True/ False question

1. Identify the reasons behind events, behaviors, or outcomes is called Descriptive analytics

   Answer: False, it is called diagnostic analytics

2. To predict a future stock performance, we can use historical data and combine it with statistical models and machine learning algorithms to forecast future stock movements.

   Answer: True

3. The reason why a company's net profit decreased can be solved with a diagnostic analytics

   Answer: True.

4. Estimate future outcomes and trends based on patterns found in existing data is called predictive analytics

   Answer: True

5. We can predict future effects with a similar condition in the past.

   Answer: True

6. Supervised learning is used to find a hidden patterns within new data

   Answer: False, Unsupervised learning, is used to find a hidden patterns within new data

7. E-commerce use big data just to gather user's information

   Answer: False, Ecommerce also uses big data to give a great recommendation.

8. Data quality doesn't affect the accuracy of the prediction in linear regression

   Answer: False, Companies compete against each other to see who can use the data they have the best.

9. Nowadays, consent of a formality has turned into a formality

Answer: True, because of legal complex words, users often don't understand what's written on the agreement.

10. Differential privacy is giving a "noise" or adding a little bit of a false number to the data so others wouldn't find out who the data is referring to.

    Answers: True

Multiple Choice Questions

Which of the following best defines Big Data?

A. Small and simple datasets that can be processed easily using traditional methods

B. Large and complex datasets that require advanced technologies to process and analyze

C. Data limited to business transactions only

D. Organized data stored in a single computer system

In the 5 V's of Big Data, which characteristic refers to the speed at which data is generated and processed?

A. Volume

B. Variety

C. Velocity

D. Veracity

Which of the following is an example of human-generated data?

A. Sensor data from a smart thermostat

B. Tire pressure readings from a car

C. Social media posts and online purchases

D. Temperature logs from industrial machines

What type of data includes text, images, videos, and social media posts that do not follow a fixed structure?

A. Structured data

B. Semi-structured data

C. Unstructured data

D. Categorical data

Which of the following is an example of semi-structured data?

A. SQL tables in a customer database

B. JSON or XML files with key-value pairs

C. Raw video files from surveillance cameras

D. Numeric transaction records from a bank

What is the main purpose of big data architecture?

A. Manage the ingestions, process, and analysis of big data
B. Create a website for a big companies
C. Predict the future
D. To protect data from hackers.

What are two main models for data processing

A. Batch Processing and Stream Processing
B. Batch processing and Real Time processing
C. Basic processing and Layered processing
D. Offline processing and online processing

What is a Database?

A. A phase to process a data
B. A way to protect a data
C. A method to predict the future using a data
D. A structured collection of data that is stored to be managed

Which of these statements isn't a benefit of data visualization?

A. Helping non-technical users understand data trends
B. Allowing managers to make data-driven decisions
C. Revealing insights that raw numbers cannot show
D. Replacing the need for any data analysis entirely

Which of the following best describes semi-structured data?

A. Data that is completely unorganized and lacks any identifiable structure or format.
B. Data that fits perfectly into relational tables with a fixed schema.

C. **Data that doesn't fit perfectly into tables but still has some organizational structure, such as tags or keys, making it easier to analyze.**
D. Data that can only be stored in spreadsheets or relational databases.

Which of the following best describes data analytics?

A. The process of storing data for long-term archival purpose
B. **The process of collecting, organizing, and interpreting data to extract insights**
C. The act of using computers to automate manual tasks
D. The process of visualizing only future data

Which of the following best summarizes the relationship among the four types of analytics?

A. Each type operates independently with no overlap
B. **Each level builds on insights from the previous one to increase decision-making value**
C. They all use the same data and methods for different industries
D. They are interchangeable and serve the same function

Which of the following tools is most commonly used in *diagnostic analytics* to find correlations?

A. Decision trees
B. Time series forecasting
C. **Linear regression**
D. Monte Carlo simulation

If a company uses data to find out why its profit dropped despite high sales, it is performing:

A. Descriptive analytics
B. **Diagnostic analytics**
C. Predictive analytics
D. Prescriptive analytics

A financial analyst uses past data on interest rates and GDP to estimate next quarter's stock performance. This is an example of:

A. Descriptive analytics

B. Diagnostic analytics
C. Predictive analytics
D. Prescriptive analytics

Why is random forest generally considered an improvement over a decision tree in supervised learning?

