School Connect Online

SCO INTERNATIONAL OLYMPIAD

CLASS 7 ARTIFICIAL INTELLIGENCE OLYMPIAD

Official Rebranded Question Paper for students, teachers, schools, and parents

Designed from Class 7 Artificial Intelligence learning pathways and aligned with SCO's guided preparation, practice, reporting, and future-ready academic growth.

- age-fit AI literacy for middle-school learners globally
- objective MCQ practice across machine learning, datasets, privacy, AI applications, and ethics
- answer key and explanations for guided revision and classroom discussion

Maths	English	Science	Mental Ability	Finance Knowledge
AI	Entrepreneurship	GK	Coding	Life Skills

SCO International Artificial Intelligence Olympiad - Class 7

Official Rebranded Question Paper | Question Paper Set S | 2025-26

Total Questions	Time	Question Type	Sections
50	60 minutes	Objective MCQ	4

Guidelines for the Candidate

1. Before the exam begins, candidates may use the additional time given by the invigilator to complete OMR or identity information.

2. Clearly write name, school code, class, registration ID, and contact number wherever required.

3. This paper contains 50 objective-type questions. Each question has only one correct answer.

4. All questions should be attempted. There is no negative marking unless separately notified by the exam authority.

5. Calculators, mobile phones, smart watches, books, notes, and external help are not allowed.

6. Use only the permitted pen or pencil to mark answers on the OMR sheet or answer sheet.

7. At the conclusion of the test, hand over the OMR sheet or answer sheet to the invigilator.

Name:	Registration ID: .....
School:	Contact No.: .....

Section 1 - Machine Learning Basics

Q.1

Which of the following best describes supervised learning?

A. A method where a computer learns by using examples that already have correct answers.

B. A method where a computer groups similar items without any hints.

C. A method where a computer learns by playing games with itself.

D. A method where a computer only follows fixed rules without learning.

Answer: A

Explanation: Supervised learning uses labeled examples (data with correct answers) so the computer can learn the relationship between input and output. This is similar to studying with a teacher's help.

Q.2

In unsupervised learning, the computer is given data without labels. What does it try to do with that data?

A. It tries to memorize the data exactly.

B. It tries to find hidden patterns or groups in the data.

C. It ignores the data entirely.

D. It automatically generates labels from the teacher.

Answer: B

Explanation: Unsupervised learning works with unlabeled data and helps the computer find patterns or clusters, grouping similar items together without knowing the "right" answers beforehand.

Q.3

Which example is a clear case of supervised learning?

A. Grouping animals by their colors from a list of pictures.

B. Predicting tomorrow's weather using historical weather data with labels such as "rainy" or "sunny."

C. Sorting fruits into different baskets without any instructions.

D. Clustering songs by genre without knowing the genre names.

Answer: B

Explanation: Predicting tomorrow's weather using historical data with labels is supervised learning because the computer learns from past examples that include the outcome (rainy, sunny, etc.).

Q.4

Imagine you have a pile of handwritten numbers and you want a computer to recognize them. If you show the computer each number along with its correct label (e.g., "5"), what type of learning is that?

A. Unsupervised learning

B. Supervised learning

C. Reinforcement learning

D. Deep learning without labels

Answer: B

Explanation: When you provide both the handwritten number and its label, the computer learns from these examples. This is a classic example of supervised learning.

Q.5

Which of the following best describes unsupervised learning using a real-life example?

A. A teacher giving students a set of math problems with answers.

B. Sorting a basket of mixed fruits into groups based on similarity without knowing the names of the fruits.

C. Reading a book with an answer key at the end.

D. Using a calculator to solve a math problem.

Answer: B

Explanation: Sorting fruits into groups based on similarities without knowing the names is an example of unsupervised learning because the computer (or person) discovers the grouping on its own without given labels.

Q.6

Which statement correctly explains the difference between supervised and unsupervised learning?

A. Supervised learning uses labeled data, while unsupervised learning uses unlabeled data.

B. Supervised learning is only used for pictures, and unsupervised is only used for text.

C. Supervised learning is slower than unsupervised learning in all cases.

D. Unsupervised learning always provides a correct answer, while supervised learning does not.

Answer: A

Explanation: The key difference is that supervised learning uses data with known answers (labels) to guide the training process, while unsupervised learning works with data that has no labels, aiming to discover underlying patterns.

Q.7

A company wants to recommend movies to users based on the movies they have watched before. If the computer uses information about users' past movie ratings (labels) to suggest new movies, this is an example of:

A. Unsupervised learning

B. Supervised learning

C. Both supervised and unsupervised learning

D. Reinforcement learning

Answer: B

Explanation: The system uses historical data that includes users' ratings (the labels) to predict which movies they might like. This use of labeled data makes it a supervised learning problem.

Q.8

In a classroom project, students are asked to group a collection of different animal pictures into categories without any hints. Which type of machine learning does this activity represent?

A. Supervised learning

B. Unsupervised learning

C. Reinforcement learning

D. Semi-supervised learning

Answer: B

Explanation: Since the students are grouping the pictures without any pre-assigned categories or labels, they are performing unsupervised learning—finding patterns on their own.

