Indicators of Heart Disease and Prediction: A Machine Learning-Based Approach

Arvin Castelo
Master of Data Analytics
Western University
London, Canada
ccastelo@uwo.ca

Xinyi Li
Master of Data Analytics
Western University
London, Canada
xli3322@uwo.ca

Yue Wan
Master of Data Analytics
Western University
London, Canada
ywan5974@uwo.ca

Miles Xi
Master of Data Analytics
Western University
London, Canada
zxi27@uwo.ca

Ren Yu
Master of Data Analytics
Western University
London, Canada
ryu86@uwo.ca

***Abstract*—Heart disease is the leading cause of death in the USA, affecting all demographics and imposing a significant economic burden. Machine Learning (ML) is increasingly applied in healthcare, addressing the complex, multifaceted nature of disease diagnosis. In this project, we aim to seek the potential of ML to estimate heart attack risk using CDC's 2022 health survey data. Specifically, we intend to identify key risk factors that predict potential heart attack. We analyze a dataset of 445,132 records with 40 features, starting with data visualization and data preprocessing and addressing missing values with imputation. We use a range of ML models including Logistic Regression, Elastic Net, Random Forest, XGBoost. For model evaluation, we prioritize F1 score, to balance false negatives and false positives in diagnosis. Our results achieve a high F1 score, accuracy and recall, effectively identifying individuals at higher risk of heart attack based on their health indicators and lifestyle factors.**

***Index Terms*—machine learning algorithms, logistic regression, random forest, XGBoost**

I. Background

According to the CDC, heart disease is the leading cause of death in the USA, and is affecting men, women, adults, and even young adults. Further, a journal [1] estimated that the expenses related to cardiovascular disease and stroke will increase to 4.6 % of USA's GDP by 2050 from 2.7% in 2020. These numbers show the significant impact of the disease to the country.

We realized that accurate prediction of heart disease helps identify high-risk individuals before severe health consequences arise. Early prediction of heart attacks can greatly enhance preventive care and save lives. We are convinced that the use of machine learning (ML) techniques can be a promising approach for this prediction.

According to Machine-Learning-Based Disease Diagnosis [2], machine learning is popular in the field of healthcare. Since medical issues can be complicated and intractable in many cases, how machine learning methods are applied to make diagnosis is our concern.

In this project, we aim to apply various ML models to identify influential indicators of heart attacks, including factors like BMI, age range, and health history. By applying these models, we seek to improve the accuracy of heart attack prediction, ultimately assisting in better clinical decision-making and patient outcomes.

Previous research has been made specifically on predicting heart disease using ML models. Some common methods include Decision Tree, Random Forest, K-Nearest Neighbor, Naive Bayes, Support Vector Machine (SVM), and Artificial Neural Network (ANN).

According to Suleiman et al. [3], Random Forest is used on training data to help in the prediction of heart disease. This algorithm can be beneficial in handling complicated dataset and minimizing overfitting.

Ghosh et al. [4] experimented on Heart Disease Prediction using ML methods with Elastic Net feature selection. This method can handle correlations among the predictors better than Lasso and the researchers aimed to investigate all the risks that can lead to a heart disease.

According to Yang et al. [5], the method of Logistic Regression is useful and worthy in the field of heart disease prediction. Researchers use this method to determine the significant characteristics and incidence probability of heart disease.

Hajiarbabi et al [6]. worked on the XGBoost, Random Forest, Ensemble Learning, and Neural Network methods, and performing better than classic machine learning models. Researchers make comparisons and propose to identify the most efficient methods for heart disease detection.

Rasheed et al [7]. examined the utility of GridSearchCV and concluded that this method works in enhancing model accuracy in the field of heart disease prediction. This article also reveals that Random Forest is the preferred algorithm for predicting heart disease, demonstrating its potential for practical use in healthcare settings.

II. Visualization and Exploratory Data Analysis

Exploratory data analysis allows us to have an understanding of the main characteristics of the data. We have done this by plotting a graph of correlations, a correlogram, between numeric variables. In addition, a histogram is plotted for every feature and the response variable, allowing us to gain an intuitive understanding of the distribution of each variable. Results are shown in Figure 8 of the Appendix.

