Heaven's Light is Our Guide

Computer Science & Engineering

Rajshahi University of Engineering & Technology

Course No: CSE 4204

Course Name: Sessional based on CSE 4203

Experiment No: 1

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

Submitted to

Rizoan Toufiq
Assistant Professor
Dept. of Computer Science & Engineering
Rajshahi University of Engineering & Technology

Submitted by

Mondol Mridul Provakar
Roll: 1803062
Dept. of Computer Science & Engineering
Rajshahi University of Engineering & Technology

Date of Submission: 04/11/2023

1 K-nearest Neighbor Classification algorithm

- Begin by defining the value of k, which represents the number of nearest neighbors to consider.
- Next, gather and organize the data that will be used for the analysis. This data should include a set of labeled training examples and a set of unlabeled test examples.
- For each test example, calculate the distance between the test example and each training example using a distance metric, such as Euclidean distance.
- Sort the training examples by their distance to the test example, with the closest training examples at the top of the list.
- Select the k training examples that are closest to the test example.
- Determine the majority label among the k training examples and assign that label to the test example.
- Repeat steps 3-6 for each test example, then evaluate the accuracy of the model by comparing the predicted labels to the true labels.
- If necessary, adjust the value of k or other parameters to improve the accuracy of the model.
- Once the algorithm is deemed accurate, it can be used to classify new examples.

2 Breast Cancer Dataset

Breast cancer is the most common cancer amongst women in the world. It accounts for 25% of all cancer cases, and affected over 2.1 Million people in 2015 alone. It starts when cells in the breast begin to grow out of control. These cells usually form tumors that can be seen via X-ray or felt as lumps in the breast area. The key challenges against it's detection is how to classify tumors into malignant (cancerous) or benign(non cancerous). The dataset has following characteristics:

- **Size:** The dataset consists of a total of 569 rows and 6 columns.
- **Data Types:** All values are numerical.
- **Features:** Mean_radius, Mean_texture, Mean_perimeter, Mean_area, Mean_smoothness.
- **Target Variable:** Diagnosis.

2.1 Exploratory Data Analysis

- **Correlation Heatmap:** A Heatmap was generated to visualize the correlations between continuous attributes. There were high correlations among the features: Mean_perimeter and Mean_area with Mean_radius. So this two column is removed from the dataset.

  Final Dataset Size: 569 rows and 4 columns.

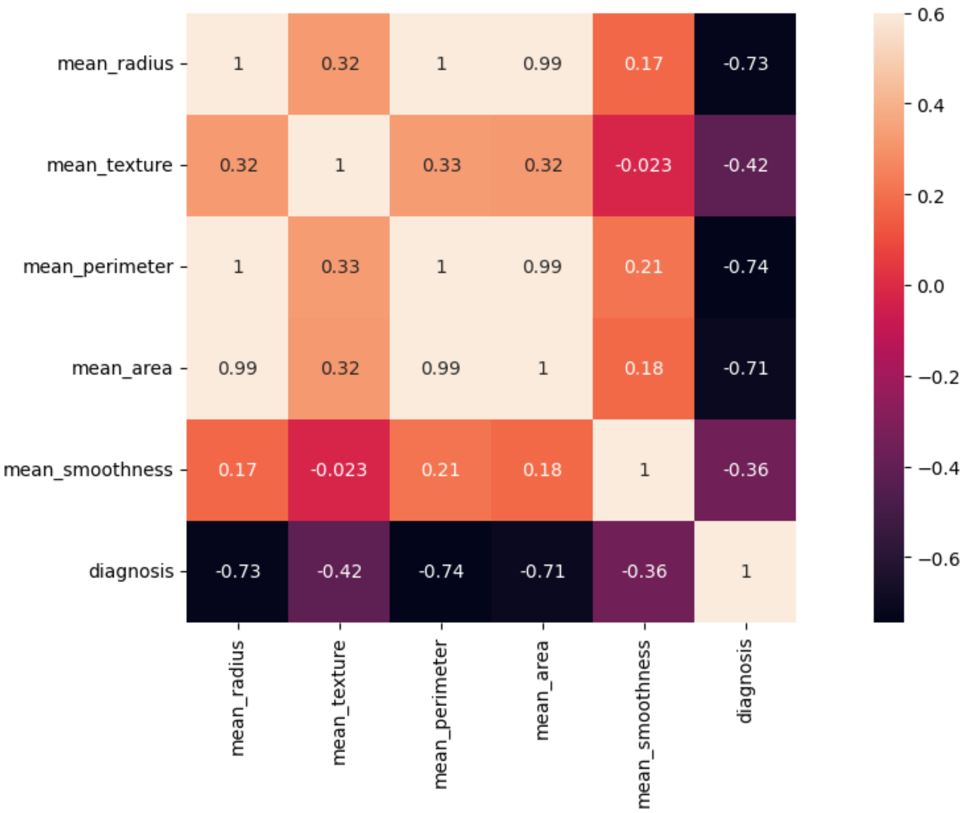


Figure 1: Correlation Heatmap

2.2 Training and Test Dataset Ratio

In this analysis, the dataset is divided into a training set and a test set with an 80/20 split. This means that 80% of the data is used for training the K-Nearest Neighbors (KNN) classifier, while the remaining 20% is reserved for testing and evaluating the classifier's performance.
After train & test split:

Training dataset size: 455 rows and 4 columns.
Test dataset Size: 114 rows and 4 columns.