Q.9

Which scenario demonstrates a potential disadvantage of supervised learning if the provided labels are incorrect?

A. The computer might learn wrong associations and make incorrect predictions.

B. The computer will ignore the incorrect labels and learn the true pattern.

C. The computer will automatically correct the wrong labels during training.

D. The computer will switch to unsupervised learning automatically.

Answer: A

Explanation: If the labels are incorrect in supervised learning, the computer will learn from these mistakes and eventually make wrong predictions, because it relies heavily on the quality of the provided data.

Q.10

Why is it often easier for a teacher to explain supervised learning compared to unsupervised learning?

A. Because supervised learning has a clear “right answer” for each example, making it easier to show correct and incorrect results.

B. Because unsupervised learning does not work at all.

C. Because supervised learning does not require any examples.

D. Because unsupervised learning is only used in very advanced topics.

Answer: A

Explanation: Supervised learning uses labeled data, which means there is a clear answer that the teacher can show, making it easier to demonstrate how the learning process works. Unsupervised learning, on the other hand, involves discovering hidden patterns without predefined answers.

Section 2 - Data Collection, Quality, Privacy and AI Workflow

Q.11

A data scientist is training a supervised learning model on a dataset of images that have been manually labeled. However, some labels are noisy (incorrect). Which of the following strategies is most effective in mitigating the impact of these noisy labels?

A. Increase the model complexity to better fit the noisy data.

B. Use data augmentation techniques to create more training samples without addressing the noise.

C. Implement robust loss functions or label cleaning methods to reduce the influence of incorrect labels.

D. Switch to an unsupervised learning approach to ignore labels completely.

Answer: C

Explanation: Robust loss functions (e.g., Huber loss) or label cleaning (e.g., using a small trusted subset for re-labeling) can mitigate the negative impact of noisy labels in supervised learning. Increasing model complexity (A) can lead to overfitting on noise, while unsupervised learning (D) discards valuable label information.

Q.12

Consider the following Python code that uses scikit-learn to perform both supervised and unsupervised learning on a simple dataset:

```
from sklearn.datasets import make_blobs

from sklearn.cluster import KMeans

from sklearn.linear_model import LogisticRegression

import numpy as np

# Create a dataset with 3 clusters

X, y_true = make_blobs(n_samples=150, centers=3, random_state=42)

# Unsupervised: KMeans clustering (ignores y_true)

kmeans = KMeans(n_clusters=3, random_state=42)

cluster_labels = kmeans.fit_predict(X)

# Supervised: Logistic Regression using the true labels

clf = LogisticRegression(multi_class='multinomial', max_iter=200)

clf.fit(X, y_true)

y_pred = clf.predict(X)
```

Which of the following statements is true regarding the two approaches used in this code?

A. KMeans clustering and Logistic Regression produce identical labels because the data is well separated.

B. Logistic Regression uses the true labels to achieve better accuracy, while KMeans groups the data based solely on feature similarities.

C. Both methods use the same internal algorithm, so the performance differences are negligible.

D. KMeans clustering outperforms Logistic Regression in supervised accuracy since it clusters data based on centroids.

Answer: B

Explanation: Logistic Regression is a supervised learning method that utilizes true labels (y_true) during training, which typically leads to higher accuracy when labels are correct. KMeans, an unsupervised method, clusters based solely on feature similarities without label guidance. They are inherently different, and B correctly distinguishes them.

Q.13

A mobile app collects user behavior data. Which of the following is an example of explicit data collection, and why is explicit data valuable in supervised learning contexts?

A. Collecting the user's GPS coordinates in the background.

B. Asking users to fill out a survey about their app preferences.

C. Tracking button clicks without user awareness.

D. Logging the time spent on each screen automatically.

Answer: B

Explanation: Explicit data collection involves directly asking the user for information (such as survey responses), which tends to be more accurate and reliable. This type of data is particularly useful in supervised learning because it can serve as a clear label or feature, whereas implicit data (A, C, D) is gathered passively and may require further processing to interpret correctly.

Q.14

A company is using a large dataset containing personal information to train a recommendation model. Which of the following measures best ensures data privacy while maintaining data utility?

A. Encrypt all data during training and storage without any further processing.

B. Remove all personally identifiable information (PII) and apply differential privacy techniques before model training.

C. Share raw data with the model developers to improve model performance.

D. Increase the dataset size to dilute the effect of individual data points.

Answer: B

Explanation: Removing PII and applying differential privacy techniques add layers of protection to ensure that individual data cannot be re-identified while preserving the overall patterns needed for training. Option A secures data but does not anonymize it, while options C and D are inadequate for ensuring privacy.

Q.15

Consider the following Python code that applies KMeans clustering to a dataset. Which metric would best help evaluate the quality of clustering when labels are not available?

```
from sklearn.datasets import make_blobs

from sklearn.cluster import KMeans

from sklearn.metrics import silhouette_score
```

X, _ = make_blobs(n_samples=200, centers=4, random_state=10)

```
kmeans = KMeans(n_clusters=4, random_state=10)

cluster_labels = kmeans.fit_predict(X)

score = silhouette_score(X, cluster_labels)

print(round(score, 3))
```

A. Mean Squared Error (MSE)

B. Silhouette Score

C. Accuracy

D. R-squared value

Answer: B

Explanation: The silhouette score is a common metric for evaluating the quality of clustering when true labels are unknown. It measures how similar an object is to its own cluster compared to other clusters. MSE, accuracy, and R-squared are metrics used for regression or supervised learning tasks.

