Hearen's Light is Our Guide

Computer Science & Engineering
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.

Course Outcomes: CO1

Learning Domain with Level: Cognitive (Applying, Analyzing, Evaluating & Creating)

Contents

1 Dataset

I've Selected Diabetes Dataset from Kaggle [1] .This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective is to predict based on diagnostic measurements whether a patient has diabetes.

1.1 Dataset Analysis

The Dataset have 8 attributes and 768 instances. The Features are :

1. **Pregnancies:** Number of times pregnant
2. **Glucose:** Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. **BloodPressure:** Diastolic blood pressure (mm Hg)
4. **Insulin:** 2-Hour serum insulin (mu U/ml)
5. **BMI:** Body mass index (weight in $kg/(heightinm)^2$)
6. **DiabetesPedigreeFunction:** Diabetes pedigree function
7. **Age:** Age (years)
8. **Outcome:** Class variable (0 or 1)

Class Distribution: class value 1 is interpreted as "tested positive for diabetes".

1.2 Training-Test Ratio

60% of data were used for training and 40% of data for validating .

2 Feature Selection

K - Nearest Algorithm was applied for only 3 selected features. At First the Correlation Matrix was piloted . After the correlation matrix histogram of every features was plotted. Based on these two , a 3D graph was plotted to get a idea. As KNN gives decision based on its neighbour values.

2.1 Correlation Matrix

For Feature selection a correlation matrix was printed for understating the correlation between different features . From the correlation matrix it is shown that the 'Glucose' attribute is highly correlated with the class variable. Then 'BMI' and age attributes are also more correlates with class variable than others . The correlation matrix is shown below:

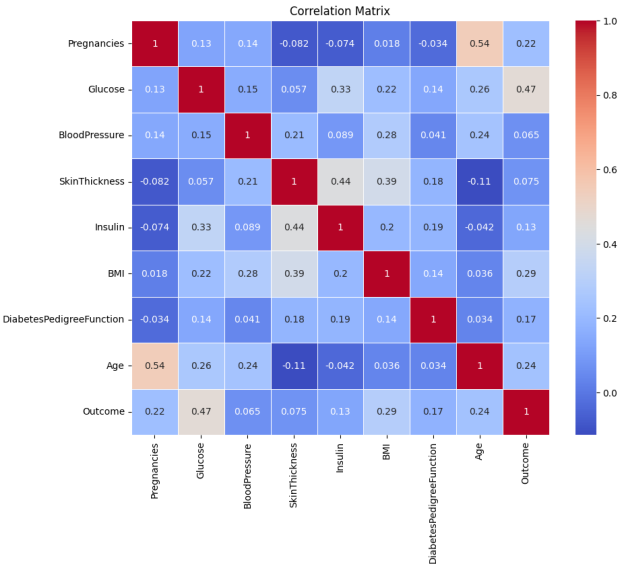

Figure 1: Correlation Matrix

2.2 Histogram

Histogram of every feature is shown below:

