U.S.N.										

B.M.S. College of Engineering, Bengaluru-560019

Autonomous Institute Affiliated to VTU

January / February 2025 Semester End Main Examinations

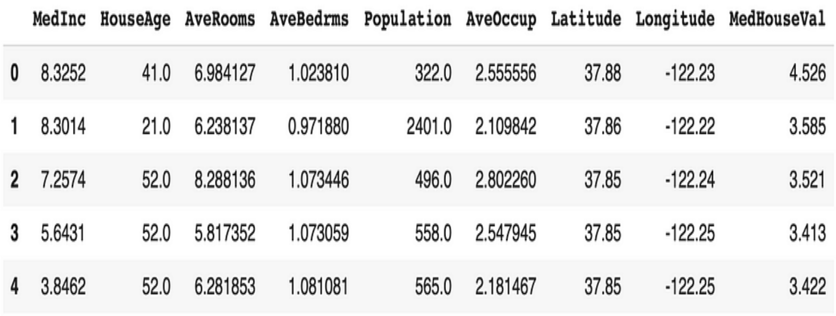
Programme: B.E.
Branch: Information Science and Engineering
Course Code: 22IS6PCMLG / 20IS5PCMLG
Course: Machine Learning

Semester: VI/V
Duration: 3 hrs.
Max Marks: 100

Instructions: 1. Answer any FIVE full questions, choosing one full question from each unit.
2. Missing data, if any, may be suitably assumed.

Important Note: Completing your answers, compulsorily draw diagonal cross lines on the remaining blank pages. Revealing of identification, appeal to evaluator will be treated as malpractice.

		UNIT - I	***CO***	***PO***	**Marks**
1	a)	What type of machine learning task would you apply for the below scenarios and justify your answer? I. Estimating the monthly electricity usage of a household II. Identifying whether a tweet is expressing positive or negative sentiment III. Training an autonomous car to drive in various traffic conditions IV. Predicting if a customer will churn next month V. Categorizing news articles into different topics	*CO1*		**05**
	b)	What role does a validation set play in machine learning, when is it essential, and how should it be utilized for optimal model training?	*CO2*	*PO1*	**05**
	c)	Main task of machine learning is to select a learning algorithm and train it on some data, the two things that can go wrong are “bad algorithm” and “bad data.” Explain the main challenges with respect to “bad algorithm” and “bad data.”	*CO2*	*PO1*	**10**
		OR			
2	a)	With a suitable example explain how One-hot encoding and Label encoding affect the dimensionality of the given dataset.	*CO1*		**05**
	b)	Imagine, you are given a dataset consisting of variables having more than 40%missing values. Let’s say, out of 100 variables, 45 variables have missing values, which is higher than 40%. Illustrate different ways of handling missing values.	*CO2*	*PO1*	**05**
	c)	Consider the California census housing data set. Write a python code for the following	*CO3*	*PO2*	**10**



	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude	MedHouseVal
0	8.3252	41.0	6.984127	1.023810	322.0	2.555556	37.88	-122.23	4.526
1	8.3014	21.0	6.238137	0.971880	2401.0	2.109842	37.86	-122.22	3.585
2	7.2574	52.0	8.288136	1.073446	496.0	2.802260	37.85	-122.24	3.521
3	5.6431	52.0	5.817352	1.073059	558.0	2.547945	37.85	-122.25	3.413
4	3.8462	52.0	6.281853	1.081081	565.0	2.181467	37.85	-122.25	3.422

I. Histogram of median house values

II. Scatter plot of median income vs. median house value

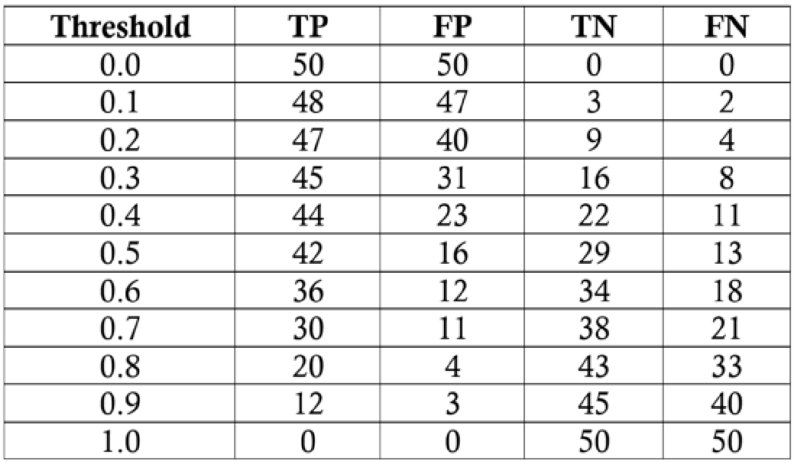
III. correlation of features against “MedHouseVal”

