A Prediction of Heart Failure Based on Variable Selection Approach in Machine Learning Model

Tanatchaya Tueyot[1]*, Jiraporn Janthiwonk[1], Panomkorn Boonjen[1], Sirilak Srichaichana[1], Sirikorn Sunthep[1] and Chalermrat Nontapa

[1]Data Science Research Center, Faculty of Science, Chiang Mai University, Chiang Mai, 50200, Thailand
[2]Department of Statistics, Faculty of Science, Chiang Mai University, Chiang Mai, 50200, Thailand
*Corresponding author email: tanatchaya_tueyot@cmu.ac.th

ABSTRACT

This research focuses on the study of heart failure, a condition of great significance due to its high prevalence worldwide and its role as a major cause of mortality in Thailand, with an increasing trend each year. The study aims to examine factors contributing to heart failure by analyzing three models including XGBoost, CatBoost and LightGBM. The goal is to compare the performance of these models and identify the most suitable one. This comparison will help assess the models' effectiveness in predicting heart failure. The findings will provide valuable information for improving health care plans and reducing risks in high-risk populations. The results revealed that the CatBoost model was the most appropriate, with an accuracy of 88.59%.

INTRODUCTION

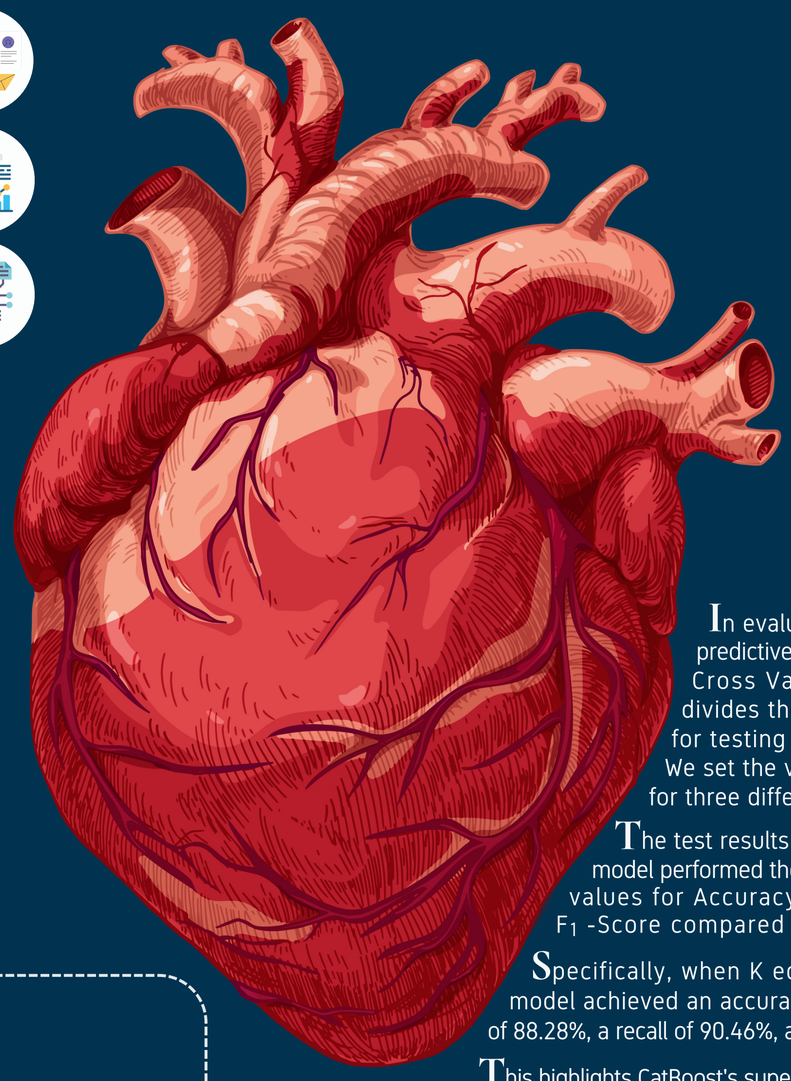
When discussing heart failure, people in many countries around the world, including Thailand, are currently facing this condition, which is a serious and potentially life-threatening disease. In this study, researchers examined the factors contributing to heart failure occurrence using models such as XGBoost, CatBoost and LightGBM to compare and identify the most suitable and accurate model for predicting heart failure. These models play a crucial role in improving and enhancing efficiency by utilizing various related data and factors to help identify trends in heart failure occurrence, ultimately increasing the chances of providing timely and effective heart health care according to established standards.

