41145

Page No.
Date

LP1 - BDA2

* Title: Naive Bayes Classification

* Problem Statement:
Download Pima Indians Dataset. Use Naive Bayes Algo. for classification. Load the data from the CSV file and split it into training and test data sets. Summarize the properties in the training dataset so that we can calculate probabilities and make predictions.
Classify samples from a test dataset and a summarized train. data set.

* Objectives:
Understand Naive Bayes Algorithm for classification, and use it on Pima Indians Dataset.

* S/W and H/W requirements:
i) 64 bit processor
ii) RAM
iii) Linux OS
iv) Python 3

* Theory:
i) Naive Bayes Classifiers are a family of simple, probabilistic classifiers.
ii) They are based on Baye's theorem, which describes the probability of a certain event occuring, based on the prior knowledge of conditions that might be related to

the event.

Baye's Theorem : $P(A/B) = \frac{P(B/A) \cdot P(A)}{P(B)}$

where, A, B: events

$P(A/B)$ → conditional probability, the likelihood of event A occuring, knowing that B occurs.

$P(B/A)$ → conditional probability, the likelihood of event B occuring, knowing that A occurs.

$P(A)$ and $P(B)$ → marginal probabilities.

This model assigns Class labels (in this case, 'Diabetic' / 'Non-diabetic') to problem instances, represented as vectors of features 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 →

"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

41145

Page No.
Date

between the features".

* Conclusion:

From this assignment, I was able to understand the Naive Bayes Algorithm and hence apply the same to the Pima Indians dataset.