IV. Differentiate between positive and negative correlation.

What is the need of Experimenting with Attribute Combinations?

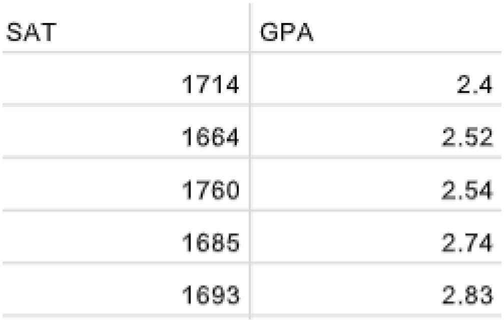
UNIT - II

3 a) Consider a “Email Spam Filtering dataset”. Our task is to detect spam emails. So spam emails are marked as 1 and not spam emails are marked as 0. Let’s say our model prediction looks like this. Define and calculate the **accuracy**, **recall, precision, F1 score and false-positive rate** at a threshold of **0.5.** *CO3* *PO2* **10**



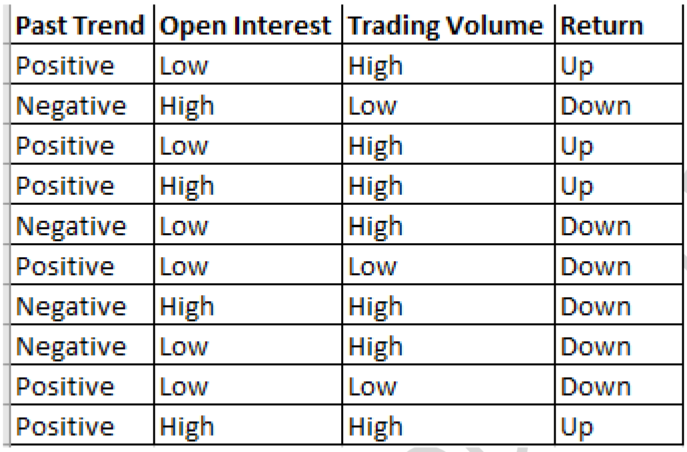
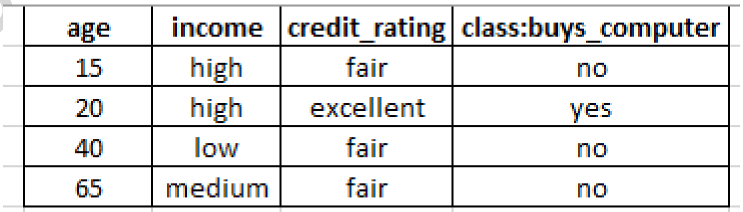
Threshold	TP	FP	TN	FN
0.0	50	50	0	0
0.1	48	47	3	2
0.2	47	40	9	4
0.3	45	31	16	8
0.4	44	23	22	11
0.5	42	16	29	13
0.6	36	12	34	18
0.7	30	11	38	21
0.8	20	4	43	33
0.9	12	3	45	40
1.0	0	0	50	50

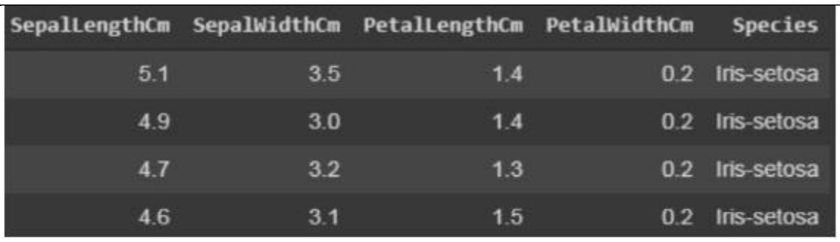
b) Consider the below dataset which contains information about students including attributes like **SAT and GPA**. Build an appropriate machine Learning model using the Scikit-learn library to predict the GPA of students based on these attributes. *CO2* *PO1* **10**