2.3 Balancing the dataset

Positive class: 357
Negative class: 212
The dataset is imbalanced. To balance it, SMOTE is used. The sampling streategy is set as 'auto'. After applying, the size of both positive & negative class is 357. K-NN algorithm is applied separately on both balanced and imbalanced dataset.
After dataset balancing:

Training dataset size: 571 rows and 4 columns.
Test dataset Size: 143 rows and 4 columns.

3 K-Nearest Neighbor (KNN) Classifier

The K-Nearest Neighbors (KNN) classifier is a simple and popular machine learning algorithm used for classification and regression tasks. It's a type of instance-based or lazy learning algorithm. KNN is often used for its simplicity and effectiveness, especially when dealing with small to medium-sized datasets.

3.1 K-NN Implementation

In this analysis, a custom K-NN class was developed to perform classification tasks on the Abalone dataset. The custom KNN class includes the following key functionalities:

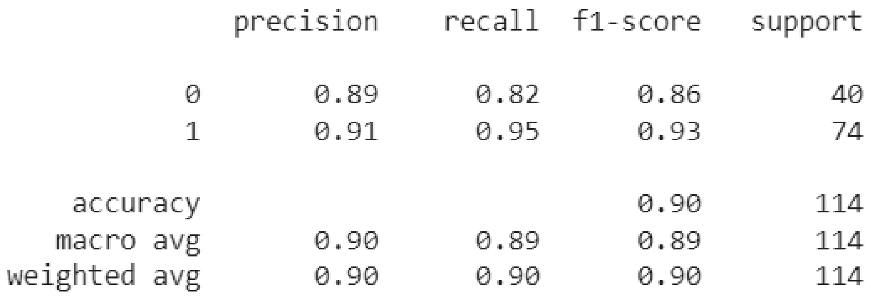
- The data is loaded
- K is initialized to chosen number of neighbors
- For each example in the data
    1. The distance between the query example and the current example is calculated from the data.
    2. The distance and the index of the example is added to an ordered collection
- The ordered collection of distances and indices is sorted from smallest to largest (in ascending order) by the distances
- The first K entries is picked from the sorted collection
- The labels is got of the selected K entries
- The mode of the K labels is returned

The accuracy of the dataset is calculated with different k values. For this dataset, highest accuracy is obtained by choosing the value as 5.

3.2 Model Evaluation

For imbalanced dataset

- **Accuracy Score:** 90%
- **Classification Report:**



```
              precision    recall  f1-score   support

           0       0.89      0.82      0.86        40
           1       0.91      0.95      0.93        74

    accuracy                           0.90       114
   macro avg       0.90      0.89      0.89       114
weighted avg       0.90      0.90      0.90       114
```

Figure 2: Classification Report

- **Confusion Matrix:**

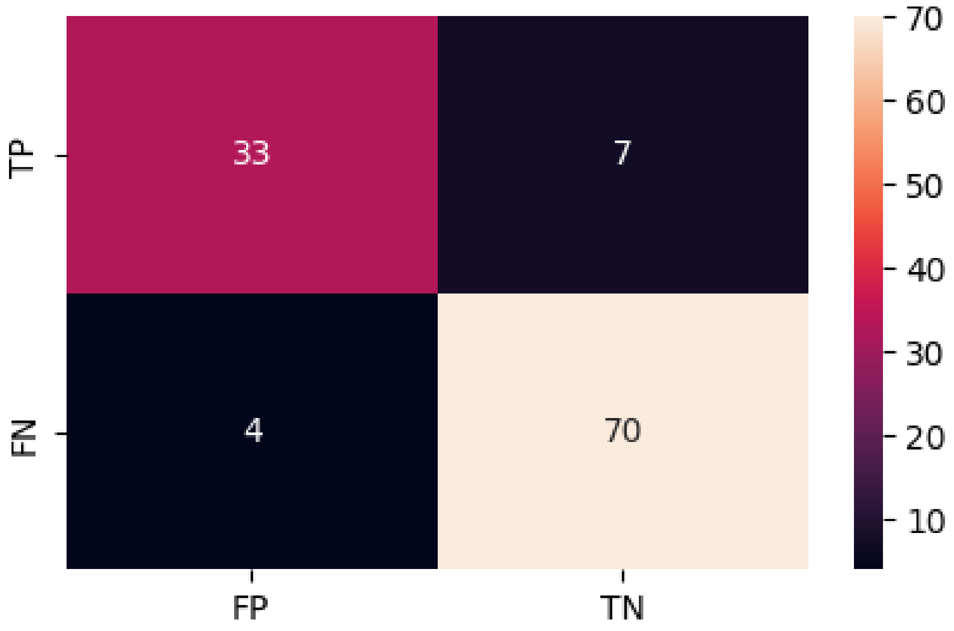
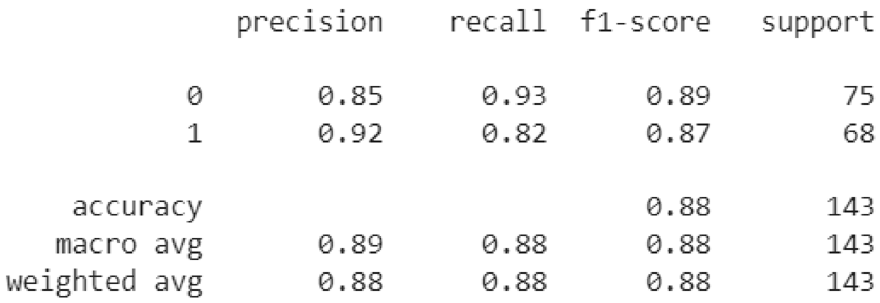


