Project-4 K-Nearest Neighbors Report

Satyanarayana Vinay Achanta (A02395874)
Levi Nielson (A02253509)

Part I Predicting Heart Disease

Introduction

The purpose of this analysis is to determine the optimal combination of attributes and the number of nearest neighbors (k) for predicting heart disease. Predicting heart disease is important as it helps in early detection, and treatment which can save lives, reduce healthcare costs could potentially inform healthcare professionals in their decision-making process. The Cleveland Heart Disease dataset was used for this project which contains 303 instances and 14 attributes. The attributes present are age, sex, blood pressure, and cholesterol levels, which were preprocessed by removing highly correlated columns, and missing values, and the target variable 'num' was renamed as 'disease' with binary values indicating the presence or absence of heart disease. The K-nearest neighbors algorithm was used to predict the presence or absence of heart disease.

Our method involves brute-forcing the problem by testing different combinations of attributes and numbers of neighbors. We will standardize the data to ensure that the models are not affected by large or small values. We will evaluate the accuracy of our model by measuring the recall, precision, and F1 score. Our results will provide insight into which attributes and a number of neighbors produce the most accurate predictions of heart disease.

GitHub Link
Presentation Slides Link

Methods

Our method mainly involved brute-forcing the problem. The data was preprocessed by dropping irrelevant columns that we thought weren't relevant to creating a solid model, more columns also make the combination selection time taking. Modifying the target variable to binary form (0 or 1) to indicate the presence or absence of heart disease. We preprocessed the data by removing symbols and dropping Nan. Then we performed standardization of several of the columns to ensure that when our models read data that large or small values didn't too easily sway them.

For each possible combination of attributes and amounts of neighbors ranging from 1 to 10, we created nearest neighbor models and found out their recall, precision, and F1 score. The K-fold cross-validation technique was used to evaluate the performance of the model. The big downside is that it takes roughly half an hour to run the analysis. After finding the best attributes and neighbor counts, the program is much quicker, meaning subsequent executions need not involve the lengthy exhaustive search.

Results

The optimal combination of attributes and k-value significantly improved the accuracy of the model, with an F1 value of 0.7986577181208053. The K-nearest neighbors' algorithm can be an effective tool for predicting heart disease. For the heart disease prediction dataset, the best combination of attributes and K-value which produced the highest F1-score were:

- Attributes: sex, cp_s, slope_s
- K-values: 4 and 7
- Precision: 0.7484276729559748
- Recall: 0.8561151079136691
- F1-score: 0.7986577181208053

This means that the model achieved a good balance between precision and recall, with an F1-score of 0.7987, which is a measure of the model's overall accuracy. The selected attributes were sex, cp_s, and slope_s, which were determined to be the best indicators of heart disease in an individual based on the analysis of the dataset. The best K-values were found to be 4 and 7, indicating that the model performed well.

Part II Part-2 Predicting Diabetes

Introduction

The purpose of this analysis is to determine the optimal combination of attributes and the number of nearest neighbors (k) for predicting Diabetes. We explored the National Institute of Diabetes and Digestive and Kidney Diseases dataset. This analysis is important because it could help healthcare providers identify people who are at risk of getting diabetes and develop prevention that will save many lives. The K-nearest neighbors' algorithm was used to predict the presence or absence of Diabetes. We have applied a similar algorithm as used in the heart disease dataset. We start by standardizing the data and removing all the symbols and missing values.

Our method involves brute-forcing the problem by testing different combinations of attributes and numbers of neighbors. We will standardize the data to ensure that the models are not affected by large or small values. We will evaluate the accuracy of our model by measuring the recall, precision, and F1 score. Our results will provide insight into which attributes and several neighbors produce the most accurate predictions of heart disease.

Dataset

The diabetes dataset is from the National Institute of Diabetes and Digestive and Kidney Diseases which we took from Kaggle. The dataset consists of women above the age of 21. The attributes present in the dataset are Pregnancies, blood pressure, skin thickness, insulin, diabetes pedigree function, age, and outcome which was a value of 0 or 1, indicating whether the subject had diabetes. We thought that an individual's blood glucose, blood pressure, insulin, and age were the most important predictors of diabetes.

Methods

Our method was similar to our method for heart disease in the Cleveland dataset. We started off by standardizing the data, then split the data into training and testing datasets and ran a similar algorithm on the data to find the best model.

Results

The optimal combination of attributes and k-value significantly improved the accuracy of the model, with an F1 value of 0.673721340388007. The K-nearest neighbors' algorithm can be an effective tool for predicting diabetes. The best combination of attributes and K-value for the Diabetes prediction dataset that produced the highest F1-score was:

- Attributes: Glucose_s, Insulin_s, BMI_s, DiabetesPedigreeFunction_s, Age_s
- K-value: 7
- Precision: 0.6387959866220736
- Recall: 0.7126865671641791
- F1-score: 0.673721340388007

This means that the model achieved a good balance between precision and recall, with an F1-score of 0.6737, which is a measure of the model's overall accuracy. The selected attributes were glucose, insulin, BMI, diabetes pedigree function, and age, which were determined to be the best indicators of diabetes in an individual based on the analysis of the dataset. Overall, this analysis could potentially help healthcare providers identify individuals who are at risk of developing diabetes and develop preventive measures that could potentially save lives.