पॉलिटेक्निक
Board of Technical Education

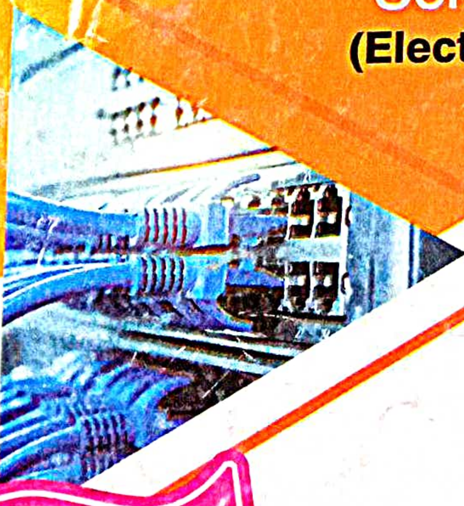
विद्या न्यूपैटर्न QUESTION BANK

सर्वश्रेष्ठ परीक्षा मार्गदर्शक

नए online परीक्षा प्रारूप के आधार पर बहुविकल्पीय प्रश्नों (MCQ) सहित

NSQF के अनुसार

Computer Science Engineering
Semester-VI
(Elective Subject)



- विगत 10 वर्षों के परीक्षा प्रश्नों का अध्यायवार समावेश
- स्वमूल्यांकन हेतु मॉडल प्रश्न–पत्र

डाटा साइंस और मशीन लर्निंग
(Data Science and Machine Learning)

नवीनतम पाठ्यक्रम
सत्र 2020-21
पर आधारित

New Syllabus

Data Science

1. **Introduction of data Science and Machine Learning**

   Fundamentals of Artificial Intelligence, need and applications of Data Science, Data Mining, data preparation, Machine Learning , Types and Applications of Machine learning

2. **Data Preprocessing, Analysis and Visualization**

   Data Pre-processing: Pre-processing Techniques- Mean Removal, Scaling, Normalization, Binarization, One Hot Encoding, Label encoding, Data Analyses: Loading and summarizing the dataset, Data Visualization:Univariate Plots, Multivariate Plots, Training Data, Test Data,Performance Measures.

3. **Statistical Inference**

   Populations and samples,Types of Statistical modelling, Types of probability distributions. Parametric and Non-Parametric Methods, Distance Metrics.

4. **Exploratory Data Analysis and the Data Science Process**

   Basic tools (plots, graphs and summary statistics) of EDA, Philosophy of EDA, The Data Science Process.

5. **Machine Learning Algorithms**

   Introduction to Supervised Learning Algorithms –Decision Tree,Linear Regression, k-Nearest Neighbours (k-NN), SVM and Introduction to Unsupervised Learning Algorithms-K-means Clustering,MeanShift Algorithm,Dimensionality Reduction Techniques, Introduction to Neural Networks.

6. **Mining Social-Network Graphs**

   Social networks as graphs, Clustering of graphs, Direct discovery of communities in graphs,Partitioning of graphs, Neighbourhood properties in graphs.

7. **Data Science and Ethical Issues**

   Discussions on privacy, security, ethics,A look back at Data Science, Next-generation data scientists.

1 डाटा साइंस और मशीन लर्निंग
Data Science and Machine Learning

बहुविकल्पीय प्रश्न (MCQ)

प्रश्न 1. निम्न में से कौन AI के जनक के रूप में जाना जाता है?

(a) एलन टर्निंग (b) जॉन मकर्ति
(c) अदा फिशर (d) एलिन न्यूवेल

उत्तर (b) जॉन मकर्ति

प्रश्न 2. न्यूवेल और साइमन द्वारा एक GPS नामक नावेल प्रोग्राम किस वर्ष में बनाया गया?

(a) 1954 (b) 1955
(c) 1956 (d) 1957

उत्तर (d) 1957

प्रश्न 3. कृत्रिम बुद्धिमत्ता (Artificial Intelligence) क्या है?

(a) मानव की बुद्धिमत्ता को कम्प्यूटर में फीड करना (b) कम्प्यूटर पर गेम खेलना
(c) मानव की बुद्धिमत्ता से प्रोग्रामिंग करना (d) एक मशीन को बुद्धिमान बनाना

उत्तर (d) एक मशीन को बुद्धिमान बनाना

प्रश्न 4. निम्नलिखित में से कौन-सा AI का प्रकार नहीं है?

