SET-4

Series S7PQR

प्रश्न-पत्र कोड
Q.P. Code **368/7**

रोल नं.
Roll No. | | | | | | | | |

परीक्षार्थी प्रश्न-पत्र कोड को उत्तर-पुस्तिका के मुख-पृष्ठ पर अवश्य लिखें।
Candidates must write the Q.P. Code on the title page of the answer-book.

- *कृपया जाँच कर लें कि इस प्रश्न-पत्र में मुद्रित पृष्ठ **27** हैं।* [ ]
- *प्रश्न-पत्र में दाहिने हाथ की ओर दिए गए प्रश्न-पत्र कोड को परीक्षार्थी उत्तर-पुस्तिका के मुख-पृष्ठ पर लिखें।*
- *कृपया जाँच कर लें कि इस प्रश्न-पत्र में **21** प्रश्न हैं।*
- ***कृपया प्रश्न का उत्तर लिखना शुरू करने से पहले, उत्तर-पुस्तिका में यथा स्थान पर प्रश्न का क्रमांक अवश्य लिखें।***
- *इस प्रश्न-पत्र को पढ़ने के लिए 15 मिनट का समय दिया गया है। प्रश्न-पत्र का वितरण पूर्वाह्न में 10.15 बजे किया जाएगा। 10.15 बजे से 10.30 बजे तक परीक्षार्थी केवल प्रश्न-पत्र को पढ़ेंगे और इस अवधि के दौरान वे उत्तर-पुस्तिका पर कोई उत्तर नहीं लिखेंगे।*
- *Please check that this question paper contains 27 printed pages.*
- *Q.P. Code given on the right hand side of the question paper should be written on the title page of the answer-book by the candidate.*
- *Please check that this question paper contains 21 questions.*
- *Please write down the Serial Number of the question in the answer-book at the given place before attempting it.*
- *15 minute time has been allotted to read this question paper. The question paper will be distributed at 10.15 a.m. From 10.15 a.m. to 10.30 a.m., the candidates will read the question paper only and will not write any answer on the answer-book during this period.*

डाटा साइन्स
DATA SCIENCE

*निर्धारित समय : **2** घण्टे* — *अधिकतम अंक : **50***

*Time allowed : **2** hours* — *Maximum Marks : **50***

सामान्य निर्देश :

(i) *कृपया निर्देशों को ध्यान से पढ़ें।*

(ii) *इस प्रश्न-पत्र में **दो** खण्डों में **21** प्रश्न हैं : **खण्ड क** और **खण्ड ख**।*

(iii) ***खण्ड क** में वस्तुनिष्ठ प्रकार के प्रश्न हैं, जबकि **खण्ड ख** में विषयपरक प्रकार के प्रश्न हैं।*

(iv) *दिए गए (5 + 16) = 21 प्रश्नों में से, उम्मीदवार को 2 घंटे के आबंटित (अधिकतम) समय में (5 + 10) = 15 प्रश्नों के उत्तर देने हैं।*

(v) *किसी विशेष खण्ड के सभी प्रश्नों को सही क्रम में करने का प्रयास किया जाना चाहिए।*

(vi) ***खण्ड क** : वस्तुनिष्ठ प्रकार के प्रश्न (24 अंक) :*

(a) *इस खण्ड में **5** प्रश्न हैं।*

(b) *कोई नकारात्मक अंकन नहीं है।*

(c) *दिए गए निर्देशों के अनुसार कीजिए।*

(d) *प्रत्येक प्रश्न/भाग के सामने आबंटित अंकों का उल्लेख किया गया है।*

(vii) ***खण्ड ख** : विषयपरक प्रकार के प्रश्न (26 अंक) :*

(a) *इस खण्ड में **16** प्रश्न हैं।*

(b) *उम्मीदवार को **10** प्रश्न करने हैं।*

(c) *दिए गए निर्देशों के अनुसार कीजिए।*

(d) *प्रत्येक प्रश्न/भाग के सामने आबंटित अंकों का उल्लेख किया गया है।*

खण्ड क

(वस्तुनिष्ठ प्रकार के प्रश्न) **(24 अंक)**

1. रोज़गार कौशल पर दिए गए ***6*** प्रश्नों में से किन्हीं ***4*** के उत्तर दीजिए। *4×1=4*

(i) एक वाक्य क्या है ?

