Logistic Regression Models 1, 2, & 3**

Logistic Regression Model 1 achieved an accuracy of 86%, with hyperparameters 'C' (inverse of regularization strength) set to 1 and 'penalty' (specification of the norm used in the

penalization) set to 'l2', showing that a bit of regularization helped the model avoid overfitting. The model made use of the StandardScaler for scaling the features.

Model 2 introduced interaction terms between features, which maintained the model's accuracy at 86%. This suggests that the relationships between the variables in our dataset could be adequately captured without explicit interaction terms.

*Figure- Classification report from Logistic Regression with best parameters found*

Model 3 took a different approach by categorizing the 'age' feature into bins and achieved a slightly lower accuracy of 85%. This indicates that the linear nature of logistic regression might be handling the 'age' feature better without binning, capturing the nuances of changes in age.

*Figure- Classification report from Logistic Regression with best parameters found*

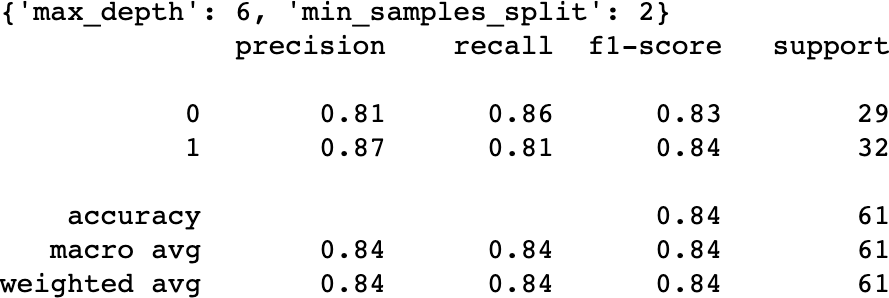
**Decision Tree Model 1**

The model achieved an overall accuracy of 84% and had the best results with the hyperparameters.

Stakeholders will appreciate the simplicity of decision trees, as they are easily interpretable and can give insights into which features are important for making predictions. However, the model's relatively lower accuracy than others may be a downside.

Decision trees are easy to understand and interpret and can handle categorical and numerical data. But at the same time, they can easily overfit or underfit if not properly tuned and are sensitive to changes in the data, potentially leading to different splits and a different overall

tree structure.



*Figure- Classification report from Decision tree with best parameters found*

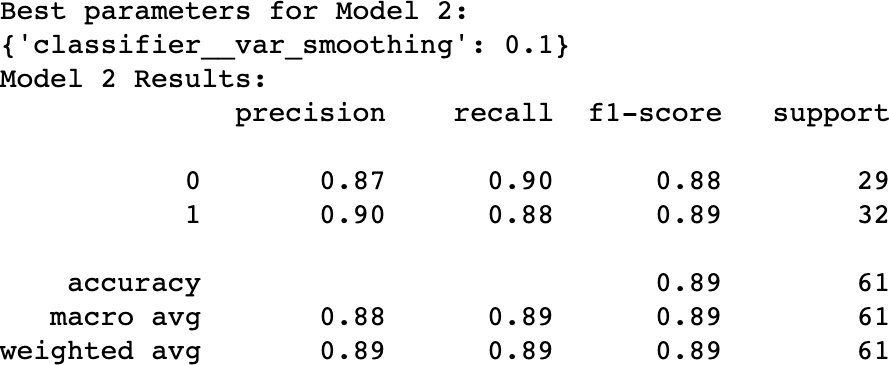
**Naive Bayes Models 1 & 2**

Both Naive Bayes models performed well, with 87% and 89% accuracy, respectively. The models worked best with the features scaled using StandardScaler, and in Model 2, age was binned into three categories.

The stakeholders will appreciate these models' high accuracy, although they might find the underlying calculations less intuitive than those of decision trees.

One of the advantages of Naive Bayes models is that they are easy to implement, efficient, and work well with large datasets and high-dimensional data.

On the other hand, they assume that all features are independent, which might not be true and could affect performance.



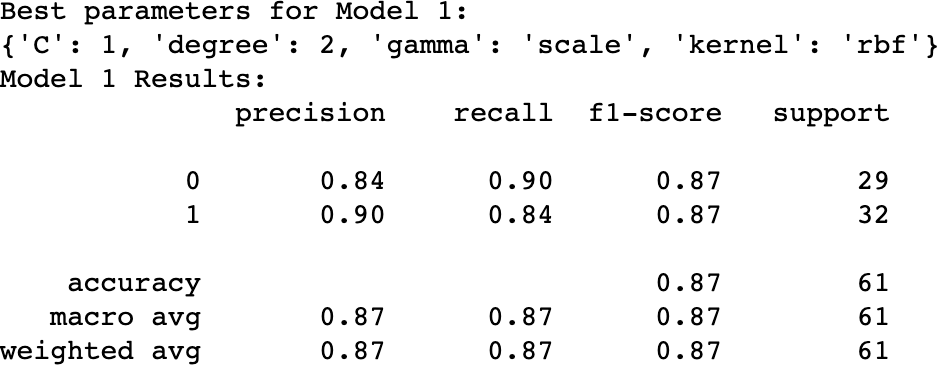
*Figure -Classification report from Naïve Byes with best parameters found*

**Support Vector Machine Models 1, 2, & 3**

The SVM models also had high accuracies ranging from 85% to 87%. They all used the StandardScaler for feature scaling. The SVM models also experimented with creating new feature interactions, which appeared to maintain the accuracy of the models.