A. It uses fewer data points and less computation to make predictions
B. It replaces labelled data with unlabeled data to increase flexibility
C. It combines multiple trees trained on random subsets of data to reduce overfitting and improve generalization
D. It eliminates the need for human-labelled training data altogether

Which statement best captures the fundamental difference between classification and clustering?

A. Classification deals with continuous data, while clustering deals only with categorical data
B. Classification uses pre-labeled data to assign categories, while clustering finds natural groupings in unlabeled data
C. Clustering predicts numerical outcomes, while classification predicts categorical ones
D. Classification focuses on data reduction, while clustering focuses on association

What main advantages does semi-supervised learning have over both supervised and unsupervised learning

A. It removes the need for large datasets altogether
B. It uses both labeled and unlabeled data to increase learning accuracy while reducing labeling effort
C. It completely replaces the need for supervised models in predictive analytics
D. It is faster but less accurate than both supervised and unsupervised learning

Why is dimensionality reduction (such as PCA) essential in complex datasets like image recognition?

A. It increases the number of features to improve model diversity
B. It simplifies data by reducing input variables while preserving the most critical information

C. It helps convert unlabeled data into labeled data for supervised learning
D. It prevents clustering algorithms from forming overlapping groups

Which of the following best demonstrates how *supervised learning* is applied in business decision-making?

A. Grouping customers with similar purchasing behavior to identify segments
B. Predicting future sales based on advertising expenditure using linear regression
C. Discovering unknown relationships between products frequently bought together
D. Reducing redundant customer data while retaining the most relevant attributes

Which of the following best explains how Amazon uses Big Data to increase sales?

A. By collecting customer emails and sending general promotions
B. By using collaborative filtering to recommend products based on user history
C. By manually adjusting prices based on customer feedback
D. By reducing the number of products available to simplify choices

How does Walmart primarily use IoT and weather data in its operations?

A. To predict customer mood and shopping frequency
B. To determine product design preferences
C. To reroute shipments efficiently and optimize supply chains
D. To personalize homepage displays for online users

What is one key benefit Alibaba gains from applying predictive analytics to its logistics system

A. It eliminates the need for warehouses
B. It reduces delivery times and logistics costs through route optimization
C. It allows manual tracking of shipments
D. It creates more advertisements for users

In the media industry, how does Netflix use Big Data to decide which shows to produce?

A. By choosing content randomly to test audience reactions

B. By monitoring user reviews and ratings after production
C. By analyzing global viewing trends to predict popular genres and themes
D. By hiring focus groups in each region

Spotify's "Discover Weekly" playlist is an example of Big Data applied to:

A. Fraud detection using payment history
B. Predictive analytics for supply chain efficiency
C. Machine learning-based music personalization
D. Manual playlist curation by human editors

What would happen if only a few groups or companies hold the major field of the data?

A. Allows everyone to access data equally
B. Small researchers can develop well
C. Limit the innovation and competition for smaller institutions
D. It will help people living in the rural area to compete with the big cities

What do most of the "Third parties" do when they obtain many people's personal data?

A. Delete it
B. Share it to the social media to get credits for their action
C. Use it only for public research
D. Analyze to gain profit

What is the major weakness for the "Terms of service" Agreement?

A. It is written too complex legal words
B. The font size is too little
C. There are too many lines
D. The font type is hard to read

Why should big companies increase transparency?

A. To increase trust of the public
B. To gain credits
C. To analyse market
D. So that hacker could interrupt their system

According to Montjoye, how many location points does it take to figure someone out?

A. 1
B. 2
C. 3
D. 4

Open Ended Question

1. Explain What is Big data in general

   Sample Answer: Massive amount of data that is beyond the ability of traditional data processing systems. It can contain public data,knowledge data, etc.These datasets can be applied to many things. We can use it to uncover patterns, predict the future, and help organisations or companies to make a great decision. Nowadays, almost everyone depends on big data such as the government, company, or even a single person.

