# Heart Disease Prediction Using Machine Learning

Denise Patricia B. Manalo
*College of Engineering, Architecture and Fine Arts (CEAFA)*
*Electrical and Computer Engineering Department*
*Batangas State University-Alangilan Campus*
Batangas City, Philippines
denisepatricia.manalo@g.batstate-u.edu.ph

## I. INTRODUCTION

Heart is a vital organ in the human body. If it fails to function properly, the brain and other organs will stop working, and a person can die within minutes. Diabetes, high blood pressure, high cholesterol, abnormal pulse rate, and a variety of other risk factors make it difficult to detect heart disease. Thus, it is one of the world's leading causes of death today.

According to the World Health Organization, heart diseases claim the lives of 17.7 million people each year, accounting for 31% of all deaths worldwide [1]. As a result, the provision of high-quality services at reasonable prices is a major challenge for healthcare organizations (hospitals, medical centers) [2]. Quality service entails correctly diagnosing patients and providing effective treatments. Poor clinical decisions can have disastrous consequences, which is unacceptably dangerous [3]. Whereas, they can improve their performance through the use of appropriate computer-based information and/or decision support systems.

Nowadays, the majority of hospitals use some type of hospital information system to manage their healthcare or patient data [4]. Typically, these systems generate massive amounts of data in the form of numbers, text, charts, and images. Unfortunately, these data are rarely used to assist clinicians in making clinical decisions.

Furthermore, the study used data mining techniques for machine learning. The two most common modelling objectives are classification and prediction. Classification models predict discrete, unordered categorical labels, whereas prediction models predict continuous-valued functions. Classification algorithms are used in Decision Trees and Neural Networks, whereas prediction algorithms are used in Regression.

Motivated by the global increase in heart disease mortality each year and the availability of massive amounts of patient data from which to extract useful knowledge, the researcher care to use data mining techniques to assist health care professionals in diagnosing heart disease. And simply transform data into actionable information that enables healthcare professionals to make intelligent clinical decisions.

Thus, the study covers the classification of the target variable using different machine learning algorithm and disclose which algorithm is suitable for the dataset. It fails to apply the machine learning algorithm first but rather proceed directly to hyperparameter tuning for better accuracy.

## II. METHODOLOGY

This section describes the procedural step used in conducting the major aspect of the study. The study focuses on machine learning techniques and modelling in order to attain the desired accuracy score results as shown in Figure 1.1. Whereas, taking into consideration different features of the machine learning technique.

### *A. Raw Dataset of Attributes of the Heart Disease*

The study gathered raw dataset in Kaggle [5] consisting of 14 columns and the target is the class variable which is affected by other 12 columns. There are over 4,000 records and 15 attributes in total. Each characteristic has the potential to be a risk factor. Risk factors include demographic, behavioral, and medical factors. It includes demographic dataset of nominal values which includes: sex, current smoker, blood pressure medication, prevalent stroke, prevalent hypertensive, and diabetes. Whereas, ordinal values includes: systolic blood pressure, cholesterol level, body mass index, heart rate, the number of cigarettes that the person smoked on average in one day, and glucose level.

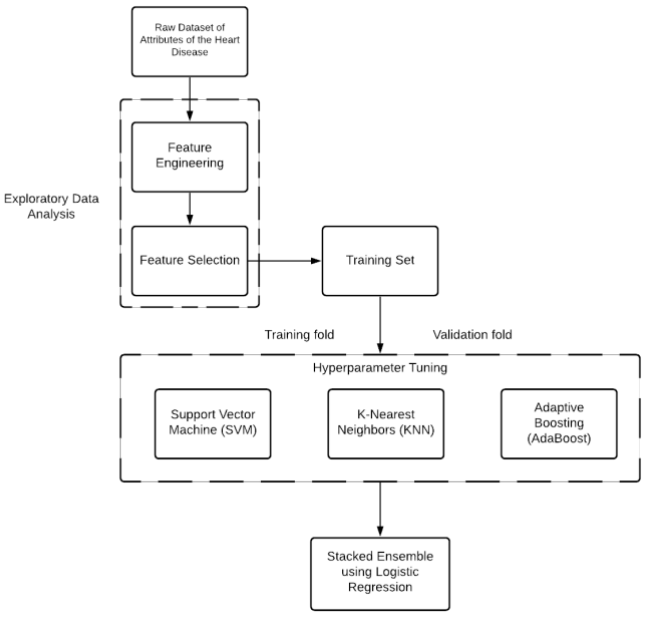


Figure 1.1 Heart Disease Prediction Methodology

### *B. Feature Engineering*

The Exploratory Data Analysis (EDA) starts in this section whose primary goal is to examine the dataset for distribution, outliers, and anomalies. It entails analyzing the data to determine its distribution, main characteristics, patterns, and visualizations[6]. It also includes tools for generating

hypotheses by visualizing and comprehending data using graphical representation as shown in Figure 1.2.

