Project 4 - kNN - Health Data

Matt Burns, Rebecca Lamoreaux

October 7th, 2024

Github Repo: https://github.com/m6urns/project4-knn-cs5830/tree/fix
Presentation:
https://docs.google.com/presentation/d/1GHoWi-Bbv0ZF7ZPxAs21K-r-QTU8pOWgTsqL7XIBsbs/edit?usp=sharing

Part 1

Introduction

Health insights extracted from historical health datasets are extremely relevant in the current healthcare system of the United States, given its limited resources and budgets. Future outlooks on healthcare indicate that the extraction of insights from healthcare data will only continue to grow in relevancy, with a hopeful goal of being able to extract per-patient insights. In this report, we intend to examine a dataset collecting information on heart attacks from patients in the United States healthcare system.

To make predictions regarding the occurrence of heart attacks, we will utilize a machine learning method called K Nearest Neighbor (KNN), in conjunction with methods for selecting optimal model parameters and disease factors for our model. We found that the KNN model was effective for classification in the case of heart attack data.

Methods

We utilized a number of methods in our analysis of the Cleveland heart attack dataset, these methods can be separated into three primary groups which combined contributed to the design of the final K Nearest Neighbor model. These three groups are as follows:

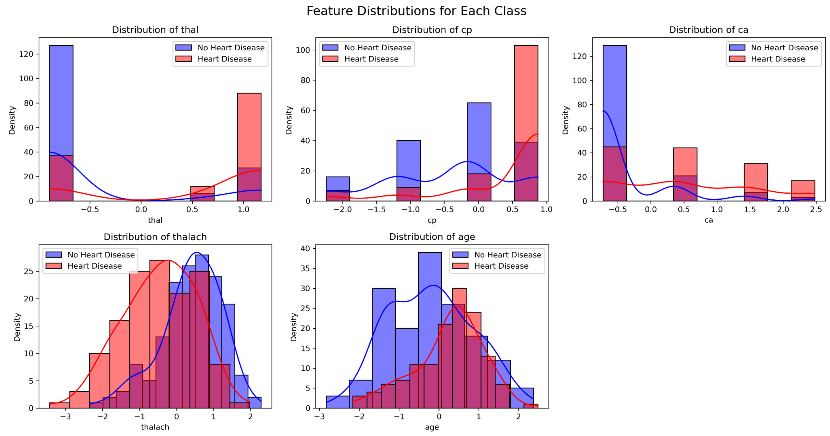
- Feature Selection
    - We utilized two methods to perform feature selection for our model implementation. Feature selection is the analysis and selection of significant features in a dataset. Feature selection is needed because not all features in a dataset are valuable for predicting an outcome. We first performed feature selection by performing a feature importance analysis, this provided us a list of features in the dataset, and an estimate of how important each feature is to predicting whether an individual has had a heart attack. After experimenting with this feature selection method we determined that we would need to perform further manual selection through examining feature distributions for linear separability. This is a useful analysis and helped us to eliminate features with poor separability, and allow us to reduce the dimensionality of our model.

- K parameter selection
    - In the K Nearest Neighbor model, a key parameter for the model is K, this parameter represents the number of neighbors to examine before labeling an example a part of a group. If the majority of an example neighbor are part of one group, that example will also be labeled as part of the group. We utilized the elbow method for determining the optimum K value for our model. This method consists of plotting the F1 estimator of the model across an increasing value of K (neighbor to consider for classification). A selection for K is made based on where the value of F1 begins to decrease and eventually level out.
- 10-fold Cross Validation
    - We utilized 10-fold cross validation of our model. 10-fold cross validation is an important part of training a model, it can help to reduce model overfitting, where a model too closely approximates the models training data. 10-fold cross validation work by splitting the training data into 10 groups or folds, each group is then utilized as a testing group, while a model is trained on the data in the other 9 groups. The final mean model estimates of F1, precision, and recall are then reported.