III. Data Preprocessing

A. Data cleaning

Observations with missing HadHeartAttack (the response variable) values were removed (3,065 observations). The dataset size decreased from 445,132 to 442,067.

The State variable was dropped during preprocessing to address the high cardinality issue. Including it would have resulted in numerous dummy variables, significantly increasing computational cost and the risk of overfitting due to the curse of dimensionality. Dropping it also simplified the model, making it more interpretable without sacrificing performance.

B. Data splitting

After the data cleaning step, the dataset had 442,067 observations and 40 variables (including the response), but it still contained 882,346 missing values. However, directly imputing the entire dataset is not advisable since the missing values in the would-be validation and test sets will be influenced by the training set, causing data leakage.

To avoid this, the dataset was first split into three sets. Since the entire clean dataset is still sufficiently big, the group decided to split the dataset into 80% training (353,653 observations), 10% validation (44,207), and 10% test (44,207) sets. The overall proportion of the response variable is maintained across these sets (around 5.7%).

C. Data Imputation

To handle the missing data, the training, validation, and test sets were independently imputed using the mice (Multivariate Imputation by Chained Equations) package in R. This is a widely used tool for performing multiple imputations of missing data. We created single imputed datasets to simplify the workflow, reduce the resource allocation requirement, and reduce computation time. As a limitation, considering the process' complexity and the dataset size, the group decided to use only one (the first) imputed dataset.

The mice package offers an intuitive and flexible framework for imputing missing values using various methods, including regression, predictive mean matching, and others. It uses the Fully Conditional Specification (FCS) algorithm, where missing values are imputed based on the relationships between variables in the dataset.

IV. Analysis and Modeling

The first model we used was Logistic Regression, which showed an accuracy of 0.9381 and an F1 score of 0.4876. To achieve a better performance and reduce overfitting, Penalized Logistic Regression was also tried.

Elastic net combines the penalties of both Lasso and Ridge, together with the stochastic gradient descent algorithm that reduces computational cost. We employed Elastic Net's penalty and tuned it using the RandomizedSearchCV in scikit-learn. Since the best 'l1-ratio' returned by the RandomSearchCV function was 0, the Ridge penalty was sufficient. Thus, we applied a Grid Search with L2 penalty term to get deterministic results. Moreover, we adjusted the decision threshold for the fine-tuned Ridge Regression model. The final result showed an accuracy of 0.9386 and F1 score of 0.4895.

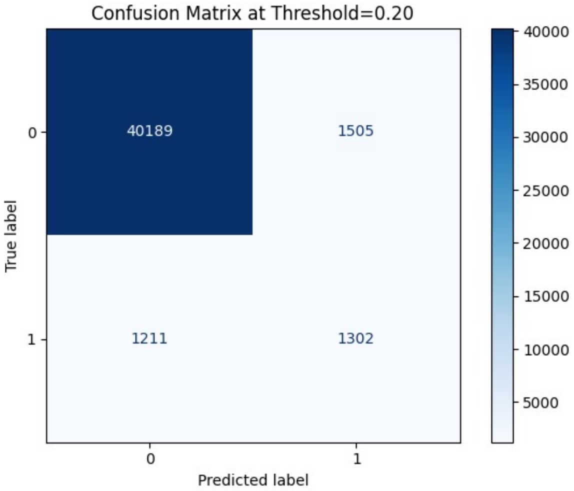

Fig. 1: Logistic Regression Confusion Matrix.

As shown in Figure 2, the important predictors for having a heart attack were if the person has had angina, age, gender, and if the person had a recent chest scan. All of which are categorical or dummy variables.

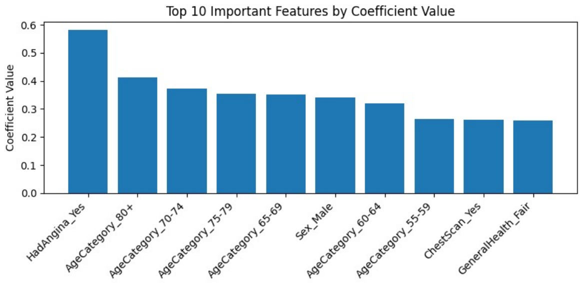

Fig. 2: Logistic Regression Features Coefficients.