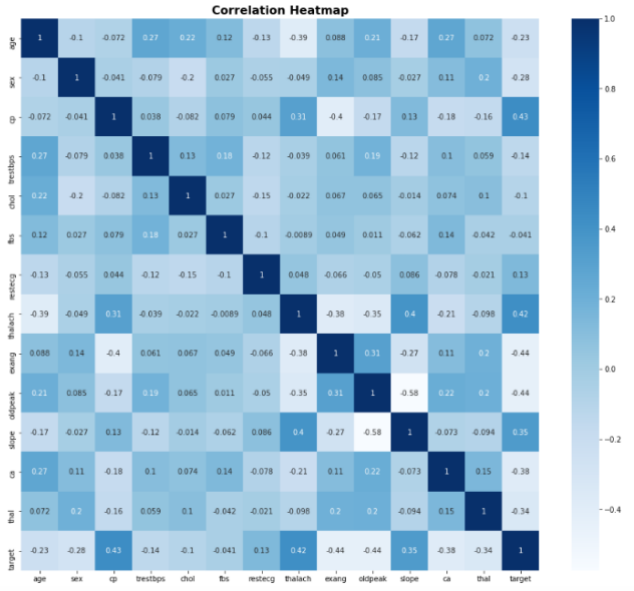


Figure 1.2 Correlation Heatmap

### C. *Feature Selection*

After familiarizing and visualizing the dataset. We use some techniques in this step to identify the important and irrelevant features to feed into our model.

### D. *Training Set*

To select a stable feature set which improves classification accuracy, we use ensemble based feature selection, drop the target column and use the splitting of training and validation datasets in this section. The validation fold is consists of 30% of the training dataset while the training fold is consists of 70% of the dataset.

### E. *Hyperparameter Tuning*

To enhance the performance of our model we make use of hyperparameter tuning using Grid Seach and three (3) base classifier models namely Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Adaptive Boost (AdaBoost) and define the range of hyperparameter values corresponding to the model to be searched over for hyperparameter tuning.

### F. *Stacked Ensemble using Logistic Regression*

Stacking is a model ensembling technique that involves combining information from multiple predictive models and using them as features to generate a new model [7]. It is a method for forming combination of various indicators to improve prediction accuracy[8]. Therefore, the study used logistic regression as meta-model for it is relevant in classification tasks like predicting a class label.

## III. Results And Discussions

This section presents the data collected from the study on this major aspect of the research problem and interpretation of the findings of study.

### A. *Confusion Matrix*

The performance of the classification model is summarized in the error matrix or what we called as confusion matrix. The number of correct and incorrect predictions are summarized with count values and broken down by each class.

Figure 1.3 shows the confusion matrix of the base classifiers in testing the dataset. It shows that majority of the outcome of the model correctly predicts the positive and negative class with its actual values in the X-axis and predicted values in the Y-axis.

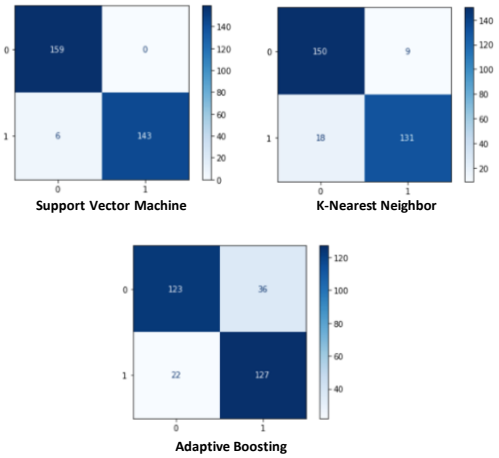


Figure 1.3 Confusion Matrix of Base Classifiers

The confusion matrix entails that our prediction model is relevant in detecting a potential heart disease with as close as 100% accurate as possible for diagnosis.