(a) Reactive machines (b) Self- awareness
(c) Theory of mind (d) Unlimited memory

उत्तर (d) Unlimited memory

प्रश्न 5. डेटा साइंस के लिए अतिआवश्यक language कौन-सी है?

(a) Rusy (b) Java
(c) R (d) C++

उत्तर (c) R

प्रश्न 6. डेटा को निम्न में से किसके लिए उपयोग किया जाता है?

(a) Communication के लिए (b) Interpretation के लिए
(c) Processing के लिए (d) इनमें से कोई नहीं

उत्तर (d) इनमें से कोई नहीं

प्रश्न 7. डेटा माइनिंग analysis कितने प्रकार के होते हैं?

(a) दो (b) तीन
(c) चार (d) पाँच

उत्तर (c) चार

प्रश्न 8. निम्न में से किसमें AI तकनीक का उपयोग होता है?

(a) रेडियो (b) टीवी
(c) गूगल मैप (d) समाचार-पत्र

उत्तर (c) गूगल मैप

प्रश्न 9. निम्नलिखित में से कौन-सा उत्पाद Amazon द्वारा लॉंच किया गया है?

(a) Tesla (b) Echo

(c) Samsung (d) Nest

उत्तर (b) Echo

प्रश्न 10. पहली AI प्रोग्रामिंग भाषा कौन-सी है?

(a) Fortran (b) Basic

(c) Lisp (d) JPL

उत्तर (c) Lisp

खण्ड 'अ' : अतिलघु उत्तरीय प्रश्न

प्रश्न 1. आर्टिफिशियल इंटेलिजेंस क्या है?

उत्तर आर्टिफिशियल इंटेलिजेंस दो शब्दों से मिलकर बना है—artificial और intelligence, जहाँ Artificial परिभाषित करता है—''मानव निर्मित'' और Intelligent परिभाषित करता है—''सोच शक्ति'। इसलिए AI का मतलब है ''एक मानव निर्मित सोच शक्ति।''

प्रश्न 2. Strong AI और weak AI में क्या अंतर है?

उत्तर Weak AI किसी Strong AI या जनरल आर्टिफिशियल इंटेलिजेंस के विपरीत किसी विशेष समस्या को पूरा करने के लिये होती है। यह मशीन अपना काम करने में बहुत स्मार्ट नहीं होती है। स्ट्रांग AI ऐसी मशीनें बनाता है जो वास्तव में इंसान की तरह सोच और कार्य कर सकती हैं।

प्रश्न 3. AI के 4 प्रकार कौन-से हैं?

उत्तर 1. Reactive Machines AI 2. Limited Memory AI

3. Theory of Mind AI 4. Self-aware AI

प्रश्न 4. AI के उदाहरण क्या हैं?

उत्तर AI के निम्नलिखित उदाहरण हैं—

(i) Chatbots (ii) Facial recognition

(iii) Image tagging (iv) Natural language processing

(v) Sales prediction (vi) Self-driving cars

(vii) Sentiment analysis

प्रश्न 5. डेटा क्या है?

उत्तर Data facts, concepts या instructions का एक formalized manner में, एक representation होता है जोकि communication, interpretation या processing के लिए उपयोग किया जाता है।

प्रश्न 6. डेटा माइनिंग से आपका क्या तात्पर्य है?

उत्तर ''डेटा माइनिंग OLAP, OLTP, machine learning के तरीकों का उपयोग करके एक बड़े डेटा सेट में छिपे हुए pattern की खोज के लिए एक computation process है।''

खण्ड 'ब' : लघु एवं दीर्घ उत्तरीय प्रश्न

प्रश्न 1. कृत्रिम बुद्धिमत्ता (Artificial Intelligence) क्या है?

उत्तर Artificial Intelligence (AI) या ''कृत्रिम बुद्धिमत्ता'' कम्प्यूटर साइंस की एक शाखा है, जो ऐसी मशीनों को विकसित कर रही है जो मनुष्यों की तरह सोच सके और कार्य कर सके; जैसे—आवाज की पहचान, समस्या को सुलझाना, लर्निंग और प्लानिंग, यह मनुष्यों और जानवरों के द्वारा प्रदर्शित natural intelligence के विपरीत machines द्वारा प्रदर्शित intelligence है।