(A) विचारों का एक समूह जो एक संपूर्ण अनुच्छेद बनाता है।

(B) शब्दों का एक समूह जो किसी संपूर्ण विचार या अर्थ का संचार करता है।

(C) नियमों का एक समूह जिसका हमें सही ढंग से लिखने के लिए पालन करना चाहिए।

(D) शब्दों का एक समूह जिसमें मूल विराम चिह्न होते हैं।

General Instructions :

(i) *Please read the instructions carefully.*

(ii) *This question paper consists of **21** questions in **two** Sections : **Section A** and **Section B**.*

(iii) ***Section A** has Objective Type Questions, whereas **Section B** contains Subjective Type Questions.*

(iv) *Out of the given (5 + 16) = 21 questions, a candidate has to answer (5 + 10) = 15 questions in the allotted (maximum) time of 2 hours.*

(v) *All questions of a particular section must be attempted in the correct order.*

(vi) ***Section A** : Objective Type Questions (24 marks) :*

(a) *This section has **5** questions.*

(b) *There is no negative marking.*

(c) *Do as per the instructions given.*

(d) *Marks allotted are mentioned against each question/part.*

(vii) ***Section B** : Subjective Type Questions (26 marks) :*

(a) *This section has **16** questions.*

(b) *A candidate has to do **10** questions.*

(c) *Do as per the instructions given.*

(d) *Marks allotted are mentioned against each question/part.*

SECTION A

(Objective Type Questions) **(24 marks)**

1. Answer any ***4*** out of the given ***6*** questions on Employability Skills. *4×1=4*

(i) What is a sentence ?

(A) A group of ideas that form a complete paragraph.

(B) A group of words that communicates a complete thought or meaning.

(C) A set of rules that we must follow to write correctly.

(D) A set of words that contains basic punctuation marks.

(ii) "तनाव" शब्द को परिभाषित कीजिए।

(iii) क्लस्टर 'ए' व्यक्तित्व विकार उन व्यक्तियों या लोगों से संबंधित हैं जो स्वभाव से _____ होते हैं।

(A) संदिग्ध

(B) भावुक

(C) आवेगी

(D) चिंतित

(iv) आईसीटी का अर्थ है :

(A) इनोवेशन एंड कोऑर्डिनेशन टीम

(B) इंफॉर्मेशन एंड कम्युनिकेशन टेक्नोलॉजी

(C) इंटीग्रेटेड एंड क्लीयर टेक्नोलॉजी

(D) इंस्पायर्ड एंड क्रिएटिव टीम

(ii) Define the term “stress”.

(iii) Cluster ‘A’ personality disorders are associated with individuals who are _____ in nature.

(A) Suspicious

(B) Emotional

(C) Impulsive

(D) Anxious

(iv) ICT stands for :

(A) Innovation and Coordination Team

(B) Information and Communication Technology

(C) Integrated and Clear Technology

(D) Inspired and Creative Team

(v) निम्नलिखित में से कौन-सा स्प्रेडशीट का घटक ***नहीं*** है ?

(A) सेल

(B) वर्कशीट

(C) वर्कबुक

(D) वर्कएरिया

(vi) भारत सरकार की उस प्रमुख पहल का नाम बताइए जिसका उद्देश्य स्टार्ट-अप व्यवसायों के विकास के लिए एक पारिस्थितिकी तंत्र का निर्माण करना है।

2. दिए गए ***6*** प्रश्नों में से किन्हीं ***5*** के उत्तर दीजिए। *5×1=5*

(i) निम्नलिखित में से कौन-सा विकल्प डेटा गवर्नेंस (Data Governance) का सबसे अच्छा वर्णन करता है ?

(A) यह एक सॉफ्टवेयर सिस्टम है जो डिजिटल डेटा को स्वचालित रूप से व्यवस्थित और स्टोर (store) करने के लिए उपयोग किया जाता है।

(B) यह डेटा को प्रभावी ढंग से प्रबंधित करने के लिए लोगों, नीतियों (policies), प्रक्रियाओं (processes) और तकनीकों का एक ढाँचा (framework) है।

(C) यह एक सरकारी विभाग है जो नेटवर्क प्रणालियों की निगरानी और सुरक्षा के लिए जिम्मेदार है।

(D) यह एक प्रोग्रामिंग भाषा है जो क्लाउड डेटाबेस (cloud databases) की सुरक्षा और प्रबंधन के लिए डिज़ाइन की गई है।

(v) Which of the following is ***not*** a component of a spreadsheet ?

(A) Cell

(B) Worksheet

(C) Workbook

(D) Workarea

(vi) Name the flagship initiative of the Government of India that is intended to build an ecosystem for the growth of start-up businesses.

2. Answer any ***5*** out of the given ***6*** questions. *5×1=5*

(i) Which of the following options best describes Data Governance ?

(A) It is a software system used to organize and store digital data automatically.

(B) It is a framework of people, policies, processes, and technologies for managing data effectively.

(C) It is a government department responsible for monitoring and securing network systems.

(D) It is a programming language designed for protecting and managing cloud databases.

(ii) कौन-सा ग्राफ़ विशेष रूप से किसी चर (variable) की क्वांटाइल श्रेणियों (quantile ranges) को दिखाने और आउटलायर्स (outliers) की उपस्थिति का पता लगाने में मदद करने के लिए उपयोग किया जाता है ?