### B. *Curve Plot*

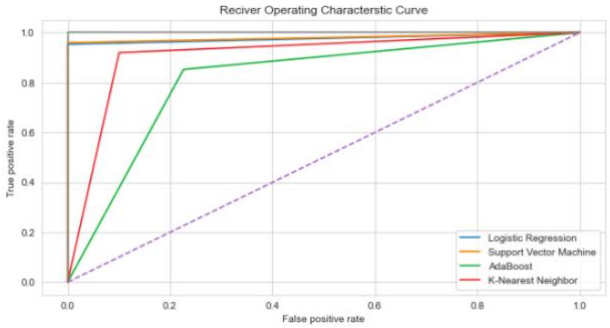


Figure 1.4 AUC-ROC Curve

It is evident from the plot that Support Vector Machine (SVM) has the highest ROC Curve compared to the other based classifier models and stacked model. Therefore, we can interpret that SVM did a great job of classifying the positive class in the dataset since the higher Y-axis value indicates a higher number of True positives than False negatives.

### C. *Test Dataset Results*

Table 1.1 shows the accuracy scores in percentage of the tuned base classifier models during training and validation of the dataset.

| | Base Classifier Model | Training Accuracy % | Testing Accuracy % |
|---|---|---|---|
| 0 | Tuned Support Vector Machine | 100.000000 | 98.051948 |
| 1 | Tuned K-nearest neighbors | 84.239888 | 91.233766 |
| 2 | Tuned AdaBoost Classifer | 90.794979 | 81.168831 |

Table 1.1 Training and Testing Dataset Results of Base Classifier Models

It is manifested that SVM has the highest training and testing accuracy results with 100% and 98% training and validation scores. Whereas, Adaptive Boost Classifier scores an approximate 91% of training accuracy score and 81% validation score. And lastly, KNN Classifier with 84% and 91% training and validation scores, respectively.

| | Stacked Model | Training Accuracy % | Testing Accuracy % |
|---|---|---|---|
| 0 | Stacked Model: Logistic Regression | 95.39749 | 97.727273 |

Table 1.2 Training and Testing Dataset Result of Stacked Model

Furthermore, the study uses Logistic Regression as classifier for stacked modelling from the given base classifier models with 95% and 98% training and validation scores, respectively.

## IV. Conclusion

A heart disease prediction using machine learning used two most common modelling objectives which are classification and prediction. Classification algorithms were used in hyperparameter tuning while prediction algorithms were used in regression. The study used three (3) base classifiers for classification algorithm namely Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Adaptive Boosting (AdaBoost).

Overall, in training and testing these classifiers, SVM got the highest training and testing scores compared to the other two based classifiers. With that, the results shows that training the dataset using hyperparameter tuning demonstrates better results in terms of scores in predicting a potential heart disease. Also, combining the various indicators used by these models and stacked them in the ensemble logistic regression improves the prediction with 98% accuracy score.

This made the model pertinent in predicting a potential heart disease using the information of the patient data with the accuracy score trained and tested using the machine learning.

## V. Recommendation

The future work of the study includes training and testing other classification techniques in order to compare the accuracy results that would be generated in the future results. Also, using a different type of ensemble modelling can be considered for further discussion.

## References


[1] Global Atlas on Cardiovascular Disease Prevention and Control. Geneva, Switzerland: World Health Organization, 2011

[2] Palaniappan, S., & Awang, R. (2008). Intelligent heart disease prediction system using data mining techniques. *AICCSA 08 - 6th IEEE/ACS International Conference on Computer Systems and Applications*,108–115. https://doi.org/ 10.1109 /AICCSA. 2008. 4493524

[3] Kanakaraddi, S. G., Gull, K. C., Bali, J., Chikaraddi, A. K., & Giraddi, S. (2021). Disease prediction using data mining and machine learning techniques. *Lecture Notes on Data Engineering and Communications Technologies*, *64*, 71–92. https://doi.org/10.1007/978-981-16-0538-3_4

[4] Kanakaraddi, S. G., Gull, K. C., Bali, J., Chikaraddi, A. K., & Giraddi, S. (2021). Disease prediction using data mining and machine learning techniques. *Lecture Notes on Data Engineering and Communications Technologies*, *64*, 71–92. https://doi.org/10.1007/978-981-16-0538-3_4