SAT	GPA
1714	2.4
1664	2.52
1760	2.54
1685	2.74
1693	2.83

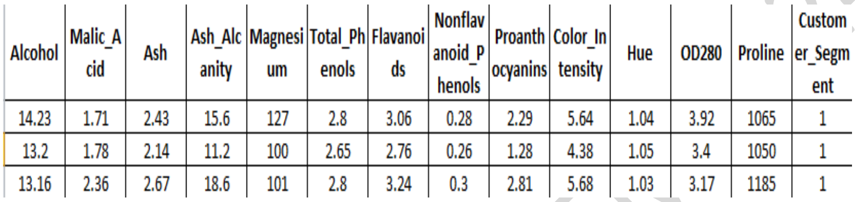
OR

<table>
<tr><td></td><td>4</td><td>a)</td><td>The HR team of a company needs to verify the salary details of a potential employee based on their job level. They have access to salary information for 10 positions, indexed by their levels (1 through 10). Using this build a <b>polynomial regression model</b> using SKlearn library to predict salaries accurately and predict the salary of employee whose position is level 6.5</td><td><i>CO3</i></td><td><i>PO2</i></td><td><b>10</b></td></tr>
<tr><td></td><td></td><td>b)</td><td>Elaborate key advantages of using Gradient Descent for optimization tasks and Illustrate with an example of three kinds of Gradient Descent algorithm?</td><td><i>CO1</i></td><td></td><td><b>10</b></td></tr>
<tr><td></td><td></td><td></td><td><b>UNIT - III</b></td><td></td><td></td><td></td></tr>
<tr><td></td><td>5</td><td>a)</td><td>What role does the Gini index play in minimizing misclassification errors in decision tree models?Calculate Gini Index for <b>past trend, open interest and trading volume .</b>

<table>
<tr><th>Past Trend</th><th>Open Interest</th><th>Trading Volume</th><th>Return</th></tr>
<tr><td>Positive</td><td>Low</td><td>High</td><td>Up</td></tr>
<tr><td>Negative</td><td>High</td><td>Low</td><td>Down</td></tr>
<tr><td>Positive</td><td>Low</td><td>High</td><td>Up</td></tr>
<tr><td>Positive</td><td>High</td><td>High</td><td>Up</td></tr>
<tr><td>Negative</td><td>Low</td><td>High</td><td>Down</td></tr>
<tr><td>Positive</td><td>Low</td><td>Low</td><td>Down</td></tr>
<tr><td>Negative</td><td>High</td><td>High</td><td>Down</td></tr>
<tr><td>Negative</td><td>Low</td><td>High</td><td>Down</td></tr>
<tr><td>Positive</td><td>Low</td><td>Low</td><td>Down</td></tr>
<tr><td>Positive</td><td>High</td><td>High</td><td>Up</td></tr>
</table></td><td><i>CO2</i></td><td><i>PO1</i></td><td><b>10</b></td></tr>
<tr><td></td><td></td><td>b)</td><td>Design a Decision Tree model using SKlearn library for the petrol consumption dataset that has 4 features- petrol_tax, average_income,paved_highways and population_delivery which predicts petrolconsumption. Apply necessary pre-processing steps and performance measures. Predict the petrol consumptionfor values - [9,3471,1250,0.58].</td><td><i>CO3</i></td><td><i>PO2</i></td><td><b>10</b></td></tr>
<tr><td></td><td></td><td></td><td><b>OR</b></td><td></td><td></td><td></td></tr>
<tr><td></td><td>6</td><td>a)</td><td>Consider class-labeled training tuples from the <b>AllElectronics</b> Customer Database as given in table below. Calculate the weighted <b>Gini(income)</b> and <b>Gini(age)</b> columns.

