PES University, Bengaluru
(Established under Karnataka Act No. 16 of 2013)
M TECH DATA SCIENCE AND MACHINE LEARNING - SEMESTER II
UE20CS931 - MACHINE LEARNING - II
Time: 3 Hrs
Max Marks: 100

Instructions:

1. Answer all the questions.
2. Section A should be handwritten in the answer script provided.
3. Section B and C are coding questions which have to be answered in the system and uploaded in the learning portal.
4. Smartly use GridSearchCV as it might impact the system performance.
5. Write appropriate inferences.

Section A (30 Marks)

1. **(a)** Explain the working of the Support Vector Machine (SVM) algorithm with an example. *(5 Marks)*
2. **(b)** Define Recall, Precision, and F1 Score. Explain their significance in imbalanced classification problems. *(5 Marks)*
3. **(c)** Differentiate between Bagging, Boosting, and Stacking with examples. *(5 Marks)*
4. **(d)** What are the assumptions of Linear Regression? How do they impact the model's performance? *(5 Marks)*
5. **(e)** Explain the concept of Feature Engineering. Provide three common feature engineering techniques with examples. *(5 Marks)*
6. **(f)** Compare K-Means and DBSCAN clustering algorithms. When should each be used? *(5 Marks)*

Section B (30 Marks)

A retail company wants to predict whether a customer will buy a product based on demographic and behavioral features. The dataset contains the following attributes:

- **Customer_ID**: Unique identifier
- **Age**: Age of the customer
- **Gender**: Male/Female
- **Annual_Income**: Annual income in USD

- **Spending_Score**: A score from 1-100 measuring spending behavior
- **Product_Category**: The category of product the customer is interested in
- **Purchased**: Target variable (1 = Purchased, 0 = Not Purchased)

7. **(a)** Load the dataset and summarize key observations: *(5 Marks)*
    - Find the number of rows and columns.
    - Identify numerical and categorical variables.
    - Compute basic descriptive statistics.
8. **(b)** Identify and fix defects in the dataset: *(10 Marks)*
    - Handle missing values, if any.
    - Identify and treat outliers.
    - Examine class imbalance in the target variable and suggest a way to handle it.
9. **(c)** Perform exploratory data analysis (EDA): *(5 Marks)*
    - Plot relevant visualizations for categorical variables.
    - Identify features most correlated with the target variable.
    - Perform hypothesis testing to determine variable importance.
10. **(d)** Split the dataset into train and test sets (70:30). *(5 Marks)*
11. **(e)** Fit a baseline classification model (e.g., Logistic Regression) and evaluate its performance. *(5 Marks)*

Section C (40 Marks)

12. **(a)** Improve the model accuracy: *(20 Marks)*

- Implement hyperparameter tuning using GridSearchCV.
- Apply feature selection techniques.
- Compare multiple models (e.g., Decision Tree, Random Forest, Gradient Boosting) and choose the best one.

13. **(b)** Final model evaluation and business interpretation: *(20 Marks)*

- Summarize overall model performance using evaluation metrics.
- Explain which features had the most influence on predictions.
- Discuss any key risks or limitations of the model.
- Provide a business-level interpretation of the results.

End of Question Paper