(A) बॉक्स प्लॉट (Box plot)

(B) स्कैटर प्लॉट (Scatter plot)

(C) फ़्रीक्वेंसी पॉलीगॉन (Frequency polygon)

(D) पाई चार्ट (Pie chart)

(iii) एक डिसीजन ट्री (decision tree) में लीफ नोड्स (leaf nodes) क्या दर्शाते हैं ?

(A) ट्री (tree) का प्रारंभिक बिंदु

(B) विफलता (failure) की संभावना

(C) प्रत्येक निर्णय (decision) के परिणाम (result)

(D) एकत्रित किया गया शोध डेटा

(iv) K-NN एल्गोरिदम में K के लिए विषम (odd) मान चुनने की सलाह आमतौर पर क्यों दी जाती है ?

(A) यह पूर्वानुमान (prediction) प्रक्रिया के दौरान आवश्यक मेमोरी (memory) की मात्रा को कम करता है।

(B) यह पड़ोसी डेटा बिंदुओं (data points) के बीच की दूरी की गणना को गति देता है।

(C) यह बहुमत मतदान (majority vote) करते समय एल्गोरिदम (algorithm) की सहायता करता है।

(D) यह सुनिश्चित करता है कि निर्णय सीमा (decision boundary) पूरी तरह से रैखिक (linear) बनी रहे।

(ii) Which graph is specifically used to show the quantile ranges of a variable and help detect the presence of outliers ?

(A) Box plot

(B) Scatter plot

(C) Frequency polygon

(D) Pie chart

(iii) What do the leaf nodes represent in a decision tree ?

(A) The starting point of the tree

(B) The probability of failure

(C) The results of each decision

(D) The research data collected

(iv) Why is it generally recommended to select odd values for K in the K-NN algorithm ?

(A) It reduces the amount of memory required during the prediction process.

(B) It speeds up the calculation of distances between neighbouring data points.

(C) It helps the algorithm while taking a majority vote.

(D) It ensures that the decision boundary remains completely linear.

(v) रैखिक रिग्रेशन (linear regression) को "रैखिक" (linear) के रूप में क्यों वर्गीकृत किया गया है ?

(A) क्योंकि डेटा बिंदु (data points) हमेशा पूरी तरह से एक सीधी रेखा में होते हैं।

(B) क्योंकि स्वतंत्र चर (independent variable) के साथ आश्रित चर (dependent variable) का संबंध कुछ मापदंडों (parameters) का एक रैखिक फंक्शन (linear function) होता है।

(C) क्योंकि यह केवल सकारात्मक संबंधों (relationships) को ही मॉडल कर सकता है।

(D) क्योंकि यह केवल सकारात्मक पूर्णांक मानों के पूर्वानुमान (predicting) तक ही सीमित है।

(vi) निम्नलिखित कथन सत्य है या असत्य बताइए :

अनसुपरवाइज्ड लर्निंग (Unsupervised Learning) का उपयोग एक्सप्लोरेटरी डेटा एनालिसिस (Exploratory Data Analysis) के लिए नहीं किया जा सकता है।

3. दिए गए ***6*** प्रश्नों में से किन्हीं ***5*** के उत्तर दीजिए। *5×1=5*

(i) डेटा को संभालने (handling) के लिए नैतिक दिशानिर्देशों (ethical guidelines) के संबंध में निम्नलिखित दो कथनों पर विचार कीजिए :

कथन I : हमें ऐसे मशीन लर्निंग मॉडल डिज़ाइन करने चाहिए जो मजबूत (robust) और निष्पक्ष हों।

कथन II : हमें ऐसी तकनीकों और डेटा आर्किटेक्चर (data architecture) का उपयोग करना चाहिए जिनमें अधिकतम अतिक्रमण (intrusion) संभव हो।

विकल्प :

(A) कथन I और कथन II दोनों सही हैं।

(B) कथन I सही है, लेकिन कथन II ग़लत है।

(C) कथन I ग़लत है, लेकिन कथन II सही है।

(D) कथन I और कथन II दोनों ग़लत हैं।

(v) Why is linear regression classified as "linear" ?

(A) Because the data points always lie perfectly in a straight line.

(B) Because the relationship of the dependent variable to the independent variable is a linear function of some parameters.

(C) Because it can only model positive relationships.

(D) Because it is limited to predicting positive integer values.

(vi) State whether the following statement is True or False :

Unsupervised Learning cannot be used for Exploratory Data Analysis.

3. Answer any ***5*** out of the given ***6*** questions. *5×1=5*

(i) Consider the following two statements regarding ethical guidelines for handling data :

Statement I : We should design machine learning models that are robust and impartial.

