Project 4 Report

Note: all these scores are based on the time that we run them. It may be different at another run time (because we did not add a seed or set the random state!)

Part 1:

Introduction

The primary goal of this analysis is to create a K-Nearest Neighbors model with the best k and features to predict heart disease based on various individual features, including age, sex, resting blood pressure, etc from the dataset. The stakeholders include heart disease patients seeking quick check-ups, doctors for accurate diagnosis, and researchers for analyzing the effect of each factor. Features are selected based on their correlation, distribution, slope, and p-value from T-tests. Notably, 'thalach'[1], 'exang'[2], 'oldpeak'[3], 'ca'[4], and 'thal'[5] are identified as having the greatest impact. The model undergoes training with these features and is tested using a 10-fold cross-validation method to determine the best value for neighbors (k) in the range of 1-50. The best k value, determined to be 32, shows promising performance metrics. These metrics include precision, recall, F-score, and accuracy. With cross-validation, the model got a recall of 0.759, precision of 0.844, F-score of 0.795, and accuracy of 0.827. Without cross-validation, the precision is 0.913, recall is 0.777, F-score is 0.84, and accuracy is 0.866. This analysis holds significant potential in aiding timely prevention and treatment strategies for at-risk individuals. Our surprising results can be found here. Also, you may see our GitHub repository.

Methods

When selecting features for our K-Nearest-Neighbors model, we initially identified those with the highest correlation to heart disease existence (class label “1”), plotting their correlation plot against disease presence and finding the slope of them after standardizing the data for equal weighting. The top four and five features with the highest slopes were then chosen. Distributions plot and T-tests (p-value) results confirmed their difference is significant (selected visually by looking at the results and numbers).

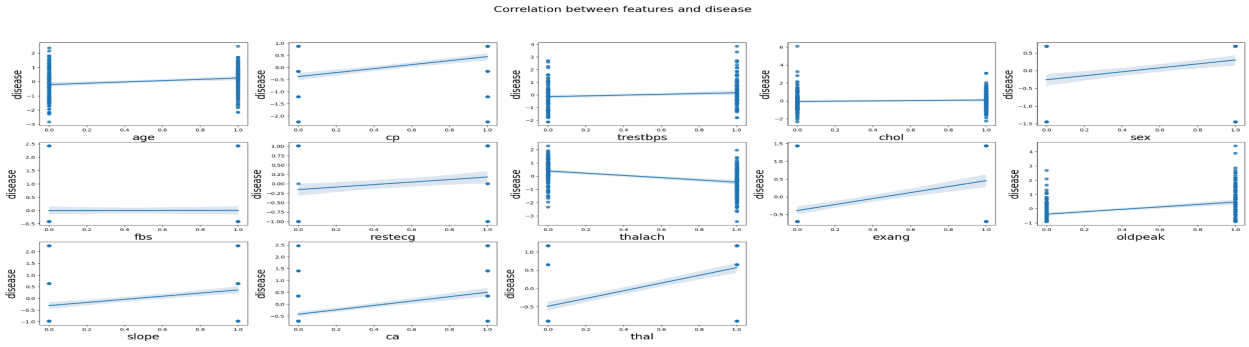
To determine the optimal k value, we employed 10-fold cross-validation (each time the dataset will be shuffled randomly to get different results), calculating average f-scores for k values ranging from 1 to 50. The k value with the highest f-score for disease presence (label "1") was selected. This process was iterated 1000 times, resulting in the best k value of 32. Due to multiprocessing[6] limitations, an iterative approach was used in the Jupyter Notebook, while parallelized processing[7] was implemented separately in another Python file.

