Department of Computer Science & Engineering

Final Year B. Tech. (CSE) – I: 2022-23

5CS462: PE5 - Data Mining Lab

Assignment No. 5

Name: Sanket Mote

PRN: 2019BTECS00113

Batch: B8

Title: Design and implement the following classifiers:

a) Regression classifier.

b) Naive Bayesian Classifier.

c) k-NN classifier

d) Three-layer Artificial Neural Network (ANN) classifier (use back propagation). Plot error graph (iteration vs error).

Objective/Aim:

1. **To implement data analysis tool using python programming language.**
2. **To design the Regression, Naïve Bayes, k-NN, Three-layer Artificial Neural Network (ANN) classifier.**

Theory:

A) Regression classifier:

Regression algorithms predict a continuous value based on the input variables. The main goal of regression problems is to estimate a mapping function based on the input and output variables. If your target variable is a quantity like income, scores, height or weight, or the probability of a binary category (like the probability of rain in particular regions), then you should use the regression model. However, there are various types of regressions used by data scientists and ML engineers based on different scenarios. The different types of regression algorithms include:

1. Simple linear regression

With simple linear regression, you can estimate the relationship between one independent variable and another dependent variable using a straight line, given both variables are quantitative.

2. Multiple linear regression

An extension of simple linear regression, multiple regression can predict the values of a dependent variable based on the values of two or more independent variables.

3. Polynomial regression

The main aim of polynomial regression is to model or find a nonlinear relationship between dependent and independent variables.

B) Naïve Bayesian Classifier:

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other

Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation:

$$\mathrm{P(A|B)} = \frac{P(B|A)P(A)}{P(B)}$$

where A and B are events and P(B) ≠ 0.

Basically, we are trying to find probability of event A, given the event B is true. Event B is also termed as evidence.

P(A) is the priori of A (the prior probability, i.e. Probability of event before evidence is seen). The evidence is an attribute value of an unknown instance(here, it is event B).

P(A|B) is a posteriori probability of B, i.e. probability of event after evidence is seen.

C) k-NN classifier:

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.

K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.

It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

Example: Suppose, we have an image of a creature that looks similar to cat and dog, but we want to know either it is a cat or dog. So for this identification, we can use the KNN algorithm, as it works on a similarity measure. Our KNN model will find the similar features of the new data set to the cats and dogs images and based on the most similar features it will put it in either cat or dog category.

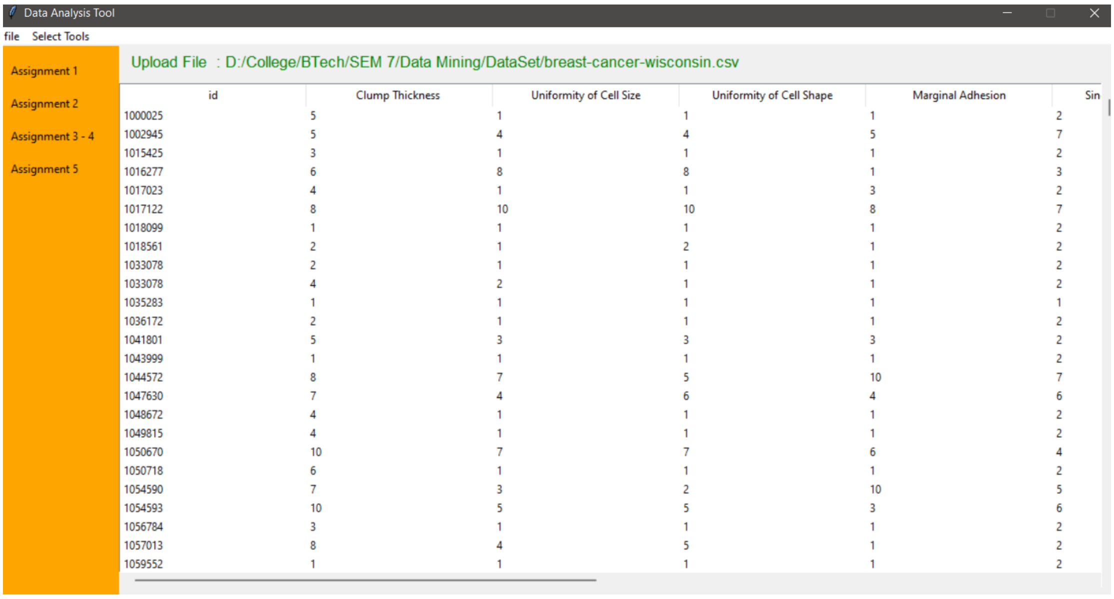
D) Artificial Neural Network (ANN) classifier:

Classification ANNs seek to classify an observation as belonging to some discrete class as a function of the inputs. The input features (independent variables) can be categorical or numeric types, however, we require a categorical feature as the dependent variable**.**

Procedure:

