Machine learning for predicting risk from heart failure

Yingxu Wang

***Abstract*—Creating a model to predict mortality from heart failure involves several steps, including data collection, preprocessing, feature selection, model training, and evaluation**

I. INTRODUCTION

Cardiovascular disease (CVD) is the leading cause of death worldwide, claiming an estimated 17.9 million lives annually and accounting for 31% of global deaths. Heart failure is a common event caused by CVD, and this dataset contains 12 features that can be used to predict mortality from heart failure. Most cardiovascular diseases can be prevented through population-wide strategies targeting behavioral risk factors such as smoking, unhealthy diet and obesity, physical inactivity, and harmful use of alcohol.

People with cardiovascular disease or an increased risk of cardiovascular disease (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidemia, or an established disease) require early detection and management, and machine learning models can provide beneficial help.[1]

II. DATA ANALYZE

This section discusses data analysis methods to examine the processed data set.

A. *Correlation Matrix of the dataset*

This heatmap visualizes the correlation coefficients between various clinical and demographic variables relevant to patient outcome studies

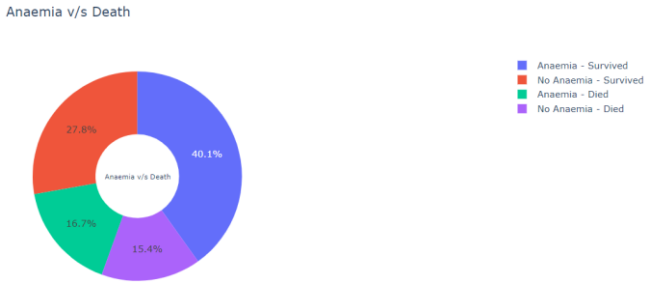
-There is a strongly negative correlation between 'DEATH_EVENT' and 'TIME', A negative association may mean that patients who survive longer after initial observation or treatment are less likely to experience the event in question.

-There is a positive correlation between "DEATH_EVENT" and "Serum Creatinine", which means that higher serum creatinine levels increase the risk of death.

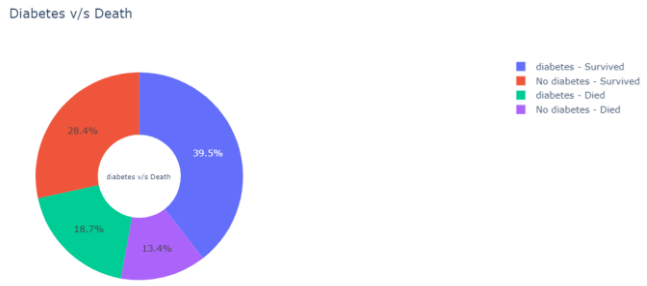
-There is a moderate positive correlation between "DEATH_EVENT" and "Age", which implies the older the patient is, the higher the risk of death may be.

B. *The relationship between the target and each feature*

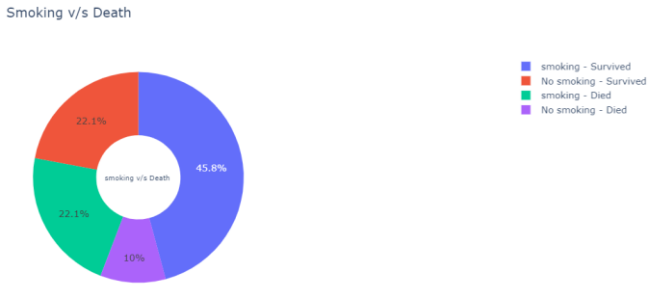
Each pie chart shows the percentage of patients with a specific health condition who survived and died.