Statement II : We should use technologies and data architecture that has maximum possible intrusion.

Options :

(A) Both Statement I and Statement II are correct.

(B) Statement I is correct, but Statement II is incorrect.

(C) Statement I is incorrect, but Statement II is correct.

(D) Both Statement I and Statement II are incorrect.

(ii) डेटा साइंस में डेटा क्लीनिंग (data cleaning) क्यों महत्त्वपूर्ण है ?

(A) यह इंटरनेट सिस्टम की गति में सुधार करता है।

(B) यह एल्गोरिदम को बेहतर निष्कर्ष (insights) देने में मदद करता है।

(C) यह डेटासेट से महत्त्वपूर्ण डेटा को हटा देता है।

(D) यह डेटासेट में डुप्लिकेट (duplicate) रिकॉर्ड बनाता है।

(iii) एक क्लासिफिकेशन ट्री (classification tree) में, अनदेखे डेटा पर पूर्वानुमान (predictions) कैसे लगाए जाते हैं ?

(A) अवलोकनों (observations) के माध्य (mean) का उपयोग करके

(B) अवलोकनों के माध्यिका (median) का उपयोग करके

(C) अवलोकनों के रैंडम सैंपलिंग (random sampling) का उपयोग करके

(D) अवलोकनों के बहुलक (mode) का उपयोग करके

(iv) के-नियरेस्ट नेबर्स (K-Nearest Neighbors – KNN) एल्गोरिदम में, विभिन्न वर्गों (classes) को अलग करने वाली ज्यामितीय सीमा (geometric boundary) का वर्णन करने के लिए किस शब्द का उपयोग किया जाता है ?

(A) नेबर बाउंड्री (Neighbor boundary)

(B) क्लास बाउंड्री लाइन (Class boundary line)

(C) डिसीजन सरफेस (Decision surface)

(D) डायमेंशनल सेपरेटर (Dimensional separator)

(v) लाइन ऑफ बेस्ट फिट (line of best fit) और डेटा बिंदुओं (data points) के बीच की ऊर्ध्वाधर दूरी (vertical distance) को न्यूनतम करने का प्रयास करने की बुनियादी प्रक्रिया को किस रूप में जाना जाता है ?

(A) डेटासेट को वर्गीकृत करना (Categorizing the dataset)

(B) आयाम में कमी करना (Dimension reduction)

(C) ट्री नोड्स को विभाजित करना (Splitting the tree nodes)

(D) लाइन को डेटा में फिट करना (Fitting the line to the data)

(ii) Why is data cleaning important in data science ?

(A) It improves the speed of Internet systems.

(B) It helps algorithms produce better insights.

(C) It removes important data from datasets.

(D) It creates duplicate records in datasets.

(iii) In a classification tree, how are predictions on unseen data made ?

(A) Using the mean of the observations

(B) Using the median of the observations

(C) Using a random sampling of the observations

(D) Using the mode of the observations

(iv) In the K-Nearest Neighbors (K-NN) algorithm, what is the term used to describe the geometric boundary that separates different classes ?

(A) Neighbor boundary

(B) Class boundary line

(C) Decision surface

(D) Dimensional separator

(v) The basic process of trying to reduce the vertical distance between the line of best fit and the data points to make it minimum is known as :

(A) Categorizing the dataset

(B) Dimension reduction

(C) Splitting the tree nodes

(D) Fitting the line to the data

(vi) के-मीन्स क्लस्टरिंग एल्गोरिदम (K-means clustering algorithm) में सेंट्रोइड्स (centroids) को इनिशियलाइज़ (initialize) करने के लिए, विधि में डेटा बिंदुओं को _____ करना और फिर रैंडम तरीके से (randomly) _____ डेटा बिंदुओं का चयन करना शामिल है।

(A) नॉर्मलाइज़िंग (normalizing), दो (two)

(B) शफल करना (shuffling), K

(C) फ़िल्टर करना (filtering), रैंडम (random)

(D) क्लस्टर करना (clustering), सभी (all)

4. दिए गए ***6*** प्रश्नों में से किन्हीं ***5*** के उत्तर दीजिए। *5×1=5*

(i) डेटा गोपनीयता अधिकारों (data privacy rights) के संदर्भ में निम्नलिखित में से कौन-सा कथन सही है ?