Q.16

In a supervised learning task, a student builds a classifier to predict whether a news article is "reliable" or "unreliable." They train the model on 90% of the dataset and test it on the remaining 10%. Which potential risk does this data splitting help mitigate, and why is it important?

A. It mitigates the risk of underfitting by increasing the training set size.

B. It helps prevent overfitting by testing the model on unseen data, ensuring it generalizes well.

C. It eliminates the need for hyperparameter tuning.

D. It increases the bias in the model predictions.

Answer: B

Explanation: Splitting the dataset into training and testing sets helps ensure that the model is evaluated on unseen data, which is crucial for detecting overfitting. Overfitting occurs when a model learns the training data too well and fails to generalize to new data. This validation step is a key part of building robust supervised models.

Q.17

A smart thermostat collects temperature, humidity, and user-set preferences to optimize energy usage. Which of the following correctly identifies the types of data collected, and why is it important to differentiate them in a privacy context?

A. Sensor data (automatic) and explicit data (user-provided); because sensor data might reveal patterns without user consent while explicit data is provided voluntarily.

B. Only sensor data is collected, so privacy concerns do not apply.

C. Only user-provided data is collected, ensuring complete privacy protection.

D. Both types of data are equally sensitive and need no differentiation.

Answer: A

Explanation: The thermostat collects sensor data automatically and explicit data from user inputs. Sensor data may reveal behavioral patterns that users are unaware of, while explicit data is provided with consent. Differentiating these types is crucial for implementing appropriate privacy measures such as anonymization and obtaining informed consent where necessary.

Q.18

An online learning platform uses both test scores and behavioral data (e.g., time spent on assignments, click patterns) to tailor educational content to each student. What is a key privacy concern in collecting such data, and which strategy should be implemented to address it?

A. There is no privacy concern since the data is used for educational purposes.

B. The risk of student profiling and potential misuse; it can be addressed by anonymizing data and ensuring compliance with data protection laws (e.g., FERPA or GDPR).

C. The risk of data loss during transmission; this is addressed by increasing the data storage capacity.

D. The risk of students cheating; this is mitigated by time-limiting assignments.

Answer: B

Explanation: Collecting detailed behavioral and academic data can lead to profiling that may be misused. Anonymization and adherence to data protection regulations are crucial to protecting students' privacy while still enabling personalized learning. Options A, C, and D do not directly address the privacy issue.

Q.19

A research team is analyzing customer reviews to understand product satisfaction. They use a supervised learning model to predict sentiment based on labeled reviews and an unsupervised clustering method to discover topics in unlabeled reviews. Which outcome is most likely, and why might both methods be valuable?

A. The supervised model will always outperform unsupervised methods, so clustering is unnecessary.

B. The supervised model provides accurate sentiment predictions, while clustering reveals hidden topics and patterns that are not evident from sentiment alone; together, they offer a comprehensive understanding of customer opinions.

C. Unsupervised clustering is only useful for small datasets, while supervised learning scales better.

D. Both methods will yield identical results if the data quality is high.

Answer: B

Explanation: Supervised learning, with labeled data, can accurately predict sentiment, while unsupervised clustering can uncover underlying topics and themes within the reviews. Combining both approaches allows for a richer analysis that captures both sentiment and nuanced patterns in customer feedback.

Q.20

Consider the following Python code snippet that anonymizes a list of user email addresses by hashing them. Why is this approach beneficial for data privacy, and what is a potential limitation of this method?

```
import hashlib

def anonymize_emails(emails):

    return [hashlib.sha256(email.encode()).hexdigest() for email in emails]

emails = ["user1@example.com", "user2@example.com", "user3@example.com"]

anonymized = anonymize_emails(emails)

print(anonymized)
```

A. It ensures that original email addresses are completely hidden, protecting privacy; however, if the same email appears again, the same hash is produced, which could allow linking across datasets.

B. It improves email deliverability; however, it may lead to data duplication.

C. It speeds up data retrieval; however, it increases the storage space required.

D. It automatically deletes the original emails; however, it may reduce model accuracy.

Answer: A

Explanation: Hashing email addresses converts them into a fixed-length string that does not reveal the original information, thus protecting privacy. However, because hashing is deterministic, the same input always yields the same hash, which might allow an adversary to link records across different datasets if they have access to the hashed values. Options B, C, and D do not correctly address the primary benefits and limitations of this method.

Section 3 - AI Applications in Healthcare, Education and Environment

Q.21

A major hospital deploys an AI system that analyzes chest X-ray images to identify early signs of pneumonia and other lung conditions. The system uses a deep learning model, specifically a convolutional neural network (CNN) pre-trained on a large medical image dataset, then fine-tuned on the hospital's own images. This approach helps reduce

diagnostic errors and speeds up the triage process. What is the primary advantage of using transfer learning in this healthcare context?

