Nick Patrick, Dillon Murphy, Zach Weinfeld

CSC 369 Final Project

March 13 2024

Report

For our project, we found a dataset on Kaggle that contained data on over 300,000 subjects and whether or not they were diagnosed with heart disease. Our goal was to implement a classification algorithm that could classify whether or not a new patient has heart disease based on predictor variables. Our first attempt was to utilize logistic regression. This approach did not do very well, only achieving a macro F1 score of just over 0.5. Next, we implemented a K-Nearest-Neighbors algorithm. This model computes the Euclidean distance from each point, and uses the closest neighbors to classify the data. At first, our KNN model only performed marginally better than logistic regression, achieving a macro F1 score around 0.58. Next, we tried tuning the two hyperparameters: K (the number of neighbors) and N (the number of features used). First, we found the optimal K to be 14, meaning that the model only uses the 14 nearest neighbors to classify the new row. Next, we found that the optimal amount of features was 13. This means that the model was optimal when we removed the 10 of the 23 features with the lowest correlation with the target variable. With these parameters, we achieved a maximum macro F1 score of 0.71. It is important to note that this was achieved using a subset of the data of only 10,000 rows (it would have taken an unreasonable amount of time to tune these hyperparameters on the full dataset). When we performed KNN on our entire dataset using the optimal hyperparameters, the macro F1 score decreased to 0.5654. Overall our model performed mediocrely on the entire dataset, even after tuning the hyperparameters on a subset of the data. We believe the main reason for this is related to the quality of the data.