(A) व्यक्तियों का इस बात पर नियंत्रण होता है कि उनका व्यक्तिगत डेटा कैसे एकत्र और उपयोग किया जाता है।

(B) व्यक्तियों को अपने व्यक्तिगत डेटा को ऑनलाइन एन्क्रिप्टेड (encrypted) रखने के लिए मासिक शुल्क का भुगतान करना होता है।

(C) व्यक्तियों को अपनी फाइलों तक पहुँचने (access) के लिए एक वैश्विक प्राधिकरण (global authority) के साथ पंजीकरण करना होता है।

(D) कंपनियाँ स्वचालित रूप से उनके द्वारा एकत्र किए गए सभी व्यक्तिगत डेटा की स्वामी बन जाती हैं।

(ii) निम्नलिखित कथनों पर विचार कीजिए :

कथन I : लीनियर रिग्रेशन मॉडल (linear regression models), डिसीजन ट्री मॉडल (decision tree models) की तुलना में आउटलायर्स (outliers) के प्रति अधिक मज़बूत (robust) होते हैं।

कथन II : सभी आउटलायर्स अमान्य डेटा बिंदु (invalid data points) होते हैं और मॉडल के प्रदर्शन को बेहतर बनाने के लिए उन्हें हमेशा हटाया जाना चाहिए।

विकल्प :

(A) कथन I और कथन II दोनों सही हैं।

(B) कथन I सही है, लेकिन कथन II ग़लत है।

(C) कथन I ग़लत है, लेकिन कथन II सही है।

(D) कथन I और कथन II दोनों ग़लत हैं।

(vi) To initialize centroids in the K-means clustering algorithm, the method involves _____ the data points and then randomly selecting _____ data points.

(A) normalizing, two

(B) shuffling, K

(C) filtering, random

(D) clustering, all

4. Answer any ***5*** out of the given ***6*** questions. *5×1=5*

(i) Which of the following statements is correct in the context of data privacy rights ?

(A) Individuals have control over how their personal data is collected and used.

(B) Individuals must pay a monthly fee to keep their personal data encrypted online.

(C) Individuals need to register with a global authority to access their own files.

(D) Companies automatically become the owners of all personal data they collect.

(ii) Consider the following statements :

Statement I : Linear regression models are more robust to outliers than decision tree models.

Statement II : All outliers are invalid data points and must always be removed to improve model performance.

Options :

(A) Both Statement I and Statement II are correct.

(B) Statement I is correct, but Statement II is incorrect.

(C) Statement I is incorrect, but Statement II is correct.

(D) Both Statement I and Statement II are incorrect.

(iii) एक असंतुलित डेटासेट (imbalanced dataset) पर प्रशिक्षित होने पर K-NN आमतौर पर कैसा व्यवहार करता है ?

(A) मॉडल दूरी मेट्रिक्स (distance metrics) का उपयोग करके डेटासेट को स्वचालित रूप से संतुलित करता है।

(B) मॉडल बहुसंख्यक (majority) और अल्पसंख्यक (minority) दोनों वर्गों (classes) पर समान रूप से अच्छा प्रदर्शन करता है।

(C) मॉडल उस वर्ग (class) के प्रति पक्षपाती (biased) हो जाता है जिसमें अधिकांश प्रशिक्षण डेटा (training data) होता है।

(D) मॉडल केवल मल्टी-क्लास समस्याओं के साथ बहुमत मतदान (majority voting) करने में विफल रहता है।

(iv) निम्नलिखित कथन सत्य है या असत्य बताइए :

आश्रित (dependent) और स्वतंत्र (independent) चरों के बीच के संबंध के प्रकार के आधार पर, आश्रित चर में किया जाने वाला समायोजन केवल धनात्मक (positive) या ऋणात्मक (negative) हो सकता है, लेकिन कभी भी शून्य नहीं हो सकता।

(v) ग़ैर-रेखीय फंक्शन्स (non-linear functions) में _____ फंक्शन्स (functions) और _____ फंक्शन्स (functions) शामिल हैं।

(A) रेखीय (linear), स्थिर (constant)

(B) घातांकीय (exponential), लघुगणकीय (logarithmic)

(C) सरल (simple), एकाधिक (multiple)

(D) एडिटिव (additive), सब्ट्रैक्टिव (subtractive)

(vi) निम्नलिखित कथन सत्य है या असत्य बताइए :

अनसुपरवाइज्ड लर्निंग (unsupervised learning) में, एल्गोरिदम (algorithms) को डेटा को प्रोसेस (process) करने और समानताओं की खोज करने के लिए निरंतर मानवीय हस्तक्षेप की आवश्यकता होती है।

(iii) How does K-NN typically behave when trained on an imbalanced dataset ?

(A) The model automatically balances the dataset using distance metrics.

(B) The model performs equally well on both majority and minority classes.

(C) The model becomes biased towards the class having the majority of training data.

(D) The model fails to perform majority voting with only multiclass problems.

(iv) State whether the following statement is True or False :

Depending on the kind of relationship between the dependent and independent variables, the adjustment to the dependent variable can only be positive or negative, but never zero.

(v) Non-linear functions include _____ functions and _____ functions.

(A) linear, constant

(B) exponential, logarithmic

(C) simple, multiple

(D) additive, subtractive

(vi) State whether the following statement is True or False :

In unsupervised learning, algorithms require continuous human intervention to process data and discover similarities.