Results

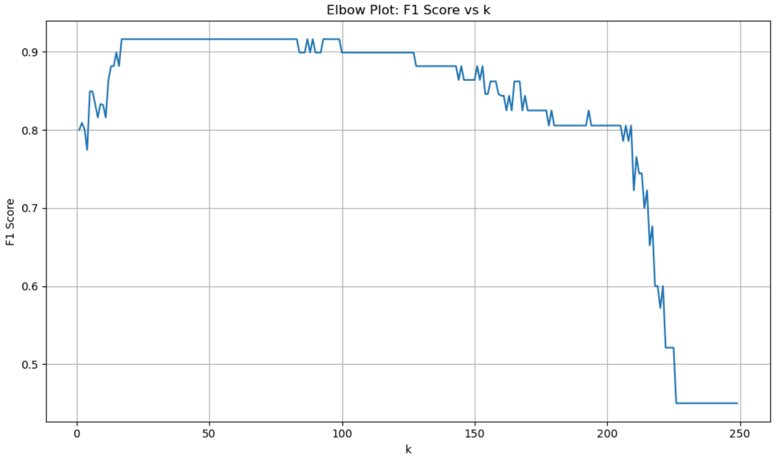
We were able to produce a K Nearest Neighbor model trained on the provided training heart attack dataset, the performance estimators of this model can be seen in the table titled KNN Heart Attack Model Estimators. We were able to create this model leveraging a number of methods as described above, we will go into more detail on the outcomes of each of these methods and how we made decisions regarding selecting K values and feature selections.

KNN Heart Attack Model Estimators	
F1	0.848
Precision	0.933
Recall	0.778



For feature selection we opted to utilize two methods in conjunction. After finding that examining only feature importance was too little information, we elected to examine the features we had selected for linear separability. The results of this analysis can be seen in the left of this paragraph in the figure labeled Feature Distributions for Each Class. By examining these graphs we were able to eliminate features that had limited relevance on our disease outcome, and were not linearly separable. This allowed us to reduce the dimensionality of the model, and improved its classification performance. We selected five features for our model: thal, cp, ca, thalach, age. These were features with high feature importance and good visual separation of their distributions.

After performing feature selection we next moved onto selecting a K value for our model. As discussed we utilized the elbow method for this analysis. We determined through visual inspection that a good estimation of this value for K was 150. We further validated these choices of features and K through experimentation with our model. We also tested other K values, and parameters with the model, and settled on these as the optimal values.



The model we created works relatively well on this small heart attack dataset. We hypothesize that this dataset is modeled relatively well by the KNN model, with features that exhibit good linear separation. We were interested to see how well this model applied to a more complex dataset, with more factors to consider.

Part 2

Introduction

As healthcare data analysis becomes increasingly important in the United States' resource-constrained healthcare system, we turn our attention to another critical health issue: diabetes. This report examines a dataset pertaining to factors relating to diabetes, collected from patients within the U.S. healthcare system.

Similar to our approach with the heart attack dataset, we will employ the K Nearest Neighbor machine learning method to predict the occurrence of diabetes, optimizing model parameters and selecting relevant disease factors. However, we were less successful in the classification of the diabetes dataset, perhaps indicating a more complex set of underlying factors. This suggests that a more sophisticated model than K Nearest Neighbor might be necessary to better capture the intricacies of diabetes prediction.

Dataset

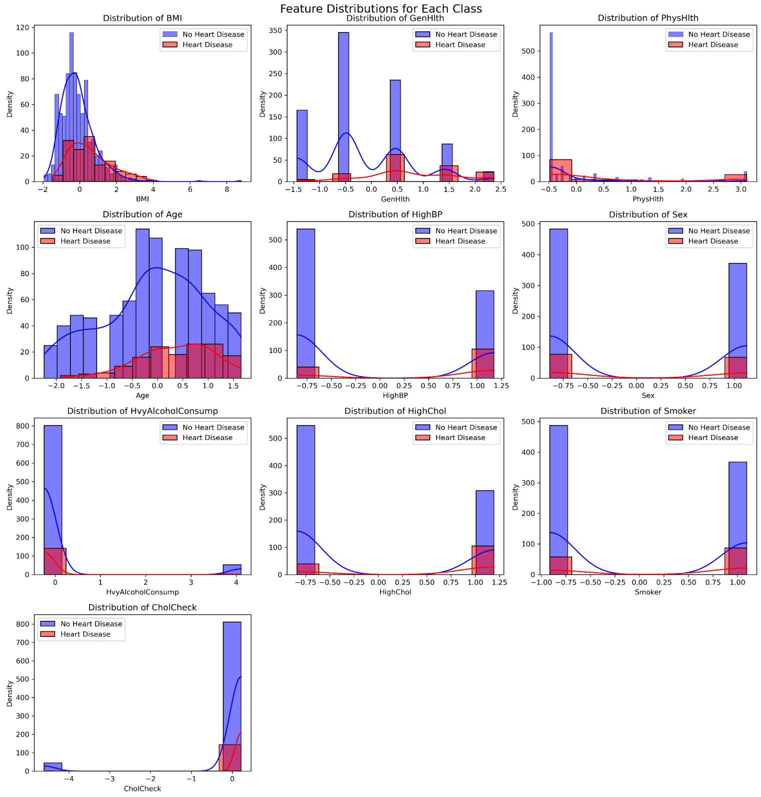
The data set for the second part of our analysis was sourced from the CDC diabetes dataset. This dataset was straight forward to work with, as it already came with a binary column for diabetes or no diabetes. The dataset consisted of 21 features and the one binary disease column. There are a total of 253,680 examples in this dataset. Because of the size of the dataset we elected to train and test our KNN model on a 1000 example random subsample of the dataset due to limited access to compute time. This dataset also allowed us to follow an identical procedure to our Part 1 model.

