KUMARAGURU
college of technology
character is life

B.TECH DEGREE EXAMINATIONS: NOV/DEC 2022

(Regulation 2018)

Third Semester

ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

U18AII3204: Applied Machine Learning

COURSE OUTCOMES

CO1: Understand different methodologies to create application using statistical models

CO2: Design the test procedures to assess the efficacy of the developed model.

CO3: Identify and apply appropriate machine learning models for analyzing the data for a variety of problems.

CO4: Implement different algorithms for business intelligence

Time: Three Hours   **Maximum Marks: 100**

Answer all the Questions:-
PART A (10 x 2 = 20 Marks)
(Answer not more than 40 words)

1.	Define machine learning.	CO1	[K_1]
2.	List the methods of data pre-processing?	CO1	[K_1]
3.	Interpret the shape of the multiple regression models, if there are three variables?	CO1	[K_2]
4.	State the reasons for overfitting.	CO2	[K_2]
5.	List the use of cost function in machine learning algorithm.	CO2	[K_1]
6.	Discuss on Ridge Regularization?	CO2	[K_2]
7.	Explain the importance of p-value in machine learning?	CO3	[K_2]
8.	Describe how cross-validation is applied?	CO3	[K_2]
9.	Compare Bagging and Boosting.	CO4	[K_2]
10.	Explain the difference between the CART and ID3 Algorithms.	CO4	[K_2]

Answer any FIVE Questions:-
PART B (5 x 16 = 80 Marks)
(Answer not more than 400 words)

11.	a)	Develop a multi-variable regression problem using SciKit-Learn.	6	CO1	[K_3]
	b)	A regression analysis done for the relationship between hours spend (x) in study and the scores obtained (y) in Predictive Data analytics course for 12 students yielded the following output in OLS statsmodel.	10	CO1	[K_3]

```
                        OLS Regression Results
==============================================================================
Dep. Variable:                    y   R-squared:                       0.338
Model:                          OLS   Adj. R-squared:                  0.272
Method:               Least Squares   F-statistic:                     5.113
Date:              Tue, 30 Jan 2018   Prob (F-statistic):             0.0473
Time:                      14:21:22   Log-Likelihood:                -41.442
No. Observations:                12   AIC:                             86.88
Df Residuals:                    10   BIC:                             87.85
Df Model:                         1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const          176.6364     20.546      8.597      0.000     130.858    222.415
x               -0.3572      0.158     -2.261      0.047      -0.709     -0.005
==============================================================================
Omnibus:                        1.934   Durbin-Watson:                   1.182
Prob(Omnibus):                  0.380   Jarque-Bera (JB):                1.010
Skew:                          -0.331   Prob(JB):                        0.603
Kurtosis:                       1.742   Cond. No.                     1.10e+03
==============================================================================
```

A. What is the regression equation to predict score (y) as a function of hours spends (x)?
B. Briefly interpret the coefficients of this regression model.
C. Describe the evidence for a relationship between hours spends and score obtained?
D. What does the R-squared value indicate?
E. Compute the variance influence factor and specify the significance of it.

| 12. | a) | Interpret the trade-off between Bias and Variance of a Machine Learning Model. | 6 | CO2 | [K_2] |
| | b) | Explain Support Vector Machine model and illustrate the kernel trick using linear and non-linear data. | 10 | CO2 | [K_2] |

13. a) Predict the quality of the restaurant based on the customer ratings given in a 10-point scale.

X1 (Quality of food)	X2 (service)	Y
7	7	Bad
7	4	Bad
3	4	Good
1	4	Good

8   CO2   [K_3]

Predict whether the ratings given by a new customer for the quality of food and the service offered by the restaurant as (3, 7) is Good/Bad using k-NN (k=3).

b) A multivariate logistic regression model has been modelled to predict the tendency of retailer to provide a free gift for the repeated purchases. The features used for the model are AGE, SOCIO ECONOMIC BAND A, SOCIO ECONOMIC BAND B, the average money the customer spends on each visit, and the average frequency of visits per week. This model will be used by the marketing department to decide to whom they should provide the free gift. The coefficients of the features in the model are given below. Intercept: -2.75, age: -0.04, Socio economic band A: -0.09, Socio economic band B: -0.3, Shop value: 0.04, Shop frequency: 0.8

Use this model to make predictions for each of the following query instances.

Id	age	Socio economic band A	Socio economic band B	Shop value	Shop frequency
1	30	1	0	175.19	3.21
2	28	0	1	15.28	8.92

8   CO2   [K₃]

14. a) Illustrate the evaluation of regression machine learning models.

8   CO3   [K₂]

b) Summarize the metrics for assessing classification machine learning models. Confusion Matrix.

8   CO3   [K₂]

15. a) How MAP hypothesis is determined using Bayes theorem? Explain in detail.

6   CO4   [K₂]

b) Distinguish the approaches to avoid overfitting in Decision tree learning.

10   CO4   [K₃]

16. a) Describe Random Forest Classifier Algorithm with diagram and conclude how it outperforms incorrect prediction.

10   CO4   [K₂]

b) Explain decision stumps are used rather than complete tree in Ada boost.

6   CO4   [K₂]

****