

Naive Bayes



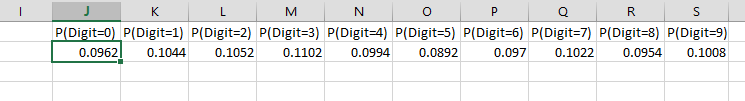
Course: ALY 6020 Predictive Analytics

Instructor: Dr. Marco Montes de Oca

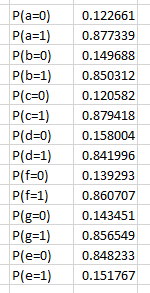
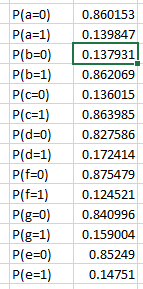
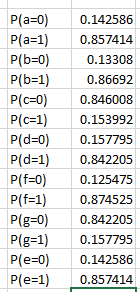
Submitted by: Nithish Saravanan

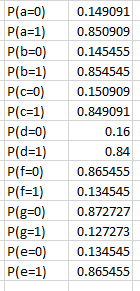
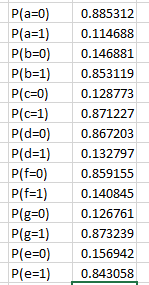
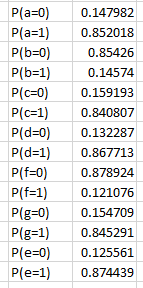
Naïve Bayes is a simple classifier which classifies using maximum posterior decision rule. This is applied in text classification and spam detection in email system. This uses the basic principles of Bayes theorem.

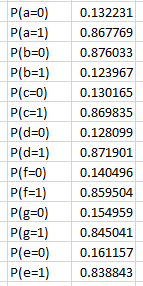
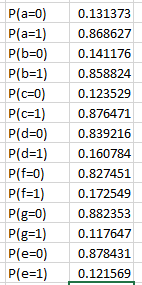
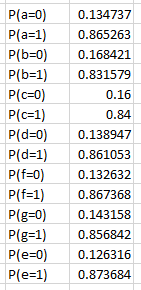
Considering the train data set of the seven segment display and finding out the probability of each digit starting from 0-9 by using =Countif and =count formula in excel we can find the probability of each digit as follows.

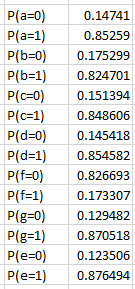


Now we will create new sheets for calculating probability of each segment ON or OFF for the digit with the available data is calculated individually for each digit and segment.

     
Fig 1: Digit-0 Fig 2: Digit 1 Fig 3: Digit 2

    
Fig 4: Digit 3 Fig 5: Digit 4 Fig 6: Digit 5

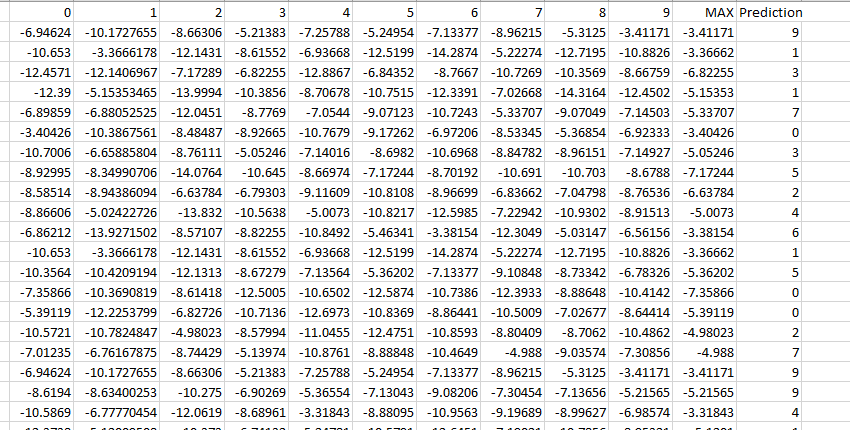
    
Fig 7: Digit 6 Fig 8: Digit 7 Fig 9: Digit 8

  
Fig 10: Digit 9

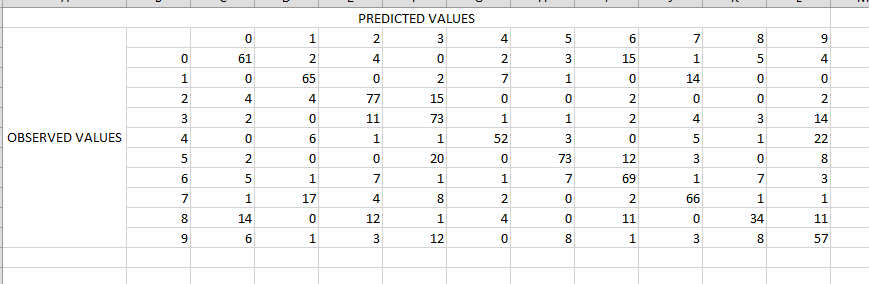
Based upon the probability values available for the train data we need to find the ln function for the test data and calculate for each digit present in the test data. Then the MAX() function of the

row is taken and corresponding index value is matched for the data that has the maximum.

Similarly it is calculated for the rest of the digits in the test dataset.



The next step is to create a confusion matrix for the observed values and predicted values. This will help us to find out whether the prediction is correct or not.



The next step is to find the accuracy of the model that is created which can be calculated using the =Sum() of the diagonal values of the digit or the values of corresponding digit in observed and predicted. Such that we get an accuracy of 62.7% for the model that is created.

Form the confusion matrix we can find out that the digits which are similar to the observed values have highest error rate as there is a slight difference in the segment value for such digits.

The values in observed and predicted cells for every digit is high in our model, so we can conclude that Naïve Bayes has 62.7% accuracy in finding the seven segment for the digits.