इसके द्वारा एक ऐसे कम्प्यूटर कंट्रोल्ड रोबोट या सॉफ्टवेयर बनाने की योजना है, जो वैसे ही सोच सके जैसे मानव मस्तिष्क सोचता है। Artificial Intelligence को इसमें परिपूर्ण बनाने के लिए उसे लगातार तैयार किया जा रहा है। इसके प्रशिक्षण में इसे मशीनों से अनुभव सिखाया जाता है, नए इनपुट के साथ तालमेल बनाने और मानव जैसे कार्यों को करने के लिए तैयार किया जाता है।

Artificial Intelligence के उपयोग से ऐसी मशीनें बन रही हैं, जो अपने एनवायरनमेंट के साथ इंटरेक्ट करके प्राप्त डेटा पर खुद बुद्धिमानी से कार्य कर सकती हैं। अगर भविष्य में AI concept और मजबूत होता है, तो यह हमारे दोस्त जैसा होगा। अगर आपको कोई समस्या आयेगी तो उसके लिए क्या करना है यह आपको खुद सोच कर बतायेगा। मूल रूप से Artificial Intelligence (AI) एक मशीन या कम्प्यूटर प्रोग्राम की सोचने और सीखने की क्षमता है। यह अवधारणा इस विचार पर आधारित है, कि मशीनों को इतना सक्षम बनाया जाए कि वह खुद किसी समस्या के बारे में इंसानों की तरह सोचे उस पर कार्य करें और उससे सीखे।

प्रश्न 2. उदाहरण के साथ आर्टिफिशियल इंटेलिजेस की व्याख्या कीजिए।

उत्तर आज AI एक बहुत ही लोकप्रिय विषय है। कई विशेषज्ञों और उद्योग के जानकारों का मानना है, कि AI या machine learning हमारा भविष्य है लेकिन अगर हम अपने चारों तरफ देखें तो हम पाएँगे की यह हमारा भविष्य नहीं बल्कि वर्तमान है। टेक्नोलॉजी के विकास के साथ आज हम किसी-न-किसी तरीके से Artificial Intelligence से जुड़े हुए हैं और इसका फायदा भी ले रहे हैं लेकिन अभी AI technology अपने पहले चरण में है।

अभी हाल में कई कंपनियों ने machine learning पर काफी निवेश किया है जिसके कारण कई AI product और Apps हमारे लिए उपलब्ध हुए हैं। वर्तमान में इस्तेमाल होने वाले AI examples निम्नलिखित हैं—

1. **Siri** Siri Apple द्वारा पेश किया गया सबसे लोकप्रिय आभासी सहायक (virtual assistant) है। हालाँकि यह सिर्फ iPhone और iPad में उपलब्ध है। यह AI का सबसे बेहतरीन उदाहरण है, इससे बस आप 'Hey Siri' बोलिए और यह आपके लिए मैसेज भेज सकता है। इंटरनेट पर इनफार्मेशन ढूँढ सकता है, फोन कॉल कर सकता है, कोई भी एप्लीकेशन ओपन कर सकता है यहाँ तक कि टाइमर सेट व कैलेंडर में इवेन्ट सेव करने जैसे कार्यों में आपकी सहायता कर सकता है।

   Siri आपकी भाषा और सवालों को समझने के लिए Machine Learning तकनीक का प्रयोग करती है। यह सबसे अनुकूल वॉइस एक्टिवेटिड कम्प्यूटर है। इससे संबंधित डिवाइस Alexa और Google Assistant भी हैं। जो समान कार्य के लिए ही प्रयोग किये जाते हैं।

2. **Tesla** न केवल Smartphones बल्कि Automobiles भी Artificial Intelligence की ओर बढ़ रहे हैं। Tesla motors अब तक उपलब्ध सबसे बेतहरीन Automobiles में से एक है। Tesla car में न केवल self driving बल्कि उत्पादक क्षमताओं और पूर्ण तकनीकी नवाचार जैसे फीचर उपलब्ध हैं। ऐसी ही न जाने कितनी self driving car और बन रही है जो आने वाले वक्त में और भी स्मार्ट हो जाएँगी।

3. **Google Map** वैसे Google कई क्षेत्र में AI का इस्तेमाल करता है लेकिन Google map में AI Technology का अच्छा इस्तेमाल हुआ है। हमको किसी भी जगह का रास्ता बताने के लिए AI मैपिंग के साथ giant's technology सड़क जानकारी को स्कैन करती है और एल्गोरिथम्स का प्रयोग करके सही रूट को हमें बताती है।