SVMs can model non-linear decision boundaries, and there are many kernels to choose from. They are also robust against overfitting, especially in high-dimensional space.

Also, SVMs are unsuitable for larger datasets as the training time can be extended. They also require feature scaling and are less easily interpretable than other models.



***Figure -*** *Classification report from SVM with best parameters found*

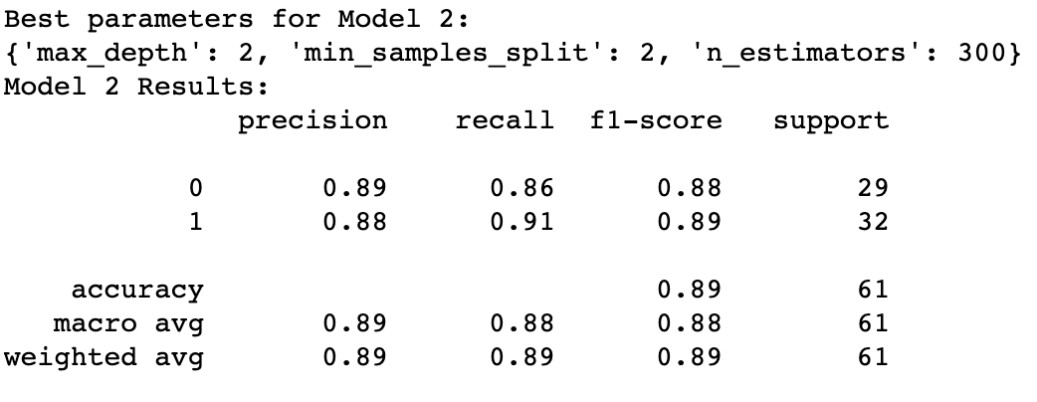
**Random Forest Models 1, 2, & 3**

Model 1: In this model, hyperparameter tuning is done using GridSearchCV with cross- validation, but there's no scaling of features. The performance metrics achieved were fairly balanced between both classes, with an overall accuracy of 87%.

Model 2: In this model, The hyperparameters were tuned similarly to Model 1. This model achieved slightly higher accuracy (89%) than Model 1, with balanced precision and recall scores for both classes.

Model 3: In this model, a feature importance study was carried out, following which interaction terms were created using PolynomialFeatures for some of the most important features. The model also incorporated class balancing and scaled features like Model 2. However, it achieved a slightly lower accuracy of 85% compared to the previous two models.

In summary, Model 2 provided the best results with an overall accuracy of 89%. Model 1 performed relatively well, too, with an overall accuracy of 87%. The addition of interaction terms in Model 3 didn't improve the model's performance, suggesting that the complexity added by these terms might have been less beneficial for this particular dataset.



***Figure -*** *Classification report from Random Forest with best parameters found*

**Comparative Analysis**

**Decision Trees:** The single decision tree model was simple and easy to interpret but fell short in accuracy (84%) compared to the other models.

**Naive Bayes:** Both Naive Bayes models performed similarly in terms of accuracy (87% and 89%). However, Model 2, which used a tuned 'var\_smoothing' hyperparameter and binned age feature, performed slightly better, suggesting some improvement with these modifications.

**Support Vector Machines:** All three SVM models achieved similar accuracy (85% to 87%). However, Model 2, which used a sigmoid kernel function, stood out by achieving a slightly higher accuracy, indicating that this kernel might be better suited to this particular dataset.

**Random Forest:** All models performed well with small variations in accuracy (85% to 89%). Model 3, despite adding interaction terms, did not show an increase in accuracy, suggesting that the ensemble nature of Random Forests might already be capturing complex relationships between features.