Our second model is a Random Forest, an ensemble learning method that uses Bagging, where decision trees are fitted to bootstrapped samples and their class predictions are aggregated. Trees tend to overfit, so random forest selects a random subset of the features at each split during training. We tune the model's parameters, including the number of trees and the number of features in a subset, according to the findings of Probst et al. [8].This model achieved an accuracy of 0.9416 and an F1 of 0.4869 on the validation set.

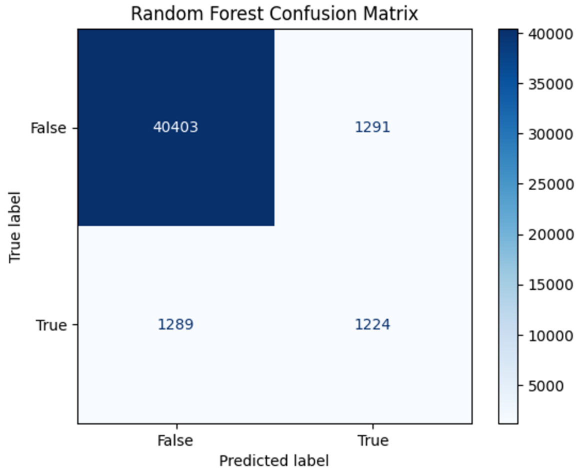

Fig. 3: Random Forest Confusion Matrix.

In addition, we inspected what features were important by checking their contribution to decrease in impurity [9].As is shown in Figure 4, whether a person had angina, their BMI, and weight were top three important predictors of heart disease.

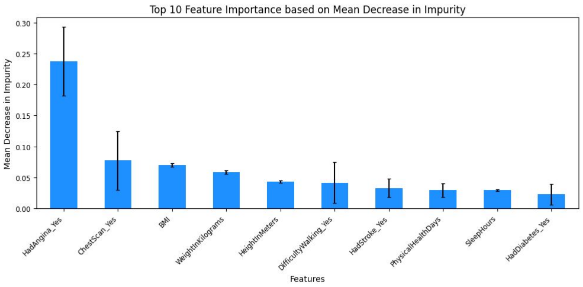

Fig. 4: Random Forest.

We had also applied models in the Boosting family, because their iterative learning process helps to capture complex patterns within lots of features. The dataset provides a rich amount of information for XGBoost to extract the interactions within the data. While XGBoost also provides class weighting and regularization that makes it an excellent and robustic model for handling imbalance data, as it could remain sensitive to the minority class. CatBoost, also has an outstanding performance on datasets with more categorical features. AdaBoost, as a member of the Boosting family, was also fitted for comparison.

The imbalance between the amounts of positive and negative observations in the dataset made it essential to adjust the weight on the positive samples; which was true since the parameter 'scale _pos_weight' had a big impact on the f1 score. After tuning other parameters (please see the appendix for graph used during tuning) we obtained the final model.

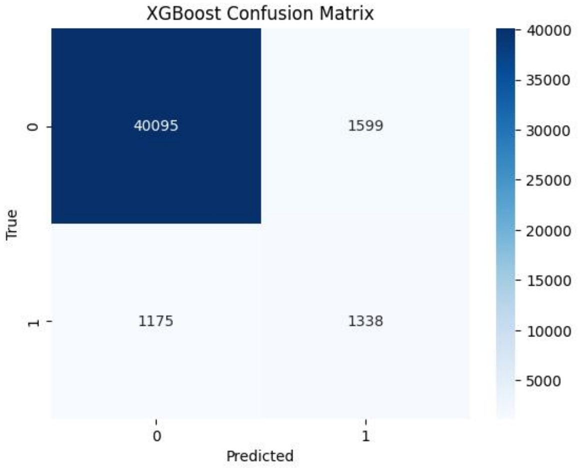

Fig. 5: XGBoost Confusion Matrix.

The top 10 important features of XGBoost suggested that an individual's angina history was most important for predicting the heart attack, followed by Chest scan, Difficult walking, and Stroke history. The top 10 important features of the CatBoost were different from the above; it showed that the age group was the most important feature. Apart from the angina history and chest scan, an individual's general health condition, BMI, and Smoke history also contributed a lot.

