Heaven's Light is Our Guide

Rajshahi University of Engineering & Technology

Course No: CSE 4204

Course Title: Sessional Based on CSE 4203

Experiment No. 1

Name of the Experiment: Implementation of Nearest Neighbor classification algorithms with and without distorted pattern.

Submitted by:
Ashraf-Ul-Alam
Roll: 1803070
Section: B
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology

Submitted to:
Rizoan Toufiq
Assistant Professor
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology

Date of Submission: 04 November 2023

Contents

1 Introduction

In this report, an analysis of liver disease prediction using the K-Nearest Neighbors (KNN) algorithm is presented. Liver disease is a significant health concern worldwide, and early detection is crucial for effective treatment. Machine learning techniques, such as KNN, can aid in predicting liver disease based on various patient attributes.

2 Dataset

2.1 Dataset Analysis

This analysis began by loading the dataset from the "indian_liver_patient.csv" [1] file. The dataset contains information on various attributes related to liver health, such as age, gender, bilirubin levels, liver enzymes, and more. The dataset has 583 entries and 11 columns.

- Age: An integer representing the patient's age.
- Gender: Categorical feature (0 for female, 1 for male).
- Total Bilirubin, Direct Bilirubin, Alkaline Phosphotase, Alamine Aminotransferase, Aspartate Aminotransferase: Numeric measurements.
- Total Proteins, Albumin, Albumin and Globulin Ratio: Numeric measurements.
- Dataset: The target variable (1 for liver disease, 2 for no liver disease).

2.2 Feature Processing

The dataset's features were examined to identify which ones were suitable for our KNN model. It was noticed that the "Gender" column contained categorical data, so the label encoding was used to convert it into numerical values (0 for female and 1 for male). This conversion is necessary because KNN requires numerical inputs.

2.3 Handling Missing Values

The missing values within the dataset were also checked and it was found that the "Albumin_and_Globulin_Ratio" column had four missing values. To address this, the missing values were filled with the mean of the column.

2.4 Correlation Analysis

To ensure that the selected features were suitable for the KNN model, a correlation analysis were performed. The correlations between different features were visualized using a heatmap

and no highly correlated data was found. This indicated that all the selected features could be used for KNN without issues.

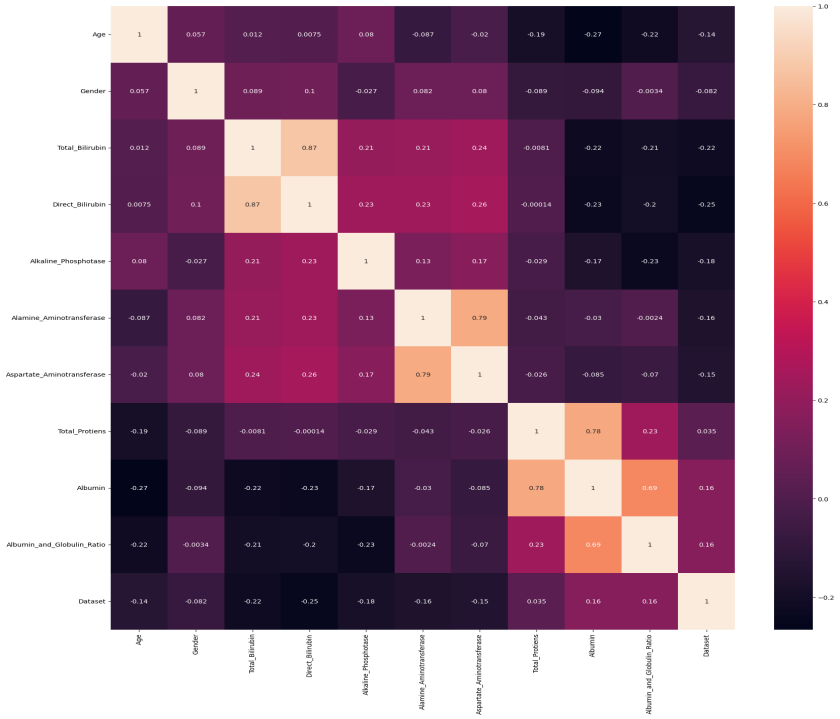

Figure 1: Correlation Heatmap

2.5 Data Type Conversion

To maintain consistency and avoid data loss, all the features were converted to the float data type.

2.6 Data Imbalance

A significant class imbalance can be observed in the target variable ("Dataset") with more instances of one class (1) compared to the other (2). To address this imbalance, Synthetic Minority Over-sampling Technique (SMOTE) was applied to oversample the minority class. The following table illustrates the class distribution before and after applying SMOTE:

Class	Count Before SMOTE	Count After SMOTE
1	416	416
2	167	416

Table 1: Class Distribution Before and After SMOTE

In the original dataset, class 1 (indicating patients with liver disease) had 416 instances, while class 2 (indicating patients without liver disease) had only 167 instances. After applying SMOTE, both classes have an equal number of instances (416), effectively addressing the class imbalance issue.

Balancing the dataset through SMOTE ensures that the machine learning model is trained on a more representative dataset, preventing bias towards the majority class and improving its ability to make accurate predictions for both classes.

2.7 Train-Test Split

The dataset was split into training and testing sets with a 70-30 ratio to evaluate the performance of KNN model.