5. दिए गए ***6*** प्रश्नों में से किन्हीं ***5*** के उत्तर दीजिए। *5×1=5*

(i) सामान्य डेटा संरक्षण विनियमन (General Data Protection Regulation) का मुख्य उद्देश्य क्या है ?

(A) ऑनलाइन विज्ञापन बढ़ाना

(B) यूरोपियन यूनियन के उपभोक्ता डेटा की सुरक्षा करना

(C) कंपनियों को उपयोगकर्ता के डेटा को बेचने की अनुमति देना

(D) यूरोपियन यूनियन के उपभोक्ता के इंटरनेट उपयोग को कम करना

(ii) निम्नलिखित कथनों पर विचार कीजिए :

कथन I : डिसीजन ट्री (Decision trees) डेटा में ग़ैर-रेखीय रिलेशनशिप (non-linear relationships) को मैप (map) करने में सक्षम नहीं हैं।

कथन II : डिसीजन ट्री (Decision Tree) में, मुख्य उद्देश्य (main objective) संपूर्ण डायग्राम (diagram) का रूट (root) होना चाहिए।

विकल्प :

(A) कथन I और कथन II दोनों सही हैं।

(B) कथन I सही है, लेकिन कथन II ग़लत है।

(C) कथन I ग़लत है, लेकिन कथन II सही है।

(D) कथन I और कथन II दोनों ग़लत हैं।

(iii) जैसे-जैसे डेटासेट (dataset) का आकार बढ़ता है, K-NN की गति तेजी से कम क्यों होती है ?

(A) क्योंकि K-NN को पूर्वानुमान (prediction) लगाने के लिए डेटासेट के सभी बिंदुओं का मूल्यांकन करने की आवश्यकता होती है।

(B) क्योंकि K-NN में K का मान निश्चित होता है।

(C) क्योंकि K-NN को हर चरण (step) पर संभाव्यता वितरण मापदंडों (probability distribution parameters) का अनुमान लगाना होता है।

(D) क्योंकि K-NN पूर्वानुमान (prediction) चरण के दौरान 10-फोल्ड क्रॉस-वैलिडेशन (10-fold cross-validation) करता है।

5. Answer any ***5*** out of the given ***6*** questions. *5×1=5*

(i) What is the main purpose of the General Data Protection Regulation ?

(A) To increase online advertising

(B) To protect the European Union's consumer data

(C) To allow companies to sell user data

(D) To reduce the European Union's consumer Internet usage

(ii) Consider the following statements :

Statement I : Decision trees are not capable of mapping non-linear relationships in data.

Statement II : In a decision tree, the main objective should be the root of the entire diagram.

Options :

(A) Both Statement I and Statement II are correct.

(B) Statement I is correct, but Statement II is incorrect.

(C) Statement I is incorrect, but Statement II is correct.

(D) Both Statement I and Statement II are incorrect.

(iii) Why does the speed of K-NN decline rapidly as the size of the dataset grows ?

(A) Because K-NN needs to evaluate all the points in the dataset to make a prediction.

(B) Because K-NN has a fixed value of K.

(C) Because K-NN must estimate probability distribution parameters at every step.

(D) Because K-NN performs 10-fold cross-validation during the prediction phase.

(iv) _____ चर शब्द का अर्थ है कि इसका मान इसकी इच्छानुसार (स्वेच्छा से) चुना जा सकता है, और _____ चर इसके मान के आधार पर समायोजित होगा।

(A) आश्रित (dependent); स्वतंत्र (independent)

(B) स्वतंत्र (independent); आश्रित (dependent)

(C) त्रुटि (error); अवशिष्ट (residual)

(D) अंतःखंड (intercept); ढाल (slope)

(v) मल्टीपल लीनियर रिग्रेशन के मूल मॉडल

$(Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \ldots + \beta_p X_{ip} + \varepsilon_i)$

में, पद ε_i क्या दर्शाता है ?

(A) iवें आश्रित चर का अंतःखंड (intercept)

(B) iवें स्वतंत्र चरों का ढाल गुणांक (slope coefficient)

(C) iवीं स्वतंत्र, समान रूप से वितरित सामान्य त्रुटि

(D) स्वतंत्र चर का iवाँ प्रेक्षण (observation)

(vi) के-मीन्स क्लस्टरिंग (K-means clustering) प्रक्रिया के दौरान, एल्गोरिदम प्रत्येक डेटा बिंदु (data point) और सभी सेंट्रॉइड्स (centroids) के बीच _____ के योग की गणना करता है।

(A) निरपेक्ष अंतर (absolute difference)

(B) वर्ग दूरी (squared distance)

(C) निर्देशांक गुणनफल (coordinate product)

(D) व्युत्क्रम मान (inverse values)

(iv) The term _____ variable means that its value can be chosen at will, and the _____ variable will adjust based on its value.