---

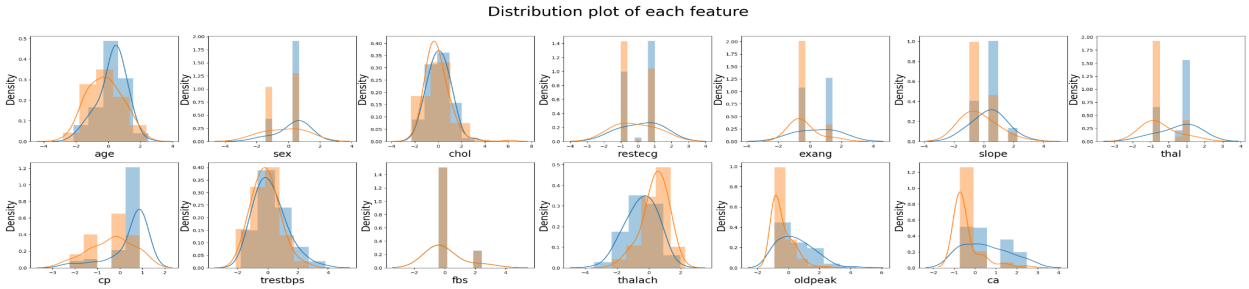
[1] Maximum heart rate achieved
[2] Exercise induced angina
[3] ST depression induced by exercise relative to rest
[4] Number of major vessels (0-3) colored by flourosopy
[5] A blood disorder called thalassemia
[6] Multiprocessing is the use of two or more central processing units within a single computer system. The term also refers to the ability of a system to support more than one processor or the ability to allocate tasks between them.
[7] a computing technique when multiple streams of calculations or data processing tasks co-occur through numerous central processing units (CPUs) working concurrently.

Results

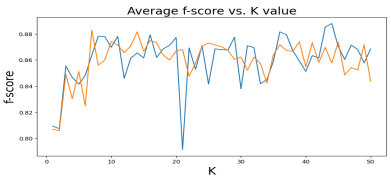
The figure below shows the plots of all the features against whether the individual has heart disease:



The figure demonstrates that the features with slopes: "thalach" (-0.211), "exang" (0.210), "oldpeak" (0.211), "ca" (0.231), and "thal" (0.262) have the strongest correlation with whether an individual has heart disease or not, thus serving as input features for the K-Nearest-Neighbors model. To confirm this, we have included a distribution plot and the T-test, which is shown below:

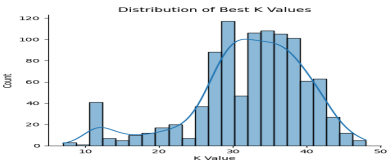


The p-values from the T-test confirm significant differences in feature distributions between healthy and sick individuals. Specifically, for "thalach", "exang", "oldpeak", "ca", and "thal", the p-values are 2.23e-14, 3.27e-14, 2.15e-14, 3.35e-17, and 1.39e-22, respectively, with corresponding T-statistics of -8.03, 7.97, 8.04, 8.97, and 10.64. These results validate the distinction between feature groups and support their suitability for feature selection.



The following figure shows the average best f-score we got from cross-validation with different k in each iteration. The figure shows the best k here is 44 without finding the mean (or mode) of them with 2 iterations!

The following plot is the distribution of the best k values when the 10-fold cross-validation is run on the data 1000 times. The distribution plot tells us that our best k value typically sits between 25 and 40. For our best k values, we will take the average best k value which is 32. Using 32 as our k value and the features “thalach”, “exang”, “oldpeak”, “ca”, and “thal”, we obtained [0.687, 0.812, 0.846, 0.833, 0.642, 0.8, 0.909, 0.846, 0.75, 0.846] with an average 0.79 for recall, [0.916, 0.928, 0.916, 0.769, 0.818, 0.888, 0.909, 0.846, 0.857, 0.846] with an average of 0.869 for precision, [0.78, 0.866, 0.879, 0.8, 0.72, 0.842, 0.909, 0.846, 0.799, 0.846] with an average f-score of 0.829, and [0.793, 0.862, 0.896, 0.827, 0.758, 0.896, 0.931, 0.862,