<table>
<tr><th>age</th><th>income</th><th>credit_rating</th><th>class:buys_computer</th></tr>
<tr><td>15</td><td>high</td><td>fair</td><td>no</td></tr>
<tr><td>20</td><td>high</td><td>excellent</td><td>yes</td></tr>
<tr><td>40</td><td>low</td><td>fair</td><td>no</td></tr>
<tr><td>65</td><td>medium</td><td>fair</td><td>no</td></tr>
</table></td><td><i>CO3</i></td><td><i>PO2</i></td><td><b>10</b></td></tr>
<tr><td></td><td></td><td>b)</td><td>Design a Decision Tree classifier for the iris dataset that has 4 features- sepal length, sepal width, petal length and petal width which classifies the tuples as iris-setosa, iris-virginica and versicolor. Apply necessary pre-processing steps and performance measures. Predict the species for values - [1.1, 0.1].</td><td><i>CO2</i></td><td><i>PO1</i></td><td><b>10</b></td></tr>
</table>



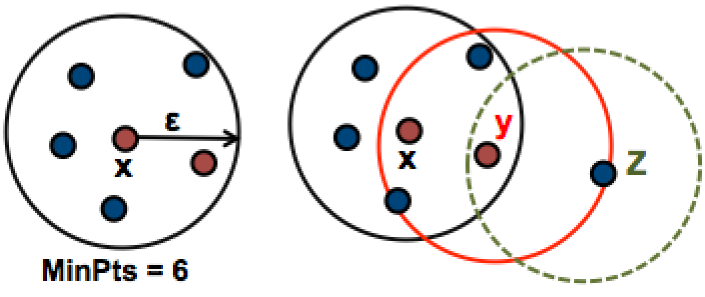
SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3.0	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa

		UNIT - IV			
7	a)	Illustrate with an example the iterative process of refining predictors in Gradient Boosting and how it leads to better predictive performance. How does Gradient Boosting differ from AdaBoost in terms of handling prediction errors?	*CO1*		**10**
	b)	Apply dimensionality reduction method on the following dataset that preserves 95% of variance. Design a suitable model that best suits the given data and write the complete python code.	*CO2*	*PO1*	**10**



Alcohol	Malic_Acid	Ash	Ash_Alcanity	Magnesium	Total_Phenols	Flavanoids	Nonflavanoid_Phenols	Proanthocyanins	Color_Intensity	Hue	OD280	Proline	Customer_Segment
14.23	1.71	2.43	15.6	127	2.8	3.06	0.28	2.29	5.64	1.04	3.92	1065	1
13.2	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28	4.38	1.05	3.4	1050	1
13.16	2.36	2.67	18.6	101	2.8	3.24	0.3	2.81	5.68	1.03	3.17	1185	1

		OR			
8	a)	Discuss the need of Ensemble models. List and explain the different types Voting methods.	*CO1*	*PO1*	**5**
	b)	What are the two main approaches for dimensionality reduction? Explain in detail with necessary diagrams.	*CO1*	*PO1*	**5**
	c)	An ensemble method which is a sequential learner where different models are generated sequentially and the mistakes of previous models are learned by their successors. This aims at identifying the dependency between models by giving the mislabeled examples with **higher weights**. Identify and illustrate the steps involved in **building a strong learner with a suitable example.**	*CO2*	*PO2*	**10**
		UNIT - V			
9	a)	After applying the DBSCAN Algorithm on a dataset, we get the following clusters (as shown below in the figure). Identify and describe the core point, border point, and noise point.	*CO2*	*PO1*	**05**



(a) (b)

		b)	Explain how clustering can be used for data preprocessing?	*CO1*		**05**
		c)	Design a model using SKlearn library to segment 500students from a high school into distinct groups based on their academic performance, demographic characteristics, and extracurricular activities. The model should determine the optimal number of clusters and analyze the resulting clusters.	*CO3*	*PO2*	**10**
			OR			
	10	a)	Apply K (=2)-Means algorithm over the data (185, 72), (170, 56), (168, 60), (179,68), (182,72), (188,77) up to two iterations and show the clusters. Initially choose first two objects as initial centroids.	*CO3*	*PO2*	**10**
		b)	Elucidate unsupervised learning? How unsupervised learning is different from supervised learning. Mention any 3 applications where you can apply unsupervised learning model.	*CO1*		**10**

******