(A) dependent; independent

(B) independent; dependent

(C) error; residual

(D) intercept; slope

(v) In the basic model of multiple linear regression,

$(Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \ldots + \beta_p X_{ip} + \varepsilon_i)$,

what does the term ε_i represent ?

(A) The ith dependent variable's intercept

(B) The slope coefficient of the ith independent variables

(C) The ith independent, identically distributed normal error

(D) The ith observation of the independent variable

(vi) During the K-means clustering process, the algorithm calculates the sum of the _____ between each data point and all centroids.

(A) absolute difference

(B) squared distance

(C) coordinate product

(D) inverse values

खण्ड ख

(विषयपरक प्रकार के प्रश्न) **(26 अंक)**

*रोज़गार कौशल पर दिए गए **5** प्रश्नों में से किन्हीं **3** के उत्तर दीजिए। प्रत्येक प्रश्न का उत्तर 20 – 30 शब्दों में दीजिए।* *3×2=6*

6. "संचार" शब्द से आप क्या समझते हैं ?

7. बृहत् पाँच कारक (बिग फाइव फैक्टर) मॉडल से संबंधित व्यक्ति के व्यक्तित्व के किन्हीं चार मानदंडों की व्याख्या कीजिए।

8. आईसीटी में डेटा सॉर्टिंग क्या है ?

9. उद्यमिता में कोई दो पर्यावरणीय बाधाएँ बताइए।

10. यूएनईपी के अनुसार "ग्रीन जॉब्स" शब्द को परिभाषित कीजिए।

*दिए गए **6** प्रश्नों में से किन्हीं **4** के उत्तर 20 – 30 शब्दों (प्रत्येक) में दीजिए।* *4×2=8*

11. (क) डेटा गवर्नेंस (Data Governance) के अंतर्गत आने वाले किन्हीं दो पहलुओं (aspects) के नाम बताइए।

(ख) स्पष्ट कीजिए कि सॉफ्टवेयर उत्पादों और डेटा के साथ काम करते समय नैतिक दिशानिर्देशों (ethical guidelines) का पालन करने की आवश्यकता क्यों होती है।

12. बाइवेरिएट एनालिसिस (bivariate analysis) क्या है ? वास्तविक जीवन के एक उदाहरण के साथ स्पष्ट कीजिए।

13. के-नियरेस्ट नेबर्स (K-NN) एल्गोरिदम के उपयोग के दो लाभ (pros) और दो नुकसान (cons) बताइए।

14. डिसीजन ट्री (decision tree) को परिभाषित कीजिए। इसकी संरचना में प्रत्येक आंतरिक नोड (internal node) और शाखा (branch) क्या दर्शाती है ?

15. (क) मूल माध्य वर्ग विचलन (root mean square deviation) गणितीय रूप से क्या निर्धारित करता है ?

(ख) मीन एब्सोल्यूट एरर (MAE) को परिभाषित कीजिए।

SECTION B

(Subjective Type Questions) **(26 Marks)**

*Answer any **3** out of the given **5** questions on Employability Skills. Answer each question in 20 – 30 words.* *3×2=6*

6. What do you understand by the term "communication" ?

7. Explain any four parameters of an individual's personality related to the Big Five Factor Model.

8. What is data sorting in ICT ?

9. State any two environmental barriers to entrepreneurship.

10. Define the term "green jobs" according to the UNEP.

*Answer any **4** out of the given **6** questions in 20 – 30 words each.* *4×2=8*

11. (a) Name any two aspects covered by Data Governance.

(b) Explain why there is a need to adhere to ethical guidelines when dealing with software products and data.

12. What is bivariate analysis ? Illustrate with a real-life example.

13. Mention two pros and two cons of using the K-Nearest Neighbors (K-NN) algorithm.

14. Define a decision tree. What does each internal node and branch represent in its structure ?

15. (a) What does Root Mean Square Deviation determine mathematically ?

(b) Define Mean Absolute Error (MAE).

16. समझाइए कि निम्नलिखित वास्तविक जीवन के अनुप्रयोगों (applications) में अनसुपरवाइज्ड लर्निंग (unsupervised learning) का उपयोग कैसे किया जाता है :

(क) कस्टमर पर्सोना (Customer Personas) को समझना

(ख) समाचार अनुभाग (Sections) बनाना

*दिए गए **5** प्रश्नों में से किन्हीं **3** के उत्तर 50 – 80 शब्दों (प्रत्येक) में दीजिए।* *3×4=12*

17. (क) हेल्थ इंश्योरेंस पोर्टेबिलिटी एंड अकाउंटेबिलिटी एक्ट (HIPAA) का उद्देश्य क्या है, और यह व्यक्तियों को डेटा पर नियंत्रण कैसे वापस देता है ?