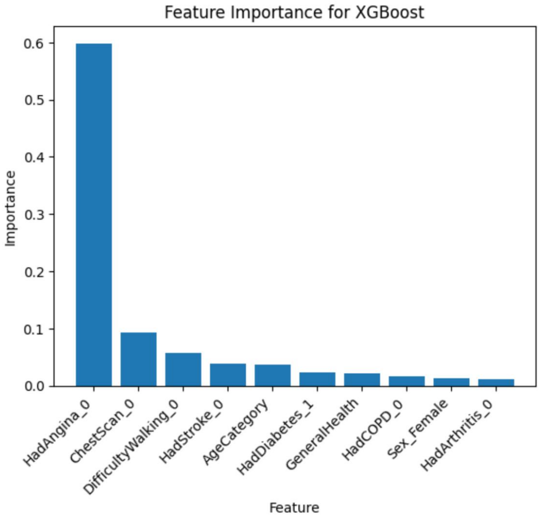

Fig. 6: XGBoost Features Coefficients.

In summary, all of our models determined that whether a person had an angina or not is the most important factor for identifying the occurrence of heart disease, while the importance of features such as Chest scan and Difficulty walking was confirmed by some of the models.

Finally, we compare the performances of these models on our validation set, focusing on metrics including Recall, Precision, and F1 score, in order to reduce false negatives in diagnosis.

TABLE I: Hyperparameters

Model	Hyperparameters
Logistic Regression	C = 0.01, penalty = ‘l2’,estimator=pipe_logistic
Random Forest	n_estimators = 100, max_depth = 50, max_features=50, min_samples_leaf = 6, max_samples = 0.8, class_weight = ’balanced’
XGBoost	scale_pos_weight = 4, max_depth = 4, eta = 0.068, min_child_weight = 235
CatBoost	learning_rate = 0.05, depth = 4, scale_pos_weight = 3.15
AdaBoost	n_estimators = 100, learning_rate = 0.9

TABLE II: Model Validation Performance

Model	Recall	Precision	Accuracy	F1
Logistic Regression	0.5185	0.4602	0.9381	0.4876
Ridge Regression	0.5181	0.4638	0.9386	0.4895
Random Forest	0.4871	0.4867	0.9416	0.4869
XGBoost	0.5324	0.4556	0.9372	0.4910
CatBoost	0.5117	0.4702	0.9395	0.4901
AdaBoost	0.2786	0.5673	0.9469	0.3736

As summarized in Table II, XGBoost is the best performing model with F1 score of 0.4910. It has the highest recall rate at 0.5324, but the lowest precision rate of 0.4556.

The best model was then refitted using the combination of training set and validation set, as the final model. At the last step, we test the model using the test set, and get the result F1 score of 0.4857, which is close to the score on the validation set before refitting. As a result, the model is not overfitting on the validation set, and it generalizes the features well.

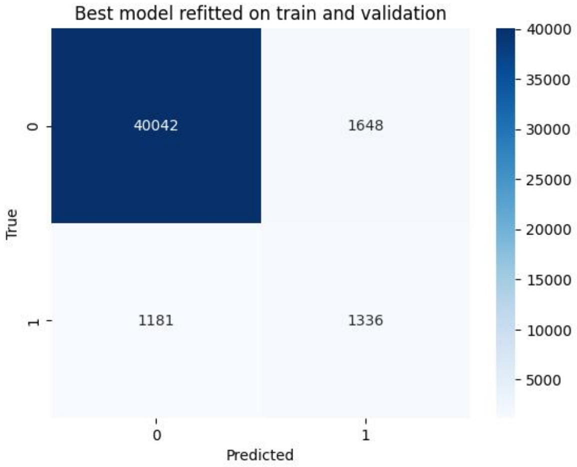

Fig. 7: Refitted XGBoost Result.

TABLE III: Refitted XGBoost Performance

Model	Recall	Precision	Accuracy	F1
Refitted XGBoost	0.5308	0.4477	0.9360	0.4857

V. Future Works

A. *CatBoost*

Catboost shows a great potential in exploring the important categorical features during modeling, although it does not have the best score. This could be due to the random fluctuation in the dataset; with more data, it might be possible for the CatBoost to give an outstanding result.