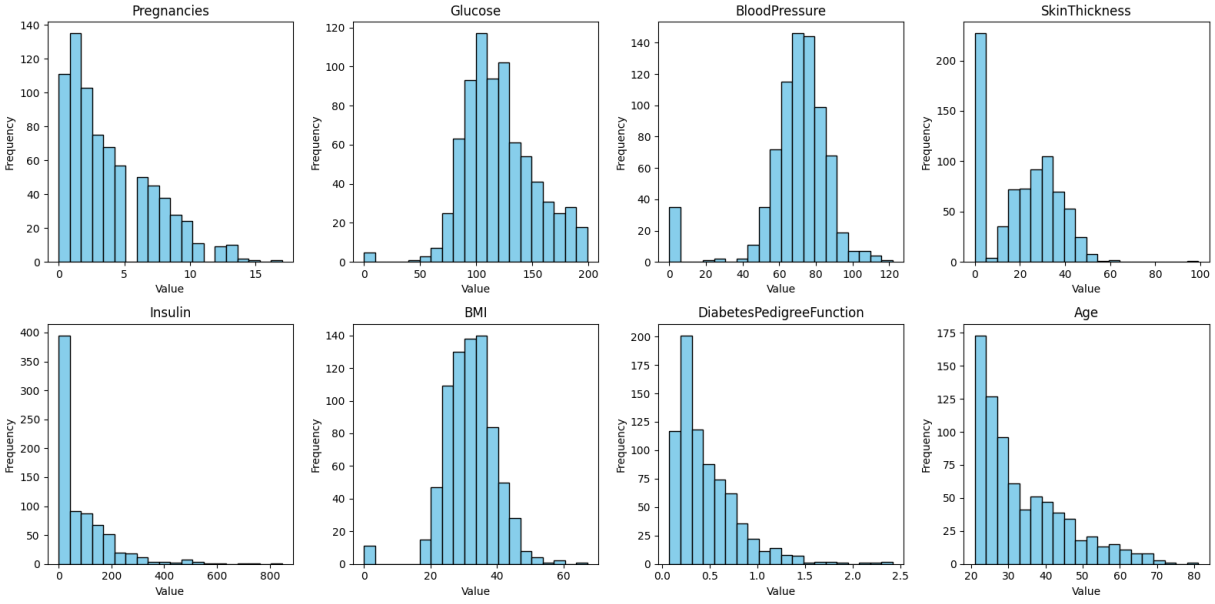

Figure 2: Histogram of every feature

Based of Correlation Matrix and Histogram It was shown that 'Glucose ' ,'BMI' , 'Age' Attributes are more correlated with output than others. So , a 3D graph was plotted taking these 3 attributes .

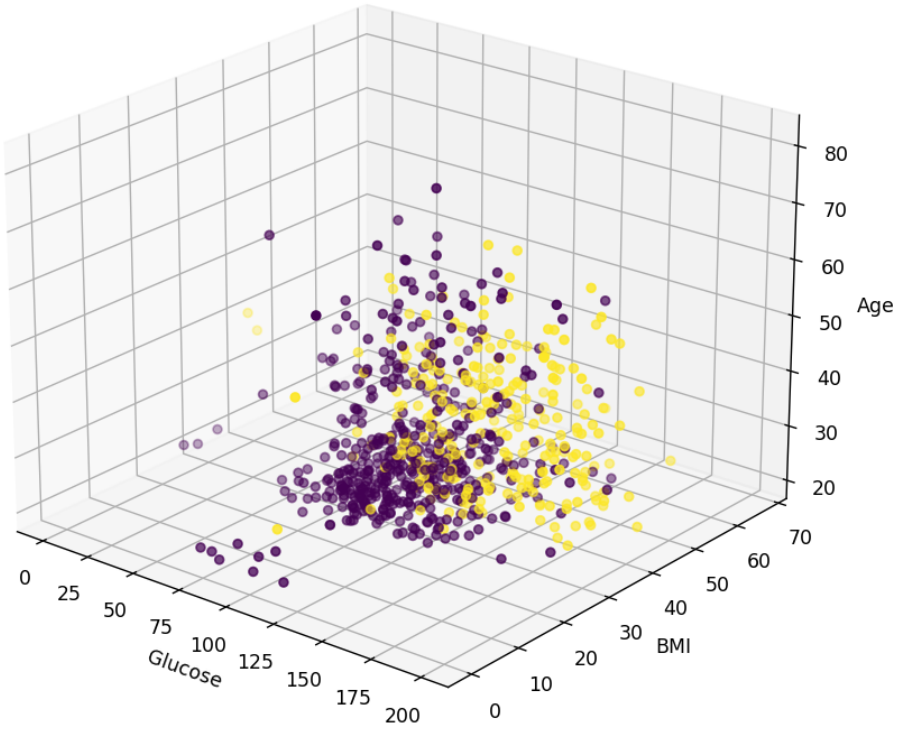

Figure 3: 3D scatter Plot of Glucose , BMI & Age

3 K-nearest Neighbor Classification algorithm

K-Nearest Neighbours is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining, and intrusion detection [2] . The Algorithm were described as follows :

1. Begin by defining the value of k, which represents the number of nearest neighbors to consider.
2. 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.
3. For each test example, calculate the distance between the test example and each training example using a distance metric, such as Euclidean distance
4. Sort the training examples by their distance to the test example, with the closest training examples at the top of the list.
5. Select the k training examples that are closest to the test example.

