Project 4 (KNN)

Code | Presentation by- Sayali Sali, Clark Farnsworth

Part I:

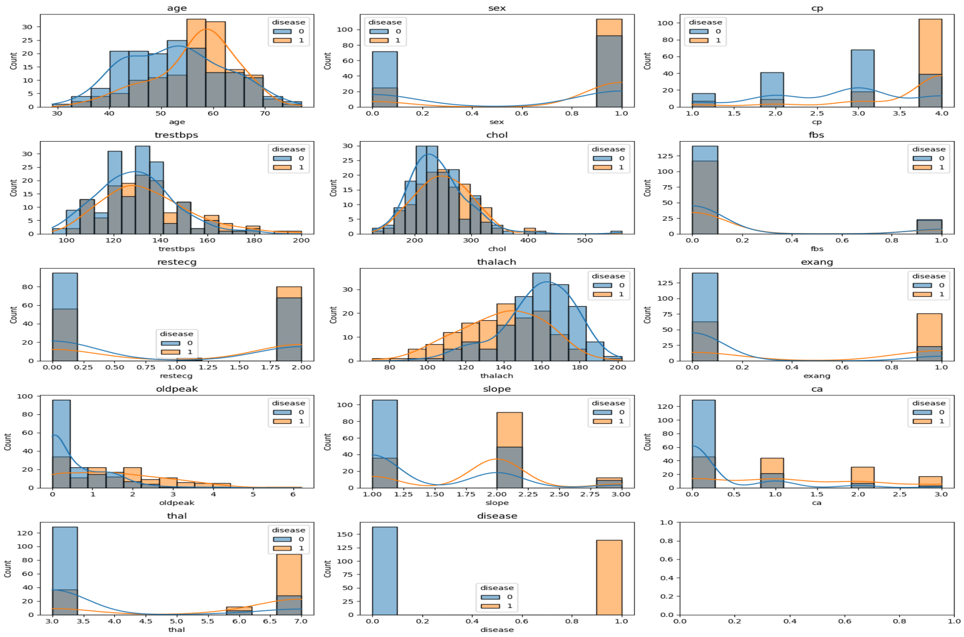
Introduction

In this study, we aimed to determine the optimal values of k and the optimal set of attributes to maximize predictive power in predicting whether a patient has heart disease or not using the Cleveland dataset. The dataset contains various clinical attributes, and the target variable "num" indicates the presence or absence of heart disease. We employed machine learning techniques, specifically k-nearest neighbors (KNN) classification, and evaluated the performance using mean precision, recall, and F1 metrics.

Methods

To determine the optimal values of k and attributes, we followed these steps:

- Data Preprocessing: We loaded the Cleveland dataset and preprocessed it by collapsing all values 1-4 of the "num" attribute into a single value to represent the presence of heart disease (1) or absence (0). We also handled missing values by imputing them with the median.Than Standardize features to bring them on one scale using StandardScaler function.
- Feature Selection: First we plot histogram for all columns and manually analyze the best attributes that can tell us more about the dataset.



Then we considered all available attributes and then used covariance and variance analysis to select potential features. We chose features with high covariance with the target variable and high variance.

- Model Training and Evaluation: We implemented a function to train the k-nearest neighbors (KNN) algorithm with various values of k and evaluated the model's performance using 10-fold cross-validation. For each iteration, we randomly split the dataset into training and testing sets. We calculated precision, recall, and F1 scores for each fold and recorded the mean of all scores. We experimented with different values of k and the set of attributes. We iteratively adjusted these parameters based on the performance metrics until we achieved satisfactory results.

Results

Our analysis revealed the following results:

- Optimal k value: After experimenting with different values of k, we found k=7.
- k=7 yielded the best performance in terms of F1 score.
- Optimal set of attributes: Based on covariance and variance analysis, we selected the following attributes: age, thal, thalach, restecg, cp, and ca.
- Performance Metrics: The mean precision, recall, and F1 score across 10-fold cross-validation are as follows:
  - Mean Precision: A precision of 0.828 means that approximately 82.8% of the instances predicted as having heart disease were indeed true positives.
  - Mean Recall: A recall of 0.792 implies that our model correctly identified approximately 79.2% of all individuals with heart disease.
  - Mean F1 Score:With an F1 score of 0.800, our model achieves a good balance between precision and recall, indicating robust performance in predicting both positive and negative instances of heart disease.
  - Mean Accuracy: A mean accuracy of 80.452% indicates that our model correctly classified approximately 80.452% of all instances in the dataset, regardless of class.

These results demonstrate the effectiveness of our approach in predicting heart disease using the Cleveland dataset. Our model achieved robust performance in terms of precision, recall, and overall F1 score, indicating its potential utility in clinical practice for risk assessment and early diagnosis of heart disease.

Part II:

Introduction: In this analysis, we applied the k-nearest neighbors (KNN) algorithm to a dataset to predict if the patient had trouble sleeping based on their number of doctor visits, physical health, mental health status, employment, stress, medication, and pain. The KNN algorithm predicts the classification of a new data point by considering the majority class among its k nearest neighbors.

Dataset:

The dataset used in this analysis is the "NPHA-doctor-visits.csv" dataset. For this subset of the original NPHA dataset we chose 14 features related to health and sleep to use for the prediction task. It contains information about patients' physical health, mental health, dental health, number of doctors visited, employment status, and factors affecting their sleep. Before performing the analysis, we renamed some columns for ease of reference and exploration.

Methods:

- Data Preprocessing: First load the dataset into a Pandas DataFrame and rename the columns to more concise names. No further cleaning was necessary as the dataset was already relatively clean.
- Feature Selection:The attributes 'num_of_visit', 'phy_health', 'men_health', ‘Employment’, ‘stress’, ‘medication’, and ‘pain’ were selected as features for predicting if the patient had trouble sleeping.
- Model Training:The use of the KNN algorithm with Euclidean distance as the metric to train the model.
- Model Evaluation: To evaluate the model's performance, I randomly sampled 200 patients from the dataset and predicted if they had trouble sleeping. I calculated precision, recall, F1-score, and support to assess the model's performance.

Results:

After training the KNN model and making predictions on the test set, we evaluated its performance using precision, recall, and F1-score metrics. The results showed varying levels of accuracy depending on the choice of attributes and the value of k. For example, when considering the number of doctors visited, physical health, mental health, employment, stress, medication, and pain as attributes, the model achieved a precision of 0.73 and recall of 0.71 for predicting if a patient had trouble sleeping.

The KNN model achieved the following results on the test data:

- Precision: 0 .73
- Recall: 0.706
- F1-score: 0.677
- Accuracy: 69%