B. Application

In the future, more improvement can be made by analyzing more medical data such as lab results and patient history. The developed final model can be integrated into clinical decision support systems for real-time heart attack risk assessment. By incorporating real-time patient data, the model could assist healthcare professionals in making timely and informed decisions.

VI. Conclusion

This study investigates the potential of machine learning (ML) to predict heart disease using the CDC's 2022 health survey data, which contains 445,132 records and 40 features. Heart disease remains the leading cause of death in the U.S., with significant economic and social impacts. By leveraging ML techniques, the project aims to identify critical predictors of heart attacks, such as BMI, age, and health history, improving diagnostic accuracy and preventive care.

After data cleaning, the dataset was reduced to 442,067 observations, split into training, validation, and test sets, and imputed using the mice package to handle missing values while avoiding data leakage. Exploratory analysis included visualizing correlations and distributions before and after imputation.

The study tested the following models: Logistic Regression, Elastic Net, Random Forest, and XGBoost. Logistic Regression and Elastic Net provided interpretability, while the complex models Random Forest and XGBoost models highlighted key predictors through feature importance. Hyperparameter tuning was conducted using GridSearchCV and RandomizedSearchCV.

Evaluation prioritized metrics like F1 score, recall, and precision to minimize diagnostic errors. Validation results showed that XGBoost model had the best potential to accurately predict heart disease risk, identifying key factors such as angina, chest scan, difficulty in walking, and stroke.

The final model, XGBoost, was refitted using the entire training and validation datasets. The resulting model was evaluated using the test data, which showed an F1 score of 0.4857, close to the validation F1 score of the model prior to refitting. This indicates that the model generalizes well on new data.

References

[1] CDC, "Heart Disease Facts," Heart Disease, Apr. 29, 2024. https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html

[2] M. M. Ahsan, S. A. Luna, and Z. Siddique, "Machine-Learning-Based Disease Diagnosis: A Comprehensive Review," Healthcare, vol. 10, no. 3, p. 541, Mar. 2022, doi: 10.3390/healthcare10030541.

[3] A. B. Suleiman, S. Luka, and M. Ibrahim, "CARDIOVASCULAR DISEASE PREDICTION USING RANDOM FOREST MACHINE LEARNING ALGORITHM," FUDMA Journal of Sciences, vol. 7, no. 6, pp. 282–289, Dec. 2023, doi: 10.33003/fjs-2023-0706-2128.

[4] S. Ghosh and M. A. Islam, "Performance Evaluation and Comparison of Heart Disease Prediction Using Machine Learning Methods with Elastic Net Feature Selection," American Journal of Applied Mathematics and Statistics, vol. 11, no. 2, pp. 35–49, Apr. 2023, doi: 10.12691/ajams-11-2-1.

[5] Y. Zhang, L. Diao, and L. Ma, "Logistic Regression Models in Predicting Heart Disease," Journal of Physics Conference Series, vol. 1769, no. 1, p. 012024, Jan. 2021, doi: 10.1088/1742-6596/1769/1/012024.

[6] M. Hajiarbabi, "Heart disease detection using machine learning methods: a comprehensive narrative review," Journal of Medical Artificial Intelligence, vol. 7, p. 21, Jun. 2024, doi: 10.21037/jmai-23-152.

[7] S. Rasheed, G. K. Kumar, D. M. Rani, M. V. V. P. Kantipudi, and A. M, "Heart Disease Prediction Using GridSearchCV and Random Forest," EAI Endorsed Transactions on Pervasive Health and Technology, vol. 10, Mar. 2024, doi: 10.4108/eet-pht.10.5523.

[8] P. Probst, M. N. Wright, and A. Boulesteix, "Hyperparameters and tuning strategies for random forest," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 9, no. 3, Jan. 2019, doi: https://doi.org/10.1002/widm.1301.

[9] "4.2. Permutation feature importance," scikit-learn, 2024. https://scikit-learn.org/1.5/modules/permutation_importance.html#permutation-importance (accessed Dec. 06, 2024).

