|  |  |  |  |
| --- | --- | --- | --- |
|  | **PES University, Bengaluru**  **(Established under Karnataka Act No. 16 of 2013)** | | **UE20CS931** |
| **March 2024: END SEMESTER ASSESSMENT (ESA)**  **M TECH DATA SCIENCE AND MACHINE LEARNING\_ SEMESTER II**  **UE20CS931- MACHINE LEARNING - II** | | | |
| Time: 3 Hrs | | Answer All Questions | Max Marks: 100 |
| **Instructions**   1. Answer all the questions. 2. Section A should be handwritten in the answer script provided. 3. Sections B and C are coding questions to be answered in the system and uploaded. 4. Smartly use GridSearchCV as it might impact the system’ performance. 5. Write appropriate inferences. | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **Section A (20 marks)** | | | |
| 1 | a) | What is Logistic Regression? Explain its working. | 4 |
| b) | Define Precision, Recall, and F1 score. State the necessary formulas. | 4 |
| c) | Explain How Random forest solves the problem of Low Bias and High Variance? | 4 |
| d) | Describe Steps involved in k-Nearest Neighbour’s algorithm. | 4 |
|  | e) | What are Bagging and Boosting? | 4 |
| **Section B (40 marks)** | | | |
| 2 | a) | Read the dataset and print/perform the following  - Shape of the data (2 mark) - Number of numerical and categorical variable (2 mark) - Descriptive stats of numerical data and write inference (2 mark) | 6 |
|  | b) | What is the distribution of hemoglobin levels (hemo) among patients with and without hypertension (htn)? Explain using visualization. | 6 |
|  | c) | Perform necessary actions to ‘fix’ defects like missing values | 6 |
|  | d) | Perform appropriate encoding on the categorical attributes. | 8 |
|  | e) | Examine the correlation and summarize the relationship between variables. Use appropriate plots to justify the same and write your inferences. | 8 |
|  | f) | Check whether the target column has balanced data or not. | 3 |
|  | g) | Split dataset into train and test and check if its a good split (70:30) | 3 |
| **Section C (40 marks)** | | | |
| 3 | a) | Make use of the imbalanced data and fit a Decision Tree and Random forest classifier Model. Compare the model performance using F1 Score and describe your observations based on output/results? | 10 |
|  | b) | Apply Sampling technique to balance the target column and check will it improve the previous model performance using balanced data. Write your observation based on results obtained. | 15 |
|  | c) | Choose any two models of your choice from Naive Bayes,KNN,Logistic regression, XGBoost and experiment with the balanced & imbalanced data. Write down your observations. | 10 |
|  | d) | From a business perspective answer the following:  - a. Which data will you choose, Balanced or Imbalanced and why?  - b. Which of the above trained models will you choose to move further as a final model and why? | 5 |