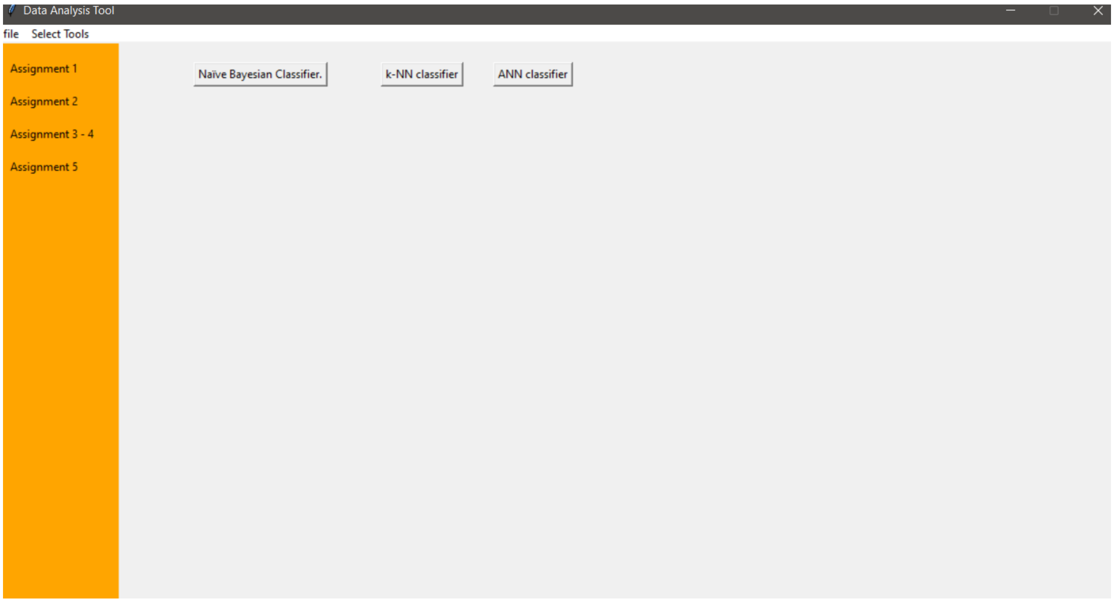
Given problem statement is solved using python programming language and specifically used tkinter module to implement GUI application and pandas module to load .csv file as dataset.

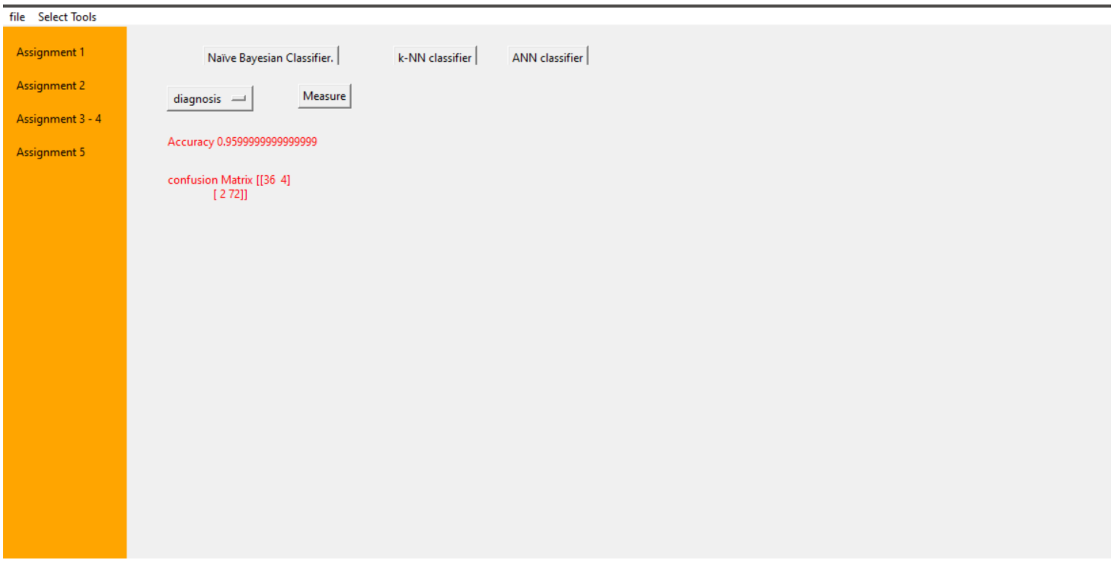
Results:



