Anmol Harsh 18CS10005
Ayudh Saxena 18CS10007

REPORT

Assignment 1
Bayesian Learning and Dimensionality Reduction

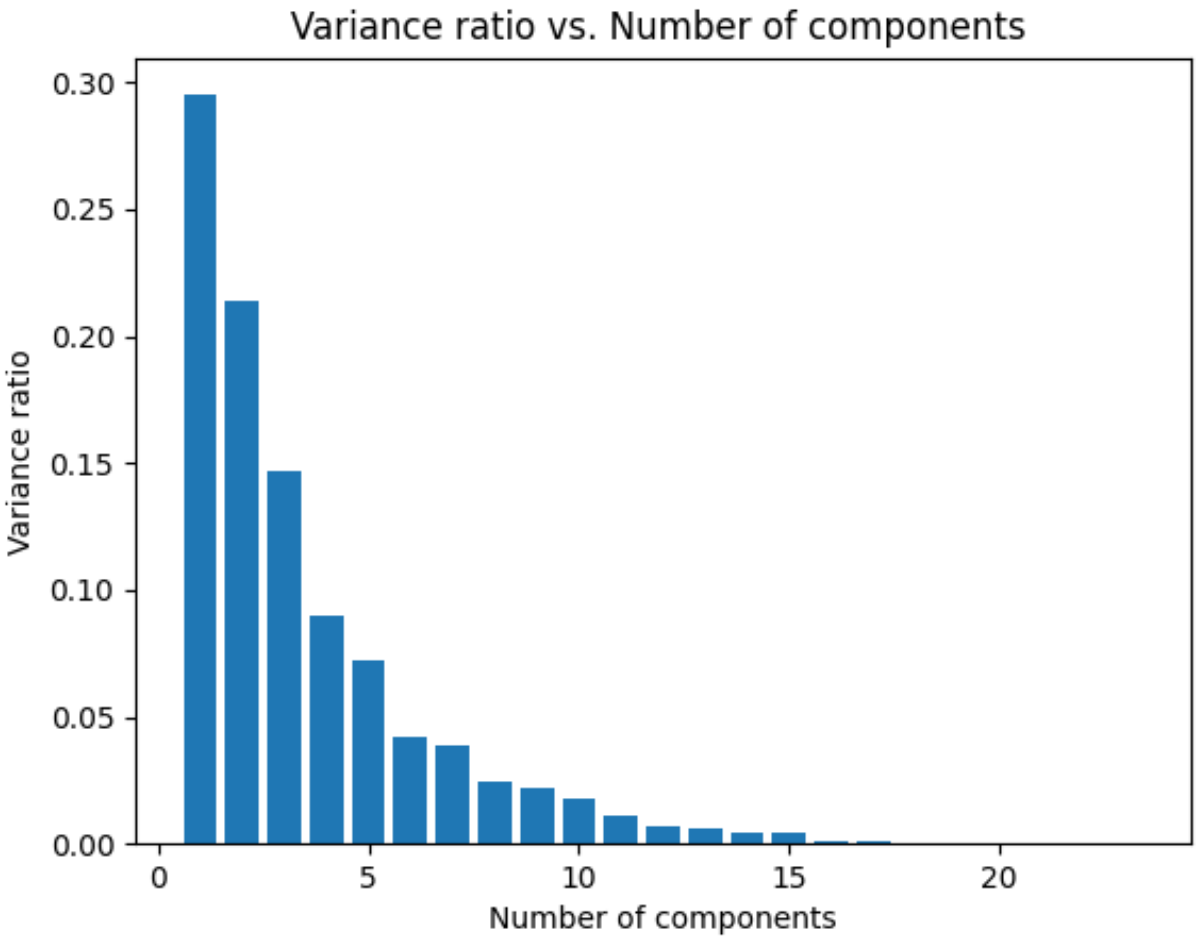
- Procedure
  - Converted the .data file to .csv and read the data using the **Pandas** library and thus converting the data-set into a **Pandas DataFrame**.
  - Defined classes for attributes, probabilities, Naïve Bayes Classifier etc.
  - Handled the missing values in the training set and the test set by assigning for :
    - **Categorical attributes** : Most frequent value
    - **Continuous attributes** : Mean value
  - Encoded the categorical attributes using **integer encoding.**
  - Implemented a Naïve Bayes classifier
    - Normalized the data set using **MinMaxScaler**() of the sklearn library.
    - Calculated the class conditional probabilities and returned a probability_class object to be used by the classifier
      - Used frequencies for categorical attributes.
      - Used **Gaussian probability** for continuous attributes.
    - Performed 5-fold Cross Validation.
  - Performed **Principal Component Analysis** using the **sklearn** library
    - Standardized the data set using **StandardScaler**() function of the sklearn library.
    - Used the 'PCA' function from **sklearn.decomposition.**
    - Passed the total number of columns of our dataset as the new number of components.
    - Took the variance ratio of each column in an array and its cumulative_sum in another.
    - Plotted a bar graph of **variance ratio of each component vs number of components** using the **matplotlib** library.
    - Components are chose up until the cumulative sum becomes greater than 0.95. This gives the new reduced number of components (= new_n).
    - Applied PCA with this new_n and extracted the new data_set
    - Applied 5-fold Cross Validation over this new data_set using the Naïve Bayes classifier constructed earlier.
  - Removed erroneous samples from the data set having large number of outliers.
  - Performed **Sequential Backward Selection** using the Naïve Bayes Classifier constructed above and then performed 5-fold Cross Validation on the new set of features obtained

- Results

```
Average validation accuracy =  100.0 %
Final test accuracy after step 1 =  100.0 %

-----------------------------------------------------------------------------------------------------------

After performing PCA :
Final number of components =  10
Performing 5-Cross Validation on the new set of components :
Average validation accuracy =  99.05759162303666 %
Test accuracy after step 2 =  99.54954954954955 %

-----------------------------------------------------------------------------------------------------------

After performing Sequential Backward Selection :
Number of features removed : 22
The final set of features are : ['iso_code']
Performing 5-Cross Validation on the new set of features
Average validation accuracy =  100.0 %
Final test accuracy after step 3 =  100.0 %
```

- Step I : Performed 5-fold Cross Validation on the **Naïve Bayes classifier** constructed
  - **Validation Accuracy** = 100%
  - **Test Accuracy** = 100%

- Step II : Performed **Principal Component Analysis** (PCA)
  - Total number of components chosen : 10



- Performed 5-fold Cross Validation on the new set of components using the Naïve Bayes Classifier constructed in Step I :
    - **Validation Accuracy** = 99.06%
    - **Test Accuracy =** 99.55%

  NOTE : The validation/test accuracy might change due to Random shuffling

- Step III : Performed **Sequential Backward Selection** with the help of Naïve Bayes Classifier :
    - Number of features removed = 22
    - Features retained = ‘iso-code’
    - Performed 5-fold Cross Validation on the new set of features :
        - **Validation Accuracy** = 100%
        - **Test Accuracy =** 100%