Results

We produced a KNN model for our second diabetes dataset using an identical procedure to our first model, with the exception that we elected to split the 1000 example set into train and test sets for a final evaluation of our model. The test results of this model can be seen in the table below labeled KNN Diabetes Model Estimates. Examination of these results shows very poor performance of the model. We will try to make some determinations about why this dataset might be less suited to analysis via modeling with KNN versus our previous dataset.

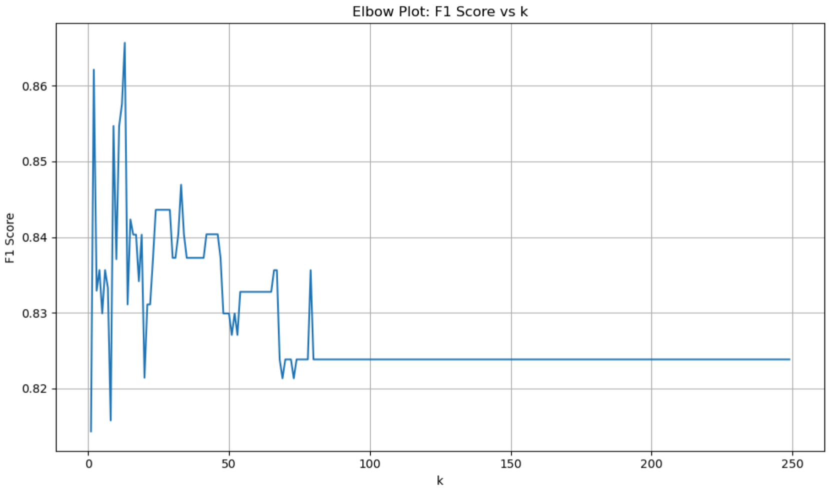
In order to make these determinations about the potential suitability or unsuitability of the model for this set we will first examine the plot of features for linear separability. As before we also examined the feature importance of each of the features in the dataset. We observed that there are many more features in the dataset, and most are correlated about the same value. With the exception of BMI and GenHlth features. We examined the features for linear separability and found very low separability for all features. This can be seen in the grid titled Feature Distributions for each class below. This grid only represents the features that were selected in our final model for this part, not the entirety of the feature set.

KNN Diabetes Model Estimators	
F1	0.429
Precision	0.500
Recall	0.375



We ultimately selected the following features for our model, despite their low separability, mostly based on prior knowledge of the factors contributing to diabetes. We selected the following: BMI, GenHlth, PhysHlth, Age, HighBP, Sex, HvyAlcoholConsump, HighChol, Smoker, CholCheck.

After making the determination that features in this dataset have low linear separability we attempted to utilize the elbow method to determine what the optimal K value would be for our model. We were able to utilize the same method for charting F1 score as in part 1, however it was difficult to find a specific point at which the F1 score began to diverge.



While there seems to be a point where this is occurring in the graph labeled Elbow Plot: F1 Score vs K, we were unable to see an improvement by setting our K value around the divergence of this plot. We found through experimentation that the best value for K for our model was to set K equal to 5. We spent time experimenting with various configurations of K values as well as selecting an array of features and were unable to come up with a combination that was more effective than our reported results for this part.

We hypothesize the poor performance on this dataset could be a result of a poorly suited model for this data. The relationships between features in this data seem to have a high degree of nonlinearities, and we believe that a more complex machine learning model might perform better on this dataset, than the relatively simple KNN model.