Appendix I

Fig. 8: Distribution of Variables.

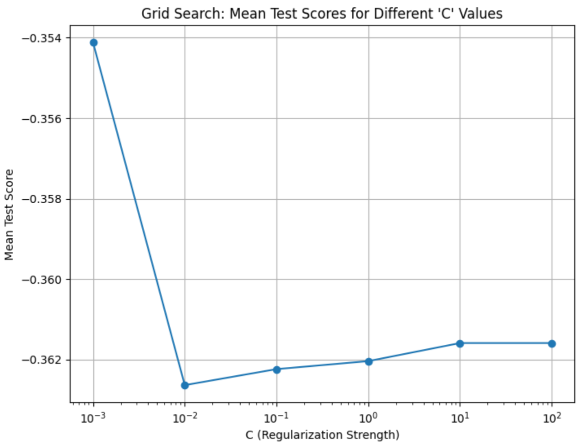

Fig. 9: Logistic Regression Grid Search Result.

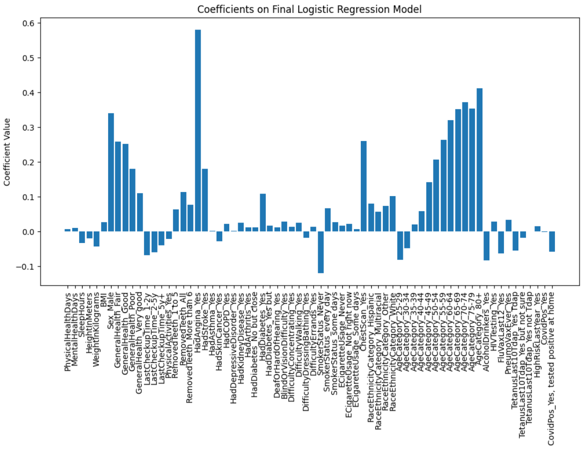

Fig. 10: F1 Score vs Scale_Pos_Weight.

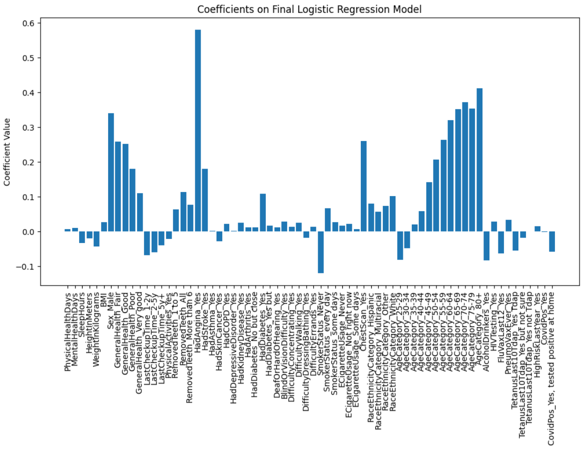

Fig. 11: F1 Score vs Scale_Pos_Weight.

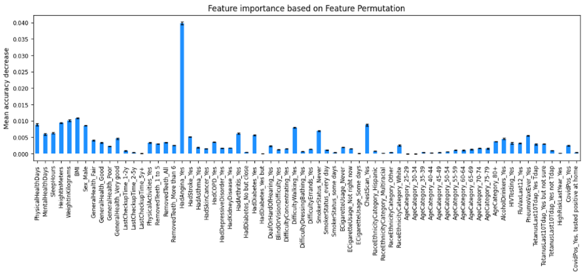

Fig. 12: Random Forest.

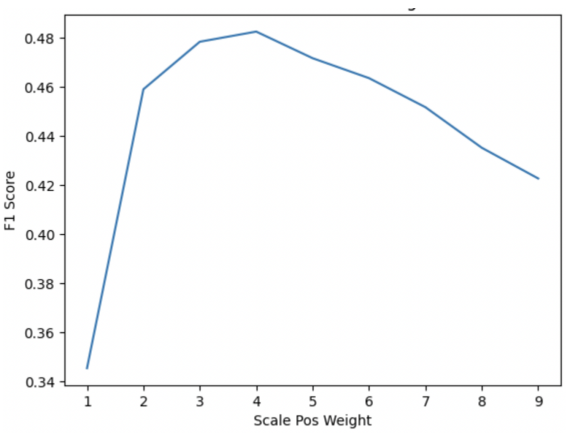

Fig. 13: F1 Score vs Scale_Pos_Weight.