OBJECTIVES

1. To study the relationships among various factors that tend to lead to heart failure.
2. To compare and build three prediction models to predict heart failure based on various factors, and ultimately select the most suitable model for accurate prediction.

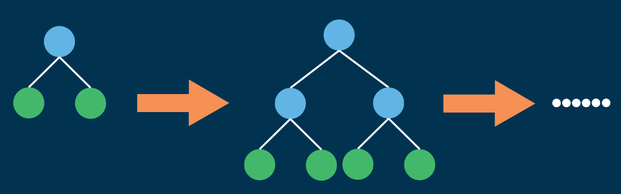
METHODOLOGY

1. Selecting a Dataset
2. Data Preprocessing
3. Data Transformation
4. Feature Selection
5. Data Splitting into Train and Test Sets
6. Model Evaluation



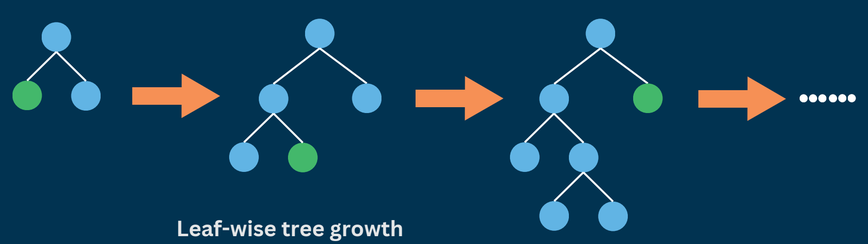
MODELS

XGBoost



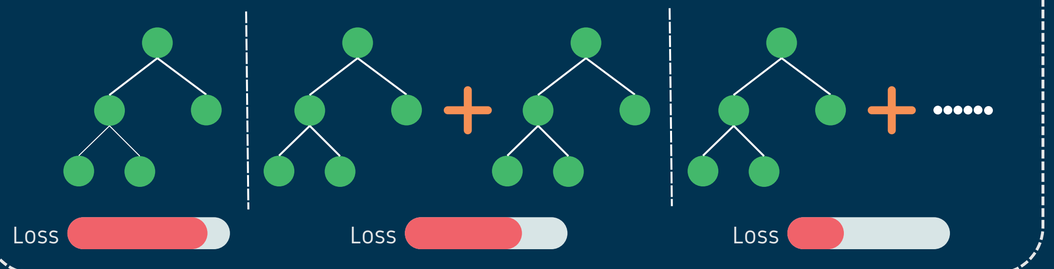
Level-wise tree growth

LightGBM



Leaf-wise tree growth

CatBoost



FACTORS

From 12 variables, with heart failure as the dependent variable, it was found that after selecting the important variables, 8 independent variables significantly affect the occurrence of heart failure.

RESULT

In evaluating the performance of predictive models, we use the K-Fold Cross Validation method, which divides the data into different sets for testing and training the models. We set the value of K to 6, 8, and 10 for three different models.

The test results showed that the CatBoost model performed the best, as it had the highest values for Accuracy, Precision, Recall, and F_1 -Score compared to the other models.

