CS5830 Project 4 Report

Gavin Murdock and Megh Kc

Our Github project folder and presentation slides can be found below.

Github project, presentation slides.

Part 1

Introduction

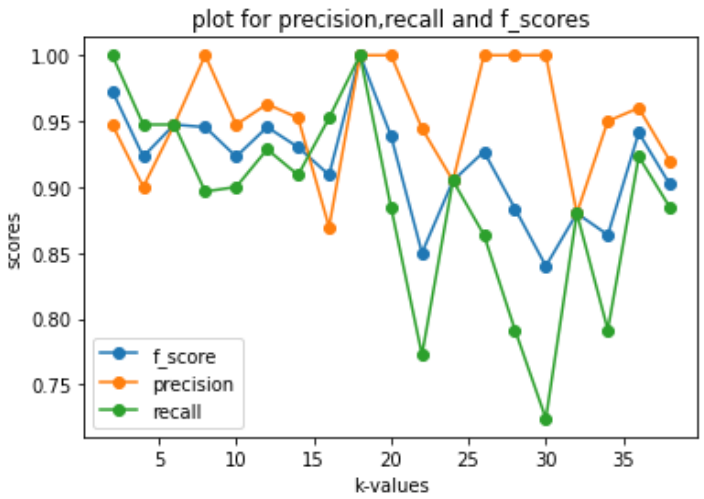
The cleveland dataset that describes the several parameters relating to the heart disease and finally includes the ‘num’ column dataset to determine whether the patient has heart disease or not. The considered variables in this dataset are age, sex, blood pressure, blood sugar level, cholesterol, chest pain type and so on. The challenge is to select the optimal set of attribute parameters from the available variables set. GridSearchCV function available in scipy library was used to obtain optimal set of attributes. Similarly, 10 fold Monte carlo cross validation for train and test dataset was performed and optimal k-value was found. k-nearest neighborhood algorithm is used to predict the heart disease in patients.

Methods

The datsets columns are standardized and cleaned ro remove any NA values. The string values (if any) are converted to numeric. To test the accuracy of the model, we tried to search optimal set of attributes using GridSearchCV function in SciPy library. Then a function is created for train test data split and return a precision, recall, f_score and support scores for each iteration. The function would take different k-values and number of iteration for cross validation. In addition, monte carlo cross validation function is devised. The average three scores (precision, recall & f_score) are plotted with respect to different k-values and optimal k values.

Results

The optimal k-value determined was k=17 that gave highest f_score. the precision recall values are also in high range.



Part 2

Introduction

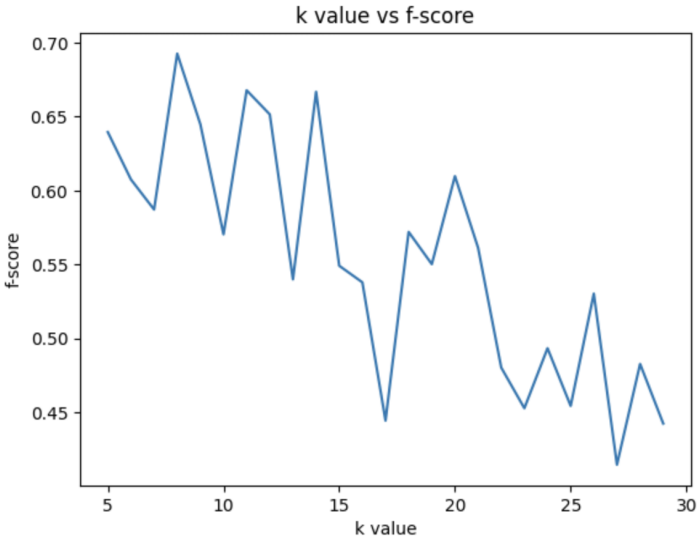
In this part of our project, we had the goal of predicting whether or not a patient has diabetes. Multivariate health data and k-nearest neighbor classifiers were used to make these predictions. Models such as these could be used by health professionals as a guide to whether or not a patient is likely to have diabetes and should undergo further testing.

Dataset

The dataset we used for this portion of the project was acquired from Kaggle.com where it was uploaded by user Houcem Benmansour. It can be found here. The dataset contained cholesterol levels, glucose levels, ages, genders, heights, weights, bmi indexes, systolic blood pressures, diastolic blood pressures, waist measurements, hip measurements, and waist to hip ratios for a variety of patients. It also contained whether each patient has diabetes. We had to adjust two features of the dataset, gender and whether the patient has diabetes, to be represented as numbers instead of words. We also had to adjust a number of columns that used commas instead of periods to represent floating point numbers. Lastly, we made standardized versions of each column in the dataset to use with our kNN models.

Methods

To determine which attributes to use for this dataset, we read a few short articles about diabetes and hand-picked which attributes we thought would be the best indicators. We picked cholesterol, cholesterol_hdl_ratio, glucose, bmi, systolic and diastolic blood pressure, and waist to hip ratio. To determine a good k value for our models, we tested all k values between 5 and 30. We tested each k value 10 times and plotted the mean f-score for each value, which can be seen below. A k value of 8 gave the highest mean f-score, so we used a k value of 8 to get our final results.



Results

To get our final results, we made ten k-nearest neighbor models with our hand-picked attributes and a k value of 8. Our models had a mean f-score of 0.728. The scores of each individual model can be seen below.

Model #	1	2	3	4	5	6	7	8	9	10
Precision	1.0	1.0	1.0	0.857	0.909	1.0	1.0	1.0	0.8	0.833
Recall	0.769	0.636	0.429	0.6	0.625	0.727	0.556	0.455	0.571	0.667
F-score	0.870	0.778	0.6	0.706	0.741	0.842	0.714	0.625	0.667	0.741