4. **Nest** Nest सबसे प्रसिद्ध और Artificial intelligence स्टार्टअप में से एक था और इसे 2014 में Google द्वारा खरीद लिया गया। नेस्ट लर्निंग थर्मोस्टेट आपके व्यवहार और दिनचर्या के आधार पर एनर्जी को बचाता है। ऐसा करने के लिए यह व्यवहार एल्गोरिथम का उपयोग करता है। यह इतनी इंटेलीजेंट मशीन है, कि सिर्फ एक हफ्ते में ही आपके लिए उपयोगी तापमान का पता लगा लेती है। अगर घर में कोई न हो तो यह ऊर्जा बचाने के लिए ऑटोमेटिकली टर्न ऑफ हो जाती है।

5. **Echo** Echo को Amazon द्वारा लाँच किया गया था। यह एक ऐसा क्रांतिकारी प्रोडक्ट है, जो आपके सवालों के जवाब दे सकता है, आपके लिए ऑडियो बुक पढ़ सकता है, आपको ट्रेफिक और मौसम का हाल बता सकता है,

लोकल बिज़नेस के बारे में जानकारी उपलब्ध करा सकता है तथा स्पोर्ट्स स्कोर भी प्रदान कर सकता है। Echo में और भी बड़े बदलाव किए जा रहे हैं जिससे यह नई सुवधिाओं को जोड़ता जा रहा है। उम्मीद है, आने वाला वक्त Echo को और भी स्मार्ट बना देगा।

प्रश्न 3. AI का इतिहास क्या है? आर्टिफिसियल इंटेलिजेंस के जनक के विषय में बताइए।

उत्तर 1950 में Artificial intelligence रिसर्च की शुरुआत हुई थी। इलेक्ट्रॉनिक कम्प्यूटर और स्टोर्ड -प्रोग्राम कम्प्यूटर के विकास के साथ ही AI के क्षेत्र में रिसर्च का कार्य शुरू हुआ। इसके बाद भी कई दशकों तक एक कम्प्यूटर किसी ह्यूमन माइंड की तरह सोच या कार्य कर पाए इसकी कोई कड़ी नहीं जुड़ पायी। आगे चलकर एक खोज जिसने AI के शुरुवात विकास को बहुत हद तक आगे बढ़ाया वह Norbert Wiener द्वारा बनाई गई थी।

उन्होंने यह सिद्ध किया कि इंसानों के सभी बुद्धिमान व्यवहार प्रतिक्रिया तन्त्र के परिणाम होते हैं। मॉडर्न AI की दिशा में एक और कदम तब बढ़ा जब लॉजिक थेओरिस्ट का निर्माण हुआ। 1955 में Newell और Simon द्वारा डिजाइन किया गया यह फर्स्ट AI प्रोग्राम माना जा सकता है।

आर्टिफिशियल इंटेलिजेंस के जनक कई शोध के बाद अंततः जिस व्यक्ति ने Artificial intelligence की नींव रखी वह थे AI के जनक John McCarthy, यह एक अमेरिकन साइंटिस्ट थे। AI के क्षेत्र में और विकास करने के लिए उन्होंने 1956 में एक सम्मेलन "The Dartmouth Summer Research Project on Artificial Intelligence" का आयोजन किया। जिसमें वे सभी लोग भाग ले सकते थे जो machine intelligence में रुचि रखते हो। इस सम्मेलन का उद्देश्य रुचि रखने वाले लोगों की प्रतिभा और विशेषज्ञता को आकर्षित करना था ताकि वह इस कार्य में McCarthy की मदद कर सके।

बाद के वर्षों में AI रिसर्च सेंटर का गठन Carnegie Mellon University के साथ-साथ Massachusetts Institute of Technology में हुआ। इसके साथ ही AI को कई चुनौतियों का सामना भी करना पड़ा। पहली चुनौती जो उनके सामने थी एक ऐसे सिस्टम का निर्माण करना जो बहुत कम खोज करके किसी समस्या को कुशलता से हल कर सके। दूसरी चुनौती ऐसे सिस्टम का निर्माण जो खुद से किसी कार्य को सीख सकता हो। Artificial intelligence के क्षेत्र में पहली सफलता तब मिली जब 1957 में Newell और Simon द्वारा एक जनरल प्रॉब्लम सॉल्वर (G.P.S.) नामक नावेल प्रोग्राम बनाया गया।

यह Wiener के फीडबैक सिद्धांत का विस्तार था। इसके जरिये सामान्य ज्ञान की समस्याओं का अधिक से अधिक समाधान किया जा सकता था। AI History में 1958 में John McCarthy द्वारा LISP लैंग्वेज का निर्माण किया गया। इसे जल्द ही कई AI रिसर्चरों द्वारा अपनाया गया था और यह आज भी उपयोग में है।

प्रश्न 4. AI के लक्ष्य क्या हैं?