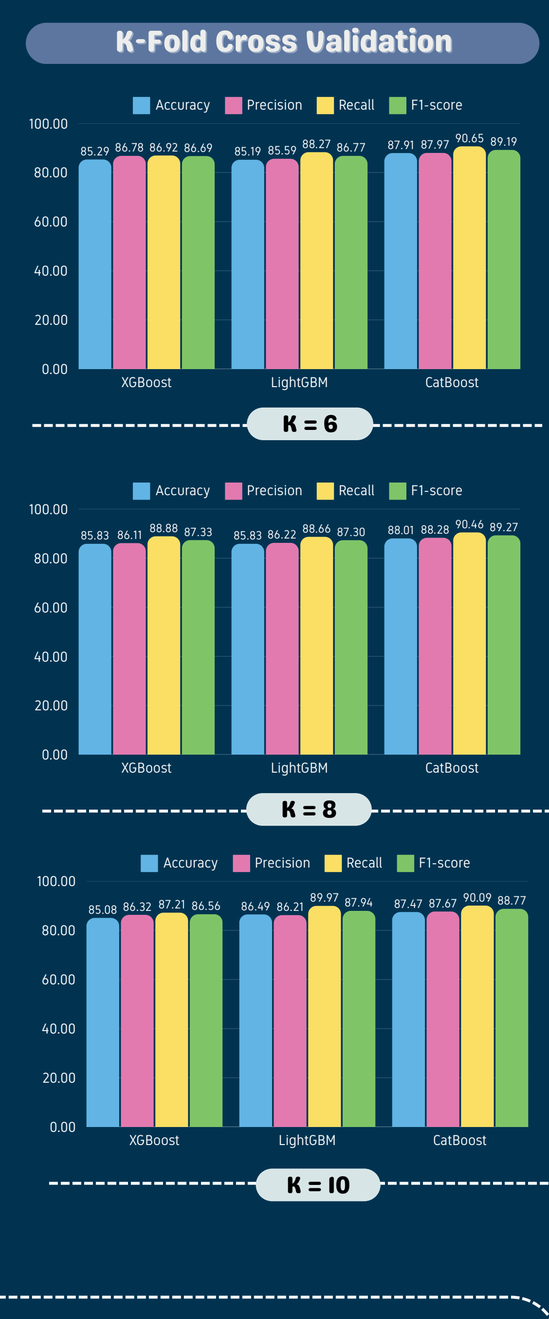
Specifically, when K equal to 8, the CatBoost model achieved an accuracy of 88.01%, a precision of 88.28%, a recall of 90.46%, and an F_1 -score of 89.27%.

This highlights CatBoost's superior performance across key evaluation metrics, further confirming it as the most suitable model for consistently providing accurate and dependable predictions in this study.

EDA

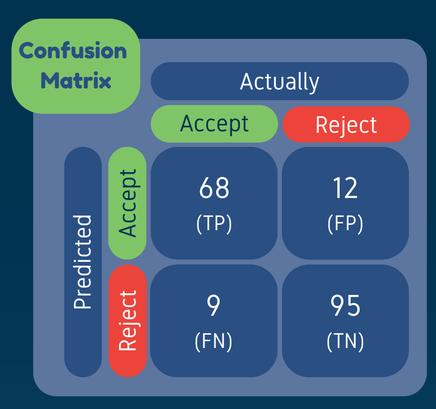
A QR code displays a graph for each variable, showing its impact on heart failure occurrence.

K-Fold Cross Validation



CONCLUSION & DISCUSSION

The results of the confusion matrix showed that CatBoost correctly predicted heart failure in 88.59% of cases, which amounts to 813 out of 918 people, while it incorrectly predicted heart failure in 11.41% of cases, which amounts to 105 people.

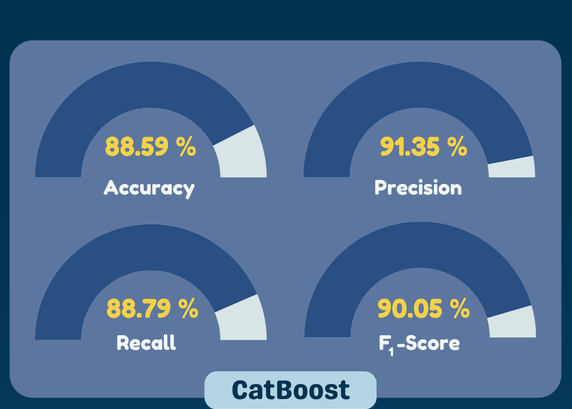


$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$F_1\text{-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$



From the study, it was found that the CatBoost model was the most suitable for predicting the occurrence of heart failure. This model achieved performance metrics with an Accuracy of 88.59%, Precision of 91.35%, Recall of 88.79%, and an F_1-Score of 90.05%. These findings align with the research of (Badik and Akar, 2024; Le Li, et al., 2024), which also demonstrated that the CatBoost model outperformed other models, showing similarly high predictive performance metrics, consistent with this study.

References

[1] J. Patel, D. TejalUpadhyay, and S. Patel, Heart disease prediction using machine learning and data mining technique, Heart Disease, vol. 7(1), pp. 129–137, 2015.
[2] Badik, Ş. T., & Akar, M. (2024). Machine learning classification models for the patients who have heart failure. Sigma Journal of Engineering and Natural Sciences, 42(1), 235-244.
[3] Li, L., Zhang, Z., Xiong, Y., Hu, Z., Liu, S., Tu, B., & Yao, Y. (2022). Prediction of hospital mortality in mechanically ventilated patients with congestive heart failure using machine learning approaches. International Journal of Cardiology, 358, 59-64.