LP1 Assignment DA2

Naive Bayes algorithm for classification on Pima Indians Diabetes dataset.

Date – 2nd September, 2020.

Assignment Number – DA2

Title:

Naive Bayes algorithm for classification on Pima Indians Diabetes dataset.

Problem Definition:

1) Download Pima Indians Diabetes dataset. Use Naive Bayes Algorithm for classification

2) Load the data from CSV file and split it into training and test datasets.

3) Summarize the properties in the training dataset so that we can calculate probabilities and make predictions.

4) Classify samples from a test dataset and a summarized training dataset.

Learning Objectives:

1) Learn Naive Bayes algorithm

2) Learn to summarize the properties in the training dataset.

3) Learn to split dataset into training and test datasets

4) Learn to classify samples from a test dataset and a summarized training dataset.

Software Packages and Hardware Apparatus Used:

- Operating System: Windows 10
- Programming Language: Python 3
- Jupyter Notebook Environment: Google Colaboratory

● Python Libraries: Numpy, Pandas, MatPlotLib, Seaborn

Related Mathematics:

Let S be the system set: S = {s; e;X; Y; Fme;DD;NDD; Fc; Sc}

where Dataset is loaded into the dataframe

s=start state

e=end state i.e. classification of samples from the test dataset

X=set of inputs

X = {X1}

Where,

X1 = Pima Indians Diabetes dataset

Where,

Y=set of outputs

1) Splitting of dataset into training and test datasets

2) Naive Bayes classifier

Fme is the set of main functions Fme = {f1,f2,f3} where

f1 = function to load dataset into dataframe

f2 = function to split dataset into training and test datasets

f3 = function to invoke Naive Bayes classifier

DD= Deterministic Data

PIMA Indians diabetes dataset

Concepts related Theory

Naive Bayes algorithm for classification:

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.

To start with, let us consider a dataset The datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variables includes:

- preg = Number of times pregnant
- plas = Plasma glucose concentration
- blood_pres = Diastolic blood pressure (mm Hg)
- skin = Triceps skin fold thickness (mm)
- insulin = 2-Hour serum insulin (mu U/ml)
- bmi = Body mass index (weight in kg/(height in m)^2)
- pedigree = Diabetes pedigree function
- age = Age (years)
- diab_class = Class variable (1: tested positive for diabetes, 0: tested negative for diabetes)

What is Naive Bayes Classifier?

Naive Bayes is a statistical classification technique based on Bayes Theorem. It is one of the simplest supervised learning algorithms. Naive Bayes classifier is the fast, accurate and reliable algorithm. Naive Bayes classifiers have high accuracy and speed on large datasets. Naive Bayes classifier assumes that the effect of a particular feature in a class is independent of other features. For example, a loan applicant is desirable or not depending on his/her income, previous loan and transaction history, age, and location. Even if these features are

interdependent, these features are still considered independently. This assumption simplifies computation, and that's why it is considered as naive. This assumption is called class conditional independence.

Formula: P(h|D)=P(D|h)P(h)/P(D)

•P(h): the probability of hypothesis h being true (regardless of the data). This is known as the prior probability of h.

•P(D): the probability of the data (regardless of the hypothesis). This is known as the prior probability.

•P(h|D): the probability of hypothesis h given the data D. This is known as posterior probability.

•P(D|h): the probability of data d given that the hypothesis h was true. This is known as posterior probability.

How Naive Bayes classifier works?

Let’s understand the working of Naive Bayes through an example. Given an example of weather conditions and playing sports. You need to calculate the probability of playing sports. Now, you need to classify whether players will play or not, based on the weather condition. First Approach (In case of a single feature) Naive Bayes classifier calculates the probability of an event in the following steps:

•Step 1: Calculate the prior probability for given class labels

•Step 2: Find Likelihood probability with each attribute for each class

•Step 3: Put these value in Bayes Formula and calculate posterior probability.

•Step 4: See which class has a higher probability, given the input belongs to the higher probability class.

Advantages

•It is not only a simple approach but also a fast and accurate method for prediction.

•Naive Bayes has very low computation cost.

•It can efficiently work on a large dataset.

•It performs well in case of discrete response variable compared to the continuous variable.

•It can be used with multiple class prediction problems.

•It also performs well in the case of text analytics problems.