**Logistic Regression:** All models performed fairly similarly (85% to 86%). The third model, where we binned the 'age' feature, led to a slight drop in accuracy, which indicates that the original, continuous age feature is better suited to this model.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Model Type*** | **Model Numb er** | **Model Description** | **Precisi on (0)** | **Precisi on (1)** | **Rec all (0)** | **Rec all (1)** | **F1**  **Sco re**  **(0)** | **F1**  **Sco re**  **(1)** | **Accur acy** |
| *Logistic Regress*  *ion* | 1 | Basic  Logistic Regression | 0.83 | 0.87 | 0.86 | 0.84 | 0.85 | 0.86 | 0.85 |
| *Logistic Regress*  *ion* | 2 | With One- Hot Encoding and Interaction  Features | 0.81 | 0.90 | 0.90 | 0.81 | 0.85 | 0.85 | 0.85 |
| *Logistic Regress*  *ion* | 3 | With Feature Selection | 0.89 | 0.88 | 0.86 | 0.91 | 0.88 | 0.89 | 0.89 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Logistic Regress*  *ion* | 4 | With Hyperparame ter Tuning | 0.86 | 0.88 | 0.86 | 0.88 | 0.86 | 0.88 | 0.87 |
| *Logistic Regress*  *ion* | 5 | With Feature Selection using SelectFromM  odel | 0.86 | 0.88 | 0.86 | 0.88 | 0.86 | 0.88 | 0.87 |
| *Decisio n Tree* | 1 | With Hyperparame ter Tuning | 0.81 | 0.87 | 0.86 | 0.81 | 0.83 | 0.84 | 0.84 |
| *Naive Bayes* | 1 | Basic Gaussian  Naive Bayes | 0.84 | 0.90 | 0.90 | 0.84 | 0.87 | 0.87 | 0.87 |
| *Naive Bayes* | 2 | With Hyperparame ter Tuning | 0.87 | 0.90 | 0.90 | 0.88 | 0.88 | 0.89 | 0.89 |
| *SVM* | 1 | With  Hyperparame ter Tuning | 0.84 | 0.90 | 0.90 | 0.84 | 0.87 | 0.87 | 0.87 |
| *SVM* | 2 | With Feature Engineering | 0.83 | 0.87 | 0.86 | 0.84 | 0.85 | 0.86 | 0.85 |
| *SVM* | 3 | With Feature Interaction | 0.83 | 0.87 | 0.86 | 0.84 | 0.85 | 0.86 | 0.85 |
| *Random Forest* | 1 | Hyperparame ter tuning with  GridSearchC V | 0.89 | 0.85 | 0.83 | 0.91 | 0.86 | 0.88 | 0.87 |
| *Random Forest* | 2 | Feature scaling and | 0.89 | 0.88 | 0.86 | 0.91 | 0.88 | 0.89 | 0.89 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Random Forest* |  | class balance adjustment |  |  |  |  |  |  |  |
| 3 | Feature importance study and interaction  terms | 0.86 | 0.85 | 0.83 | 0.88 | 0.84 | 0.86 | 0.85 |

**Conclusion**

In terms of overall accuracy, the models ranged from 84% to 89%, with Naive Bayes Model 2, Random Forest Model 2 and Logistic Regression Model 3 achieving the highest accuracy of 89%.

The models demonstrated that feature scaling, choice of kernel in SVM, and tuning hyperparameters could be vital in improving model performance. While creating interaction terms and binning features provided interesting insights, they did not significantly improve the models' performance in this case.

Comparing across the algorithms, Naive Bayes stood out for its simplicity and high performance, despite its feature independence assumption. Support Vector Machines and Random Forests also achieved high accuracy. However, the decision tree and logistic regression models, despite being simple and interpretable, couldn't achieve as high an accuracy.

# Section 5: Conclusion

With the help of this study, we learned valuable insights into how we can leverage the power of machine learning algorithms in predicting heart disease. As we can see, each of the models presented its own strengths and weaknesses; no single model was the best in all scenarios and we would recommend that model selection be based on the specific use case and data available at hand.

Our key learnings were focused on how the importance of understanding the characteristics and assumptions of each model and how they affect each model's performance. It should be noted that Feature engineering also proved to be a crucial step in the process, especially in the case of Random Forest model.

We recommend choosing between Random Forest 2, Naive Bayes 2, and Logistic Regression 3 depending on interpretability, model assumptions, computational resources, overfitting, handling of categorical variables, robustness, and update frequency. Logistic Regression is more interpretable and light on resources but may overfit. Naive Bayes is also light on resources but assumes feature independence. Random Forest handles categorical variables and overfitting well but may require more resources.

However, it's essential to note that these models should supplement and not replace the expert judgment of medical professionals.

**Future Work**

Further research could explore the use of different feature selection techniques, more complex models, or deep learning techniques for this task. Incorporating a larger and more diverse dataset could also enhance the models' performance and generalizability.

The limitations of the study, including the small dataset size, should be addressed in future work. Additionally, continuous updates and validation of the models with new data are crucial to ensure their reliability and usefulness in a clinical setting.

We believe this project has conducted in an ethical way to achieve the results and has considered keeping the data privacy as a priority and thus has worked on the public dataset to spread the awareness of health among people and what attributes contribute to the risk to heart problems, certain steps were taken to undisclose any personal information found.

## References

*UCI Machine Learning Repository: Heart Disease Data Set*. (n.d.). [https://archive.ics.uci.edu/ml/datasets/Heart+Disease](https://archive.ics.uci.edu/ml/datasets/Heart%2BDisease)