Machine Learning, Spring 2023: Project 1

KAMAL SUBEDI(G35111003)

Platform: Jupyter Notebook  
Device: MacOS M2  
Programming language: Python3, Python libraries  
Dataset: https://www.kaggle.com/uciml/pima-indians-diabetes-database

**Data Description:**

This dataset contains data on 768 women of Pima Indian heritage age 21 and above and shows whether they have diabetes or not. Using various factor like BMI, Blood pressure, and age etc, we can predict that whether they have diabetes or not. The dataset consists of several medical predictor (independent) variables and one target (dependent) variable. Independent variables include the number of pregnancies the patient has had, their BMI, Blood Pressure, DiabetesPedigreeFunction, insulin level, age and independent output variable. The outcome variable is a binary variable indicating whether the patient has diabetes or not.

**Algorithm Description**

For the K-NN classifier, we will use the medical predictor variables to predict whether or not a patient will develop diabetes. Our target variable will be the binary outcome variable indicating diabetes status.

Before building the K-NN classifier, we need to split the dataset into training and testing sets. We will use a 80-20 split, where 80% of the data will be used for training and 20% will be used for testing. We will randomly select the observations for each set.

I choose a 80-20 split because we want to have enough data to build an accurate classifier, but we also want to have enough data to evaluate its performance. Using a larger training set may result in a more accurate classifier, but it may also increase the risk of overfitting to the training data, resulting in poor performance on new, unseen data. Additionally, using a smaller testing set may increase the variability of the evaluation metrics, making it harder to draw conclusions about the classifier's performance. 80-20 split strikes a reasonable balance between these concerns. Then use, Euclidean and Manhattan distance metric to predict the neighbor and classify.

Below, we can see the scatter plot of feature variable and target variable. From the plot below, we can see that Glucose and BMI highly influence the target variable. So, we are choosing these variables as feature variable and dropping other independent variables.

A picture containing chart

Description automatically generated

Fig: Scatter plot

**Algorithm Result:**

We got the confusion matrix and accuracy value when k=5.

Confusion matrix = [[82 17]

[23 32]]

A picture containing text, clock

Description automatically generatedChart, treemap chart

Description automatically generated  
**Fig: Confusion matrix**



Chart, line chart

Description automatically generated

Fig: Graph between accuracy and K.

Here, we can see that, the accuracy is maximum is when K=5. And the K is increasing the accuracy score got deceasing after k=7.

**Runtime:**