A. It eliminates the need for any further model training.

B. It reduces the amount of labeled data required from the hospital, while achieving high accuracy.

C. It automatically interprets radiologist notes without any manual intervention.

D. It guarantees 100% accuracy in diagnosing all lung conditions.

Answer: B

Explanation: Transfer learning leverages a pre-trained model, which already understands many low-level features from a large dataset. Fine-tuning on hospital data allows the model to adapt to specific imaging characteristics, reducing the need for an enormous labeled dataset while achieving high accuracy. Options A, C, and D do not correctly capture the benefits or realistic limitations.

Q.22

A healthcare provider integrates an AI-powered clinical decision support system (CDSS) into its electronic health record (EHR) system. The CDSS analyzes patient histories, lab results, and current symptoms to recommend potential diagnoses and treatment plans. However, concerns arise about the interpretability of the model's recommendations. Which strategy is most effective for improving interpretability and fostering trust among clinicians?

A. Using a complex black-box model without any additional explanation tools.

B. Implementing explainable AI techniques such as attention mechanisms or LIME to provide visual or textual explanations of predictions.

C. Relying solely on historical data trends to explain decisions.

D. Ignoring interpretability because clinical outcomes are more important than understanding the model.

Answer: B

Explanation: Explainable AI techniques (e.g., attention maps, LIME) help provide clinicians with insight into which features influenced the model's decision. This improves transparency and trust, crucial in healthcare where decisions impact patient lives. Options A and D compromise trust, while C does not provide individualized explanations.

Q.23

An online learning platform uses AI to create personalized lesson plans by analyzing student performance, time on task, and learning preferences. The platform applies reinforcement learning to continuously adapt the content and difficulty level in real time. What ethical and practical challenge must the platform address, and which measure would be most appropriate?

A. Ensuring that the system ignores any student mistakes to avoid discouragement; measure: always provide positive reinforcement.

B. Preventing bias that may disadvantage certain students; measure: regularly audit the model's recommendations and ensure diverse training data.

C. Guaranteeing that every student receives the exact same content; measure: remove personalization.

D. Maximizing profit by showing sponsored content; measure: integrate advertisements.

Answer: B

Explanation: Personalized adaptive learning systems must ensure fairness. If the training data or algorithm is biased, it could disadvantage certain groups of students. Regular audits and diverse, representative data help mitigate bias and ensure equitable learning experiences. Options A, C, and D either ignore personalization or introduce unethical practices.

Q.24

A research center has installed IoT sensors throughout a forest to monitor temperature, humidity, and pollutant levels. An AI system uses this data to predict the likelihood of forest fires. Given that sensor data can be noisy and heterogeneous, which machine learning approach is best suited for this task and what is a key challenge?

A. Supervised learning with time-series regression, with the challenge of obtaining accurately labeled past fire events.

B. Unsupervised clustering to group sensor readings, with no challenges in labeling.

C. Reinforcement learning to control water sprinklers, with the challenge of real-time decision making only.

D. A simple linear model with no need for preprocessing.

Answer: A

Explanation: Supervised learning, especially time-series regression, is suitable for predicting continuous outcomes like fire risk. A key challenge is acquiring accurately labeled historical data (fire events) to train the model effectively. Options B and C either do not directly address prediction or focus on control rather than prediction, while D is too simplistic.

Q.25

A startup has developed an AI-powered chatbot designed to help users assess their symptoms and decide whether to seek medical attention. The chatbot uses natural language processing (NLP) to interpret user input. What is a major ethical challenge with deploying such a chatbot, and how should the company address it?

A. The chatbot may replace doctors entirely; address it by completely banning its use in healthcare.

B. The chatbot might provide inaccurate advice due to limitations in understanding complex medical symptoms; address it by integrating the chatbot with human oversight and clear disclaimers about its advisory role.

C. The chatbot will never learn new medical information; address it by freezing the model after deployment.

D. The chatbot's interface might be too complex; address it by simplifying the design only.

Answer: B

Explanation: The primary ethical challenge is ensuring the chatbot does not provide harmful or inaccurate medical advice. Integrating human oversight (e.g., triage nurses or physicians) and making clear disclaimers that the chatbot is not a substitute for professional medical advice are crucial measures. Options A, C, and D do not address the core ethical and safety issues.

Q.26

A city government deploys an AI system that uses data from air quality sensors, weather forecasts, and traffic patterns to predict pollution levels and issue health advisories. What machine learning technique is best suited for such predictions, and what is a major environmental data challenge?

A. Supervised regression to predict pollution levels, with the challenge of ensuring data quality and integration from diverse sensors.

B. Unsupervised clustering to group cities, with the challenge of standardizing pollutant measurements.

C. Reinforcement learning to control traffic lights, with the challenge of policy implementation.

D. Decision trees that ignore weather data, with the challenge of limited feature sets.

Answer: A

Explanation: Supervised regression is appropriate for predicting continuous outcomes like pollution levels. A significant challenge is integrating high-quality data from multiple sources (sensors, weather, traffic) that may have different formats and noise levels. Options B, C, and D do not address the predictive nature and data integration challenge as effectively.

Q.27

An education department uses an AI system to analyze student performance, school infrastructure, and resource availability to optimize the allocation of teaching aids and funding. The model must balance data from academic records with socio-economic indicators. What is the most critical consideration in deploying this system, and which approach best ensures fairness?