6. Determine the majority label among the k training examples and assign that label to the test example.

7. Repeat steps 3-6 for each test example, then evaluate the accuracy of the model by comparing the predicted labels to the true labels.

8. If necessary, adjust the value of k or other parameters to improve the accuracy of the model.

9. Once the algorithm is deemed accurate, it can be used to classify new examples.

4 Accuracy Analyzing

For Analyzing the accuracy, the algorithm was tested with different k value of 5 to 25 . A graph was plotted to analyze the accuracy is shown below :

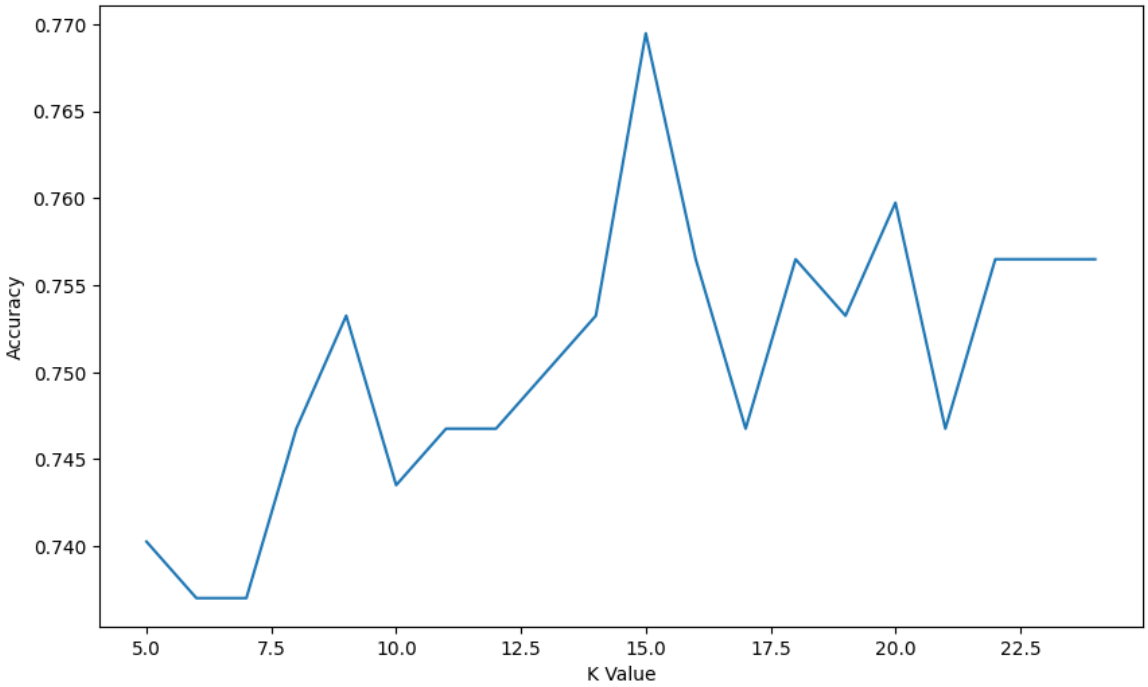

Figure 4: Accuracy analysis with different K value .

From the figure, it is clearly shown that accuracy is maximum when K = 15 which is 0.769 . So K = 15 is the optimal K value for the dataset.

5 Conclusion

For the chosen dataset the K- Nearest Neighbour algorithm did not perform as expected because there were zero values in some attributes . Also there were noise in the dataset. K-NN is sensitive to noisy data points and outliers. If the dataset contains noisy or outlier data, it can significantly affect classification accuracy. If the noises can be removed from the dataset ,it would perform better.

References

[1] Diabetes Dataset. [Online]. Available at: https://www.kaggle.com/datasets/mathchi/diabetes-data-set

[2] Geeksforgeeks. [Online]. Available at : https://www.geeksforgeeks.org/k-nearest-neighbours