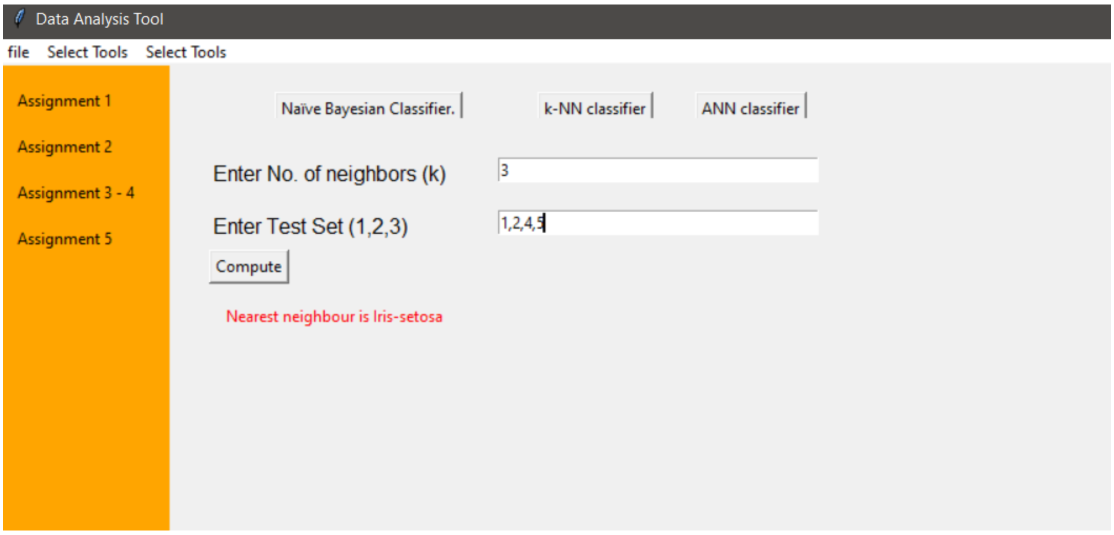
Upload File : D:/College/BTech/SEM 7/Data Mining/DataSet/breast-cancer-wisconsin.csv

id	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Sin
1000025	5	1	1	1	2
1002945	5	4	4	5	7
1015425	3	1	1	1	2
1016277	6	8	8	1	3
1017023	4	1	1	3	2
1017122	8	10	10	8	7
1018099	1	1	1	1	2
1018561	2	1	2	1	2
1033078	2	1	1	1	2
1033078	4	2	1	1	2
1035283	1	1	1	1	1
1036172	2	1	1	1	2
1041801	5	3	3	3	2
1043999	1	1	1	1	2
1044572	8	7	5	10	7
1047630	7	4	6	4	6
1048672	4	1	1	1	2
1049815	4	1	1	1	2
1050670	10	7	7	6	4
1050718	6	1	1	1	2
1054590	7	3	2	10	5
1054593	10	5	5	3	6
1056784	3	1	1	1	2
1057013	8	4	5	1	2
1059552	1	1	1	1	2





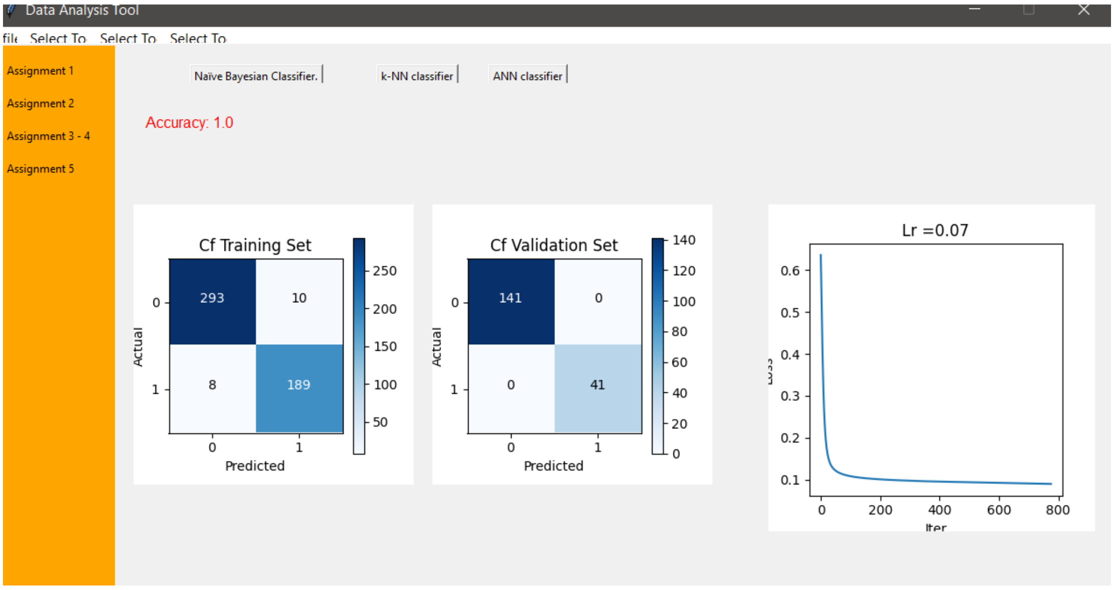
```
Accuracy 0.9599999999999999

confusion Matrix [[36  4]
          [ 2 72]]
```



Enter No. of neighbors (k) 3

Enter Test Set (1,2,3) 1,2,4,5

Nearest neighbour is Iris-setosa



Conclusion:

Successfully implemented data analysis tool (GUI) for extraction of rules from decision tree build based on information gain, gain ratio and gini index for selected attribute of given dataset.