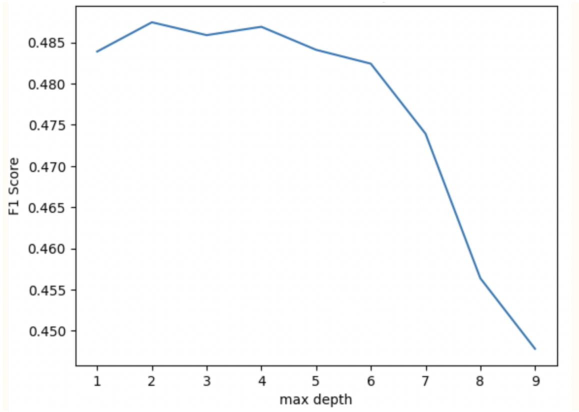

Fig. 14: F1 Score vs Max_Depth.

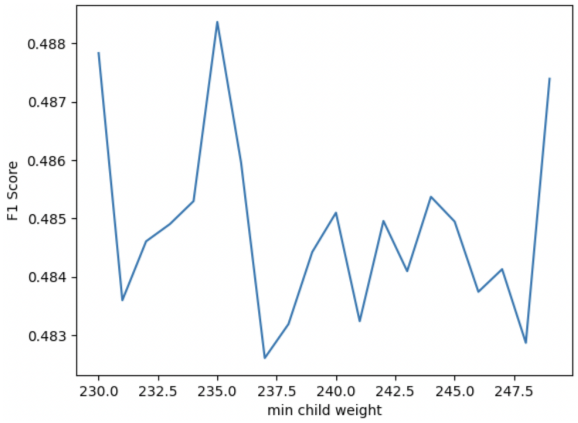

Fig. 15: F1 Score vs Min_Child_Weight.

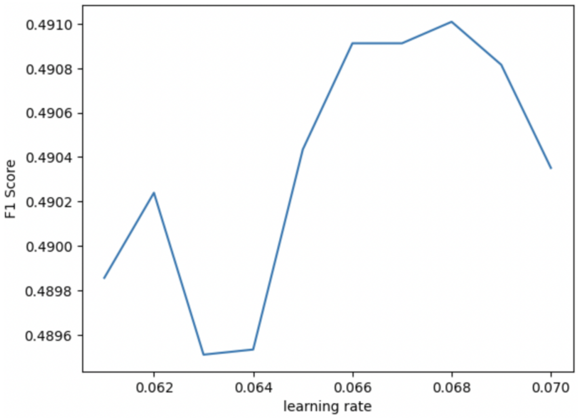

Fig. 16: F1 Score vs Min_Child_Weight.

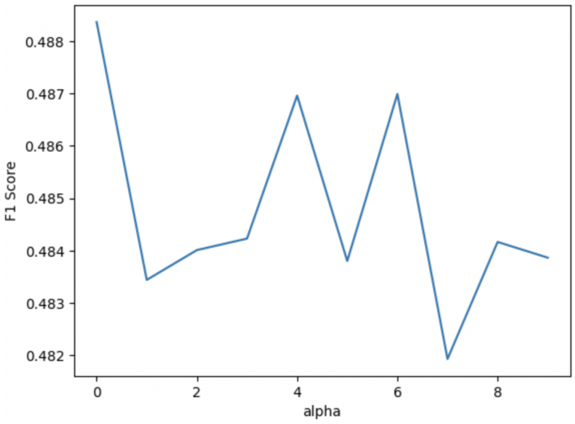

Fig. 17: F1 Score vs Alpha.

Appendix II

Memeber	
Arvin Castelo	Data preprocessing and imputation, Elastic Net implementation
Xinyi Li	Data preprocessing and imputation, Dataset search, Formatting files
Yue Wan	Logistic Regression and Elastic Net implementation
Miles Xi	Training random forest, data preprocessing, data visualization
Ren Yu	Implementing Boosting methods, Researching regarding to medical and imbalance dataset