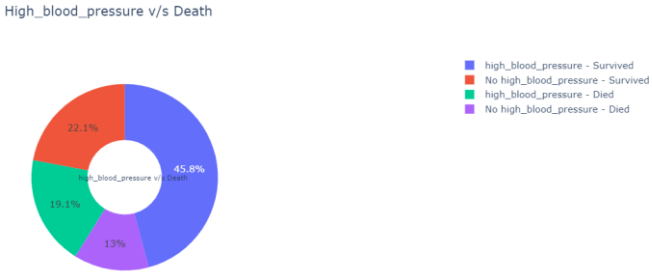
-Patients with anemia have a slightly higher mortality rate than those without anemia



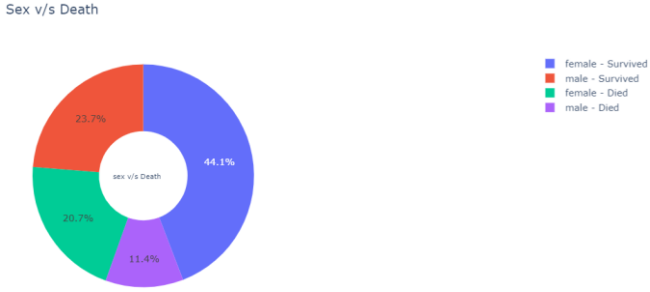
-Diabetes appears to be associated with higher mortality, but the difference is not as clear as anemia



-Smoking is associated with higher mortality rates, with nonsmokers having significantly lower rates of death



-Patients with high blood pressure have a higher mortality rate.



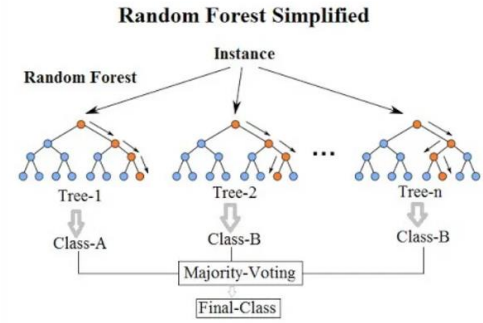
-Gender has little impact on mortality in data

III. MODELING CHOOSING

A. *Random Forest*

Random forests are an integrated learning method used primarily for classification, regression, and other tasks that work by constructing multiple decision trees and outputting patterns (e.g., classifications or average predictions) from all of the trees. This approach combines the predictive power of multiple decision trees to improve overall model accuracy and robustness

When constructing each tree, the Random Forest algorithm randomly selects samples and features from the original dataset. This "bootstrap sampling" and random feature selection provides uniqueness to each tree, which helps to reduce the variance of the model.

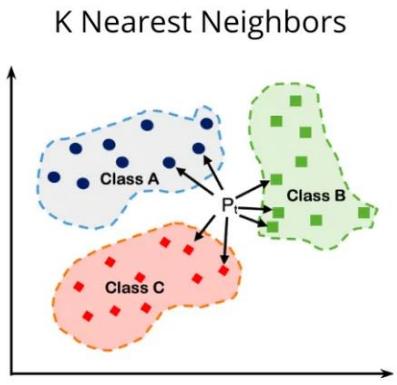


This model was chosen because it reduces the risk of overfitting by integrating multiple decision trees. It usually achieves high accuracy and maintains good generalization of the entire forest even if individual trees are overfitted on the dataset

B. *K-nearest neighbors*

KNN is a basic classification and regression method. [2]In a classification problem, the output is a category. The classification of an object is determined by the "votes" of its neighbors, and the category of an object is the most common category among its k nearest neighbors.

When given a query point, the KNN algorithm searches for the K nearest points in the dataset to the point and then predicts the attributes of the point based on the information of these nearest neighbors.



