


SRN

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

	PES University, Bengaluru (Established under Karnataka Act No. 16 of 2013)	UE20CS931
October 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
<p style="text-align: center;">Instructions</p> <ol style="list-style-type: none"> Answer all the questions. Section A should be handwritten in the answer script provided. Sections B and C are coding questions to be answered in the system and uploaded. Smartly use GridSearchCV as it might impact the system's performance. Write appropriate inferences. 		

Section A (20 marks)			
1	a)	Explain the concept of Binomial Logistic Regression and its assumptions. How does it differ from linear regression?	4
	b)	Discuss the significance of coefficients in a logistic regression model. How can they be interpreted?	4
	c)	Describe the role of proximity measures in the K-Nearest Neighbors (K-NN) algorithm. How do these measures affect model performance?	4
	d)	Define the term "purity of a node" in the context of decision trees. Explain the metrics used to measure the purity of a node.	4
	e)	Compare AdaBoost and Gradient Boosting algorithms. Highlight their differences and similarities in terms of methodology and application.	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 variables (2 mark) - Descriptive stats of numerical data and write inference (2 mark)	6
	b)	Create visualizations to explore the relationships between different numerical features in the dataset using suitable plots and share your inferences for the same.	6
	c)	Check for the correlation between various numerical features and share your inferences accordingly.	8
	d)	Perform appropriate encoding on the categorical attributes.	8
	e)	Perform the following steps on the dataset: - Check the distribution of the target column 'stroke', and comment on the class distribution (3 marks) - Segregate the dependent and the independent features. (2 marks)	5
	f)	Handle the imbalanced data using oversampling or undersampling technique, and check the distribution of the resampled target class.	5

	g)	Split the dataset into train and test data (80:20).	2
Section C (40 marks)			
3	a)	Make use of the imbalanced data and fit a Random forest classifier Model. Describe your observations based on output/results seen in the confusion matrix.	10
	b)	<p>Make use of the balanced data and fit a Random forest classifier Model and a Decision Tree Regressor model. Describe your observations based on output/results seen in the confusion matrix.</p> <p>Note:</p> <p>For each model built, follow the below approach:</p> <ul style="list-style-type: none"> - Build a base model using the balanced data - Select K features using Wrapper or Embedded Methods - Perform hyperparameter tuning on all the models to tune the hyperparameters and find the best hyperparameters. - Scale the data using the StandardScaler() method and build a model using the K selected feature and the hyperparameters, and compute its accuracy and Recall. 	20
	c)	Collectively compare the performance of all the models and find the best-performing model.	5
	d)	<p>From a business perspective,</p> <p>a. Which data will you choose, Balanced or Imbalanced and why?</p> <p>b. Based on the given problem statement, explain which metric should be considered from the confusion matrix to gauge the effectiveness of the model built.</p>	5