2. Explain Why big data is important

   Sample Answer: It helps organizations to make better, faster, and smarter decisions. By analyzing massive amounts of data, companies and governments can discover hidden patterns so that they can predict the future. It also allows businesses to improve their products, reduce costs, and increase efficiency. In fields like healthcare, Big Data can even save lives by helping doctors detect diseases earlier. For the government, it helps them create a great policy to prevent crimes or to benefit their peoples. Overall, Big Data gives people and organizations the power to turn information into valuable insights and real-world solutions.

3. Explain how does a big e commerce gain their profit using a big data

   Sample Answer:

   By analyzing the habits of the community or its users so that we can offer the products they need so that they sell and make the right decisions regarding the appropriate target market. They will analyze what people need.

   They can also analyze by looking at what most of the people search history, what they click and purchase. They can identify buying patterns and predict what products customers might want next. This allows them to create personalized recommendations, targeted advertisements, and special offers that increase sales

IMPORTANT CHAPTERS QUESTION (Fill in the blank)

1. Big Data Architecture

1. Big data architecture can be described as a **(Blueprint)** that outlines how large-scale data should be stored, processed, and analyzed.
2. The mechanism that moves data from the sources to the platform is called **(Data ingestion)**

3. Batch ingestion moves **(Chunks)** at set intervals, such as hourly, daily, or weekly.
4. **(Batch Ingestion)** is easier to maintain than steam ingestion.
5. **(Data storage)** is responsible for storing all types of data that is scalable, reliable, and accessible for processing and analysis.
6. Data that can be sorted into rows and columns like in a table is called **(Structured data)**
7. Semi structured data often uses **(Tags)**,keys or identifiers to separate elements and enforce hierarchy.
8. Data without predefined or structure is called **(Unstructured data)**
9. More than **(80%)** of data created today is unstructured
10. Processing data in groups is called **(Batch processing)**

2. Data Analytics

1. The process of collecting, organizing, and interpreting large sets of data to provide insights and patterns is called **(Data analytics)**
2. The process of interpreting historical and current data to answer the question is called **(Descriptive analytics)**
3. Type of Analytics that Identifies the reasons behind events, behaviors, or outcomes is called **(Diagnostic analytics)**
4. Type of Analytics that estimate future outcomes and trends based on patterns found in existing data is called **(Predictive analytics)**
5. Predictive analytics uses **(Historical data)** combined with statistical models and machine learning algorithms to forecast future stock movements.
6. Type of analytics that combine diagnostic analytics and predictive analytics is called **(Prescriptive analytics)**
7. The first step of doing descriptive analytics is **(Collect data)**
8. The three financial statements consist of the income statement, balance sheet, and the **(Cash flow)** Statement
9. In horizontal analysis, investors compare current financial figures with previous ones to see how key metrics have **(Changed)** overtime
10. Financial statements are periodic reports that inform investors about a company's **(Historical)** performance

3. Machine Learning

1. Branch of Artificial Intelligence (AI) that utilizes vast amounts of data (big data) to train computers to function on their own without needing explicit instructions from humans is called **(Machine Learning)**

2. The way to describe machine learning as to consider the computer as a **(Newborn Baby)**
3. Type of Machine Learning that requires a great deal of human intervention in the form of labelled data before the computer is able to perform its function is called **(Supervised learning)**
4. A method to predict a continuous (non-discrete) numerical value, such as predicting stock prices is called **(Regression)**
5. In the **(Unsupervised Learning),** the computer is forced to find hidden patterns, relationships, or structures within the data on its own.
6. When supervised learning struggles with big data and unsupervised learning lacks accuracy, **(Semi-Supervised)** learning provides a balanced approach.
7. To simplify complex datasets, the **(Dimensionality reduction)** technique is used to reduce the number of input variables while retaining the most important information.
8. A learning method that discovers relationships between variables in large datasets by identifying co-occurrences is called **(Association)**
9. **(Clustering)** is an unsupervised learning technique used to group similar data points together based on shared characteristics and patterns, without prior knowledge of categories or labels.
10. Highly accurate prediction is the strengths of **(Supervised learning)**