Bharath chennu

811225107

Report: Analysis of Diagnosis Data using K-Nearest Neighbors

Introduction: In this report, we will be analyzing a dataset that contains information about patients diagnosed with either Diagnosis A or Diagnosis B. Our goal is to use the K-Nearest Neighbors algorithm to predict whether a patient has been diagnosed with Diagnosis B or not. The steps we will follow to achieve this goal are loading the data, viewing the data, dummy encoding, cleaning the data, labeling the predictor, data partition, normalization, model fit, model train, model predict, and model evaluate.

Loading the Data: The first step in any data analysis project is to load the data into the analysis environment. We will be using a dataset that contains the following variables:

- Age: the age of the patient
- Gender: the gender of the patient
- Diagnosis: the diagnosis of the patient, either A or B

Viewing the Data: Once the data has been loaded, it's essential to examine the data's structure and characteristics. We can use the head() function to get a quick look at the data's first few rows, and the summary() function to get a summary of the data's descriptive statistics.

Dummy Encoding: Since the diagnosis variable is categorical, we need to convert it into a binary variable so that we can use it in the KNN model. We can achieve this by using dummy encoding, which will convert Diagnosis A into 0 and Diagnosis B into 1.

Cleaning the Data: After dummy encoding, we need to clean the data and get it ready for the model. We will remove any missing values and outliers that could negatively affect the model's performance.

Labelling the Predictor: Our predictor variable is Diagnosis B. We will label this variable as the response variable, and all other variables will be considered predictor variables.

Data Partition: To train and test our model's performance, we will partition the data into a training set and a test set. We will use 80% of the data for training and 20% for testing.

Normalization: To ensure that all predictor variables have equal weight in the KNN model, we will normalize the data using the Scalar function.

Model Fit: After normalizing the data, we will define the K value for the KNN model and fit the model to the training data.
Model Train: We will then build the KNN model using the training data.

Model Predict: We will predict the values of Diagnosis B using the model and the test data.

Model Evaluate: Finally, we will evaluate the model's performance using accuracy, sensitivity, specificity, and ROC curve.

Conclusion: In conclusion, we have successfully used the K-Nearest Neighbors algorithm to predict the diagnosis of patients using predictor variables such as age and gender. We have used a step-by-step approach to analyze the data, preprocess it, build the model, and evaluate its performance. Our model's accuracy, sensitivity, and specificity indicate that it can predict patient diagnoses with high accuracy.

Data source: https://www.kaggle.com/datasets/gkalpolukcu/knn-algorithm-dataset/code?resource=download