(ख) चिल्ड्रन्स ऑनलाइन प्राइवेसी एंड प्रोटेक्शन एक्ट (COPPA) किस आयु वर्ग की रक्षा करता है ? COPPA द्वारा कंपनियों पर लागू किए जाने वाली ज़िम्मेदारियों का उल्लेख कीजिए।

18. (क) एक्सप्लोरेटरी डेटा एनालिसिस (exploratory data analysis) को परिभाषित कीजिए।

(ख) मल्टीवेरिएट एनालिसिस (multivariate analysis) की किन्हीं दो विधियों के नाम लिखिए।

(ग) बाइवेरिएट एनालिसिस (bivariate analysis) के लिए उपयोग की जाने वाली दो ग्राफिकल तकनीकों का उल्लेख कीजिए।

(घ) डेटा क्लीनिंग (data cleaning) के दौरान अप्रासंगिक प्रेक्षणों (irrelevant observations) को क्यों हटाया जाता है ?

19. (क) डिसीजन ट्री (decision trees) द्वारा हल की जा सकने वाली समस्याओं के प्रकार और उनके द्वारा मैप किए जा सकने वाले संबंधों (relationships) के संदर्भ में उनकी बहुमुखी प्रतिभा (versatility) होने की व्याख्या कीजिए। कोई दो बिंदु दीजिए।

(ख) डिसीजन ट्री के आउटपुट (outputs) को ग़ैर-विश्लेषणात्मक पृष्ठभूमि (non-analytical background) के लोगों के लिए भी आसानी से समझने योग्य क्या बनाता है ?

(ग) नमूना डिसीजन ट्री (sample decision tree) का एक ब्लॉक आरेख (block diagram) बनाइए और डिसीजन नोड्स (decision nodes) तथा लीफ़ नोड्स (leaf nodes) को नामांकित (label) कीजिए।

16. Explain how unsupervised learning is used in the following real-life applications :

(a) Understanding Customer Personas

(b) Forming News Sections

*Answer any **3** out of the given **5** questions in 50 – 80 words each.* *3×4=12*

17. (a) What is the purpose of the Health Insurance Portability and Accountability Act (HIPAA), and how does it return control of data to individuals ?

(b) What age group does the Children's Online Privacy and Protection Act (COPPA) protect ? State the responsibilities COPPA imposes on companies.

18. (a) Define Exploratory Data Analysis.

(b) Name any two methods to perform multivariate analysis.

(c) Mention two graphical techniques used to perform bivariate analysis.

(d) Why are irrelevant observations removed during data cleaning ?

19. (a) Explain the versatility of decision trees in terms of the types of problems they can solve and the relationships they can map. Give any two points.

(b) What makes decision tree outputs easily understandable even to individuals from a non-analytical background ?

(c) Draw a block diagram of a sample decision tree and label the decision nodes and leaf nodes.

20. (क) मल्टीपल लीनियर रिग्रेशन (Multiple Linear Regression) क्या है ?

(ख) निम्नलिखित कथन सत्य है या असत्य बताइए :

"नॉन-लीनियर रिग्रेशन (Non-linear regression), लीनियर रिग्रेशन (linear regression) की तुलना में अधिक लचीला (flexible) होता है।"

(ग) नॉन-लीनियर रिग्रेशन (non-linear regression) के लिए सामान्य गणितीय सूत्र लिखिए।

(घ) लीनियर रिग्रेशन (linear regression) का ग्राफ़िकल निरूपण (graphical representation), नॉन-लीनियर रिग्रेशन (non-linear regression) के ग्राफ़िकल निरूपण से किस प्रकार भिन्न होता है ?

21. (क) सुपरवाइज्ड लर्निंग एल्गोरिदम (supervised learning algorithms) को प्रशिक्षित (training) करने के लिए किस प्रकार के डेटा की आवश्यकता होती है ?

(ख) संक्षेप में क्लस्टरिंग (clustering) को समझाइए और वास्तविक-जीवन का एक उदाहरण दीजिए।

(ग) मेडिकल इमेजिंग (medical imaging) में अनसुपरवाइज्ड मशीन लर्निंग (unsupervised machine learning) द्वारा प्रदान की जाने वाली किन्हीं दो सुविधाओं का उल्लेख कीजिए।

20. (a) What is Multiple Linear Regression ?

(b) State whether the following statement is True or False :

"Non-linear regression is more flexible than linear regression."

(c) Write the general mathematical formula for non-linear regression.

(d) How does the graphical representation of linear regression differ from that of non-linear regression ?

21. (a) What type of data is required for training supervised learning algorithms ?

(b) Explain clustering in brief and give one real-life example.

(c) Mention any two facilities provided by unsupervised machine learning in medical imaging.