A. Focusing solely on academic performance to ensure high test scores; using a simple ranking model.

B. Ignoring socio-economic data to avoid bias; using only standardized test scores.

C. Incorporating both academic and socio-economic data with fairness-aware algorithms and regular audits to identify and correct potential biases.

D. Allocating resources randomly to avoid any bias.

Answer: C

Explanation: For equitable resource allocation, the model must consider both academic performance and socio-economic factors. Fairness-aware algorithms and regular audits help ensure that biases do not disadvantage underrepresented groups. Options A and B neglect important factors, and D ignores data-driven decision-making.

Q.28

A non-profit organization uses AI to analyze satellite imagery and detect deforestation patterns in tropical rainforests. The system employs deep learning techniques to classify land cover types. What is a key benefit of this approach, and what technical challenge must be addressed?

A. It provides real-time deforestation alerts with perfect accuracy; the challenge is building larger satellites.

B. It enables large-scale monitoring of remote areas, but a significant challenge is managing the high dimensionality and noise in satellite imagery data.

C. It replaces field surveys entirely, with no need for human verification.

D. It simplifies data analysis by converting images to text; the challenge is text processing speed.

Answer: B

Explanation: Deep learning on satellite imagery allows for efficient, large-scale monitoring of remote areas, which is invaluable for conservation efforts. However, satellite images are high-dimensional and often noisy, requiring careful preprocessing and robust model architectures. Options A, C, and D either overstate benefits or misrepresent challenges.

Q.29

A health insurance company employs AI to create personalized healthcare plans based on an individual's medical history, lifestyle data, and genetic information. While this can lead to better health outcomes, it also raises privacy and ethical concerns. Which strategy should the company implement to ethically manage this sensitive data?

A. Use raw data to maximize model accuracy, ignoring privacy laws.

B. Anonymize the data, apply strong encryption, obtain informed consent, and regularly audit the system for bias.

C. Sell the data to pharmaceutical companies to fund research.

D. Collect data only from individuals with no pre-existing conditions.

Answer: B

Explanation: Managing sensitive healthcare data ethically requires anonymization, encryption, obtaining informed consent, and regular audits to ensure compliance with privacy standards and reduce bias. Options A, C, and D violate ethical or legal standards and could lead to misuse of data.

Q.30

A university deploys an AI-powered chatbot to assist students with administrative queries and course information. The chatbot uses NLP to understand student questions and retrieve answers from a database. Considering that the chatbot logs interactions for future improvement, what is the most important data privacy measure the university should implement?

A. Log all interactions in plain text to analyze conversation quality.

B. Anonymize and encrypt conversation logs, ensuring students' personal information is not stored or is stored only with explicit consent.

C. Share conversation logs with third-party developers for rapid improvement.

D. Disable logging entirely to maintain privacy, even if it hinders model improvement.

Answer: B

Explanation: Anonymizing and encrypting conversation logs protects student privacy while allowing the university to analyze and improve the chatbot's performance. This balances the need for data-driven improvements with the obligation to safeguard personal information. Options A and C pose privacy risks, while D sacrifices valuable insights.

Section 4 - Integrated Responsible AI Case Studies

Q.31

A large hospital uses a deep learning model to analyze chest X-ray images for early detection of lung diseases. Radiologists label each X-ray as either "normal" or "abnormal," and the model is trained on this labeled data. Which description best fits this scenario?

A. The model uses unsupervised clustering to group similar X-rays without any labels.

B. The model uses reinforcement learning to continuously improve its diagnosis through trial and error.

C. The model uses supervised classification because it learns from labeled examples provided by radiologists.

D. The model uses semi-supervised learning by combining a small set of labeled images with a larger set of unlabeled images.

Answer: C

Explanation: This scenario is a textbook example of supervised learning. The model learns from labeled data—each X-ray comes with a "normal" or "abnormal" label provided by experts—so it can later predict the labels for new images. Options A and B describe other learning paradigms, and option D is only used when labeled data is scarce.

Q.32

A popular social media platform collects users' posts, location data, and browsing history to personalize advertisements. However, there are significant privacy concerns regarding user identification. Which measure would most effectively mitigate these concerns while still enabling data analysis?

A. Storing raw data indefinitely to maximize insights.

B. Applying differential privacy techniques that add controlled noise and anonymize individual data points.

C. Collecting additional personal details to improve ad targeting.

D. Sharing raw data with advertisers to ensure transparency.

Answer: B

Explanation: Differential privacy is designed to protect individual user identities by adding noise to the dataset while preserving overall statistical properties. This approach reduces re-identification risks compared to storing or sharing raw data. Options A, C, and D either increase privacy risks or compromise user anonymity.

Q.33

An online education platform uses AI to analyze students' historical performance, time spent on activities, and response accuracy to create personalized learning plans. The system uses a combination of supervised models (trained on past graded data) and unsupervised clustering (to group similar learning styles). What is a key benefit of combining these two approaches?

A. It allows the system to ignore individual differences and focus solely on content.

B. It enables the system to both predict student performance accurately and discover underlying patterns in learning behaviors that can refine personalization.