0.793, 0.862] with an average accuracy of 0.848 in our 10-fold cross-validation. Also, just for a single run without cross-validation (random: 20% test dataset, 80% training dataset) of the split test and train data is precision: 0.904, recall: 0.791, f-score: 0.844, and accuracy: 0.883. Also, we trained the model without "exang" which the result was most of the time worse than with "exang": [0.7, 0.714, 0.823, 0.875, 1.0, 0.823, 0.571, 0.9, 0.733, 0.75] with average 0.789 for recall, [1.0, 0.909, 0.933, 0.538, 0.705, 0.933, 0.8, 0.75, 0.846, 0.5] with an average of 0.791 for precision, [0.823, 0.8, 0.874, 0.666, 0.827, 0.874, 0.666, 0.818, 0.785, 0.6] with an average of 0.773 for f-score, and [0.79, 0.827, 0.862, 0.758, 0.827, 0.862, 0.724, 0.862, 0.793, 0.724] with an average accuracy of 0.800 in our 10-fold cross-validation and without cross-validation (random: 20% testing dataset, 80% training dataset): precision: 0.875, recall: 0.75, f-score: 0.807, and accuracy: 0.833.

Part 2:

Introduction

For our second analysis, we aim to create a K-Nearest Neighbors model (with the best k and the best features) for a breast cancer dataset comprising features such as age, menopause[8] status, tumor size, involvement of nodes[9], and metastasis[10]. This model aims to predict whether a tumor is benign or malignant in a timely diagnosis to help the patients. Stakeholders include breast cancer researchers seeking a deeper understanding of each feature's impact on cancer type, doctors aiming for improved diagnosis based on breast test results, and cancer patients seeking quick check-ups with their test results. Initially, we selected a subset of features based on their correlation and the slopes of the correlation and distribution, and T-test (for finding significant differences), focusing on those with the greatest impact on malignancy. The selected features—'Menopause', 'Metastasis', 'Age', 'Tumor Size (cm)', and 'Inv-Nodes'[11]—were used to train the model with k values ranging from 1 to 50 using 10-fold cross-validation. The optimal k value was determined to be 10 through iterative testing. Training the model with this k value yielded average performance metrics in 10-fold cross-validation: recall of 0.817, precision of 0.988, F-score of 0.891, and accuracy of 0.909. Additionally, without cross-validation (utilizing a 20% testing dataset and 80% training dataset split), the model achieved a precision of 0.992, recall of 0.875, F-score of 0.933, and accuracy of 0.926.

Dataset

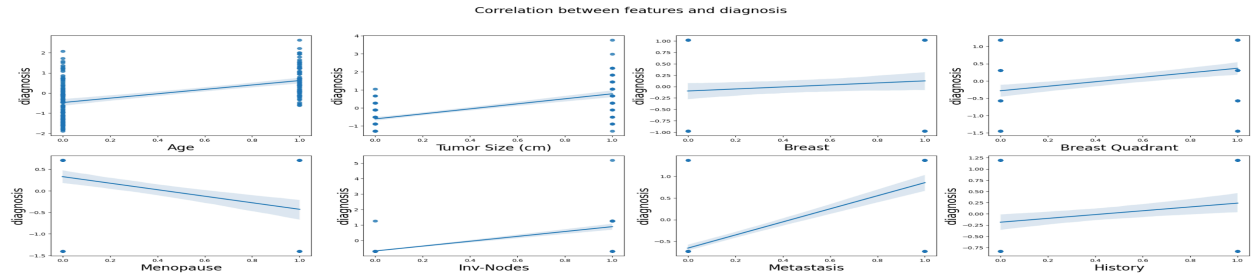
For our breast cancer dataset[12], we prepared the categorical and binary variables by mapping them to integers for plotting purposes. We adjusted the columns "Breast", "Breast Quadrant", and "Diagnosis Result", dropping rows with missing values and removing unnecessary columns like "S/N"[13] and "Year"[14]. The "Breast Quadrant" categories, including "Lower inner", "Lower outer", "Upper inner", and "Upper outer", were converted to integers (0, 1, 2, 3) for clarity. Additionally, we filtered out rows with "#" in the "Diagnosis Result" column. The dataset includes key features such as

[8] Whether the patient is pro or postmenopausal at the time diagnose
[9] The number of axillary lymph nodes that contain metastatic
[10] If the cancer has spread to another part of the body or organ.
[11] Involved nodes
[12] https://www.kaggle.com/datasets/fatemehmehrparvar/breast-cancer-prediction
[13] Unique identification for each patient.
[14] The year diagnosis was conducted.