3 Methodology

3.1 K-Nearest Neighbors (KNN) Algorithm

The K-Nearest Neighbors (KNN) [2] algorithm is a supervised machine learning technique used for classification. It makes predictions based on the majority class of its k-nearest neighbors in the feature space. Below is the implementation of the KNN algorithm:

1. **Euclidean Distance Calculation:** The distance between two data points, x_1 and x_2, is calculated using the Euclidean distance formula:

$$EuclideanDistance = \sqrt{\sum_{i=1}^{n}(x_{1i} - x_{2i})^2}$$

   If either x_1 or x_2 is a string (categorical feature), a special value like positive infinity can be returned or handled as needed.

2. **KNN Prediction:** To predict the class of a new data point (new_data_point), the following steps are performed:

   (a) Calculate the Euclidean distances between new_data_point and all data points in the training set (X_train).

   (b) Select the indices of the k nearest neighbors with the smallest distances.

   (c) Retrieve the corresponding labels (y_train) of these nearest neighbors.

   (d) Convert the labels to integers and flatten the array.

   (e) Make a prediction by selecting the class that appears most frequently among the k nearest neighbors' labels using the *argmax* function.

This KNN algorithm was applied with different values of k ranging from 3 to 24 and the accuracy for each k value was calculated, as discussed in the "Result Analysis" section.

4 Result Analysis

The KNN model was evaluated with different values of k (ranging from 3 to 24) and the accuracy for each k value was recorded. The results indicated that the model achieved the highest accuracy of 74.00% when k = 5. The visualization of the accuracy for different k values using a bar chart indicates k = 5 being the optimal choice.

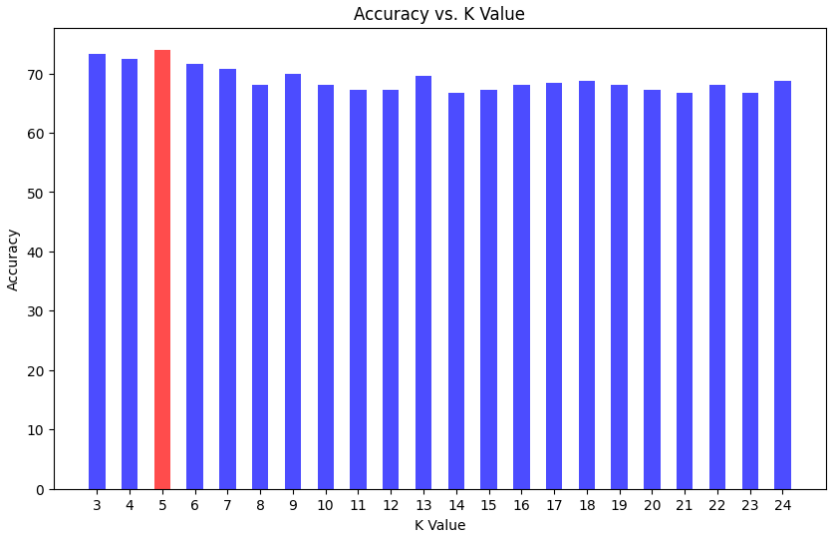

Figure 2: Comparing the accuracy for different k-values

5 Limitations of K-Nearest Neighbors Classifier

The K-Nearest Neighbors (KNN) classifier is a simple and intuitive machine learning algorithm, but it may face challenges and limitations when applied to certain types of datasets. Some of the key factors that can affect the performance of KNN and why it may not perform as expected in these scenarios:

1. **High-Dimensional Data:** KNN relies on measuring the distance between data points to make predictions. In high-dimensional spaces, the "curse of dimensionality" can lead to increased computational complexity and decreased performance. As the number of features increases, the Euclidean distance between points becomes less meaningful, making it challenging for KNN to find meaningful neighbors.

2. **Imbalanced Datasets:** KNN can struggle when dealing with imbalanced datasets where one class significantly outnumbers the other. In such cases, the classifier may be biased towards the majority class and perform poorly in predicting the minority class.

3. **Outliers:** Outliers can significantly impact the performance of KNN. Since KNN relies on the distance between points, outliers can distort the decision boundary and lead to incorrect predictions.

4. **Large Datasets:** KNN is a lazy learner, meaning it stores the entire dataset and computes distances at prediction time. For large datasets, this can be computationally expensive and slow.

5. **Irrelevant Features:** KNN treats all features equally, which can be problematic if some features are irrelevant or noisy. Irrelevant features can introduce noise into the distance calculations and reduce the classifier's accuracy.

In summary, while KNN can be a simple and effective classifier for some datasets, it may not perform as expected in cases where the data violates its underlying assumptions, such as high dimensionality, imbalanced classes, outliers, or noisy features. Careful preprocessing and feature engineering are often required to make KNN work well on specific datasets, and in some cases, alternative machine learning algorithms may be more suitable.

6 Conclusion

In this report, a comprehensive analysis of liver disease prediction using the K-Nearest Neighbors (KNN) algorithm is conducted. The analysis began by preprocessing the dataset, addressing missing values, encoding categorical features, and oversampling to handle data imbalance. The KNN model achieved the highest accuracy of 74.00% when k = 5. Early detection of liver disease is critical for effective treatment, and the KNN model can serve as a valuable tool for assisting medical professionals in this regard.

References

[1] UCI Machine Learning Repository. [Online]. Available: https://archive.ics.uci.edu/ [Accessed: November 2, 2023].

[2] Fix, Evelyn; Hodges, Joseph L. (1951), *Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties*, Report, USAF School of Aviation Medicine, Randolph Field, Texas.