C. It reduces computational costs by only using unsupervised methods.

D. It eliminates the need for human oversight in content delivery.

Answer: B

Explanation: By combining supervised learning (which uses labeled performance data to predict outcomes) with unsupervised clustering (which discovers natural groupings among students), the platform can tailor learning experiences more precisely. This hybrid approach leverages both predictive accuracy and pattern discovery, unlike the other options.

Q.34

A city deploys an AI system to predict air pollution levels using sensor data from various parts of the city, weather forecasts, and traffic patterns. The system employs supervised regression techniques to forecast pollution levels. What is one major challenge in this application, and what is a suitable strategy to address it?

A. The challenge is integrating heterogeneous data sources; a suitable strategy is to pre-process and normalize all sensor data before model training.

B. The challenge is that supervised learning cannot be applied; the strategy is to switch to unsupervised clustering.

C. The challenge is that weather forecasts are always inaccurate; the strategy is to ignore them.

D. The challenge is overfitting the model; the strategy is to use raw data without any pre-processing.

Answer: A

Explanation: Integrating data from sensors, weather, and traffic requires thorough pre-processing and normalization to ensure consistency across sources. This step is crucial for supervised regression models to learn meaningful patterns. Options B, C, and D either dismiss valid techniques or ignore critical pre-processing needs.

Q.35

A healthcare provider uses an AI system to predict patient risks by analyzing electronic health records (EHRs) that include sensitive information such as genetic data and treatment histories. What ethical measures must be taken to protect patient privacy while still enabling the system to function effectively?

A. Use all available data without modifications to maximize model accuracy.

B. Obtain informed consent, de-identify the data, and implement strict access controls along with differential privacy techniques.

C. Sell the data to external research organizations to improve research outcomes.

D. Store raw data on public servers to ensure transparency.

Answer: B

Explanation: Ethically managing sensitive healthcare data requires that patients provide informed consent, data is de-identified to prevent re-identification, and access is strictly controlled. Differential privacy can further protect individual data points during analysis. Options A, C, and D expose patients to privacy risks.

Q.36

A smart home company collects sensor data (temperature, motion, sound) and user-input preferences to optimize home automation. However, this data could potentially reveal detailed personal routines. Which method best protects user privacy while allowing the system to function effectively?

A. Collecting all raw data and storing it indefinitely.

B. Anonymizing and aggregating sensor data so that individual behavior cannot be easily isolated, and ensuring encryption during transmission.

C. Sharing raw data with third-party advertisers to increase revenue.

D. Ignoring user preferences to simplify data collection.

Answer: B

Explanation: Anonymizing and aggregating data minimizes the risk of identifying individual behavior patterns, and encryption protects data during transmission. This approach allows the smart home system to operate while safeguarding user privacy. Options A and C pose significant risks, and D compromises the personalized experience.

Q.37

A retail company wants to understand customer buying habits to improve its product recommendations. They use unsupervised learning (clustering) on transaction data to identify distinct customer segments. Which of the following statements best describes why unsupervised learning is suitable in this case?

A. It uses labeled data to predict future purchases accurately.

B. It groups customers based on inherent similarities in their purchasing patterns without pre-defined categories, allowing for the discovery of unexpected segments.

C. It requires a detailed understanding of each customer's demographic data before clustering.

D. It can only group customers into two clusters at most.

Answer: B

Explanation: Unsupervised learning like clustering does not require pre-defined labels, so it can discover natural groupings in customer transaction data. This allows the company to identify diverse segments that may not have been previously recognized. Options A, C, and D misrepresent the capabilities of unsupervised learning.

Q.38

A research team is designing an online survey to gather data on student study habits for an educational study. Which data collection method should they prioritize to ensure data quality and user privacy, and why?

A. Collect detailed personal information along with study habits, as it provides richer insights.

B. Use a voluntary, anonymous survey that only collects necessary information, ensuring that respondents are fully informed about data usage.

C. Automatically track students' online activities without their knowledge to obtain unbiased data.

D. Mandate full disclosure of all personal details to correlate study habits with academic performance.

Answer: B

Explanation: A voluntary, anonymous survey that collects only necessary information helps maintain data quality while respecting user privacy. Transparency and informed consent are critical for ethical data collection. Options A, C, and D raise serious privacy concerns.

Q.39

A city government wants to deploy an AI system that integrates data from healthcare, education, and environmental monitoring to guide public policy. What is the most critical challenge in integrating these diverse datasets, and what measure should be taken to overcome it?

A. The challenge is achieving high processing speed; the measure is to use more powerful hardware.

B. The challenge is ensuring data privacy and preventing re-identification when combining datasets; the measure is to anonymize data and implement differential privacy techniques during integration.

C. The challenge is data visualization; the measure is to use advanced graphing tools.

D. The challenge is collecting more data; the measure is to increase data sampling frequency.

Answer: B

Explanation: When integrating datasets from multiple sources, privacy risks increase—especially the risk of re-identification when linking data points. Anonymizing data and applying differential privacy help ensure that individual privacy is maintained while still allowing for comprehensive analysis. Options A, C, and D do not directly address the privacy challenges.

Q.40

A multinational corporation uses an AI system to screen job applications by analyzing resumes and social media profiles. While the system improves efficiency, it also raises ethical concerns regarding bias and privacy. Which strategy best addresses these ethical challenges?