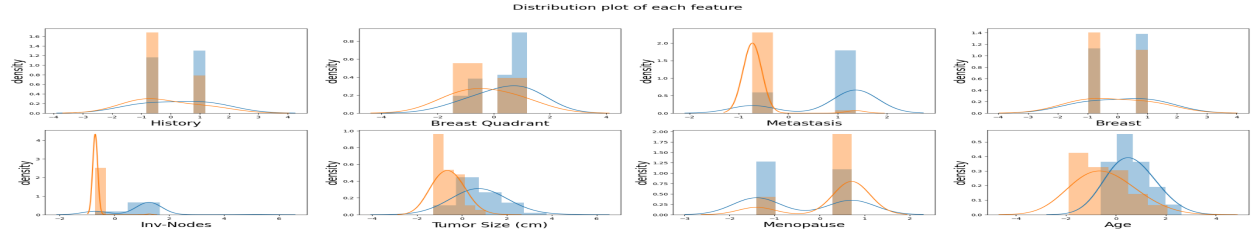
Menopause, Metastasis, Age, Tumor Size (cm), and Inv-Nodes, with “History” as an additional feature.

Results

The figure below shows the correlation between each feature and whether the tumor is malignant.

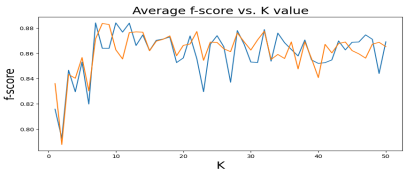


The figure suggests that the features 'menopause', 'metastasis', 'age', 'tumor size', and 'inv-nodes' are most strongly correlated with whether a patient has malignant breast cancer, with respective slopes of -0.186, 0.371, 0.269, 0.343, and 0.387. Therefore, these are the features we choose for our K-Nearest-Neighbors model. This choice is supported by the figure and the p-value of the T-test below, which shows a significant difference in the distributions for the selected features in malignancy and benign individuals.



The T-test results indicate significant differences in feature distributions: "Age": (p=4.96e-17, t=9.18), "Menopause": (p=2.81e-08, t=-5.77), "Tumor Size": (p=1.57e-30, t=13.65), "Inv-Nodes": (p=3.16e-43, t=17.76), and "Metastasis": (p=4.89e-38, t=16.07) because p-value is less than 0.05 and there is strong evidence against null hypothesis.

The following figure shows the average best f-score we got from cross-validation with different k in each iteration. The best k here is 8 without finding the mode (or mean) with 2 iterations!



The next plot shows the distribution of best k values obtained by running a 10-fold cross-validation 1000 times. Because there is one bin that dominates the others in the histogram, The most suitable value to choose for best k is the mode, which is 10 in this case. Using 10 for our k value and features “Menopause”, “Metastasis”, “Age”, “Tumor Size”, and “Inv-Nodes”, we obtain [0.833, 1.0, 0.8, 1, 0.8, 0.714, 0.571, 0.75, 0.7, 1] with average values of 0.816 for recall, [1.0, 1.0, 1.0, 0.857, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0] with an average of 0.985 for precision, [0.909, 1.0, 0.888, 0.923, 0.888, 0.833, 0.727, 0.857, 0.823, 1.0] with an average of 0.885 for f-score, and [0.9, 1.0, 0.9, 0.95, 0.95, 0.9, 0.7, 0.9, 0.85, 1.0] with an average of 0.905 for accuracy. And without cross-validation (random: 20% testing dataset, 80% training dataset): precision: 0.928, recall: 0.684, f-score: 0.787, and accuracy: 0.829.