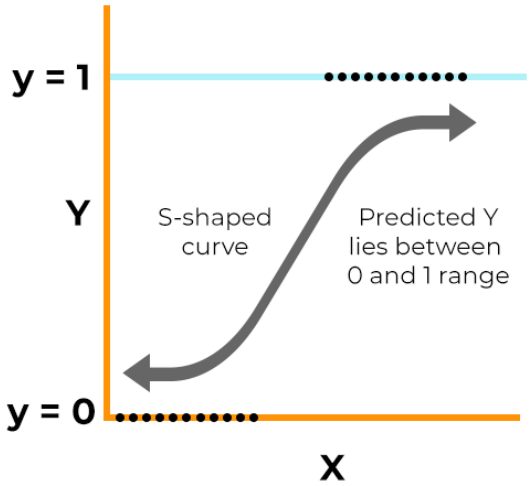
The reason for choosing this mod is that it is adaptable, can handle multi-category problems, and remains valid despite changes in data

distribution. It does not require an explicit training step, which makes it relatively easy to adapt the model to new data.

C. *Logistic Regression*

Logistic regression is a predictive analytics algorithm and statistical machine learning model for predicting the probability of a data point with two possible outcomes. It is a generalized linear regression analysis model for binary classification problems.

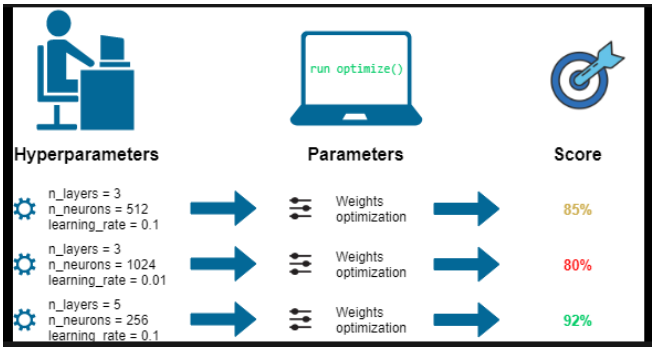
Logistic regression maps the feature space onto the binary output via a logistic function. It applies the output of linear regression to an S-shaped logistic function, producing a probabilistic output between 0 and 1.



This model was chosen because it provides probability scores for predictions, provides a quantitative basis for decision-making, and is a computationally efficient model.

D. *Hyperparameter Optimization MLP*

In machine learning, hyperparameters are parameters that are set before the learning process begins, and they control the learning process and have a significant impact on the model's performance. HPO is an automated process for selecting the optimal hyperparameters to maximize the model's performance



HPO improves model performance, especially for models like MLP, which are highly dependent on the choice of hyperparameters such as learning rate, number of layers, number of nodes per layer.

IV. DATA MODELING

A. *Splitting Datasets*

The training set accounts for 80% of the total data. This part provides a large enough sample to effectively train our model, covering various scenarios in material distribution.

The test set makes up the remaining 20%. This partitioning ensures that we have a sufficient amount of unseen data to test the model predictions and obtain reliable estimates of its performance in real-world situations.

B. *Model preprocessing*

time	DEATH_EVENT	anaemia_1	diabetes_1	high_blood_pressure_1	sex_1	smoking_1
4	1	False	False	True	True	False
6	1	False	False	False	True	False
7	1	False	False	False	True	True
7	1	True	False	False	True	False
8	1	True	True	False	False	False

One of the significant techniques that we implement is One-hot encoding. To avoid the issue known as the dummy variable trap, we employ the drop-first technique. This approach helps prevent duplication of dataset columns. We changed (sex), (smoking or not), and (diabetes or not) from the values 0 and 1 to True and False to enhance the readability of the code and the accuracy of the implementation of the verification algorithm.

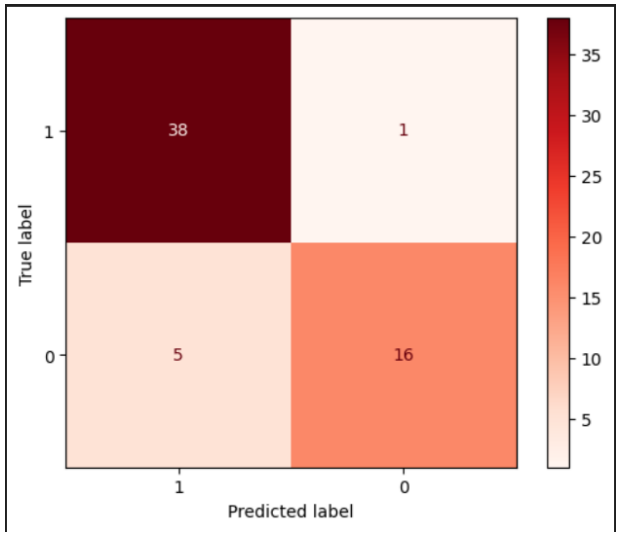
In this study, we are using four different models to process the dataset to evaluate the model's performance. The model with the best performance score will be applied to the final decision.

