Assignment - DA2

* TITLE: Naive Bayes Classification.

* PROBLEM STATEMENT: Download PIMA Indians Diabetes dataset. Use Naive Bayes Algorithm for classification. Load the data from csv file & split it into training and test database so that we can calculate probabilities and make predictions classify samples from a test dataset & 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 on parameters in dataset like Blood pressure, BMI.

* SOFTWARE & HARDWARE REQUIREMENTS: 64 bit OS (UNIX/LINUX), python 3, Jupyter, numpy, pandas, seaborn, 8GB RAM, i5 processor, 128 GB SSD.

* 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 occurring, based on prior knowledge of cond^n.

Bayes theorem is stated mathematically.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$ where A, B are events.

P(A|B) is a conditional probability the likelihood of event A occurring knowing that B is true; P(B|A) is also conditional, the likelihood of B occurring knowing that A is true.

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

are largely independent.

These models assign class labels to problem instances represented as vectors of feature values. The class labels are drawn from finite set

Principle of Naive Bayes classifier is:

'a particular feature is independent of the value of any other feature, given the class variable; each feature contributes independently to the probability of positive outcome, regardless of any possible correlation'

Abstractly, Naive Bayes is a conditional probability model, it can be trained very efficiently in a supervised learning. Despite its naive baye design & apparently oversimplified assumptions, it have proven to work well in real world settings.

About the dataset:

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

Several constraints were placed on the selection of these instances from the larger databases; in particular, all patients here at least 21 yrs old, & are females of Pima Indian heritage.

CONCLUSION:

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