Sl. No.

A10.5-R5 : MACHINE LEARNING USING PYTHON

अवधि : 03 घंटे **अधिकतम अंक : 100**
DURATION : 03 Hours **MAXIMUM MARKS : 100**

ओएमआर शीट सं. : **OMR Sheet No. :**						

रोल नं. : **Roll No. :**

उत्तर-पुस्तिका सं. : **Answer Sheet No. :**

परीक्षार्थी का नाम : **परीक्षार्थी के हस्ताक्षर :**
Name of Candidate : _____ **Signature of Candidate :** _____

परीक्षार्थियों के लिए निर्देश :	**Instructions for Candidate :**
कृपया प्रश्न-पुस्तिका, ओएमआर शीट एवं उत्तर-पुस्तिका में दिये गए निर्देशों को ध्यानपूर्वक पढ़ें।	Carefully read the instructions given on Question Paper, OMR Sheet and Answer Sheet.
प्रश्न-पुस्तिका की भाषा अंग्रेजी है। परीक्षार्थी केवल अंग्रेजी भाषा में ही उत्तर दे सकता है।	Question Paper is in English language. Candidate can answer in English language only.
इस मॉड्यूल/पेपर के **दो भाग** हैं। **भाग एक** में **चार** प्रश्न और **भाग दो** में **पाँच** प्रश्न हैं।	There are **TWO PARTS** in this Module/Paper. **PART ONE** contains **FOUR** questions and **PART TWO** contains **FIVE** questions.
भाग एक ''वैकल्पिक'' प्रकार का है जिसके कुल अंक 40 है तथा **भाग दो** ''व्यक्तिपरक'' प्रकार का है और इसके कुल अंक 60 है।	**PART ONE** is Objective type and carries 40 Marks. **PART TWO** is Subjective type and carries 60 Marks.
भाग एक के उत्तर, **ओएमआर उत्तर-पुस्तिका** पर ही दिये जाने हैं। **भाग दो** की उत्तर-पुस्तिका में **भाग एक** के उत्तर **नहीं** दिये जाने चाहिए।	**PART ONE** is to be answered in the **OMR ANSWER SHEET** only. **PART ONE** is **NOT** to be answered in the answer book for **PART TWO**.
भाग एक के लिए अधिकतम समय सीमा **एक घण्टा** निर्धारित की गई है। **भाग दो** की उत्तर-पुस्तिका, **भाग एक** की उत्तर-पुस्तिका जमा कराने के पश्चात् दी जाएगी। तथापि, निर्धारित एक घंटे से पहले **भाग एक** पूरा करने वाले परीक्षार्थी **भाग एक** की उत्तर-पुस्तिका निरीक्षक को सौंपने के तुरंत बाद, **भाग दो** की उत्तर-पुस्तिका ले सकते हैं।	Maximum time allotted for **PART ONE** is **ONE HOUR**. Answer book for **PART TWO** will be supplied at the table when the Answer Sheet for **PART ONE** is returned. However, Candidates who complete **PART ONE** earlier than one hour, can collect the answer book for **PART TWO** immediately after handing over the Answer Sheet for **PART ONE** to the Invigilator.
परीक्षार्थी, उपस्थिति-पत्रिका पर हस्ताक्षर किए बिना और अपनी उत्तर-पुस्तिका, निरीक्षक को सौंपे बिना, परीक्षा हॉल/कमरा नहीं छोड़ सकते हैं। ऐसा नहीं करने पर, परीक्षार्थी को इस मॉड्यूल/पेपर में अयोग्य घोषित कर दिया जाएगा।	**Candidate cannot leave the examination hall/room without signing on the attendance sheet and handing over his/her Answer Sheet to the invigilator. Failing in doing so, will amount to disqualification of Candidate in this Module/Paper.**
प्रश्न-पुस्तिका को खोलने के निर्देश मिलने के पश्चात् एवं उत्तर लिखना आरम्भ करने से पहले उम्मीदवार जाँच कर यह सुनिश्चित कर लें कि प्रश्न-पुस्तिका प्रत्येक दृष्टि से संपूर्ण है।	After receiving the instruction to open the booklet and before starting to answer the questions, the candidate should ensure that the Question Booklet is complete in all respect.

जब तक आपसे कहा न जाए, तब तक प्रश्न-पुस्तिका न खोलें।
DO NOT OPEN THE QUESTION BOOKLET UNTIL YOU ARE TOLD TO DO SO.

PART ONE

(Answer all the questions. Each question carries one mark)

1. Each question below gives a multiple choice of answers. Choose the most appropriate one and enter in the "OMR" answer sheet supplied with the question paper, following instructions therein. **(1x10)**

1.1 Which of the following is a widely used and effective machine learning algorithm based on the idea of bagging ?

(A) Decision Tree

(B) Regression

(C) Classification

(D) Random Forest

1.2 To find the minimum or the maximum of a function, we set the gradient to zero because :

(A) The value of the gradient at extreme of a function is always zero

(B) Depends on the type of problem

(C) Both (A) and (B)

(D) None of the above

1.3 The most widely used metrics and tools to assess a classification model are :

(A) Confusion matrix

(B) Cost-sensitive accuracy

(C) Area under the ROC curve

(D) All of the above

1.4 Which of the following is a good test dataset characteristic ?

(A) Large enough to yield meaningful results

(B) Is representative of the dataset as a whole

(C) Both (A) and (B)