Here are the models and parameters:

-Random forest:

We split data to put aside 20% for testing purposes with a random state of 50.

Since there are 300 sets of data in our dataset, 300*20% = 60



True positive: 38 => 38/60 = 63.33%

True negative: 16 => 16/60 = 26.66%

False positive: 1 => 1/60= 1.666%

False negative: 5 => 5/60= 8.333%

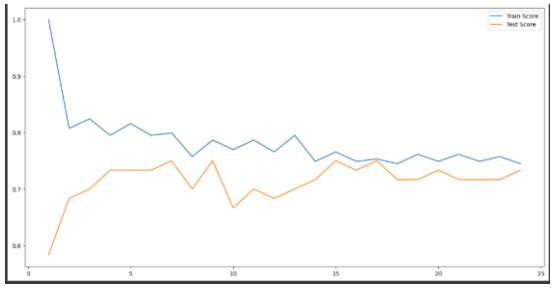
```
Accuracy: 0.9
Precision: 0.9124487004103967
Recall: 0.8681318681318682
F1 Score: 0.8844672657252888
```

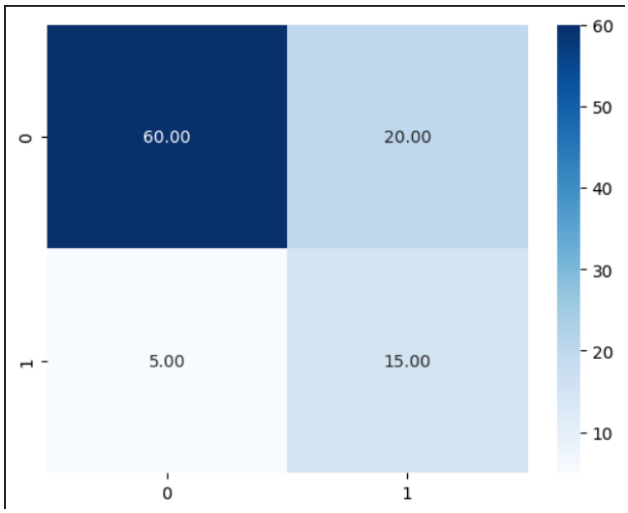
-KNN:

Initially, the KNN algorithm was run iteratively for K values from 1 to 25, and then the optimal K value was determined by finding where the test score was at its best.

```
Max Test Score: 75.0000%
k values for Max Test Score: [7, 9, 15, 17]
```

K-Nearest Neighbors Regression with the number of neighbors set to 7.

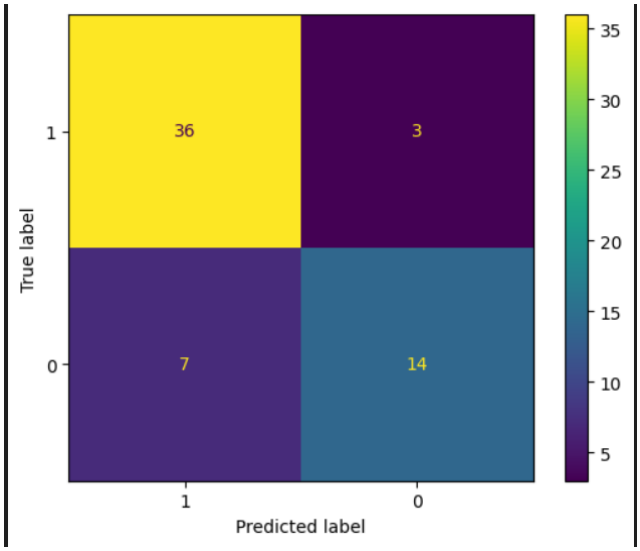





```
Accuracy Score:                 : 0.7500
KNN f1-score  : 0.5455
KNN precision : 0.4286
KNN recall    : 0.7500

              precision    recall  f1-score   support

           0       0.92      0.75      0.83        48
           1       0.43      0.75      0.55        12

    accuracy                           0.75        60
   macro avg       0.68      0.75      0.69        60
weighted avg       0.82      0.75      0.77        60
```

