****

# AMERICAN INTERNATIONAL UNIVERSITY–BANGLADESH(AIUB)

**FACULTY OF SCIENCE & TECHNOLOGY**

# DEPARTMENT OF CSE

**Introduction to Data Science**

**Fall 2022-2023**

**Section: D**

**FINAL Term Project**

**Submitted to**

**Tohedul Islam  
 Assistant Professor, CSE**

**Submitted By**

|  |  |
| --- | --- |
| **Name** | **ID** |
| **MD. ALI AHNAF** | **20-42378-1** |

**Dataset name & Description**

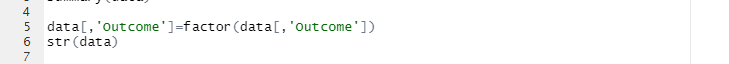
The dataset is selected from Kaggle which was the original dataset of the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether a patient has diabetes, based on certain diagnostic measurements included in the dataset. All patients here are females at least 21 years old of Pima Indian heritage.  
The source-link of the dataset: *https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database*



A picture containing text

Description automatically generated  
  
The dataset has 9 attributes/variables where the class variable is an integer. There are total of 768 observations hence the dataset has 768 of rows.

Factoring the class variable to generate the Confusion Matrix because the data and the reference value must have to be factors and have the same no. of levels.



Here, Outcome is the class variable of the dataset.

A picture containing letter

Description automatically generated

Normalizing the dataset where the values of each instance are between 0 to 1 and excluding the class variable (Outcome).

Text

Description automatically generated

Text

Description automatically generated with medium confidence

The dataset has been split into two parts, Training and Validation/test set where 80 percent of the data were selected to train the classification model and the rest for validating the performance of the model.  
Here pseudorandom samples were generated and selected for both the training and rest for test set therefore the seed was initialized early for reproducibility.

Graphical user interface, application

Description automatically generated

Text

Description automatically generated

The labels for the class (class labels) were stored in two variables for applying those for classification models.

KNN algorithm was applied, and data were predicted for the validation set where K value was chosen by calculating square root of the total number of samples/observations in the training dataset. (Square root method)

**Confusion Matrix**

Background pattern

Description automatically generated with low confidence

Graphical user interface, text, application

Description automatically generated

Here the confusion matrix was generated. In this table for predicted class variable values for test dataset, it was compared with the reference / reference datasets selected test data.

Here in the first quadrant in the confusion matrix table, true positive value is 99 which means 99 of the patients was actually non-diabetic from true class (reference data) and the model also predicted it correctly as non-diabetic.

Moving into the 2nd quadrant, the false positive value is 17 therefore 17 patients did had diabetes, but they were incorrectly predicted as non-diabetic.

Moving into the 3rd quadrant, the false negative value is 11 therefore 11 patients did not had diabetes, but they were incorrectly predicted as diabetic.

Moving into the 4th quadrant, the true negative value is 27 therefore 27 patients were diabetic, and they were correctly predicted as diabetic.  
  
The accuracy of the classifier was measured by the total sum of the diagonal value mainly the True Positive (TP) and the True Negative (TN) value divided by the number of observed test data.   
Hence, we get around 81.82% of accuracy from the KNN classification model by validating it with the unseen test/validation set.



Text

Description automatically generated