A. Using the AI system without any human oversight to ensure speed and objectivity.

B. Incorporating fairness-aware algorithms, regularly auditing the system for bias, anonymizing sensitive information, and ensuring transparency about how data is used in the recruitment process.

C. Relying solely on traditional human screening to avoid any AI-related biases.

D. Collecting additional personal data to refine the algorithm, regardless of privacy concerns.

Answer: B

Explanation: Addressing ethical challenges in AI-powered recruitment requires a multifaceted approach: using fairness-aware algorithms to mitigate bias, performing regular audits to detect and correct any discriminatory patterns, anonymizing sensitive data to protect privacy, and being transparent about the data usage policies. Options A and D exacerbate ethical issues, while C foregoes the benefits of AI.

Achievers Section - Advanced Ethical and Applied AI Reasoning

Q.41

A major hospital deploys an AI system to diagnose diabetic retinopathy from retinal images. The system is trained using a labeled dataset (supervised learning) and, in parallel, applies unsupervised clustering on the feature representations to detect novel anomalies not seen during training. What is the primary benefit of combining supervised and unsupervised learning in this scenario?

A. It completely eliminates the need for expert annotations.

B. It ensures that both known conditions are diagnosed accurately and that unusual patterns are flagged for further review.

C. It simplifies the model architecture by merging two unrelated tasks into one process.

D. It reduces computation time by avoiding separate processing steps.

Answer: B

Explanation: Using supervised learning with labeled data provides accurate diagnosis for known conditions. Unsupervised clustering, on the other hand, helps detect anomalies or new patterns that may indicate conditions not represented in the training data. Together, they offer both reliable predictions and early warnings of unknown issues. Options A, C, and D do not capture this dual advantage.

Q.42

A financial institution employs supervised models trained on historical, labeled transaction data to detect fraudulent activity. However, they are concerned about emerging fraud patterns that the model may miss. To address this, they integrate an unsupervised anomaly detection module. What is the main challenge when combining these methods, and how can it be mitigated?

A. The unsupervised model might overfit the data; mitigate by reducing model complexity.

B. Balancing the false positive rate from unsupervised detection with the high accuracy of the supervised model; mitigate by tuning decision thresholds using continuous feedback.

C. The supervised model might ignore the unsupervised output; mitigate by retraining the supervised model with synthetic anomalies.

D. The combined model will always be slower; mitigate by using parallel processing.

Answer: B

Explanation: A primary challenge in hybrid systems is balancing the sensitivity of unsupervised anomaly detection (which may yield false positives) with the reliability of supervised predictions. Fine-tuning thresholds and integrating feedback loops can help balance both systems effectively. Options A, C, and D do not directly address the integration challenge.

Q.43

A smart city initiative collects data from IoT sensors—such as traffic flows, pollution levels, and weather conditions—across various neighborhoods to train a supervised model for optimizing urban services. Due to privacy concerns, the city opts for federated learning. What is the main advantage of this approach, and what privacy risk does it minimize?

A. It centralizes data processing to boost computation speed; minimizes hardware costs.

B. It enables the model to be trained on decentralized data without transferring raw data to a central server; minimizes the risk of re-identification of sensitive location data.

C. It completely removes the need for data aggregation; minimizes the risk of model bias.

D. It simplifies model training by reducing the dataset size; minimizes the risk of overfitting.

Answer: B

Explanation: Federated learning allows local devices (or sensors) to train the model on their own data, sending only model updates (and not raw data) to a central server. This approach protects individual privacy by reducing the risk of re-identification while still benefiting from a diverse dataset. Options A, C, and D do not focus on privacy preservation.

Q.44

A university uses an AI system to evaluate student essays by comparing them to a large, labeled corpus of essays scored by expert educators. To ensure fairness—especially for non-native English speakers—the system incorporates fairness-aware algorithms. Which of the following methods would best address potential bias in this context?

A. Removing all demographic indicators from the essays before training.

B. Incorporating fairness constraints into the model's loss function and performing subgroup analysis to adjust for linguistic diversity.

C. Using unsupervised clustering to group essays and then manually scoring each cluster.

D. Training the model solely on essays from native speakers to achieve higher accuracy.

Answer: B

Explanation: Integrating fairness constraints in the model's loss function, along with subgroup analysis, helps ensure that the grading algorithm is equitable across diverse student populations. Simply removing demographic data (option A) might not remove proxy biases; option C is inefficient, and option D is clearly discriminatory.

Q.45

A non-profit uses satellite imagery to monitor deforestation patterns in tropical regions. They employ deep autoencoders to extract features from the images, followed by unsupervised clustering to identify areas of concern. Which of the following challenges is most critical for improving the accuracy of this approach, and what additional data might enhance the predictions?

A. High dimensionality and noise in the imagery; integrating ground truth data from field surveys to validate clusters.

B. The inability to process images in real time; integrating faster GPUs to speed up processing.

C. Insufficient image resolution; switching to lower-resolution images to reduce complexity.

D. Over-reliance on autoencoders; replacing them entirely with traditional clustering methods.

Answer: A

Explanation: Satellite imagery is high-dimensional and noisy. Using deep autoencoders helps reduce dimensionality, but integrating ground truth data from field surveys can further validate and refine the clustering process. Options B, C, and D either focus on unrelated technical aspects or dismiss effective technique.