-Logistic regression[3]:

Since there are 300 sets of data in our dataset, 300*20% = 60



True positive: 36 => 36/60 = 60%

True negative: 14 => 14/60 = 23.333%

False positive: 3 => 3/60 = 5%

False negative: 7 => 7/60 = 11.6666%

```
Accuracy: 0.8333333333333334
Precision: 0.8303693570451436
Recall: 0.7948717948717949
F1 Score: 0.8074454428754814
```

-Hyperparameter Optimization MLP

In this section, we trained an MLP model with three hidden layers, with the first layer comprising 32 neurons, the second layer 64 neurons, and the third layer 32 neurons.

We fit the test value of the feature value of the first MLP model into our next MLP model to meet this algorithm:[4]

Y = W2(W1*X + B1) + B2

We use Hyperparameter to optimize MLP

```
MLP(first) MSE=0.20669437095797577, MAE=0.32714965728685574
MLP(second) MSE=0.17843756622367682, MAE=0.29940266202207283
```

```
{'activation': 'tanh', 'hidden_layer_sizes': (150,), 'learning_rate_init': 0.0001}
Final Model MSE=0.13266527500296676, MAE=0.29812259894364146
```

V. CONCLUSION

In conclusion, we decide to choose three different types of models to classify our dataset. Each uses random forest, KNN, and logistic regression, respectively. We found that, from the perspective of data accuracy, the random forest classification has the highest accuracy rate. MLP has been used to implement our predictive analytics. The process of selecting the best combination of hyperparameters significantly optimizes our result.

VI. REFERENCE

[1]Larxel, “Heart failure prediction,” Kaggle, https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data (accessed Apr. 7, 2024).

[2] K. Shi, L. Li, H. Liu, J. He, N. Zhang and W. Song, "An improved KNN text classification algorithm based on density," 2011 IEEE International Conference on Cloud Computing and Intelligence Systems, Beijing, China, 2011, pp. 113-117, doi: 10.1109/CCIS.2011.6045043. keywords: {Classification algorithms;Text categorization;Training;Mathematical model;Equations;Algorithm design and analysis;Support vector machine classification;Text classification;VSM;KNN;decision function},

[3] K. He and C. He, "Housing Price Analysis Using Linear Regression and Logistic Regression: A Comprehensive Explanation Using Melbourne Real Estate Data," 2021 IEEE International Conference on Computing (ICOCO), Kuala Lumpur, Malaysia, 2021, pp. 241-246, doi: 10.1109/ICOCO53166.2021.9673533. keywords: {Training;Machine learning algorithms;Computational modeling;Linear regression;Machine learning;Predictive models;Prediction algorithms;Data Science;Housing price model;Linear regression;Logistic regression},

[4]Chai SS, Cheah WL, Goh KL, Chang YHR, Sim KY, Chin KO. A Multilayer Perceptron Neural Network Model to Classify Hypertension in Adolescents Using Anthropometric Measurements: A Cross-Sectional Study in Sarawak, Malaysia. Comput Math Methods Med. 2021 Dec 7;2021:2794888. doi: 10.1155/2021/2794888. PMID: 34917164; PMCID: PMC8670914.