Date:-

Assignment No:2.

Title:- Naive Bayes Classification.

Problem Statement:-

Download PIMA Indians Diabetes Dataset. Use Naive Bayes Algorithm for classification. Load the data from CSV file and split it into training and test datasets. Summarize the properties in the training dataset so that we can calculate probabilities in the training dataset so that we can make predictions. Classify samples from a test dataset and a summarized training dataset.

Objective:-

Understand Naive Bayes Algorithm for classification and use it on Pima Indians dataset.

Outcome:-

Predict whether the person has diabetes or not using Naive Bayes classification based on parameters in dataset like Blood Pressure, Glucose, Insulin, BMI.

Softtware & Hardware Requirements:-
64 bit OS (UNIX/LINUX), Python 3, Jypyter, numpy, pandas, seaborn, 8 GB Ram, i5 processor,

Theory:-
Naive Bayes Classifiers are a family of simple probabilistic Classifiers.
They are based on Bayes Theorem, which describes the probability of a certain event occuring, based on the prior knowledge of conditions that might be related to the event.
Bayss theorem is stated mathematically

$$P(A|B) = \frac{P(B/A)\,P(A)}{P(B)}.$$

where A, B are the events.

P(A|B) is a conditional probability, The likelihood of event A occuring knowing that B is true;
P(B|A) is also conditional, The likelihood of B occuring knowing that A is true.
P(A) and P(B) are marginal probabilities.

Naive Bayes is a technique for constructing classifiers, which applies the above theorem, with the strong (naive) assumption that the features are largely independent

These models assign class labels (in this case, 'Diabetic' or 'Non-Diabetic') to problem instances, represents as vectors of feature values. The class labels are drawn from a finite set.

A family of algorithms based on one common principle from the Naive Bayes classifier, the principle is that a particular feature is independent of the value of any other feature given the class variable each feature contributes independently to the probability of the positive outcome, regardless of any possible correlations between the features.

Abstractly Naive Bayes is a conditional probability model, and can be trained very efficiently in a supervised learning.

Despite its Naive design and apparently oversimplified assumptions, Naive Bayes classifiers have proven to work quite well in real world settings.

About the Dataset:-

The dataset is originally from the National institute of Diabetes and Digestive and kidney Diseases.

The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measures included.

Several constraints were placed on the selection of these instances from the larger database; in particular, all patients here are at least 2 years old, and are females of Pima Indian heritage.

Conclusion:-

The Naive Bayes classifier was successfully applied to the cleaned dataset, and the outcome (diabetes diagnosis) was predicted, with an accuracy of 74%.