(D) None of the above

1.5 Which of the following is one of the disadvantage of decision trees ?

(A) Factor analysis

(B) Decision trees are robust to outliers

(C) Decision trees are prone to be over fit

(D) None of the above

1.6 How do you handle missing or corrupted data in a dataset ?

(A) Drop missing rows or columns

(B) Replace missing values with mean/ median/mode

(C) Assign a unique category to missing values

(D) All of the above

1.7 What is the purpose of performing cross-validation ?

(A) To assess the predictive performance of the models

(B) To judge how the trained model performs outside the sample on test data

(C) Both (A) and (B)

(D) None of the above

SPACE FOR ROUGH WORK

1.8 Why is second order differencing in time series needed ?

(A) To remove stationary

(B) To find the maxima or minima at the local point

(C) Both (A) and (B)

(D) None of the above

1.9 When performing regression or classification, which of the following is the correct way to preprocess the data ?

(A) Normalize the data → PCA → training

(B) PCA → normalize PCA output → training

(C) Normalize the data → PCA → normalize PCA output → training

(D) None of the above

1.10 Which of the following is an example of feature extraction ?

(A) Constructing bag of words vector from an email

(B) Applying PCA projects to a large high-dimensional data

(C) Removing stop words in a sentence

(D) All of the above

2. Each statement below is either TRUE or FALSE. Choose the most appropriate one and enter your choice in the "OMR" answer sheet supplied with the question paper, following instructions therein. **(1x10)**

2.1 Logistic regression is a supervised machine learning algorithm.

2.2 Logistic regression is mainly used for Regression.

2.3 It is possible to design a logistic regression algorithm using a Neural Network Algorithm.

2.4 It is possible to apply a logistic regression algorithm on a 3-class Classification problem.

2.5 Standardization of features is required before training a Logistic Regression.

2.6 In the KNN algorithm, a small value for K provides the most flexible fit (low bias/ high variance).

2.7 Unsupervised learning involves building a statistical model for predicting, or estimating an output based upon one or more inputs.

2.8 The snowflake schema differs from the star schema in that the table holding the dimensional data are normalized..

2.9 Class is one of the core data types in Python.

2.10 Round (45.8) will run without any error in Python.

3. Match words and phrases in column X with the closest related meaning/ word(s)/phrase(s) in column Y. Enter your selection in the "OMR" answer sheet supplied with the question paper, following instructions therein. (1x10)

X		Y	
3.1	SVM	**A.**	None
3.2	print('hijk'.partition('ab'))	**B.**	'Hello'
3.3	[(a, b) for a in range(3) for b in range((a)]	**C.**	Classifier
3.4	import math def main(): math.cos(math.pi) main() print(main())	**D.**	[(1,0),(2,0),(2,1)]
3.5	s = "\t\t\t\n\nHello\n\n\n\t\t\t" s.strip()	**E.**	"Hello"
3.6	S = [['him', 'sell'], [90, 28, 43]] S[0][1][1]	**F.**	('hijk', ' ', ' ')
3.7	Cross Validation	**G.**	null hypothesis
3.8	Type I and Type II Error	**H.**	k-Fold
3.9	random forest	**I.**	Data control language
3.10	standardization	**J.**	Produces many end-results trees and merges them to get more accurate and consistent predictions.
		K.	evaluate machine learning models
		L.	process of restructuring one or more attributes
		M.	Unsupervised Learning

4. Each statement below has a blank space to fit one of the word(s) or phrase(s) in the list below. Choose the most appropriate option, enter your choice in the "OMR" answer sheet supplied with the question paper, following instructions therein. (1x10)

A.	Over fitting	**B.**	overloading	**C.**	regex	**D.**	Lemmatization
E.	Manhattan	**F.**	pyregex	**G.**	Our estimate for $P(y = 0 \mid x ; \theta)$ is 0.8	**H.**	Set of all Eigen vectors for the projection space
I.	ensemble	**J.**	re	**K.**	True, False and None are capitalized while the others are in lower case.	**L.**	Random Forest
M.	regression						

4.1 All keywords in Python are in _____.

4.2 What is pca.components_ in Sklearn _____.

4.3 Suppose you have trained a logistic regression classifier and it outputs a new example x With a prediction theta(x) = 0.2. This means _____.

4.4 Widely used and effective machine learning algorithm based on the idea of bagging is _____.

4.5 _____ is one of the disadvantage of decision trees.

4.6 _____ technique can be used for normalization in text mining.

4.7 Gradient Boosting is _____ learning algorithm.

4.8 _____ distance metric can be used in KNN.

4.9 The module in Python that supports regular expressions is _____.

4.10 Types of polymorphism _____.

PART TWO

(Answer any FOUR questions)

5. (a) How KNN is different from K- Mean ?

(b) How is a decision tree pruned ? Explain with an example.

(c) What is a difference between training set and test set ? Why do we split on the dependent variable only ? **(5+5+5)**

6. (a) How would you handle imbalanced data ?

(b) What is the loss function of SVM tries to minimize ?

(c) Discuss some pre-processing techniques used to prepare the data in python. **(4+5+6)**

7. (a) How exceptions are different than syntax errors ? Briefly discuss how different exceptions are handled in Python.

(b) Describe in brief Neural Network using Tensor Flow. **(7+8)**

8. (a) Define Cross Validation.

(b) What is multi-dimensional scaling ? **(7+8)**

9. (a) What is aggregation ? Briefly discuss different methods available in Python to perform aggregations on data.

(b) What are Sentiment Analysis and Text Classification ? **(8+7)**

- o O o -

SPACE FOR ROUGH WORK

SPACE FOR ROUGH WORK