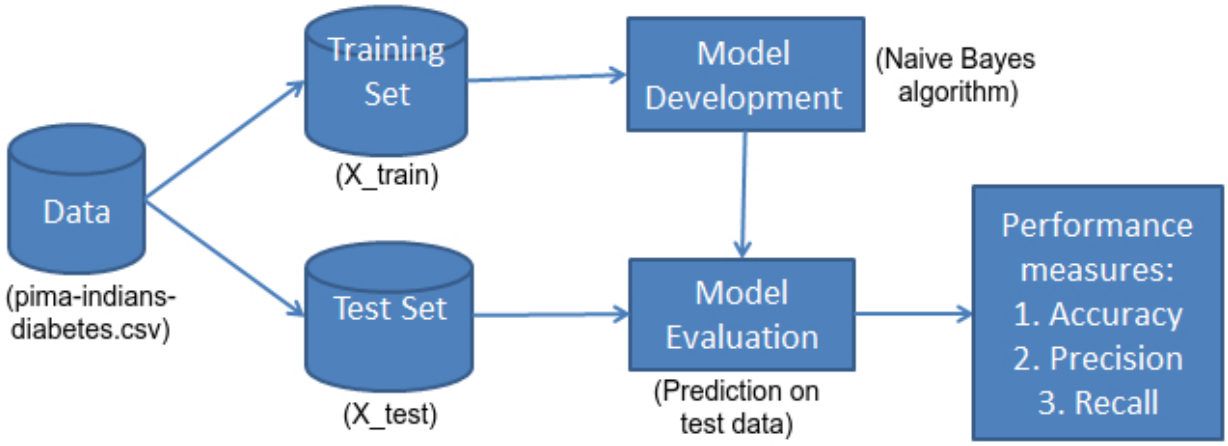
•When the assumption of independence holds, a Naive Bayes classifier performs better compared to other models like logistic regression.

Disadvantages

•The assumption of independent features. In practice, it is almost impossible that model will get a set of predictors which are entirely independent.

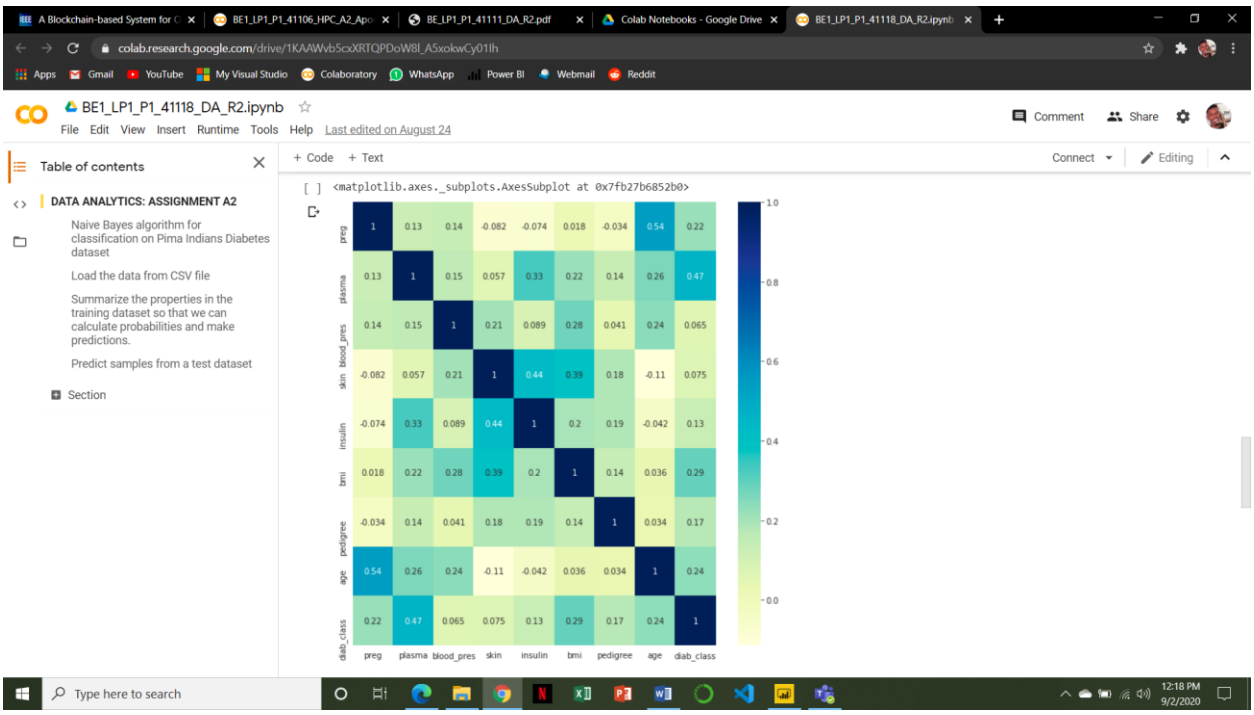
•If there is no training tuple of a particular class, this causes zero posterior probability. In this case, the model is unable to make predictions. This problem is known as Zero Probability/Frequency Problem