[5] *Heart Disease UCI | Kaggle*. (n.d.). Retrieved May 31, 2021, from https://www.kaggle.com/ronitf/heart-disease-uci

[6] Prabhu, T. N. (2019, August 10). *Exploratory data analysis in Python. | by Tanu N Prabhu | Towards Data Science*. https://towardsdatascience.com/exploratory-data-analysis-in-python-c9a77dfa39ce

[7] Chatzimparmpas, A., Martins, R. M., Kucher, K., & Kerren, A. (2021). StackGenVis: Alignment of data, algorithms, and models for stacking ensemble learning using performance metrics. *IEEE Transactions on Visualization and Computer Graphics*, *27*(2), 1547–1557. https://doi.org/10.1109/TVCG.2020.3030352

[8] Breiman, L. (2020). Stacked regressions. *Machine Learning*, *24*(1), 49–64. https://doi.org/10.1007/bf00117832


## Appendices

```
Train Result:
================================================
Accuracy Score: 100.00%
________________________________________________
Classification Report:
                0      1  accuracy  macro avg  weighted avg
precision     1.0    1.0       1.0        1.0           1.0
recall        1.0    1.0       1.0        1.0           1.0
f1-score      1.0    1.0       1.0        1.0           1.0
support     340.0  377.0       1.0      717.0         717.0
________________________________________________
Confusion Matrix:
 [[340   0]
 [  0 377]]

Test Result:
================================================
Accuracy Score: 98.05%
________________________________________________
Classification Report:
                0           1  accuracy   macro avg  weighted avg
precision    0.963636    1.000000  0.980519    0.981818      0.981228
recall       1.000000    0.959732  0.980519    0.979866      0.980519
f1-score     0.981481    0.979452  0.980519    0.980467      0.980500
support    159.000000  149.000000  0.980519  308.000000    308.000000
________________________________________________
Confusion Matrix:
 [[159   0]
 [  6 143]]
```

a. Classification Report of Support Vector Machine (SVM) Hyperparameter Tuning

```
Train Result:
================================================
Accuracy Score: 84.24%
________________________________________________
Classification Report:
                0           1  accuracy   macro avg  weighted avg
precision    0.799472    0.890533  0.842399    0.845002      0.847352
recall       0.891176    0.798408  0.842399    0.844792      0.842399
f1-score     0.842837    0.841958  0.842399    0.842398      0.842375
support    340.000000  377.000000  0.842399  717.000000    717.000000
________________________________________________
Confusion Matrix:
 [[303  37]
 [ 76 301]]

Test Result:
================================================
Accuracy Score: 91.23%
________________________________________________
Classification Report:
                0           1  accuracy   macro avg  weighted avg
precision    0.892857    0.935714  0.912338    0.914286      0.913590
recall       0.943396    0.879195  0.912338    0.911295      0.912338
f1-score     0.917431    0.906574  0.912338    0.912003      0.912179
support    159.000000  149.000000  0.912338  308.000000    308.000000
________________________________________________
Confusion Matrix:
 [[150   9]
 [ 18 131]]
```

b. Classification Report of K-Nearest Neighbor (KNN) Hyperparameter Tuning

```
Train Result:
================================================
Accuracy Score: 90.79%
________________________________________________
Classification Report:
                    0           1  accuracy   macro avg  weighted avg
precision    0.920245    0.897698   0.90795    0.908972      0.908390
recall       0.882353    0.931034   0.90795    0.906694      0.907950
f1-score     0.900901    0.914062   0.90795    0.907482      0.907821
support    340.000000  377.000000   0.90795  717.000000    717.000000
________________________________________________
Confusion Matrix:
 [[300  40]
 [ 26 351]]

Test Result:
================================================
Accuracy Score: 81.17%
________________________________________________
Classification Report:
                    0           1  accuracy   macro avg  weighted avg
precision    0.848276    0.779141  0.811688    0.813708      0.814831
recall       0.773585    0.852349  0.811688    0.812967      0.811688
f1-score     0.809211    0.814103  0.811688    0.811657      0.811577
support    159.000000  149.000000  0.811688  308.000000    308.000000
________________________________________________
Confusion Matrix:
 [[123  36]
 [ 22 127]]
```

c. Classification Report of Adaptive Boost (AdaBoost) Hyperparameter Tuning