

|  |
| --- |
| Machine Learning-2  2024 |
| |  | | --- | | PGP DSBA PROGRAM  by: ABHISHEK K HIREMATH | |



|  |  |  |
| --- | --- | --- |
| S NO | Clustering and Cleaning Ads | Page No. |
| 1.1 | Read the dataset. Describe the data briefly. Interpret the inferences for each. Initial steps like head () .info (), Data Types, etc. Null value check, Summary stats, Skewness must be discussed. | 3 - 4 |
| 1.2 | Perform EDA (Check the null values, Data types, shape, Univariate, bivariate analysis). Also check for outliers |  |
| 1.3 | Encode the data (having string values) for Modelling. Is Scaling necessary here or not?, Data Split: Split the data into train and test (70:30). The learner is expected to check and comment about the difference in scale of different features on the bases of appropriate measure for example std dev, variance, etc. Should justify whether there is a necessity for scaling. Object data should be converted into categorical/numerical data to fit in the models. (pd.categorical(). codes(), pd.get\_dummies(drop\_first=True)) Data split, ratio defined for the split, train-test split should be discussed. | 5 |
| 1.4 | Apply Logistic Regression and LDA (linear discriminant analysis). | 6 |
| 1.5 | Apply KNN Model and Naïve Bayes Model. Interpret the results. | 7 |
| 1.6 | Model Tuning, Bagging (Random Forest should be applied for Bagging), and boosting. | 7 |
| 1.7 | Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model. Final Model: Compare the models and write inference which model is best/optimized. |  |
| 1.8 | Based on these predictions, what are the insights? | 8 – 9 |
| S NO | PCA | Page No. |
| 2.1 | Find the number of characters, words, and sentences for the mentioned documents. | 10 – 11 |
| 2.2 | Remove all the stop words from all three speeches. | 12 |
| 2.3 | Which word occurs the most number of times in his inaugural address for each president? Mention the top three words. (after removing the stopwords) | 13 – 14 |
| 2.4 | Plot the word cloud of each of the speeches of the variable. (After removing the stopwords) | 15 |

|  |  |  |
| --- | --- | --- |
| S No | Figures and Tables | Page No. |
| 1 | Fig-1: Dataset head | 6 |
| 2 | Fig-2: Dataset tail | 6 |
| 3 | Fig-3: Data info | 6 |
| 4 | Fig-4: Check missing values | 7 |
| 5 | Fig-5: 5-point Summary | 7 |
| 6 | Fig-6: Univariate Analysis – age | 8 |
| 8 | Fig-7: Univariate Analysis – economic.cond.national | 9 |
| 9 | Fig-8: Univariate Analysis – economic.cond.household. | 9 |
| 10 | Fig-9: Univariate Analysis – Blair | 10 |
| 11 | Fig-10: Univariate Analysis – Hague | 10 |
| 12 | Fig-11: Univariate Analysis – Europe | 11 |
| 13 | Fig-12: Univariate Analysis – political.knowledge | 11 |
| 14 | Fig-13: Univariate Analysis – Vote | 12 |
| 15 | Fig-14: Univariate Analysis – gender | 12 |
| 16 | Fig-15: Univariate Analysis (boxplot And Histplot) | 13 |
| 17 | Fig-16: Pairplot | 15 |
| 18 | Fig-17: Heatmap | 16 |
| 19 | Fig-18: barplot and crosstab for Gender against Vote count. | 17 |
| 20 | Fig-19: Boxplot for Gender, Age against Vote. | 17 |
| 21 | Fig-20: Barplot and Crosstab for Household vs Vote. | 18 |
| 22 | **Fig-21: Bar plot and Crosstab for National** | 18 |
| 23 | **Fig-22: Encoded Data** | 20 |
| 24 | Fig-23: Best estimator for Random Forest. | 23 |
| 25 | **Fig-24: Plot misclassification error vs K.** | 23 |
| 26 | **Fig-25: Classification report on Train and Test set for LR** | 24 |
| 27 | **Fig-26: Confusion matrix on Train and Test set of LR** | 25 |
| 28 | **Fig-27: ROC curve on Train and Test set for LR** | 25 |
| 29 | **Fig-28: Free Importance and Performance metrics LR.** | 26 |
| 30 | **Fig-29: Classification Report on Train and Test set for LDA.** | 26 |
| 31 | **Fig-30: Confusion matrix on Train and Test set for LDA.** | 27 |
| 32 | **Fig-31: ROC curve for Train and Test set for LDA.** | 27 |
| 33 | **Fig-33: Confusion matrix for LDA model cutoff value 0.4** | 28 |
| 34 | **Fig-34: Performance metrics Output LDA.** | 28 |
| 35 | Fig-35: Classification report On Train and Test set of GNB | 29 |
| 36 | **Fig-36: Confusion Matrix on Train and Test set GNB.** | 29 |
| 37 | **Fig-37: ROC curve for Train and Test GNB.** | 30 |
| 38 | Fig-38: Performance Metrics Output GNB. | 31 |
| 39 | **Fig-39: Classification Report on Train and Test set KNN.** | 31 |
| 40 | **Fig-40: Confusion matrix on Train and Test set KNN.** | 31 |
| 41 | **Fig-41: ROC curve for Train and Test set KNN.** | 32 |
| 42 | Fig-42: Performance metrics output KNN. | 32 |
| 43 | Fig-43: Classification Report for Train and Test set RF. | 33 |
| 44 | Fig-44: Confusion matrix on Train and Test set RF. | 33 |
| 45 | Fig-45: ROC curve for Train and Test set RF. | 33 |
| 46 | Fig-46: Performance metrics Output RF. | 34 |
| 47 | Fig-47: Classification Report for Train and Test set BG. | 34 |
| 48 | Fig-48: Confusion matrix for Train and Test set BG. | 35 |
| 49 | Fig-49: ROC curve for Train and Test set BG. | 35 |
| 50 | Fig-50: Classification report for train Test set GB. | 36 |
| 51 | Fig-51: Confusion matrix for Train and Test set GB. | 36 |
| 52 | Fig-52: ROC curve for Train and Test set GB. | 36 |
| 53 | Fig-53: Performance Metrics Output GB. | 37 |
| 54 | Fig-54: Classification report for train and test set AB. | 37 |
| 55 | Fig-55: Confusion Matrix for Train and Test AB. | 37 |
| 56 | Fig-56: ROC curve for Train and Test set AB. | 38 |
| 57 | Fig-57: Performance metrics output AB. | 38 |
| 58 | Fig-58: Performance metrics comparison model. | 38 |
| 59 | Fig-59: Sample speech by Franklin D. Roosevelt. | 42 |
| 60 | Fig-60: Sample Speech by John F. Kennedy's. | 43 |
| 61 | Fig-61 Sample speech by Richard Nixon's | 43 |
| 1 | Table – 1 Data Dictionary | 5 |
| 2 | Table – 2 Most frequent words of each President’s speech – before stemming | 44 |