Q.46

A healthcare startup develops a wearable device that monitors heart rate, blood oxygen, and activity levels in real time. To predict cardiovascular events, the device sends data to a cloud-based model. Given the sensitivity of health data, which advanced method best ensures privacy while maintaining predictive performance?

A. Transmitting raw data over secure channels without further protection.

B. Implementing federated learning so that the model is trained on-device and only encrypted updates are shared, combined with differential privacy techniques.

C. Discarding real-time data in favor of periodic manual uploads.

D. Storing all collected data locally without encryption.

Answer: B

Explanation: Federated learning keeps raw data on the device while sharing only model updates. Combining this with differential privacy techniques ensures that individual data points remain protected, thereby balancing privacy and performance. Options A, C, and D either expose sensitive data or degrade performance.

Q.47

A retail company seeks to understand customer sentiment by analyzing online product reviews. They first use unsupervised topic modeling to discover themes and then apply supervised sentiment classification on labeled reviews. What is the main advantage of using this two-step hybrid approach?

A. It eliminates the need for human-annotated labels entirely.

B. It uncovers hidden topics and trends in the data that can inform and refine the supervised sentiment model, leading to more nuanced insights.

C. It increases the speed of analysis by bypassing data pre-processing.

D. It guarantees that sentiment classification will be perfect.

Answer: B

Explanation: Using unsupervised topic modeling allows the discovery of underlying themes in the reviews, which can then be used to enhance the accuracy and context of the supervised sentiment classifier. This hybrid approach provides richer insights than either method alone. Options A, C, and D are inaccurate or overly optimistic.

Q.48

A multinational corporation employs an AI-driven recruitment system that uses both historical hiring data and real-time applicant data to shortlist candidates. To ensure ethical use of this sensitive data, which approach should the company adopt?

A. Rely solely on historical data without updating the model with new candidate data.

B. Use fairness-aware algorithms, implement adversarial debiasing techniques, anonymize sensitive applicant information, and conduct regular bias audits.

C. Remove all performance data and rely on random candidate selection.

D. Collect additional data from third-party sources to enhance decision-making without privacy safeguards.

Answer: B

Explanation: To mitigate bias and protect privacy, the company should use fairness-aware methods (such as adversarial debiasing), anonymize sensitive data, and conduct regular audits to detect and correct any discriminatory patterns. Options A, C, and D either limit the system's effectiveness or raise ethical concerns.

Q.49

A city government plans to integrate healthcare, education, and environmental data into an AI system to inform public policy decisions. What is a significant challenge of integrating such diverse datasets, and what advanced measure should be implemented to address it?

A. The challenge is achieving high computational speed; the solution is to upgrade hardware.

B. The challenge is preserving privacy and preventing re-identification; the solution is to anonymize and aggregate data across domains and apply differential privacy techniques.

C. The challenge is data storage cost; the solution is to compress the data.

D. The challenge is lack of data formats; the solution is to convert all data into a common format like CSV.

Answer: B

Explanation: Integrating diverse datasets increases the risk of re-identification when data is linked across domains. Anonymizing and aggregating data, along with applying differential privacy techniques, can protect individual privacy while still enabling comprehensive analysis. Options A, C, and D focus on less critical technical issues.

Q.50

A tech company develops an AI-powered chatbot to assist students in an online learning platform. The chatbot collects detailed interaction logs to improve its natural language processing capabilities. However, concerns have been raised about the privacy implications of storing such logs. Which advanced measure best addresses these concerns while still allowing continuous improvement of the chatbot?

A. Store all logs in plain text for complete transparency.

B. Anonymize and encrypt interaction logs, and implement strict data retention policies along with user consent mechanisms.

C. Disable logging entirely to protect privacy, even if it hinders improvement.

D. Share logs with third-party companies to distribute the privacy risk.

Answer: B

Explanation: Anonymizing and encrypting logs protect sensitive information, while strict retention policies and user consent ensure that the data is used ethically. This approach balances continuous improvement of the chatbot with the protection of user privacy. Options A and D expose data, and C impairs the model's ability to improve.

Answer Key

Q.No.	Ans.	Q.No.	Ans.	Q.No.	Ans.	Q.No.	Ans.	Q.No.	Ans.
1	**A**	2	**B**	3	**B**	4	**B**	5	**B**
6	**A**	7	**B**	8	**B**	9	**A**	10	**A**
11	**C**	12	**B**	13	**B**	14	**B**	15	**B**
16	**B**	17	**A**	18	**B**	19	**B**	20	**A**
21	**B**	22	**B**	23	**B**	24	**A**	25	**B**
26	**A**	27	**C**	28	**B**	29	**B**	30	**B**
31	**C**	32	**B**	33	**B**	34	**A**	35	**B**
36	**B**	37	**B**	38	**B**	39	**B**	40	**B**
41	**B**	42	**B**	43	**B**	44	**B**	45	**A**
46	**B**	47	**B**	48	**B**	49	**B**	50	**B**

Space for Rough Work