Figure 3: Confusion Matrix

For balanced dataset

- **Accuracy Score:** 88%
- **Classification Report:**



```
              precision    recall  f1-score   support

           0       0.85      0.93      0.89        75
           1       0.92      0.82      0.87        68

    accuracy                           0.88       143
   macro avg       0.89      0.88      0.88       143
weighted avg       0.88      0.88      0.88       143
```

Figure 4: Classification Report

- **Confusion Matrix:**

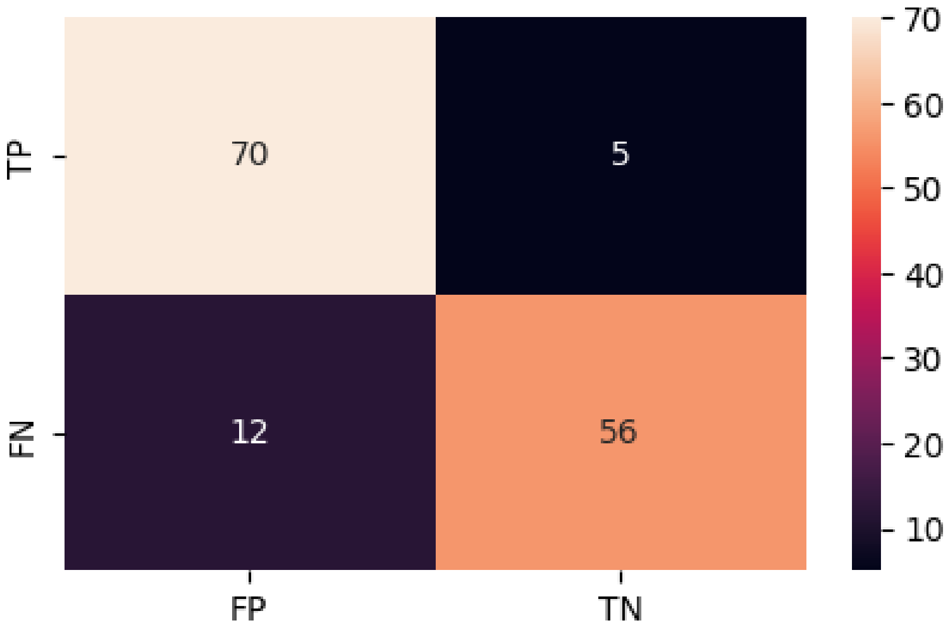


Figure 5: Confusion Matrix

4 Advantages & Disadvantages of K-NN

Advantages:

- **Easy to implement:** The complexity of the algorithm is not that high.
- **Adapts easily:** As per the working of the KNN algorithm it stores all the data in memory storage and hence whenever a new example or data point is added then the algorithm adjusts itself as per that new example and has its contribution to the future predictions as well.
- **Few hyperparameters:** The only parameters which are required in the training of a KNN algorithm are the value of k and the choice of the distance metric which we would like to choose from our evaluation metric.

Disadvantages:

- **Does not scale:** KNN algorithm is also considered a Lazy Algorithm. The main significance of this term is that this takes lots of computing power as well as data storage. This makes this algorithm both time-consuming and resource exhausting.
- **Curse of Dimensionality:** There is a term known as the peaking phenomenon according to this the KNN algorithm is affected by the curse of dimensionality which implies the algorithm faces a hard time classifying the data points properly when the dimensionality is too high.
- **Prone to Overfitting:** As the algorithm is affected due to the curse of dimensionality it is prone to the problem of overfitting as well. Hence generally feature selection as well as dimensionality reduction techniques are applied to deal with this problem.

References

[1] Neural Computing: An Introduction - R Beale and T Jackson

[2] K-NN Basics: ”https://www.geeksforgeeks.org/k-nearest-neighbours”

[3] Breast Cancer Dataset ”https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset”

[4] Source Code ”https://colab.research.google.com/drive/1UNuH_pJZDdN031X6jDCG7g0BhwO5VUfY#sc
E74GsXz04”