Flowchart

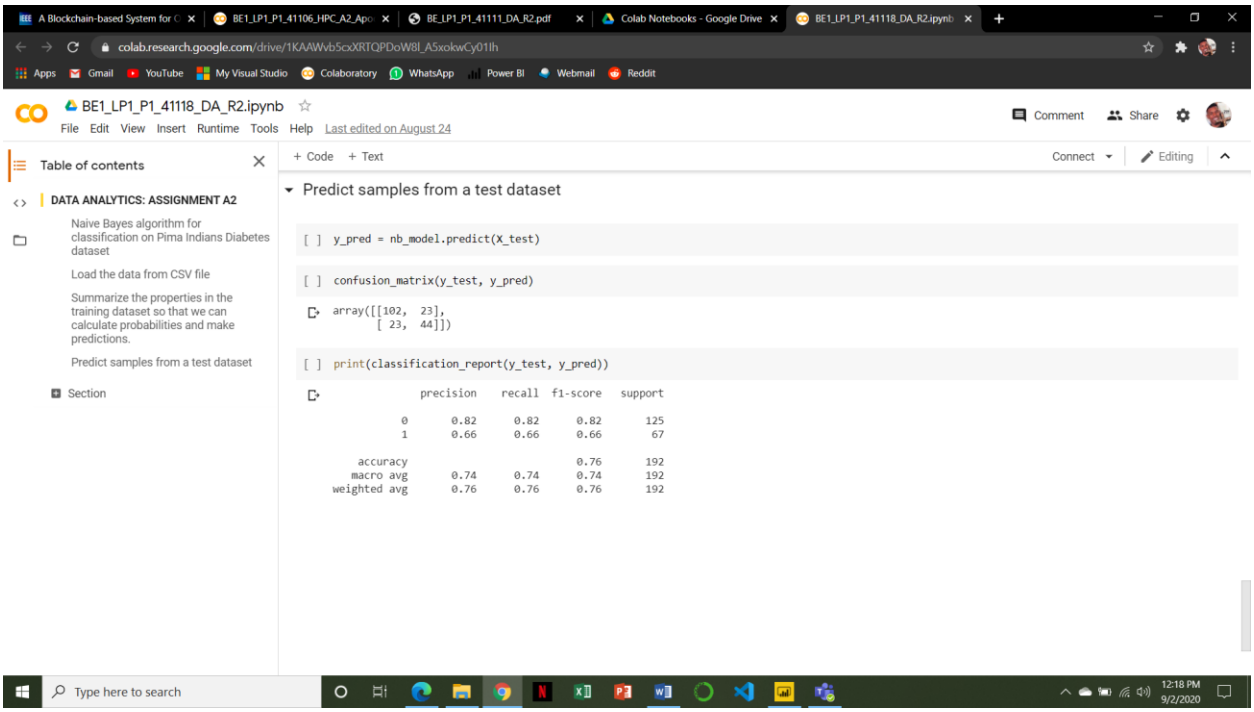


Output Screenshots –

Correlation Heatmap(Optimal viewing for colorblind people)



Confusion Matrix and Accuracy



Notebook Link –

https://colab.research.google.com/drive/1KAAWvb5cxXRTQPDoW8l_A5xokwCy01Ih?usp=sharing

Conclusion –

Hence we have successfully used Naive Bayes algorithm for classification on Pima Indians Diabetes dataset by loading, splitting the data into train and test and classifying as to whether people have diabetes or not.