उत्तर AI सम्पूर्ण दुनिया में सबसे शक्तिशाली और तेजी से बढ़ती टेक्नोलॉजी है। AI एक प्रकार की कृत्रिम चेतना (Artificial consciousness) है जो मानव के निर्देश देने पर कार्य करती है। भले ही Artificial intelligence मनुष्यों द्वारा विकसित की गई है, लेकिन इसमें कोई संदेह नहीं कि AI मनुष्यों की तुलना में अधिक कुशल, बेहतर और कम खर्च में काम करती है। इसीलिए अब कई बिजनेस इंडस्ट्री के फील्ड में AI को काम में लिया जा रहा है।

अभी कुछ हद तक AI हमारी रोजमर्रा की जिंदगी में आ चुकी है लेकिन वह दिन दूर नहीं जब हम पूरी तरह से इस टेक्नोलॉजी का उपयोग करने लगेंगे। भविष्य में ज्यादातर कार्य और कई क्षेत्र AI के ऊपर निर्भर होंगे। इसके साथ ही यह कयास भी लगाए जा रहे हैं कि यह मनुष्यों के जीवन पर बहुत बुरा प्रभाव डाल सकता है।

AI के लक्ष्य AI के लक्ष्य निम्नलिखित हैं—

(i) **निर्णय लेने की शक्ति बढ़ाना** AI का प्रथम लक्ष्य यही है, कि मनुष्यों की तरह सोचने वाली थिंकिंग मशीन को बनाया जाये जो मानव की किसी भी समस्या को खुद से डिसिजन लेकर हल कर सके। इस दिशा में AI ने कुछ उपलब्धियाँ भी हासिल की हैं। अभी हाल में एक female AI Robot (Sophia) को बनाया गया। इसके पास कुछ हद तक डिसिजन मेकिंग पावर है और यह आपके किसी भी सवाल का जवाब आसनी से दे सकती है। ऐसे ही कुछ AI Concept आपको स्मार्ट डिवाइस में भी देखने को मिलेंगे; जैसे—Google home, Siri, Alexa इत्यादि।

(ii) **कार्य में कुशलता** हम इंसान किसी कार्य को करने में काफी आलसी होते हैं, जिसके कारण हम अपने कार्यों को पूरा करने में बहुत ज्यादा समय लगाते हैं और उनमें ज्यादा गलतियाँ भी होती हैं। इंसानों की इसी आदत को देखते हुए AI researches इस दिशा में बहुत तेजी से कार्य कर रहे हैं। इनका मूल मकसद AI को ऐसा बनाना है ताकि वह किसी भी कार्य को न्यूनतम गलती के साथ तेजी से कर पाये।

(iii) **समय की बचत** AI मनुष्यों की तुलना में काफी अधिक तेजी से कार्य कर सकता है क्योंकि यह एक प्रकार की मशीन है, इसलिए यह कार्य करने में कभी नहीं थकता और हमारी तरह कभी ब्रेक भी नहीं लेता। इस विषय को देखते हुए कई ऐसी AI मशीन बनाई जा रही हैं, जो जल्द ही मनुष्यों की जगह ले लेगी।

प्रश्न 5. Artificial Techniques क्या है? ए० आई० के प्रकारों की व्याख्या भी कीजिए।

उत्तर AI तकनीक एक तरीका है, जो व्युत्पन्न ज्ञान का उपयोग करता है ताकि इसके एरर को करेक्ट करने के लिये संशोधित किया जा सके। AI technique एक सांख्यिकीय और मैथमेटिकल मॉडल के उन्नत रूपों से बने मॉडल हैं। ये मॉडल कम्प्यूटर या मशीन के लिए उन कार्यों की गणना करना सम्भव बनाते हैं जो मनुष्यों द्वारा किये जाते हैं। इसके कुछ उदाहरण हैं—

(i) आर्टिफिशियल नेचुरल नेटवर्क, (ii) हेयरिस्टिक्स, (iii) मार्कोव डिसिशन प्रोसेस, (iv) नेचुरल लैंग्वेज प्रोसेसिंग।

AI के प्रकार टेक्नोलॉजी के इस युग में, Artificial intelligence सभी इंडस्ट्रीज और कई क्षेत्रों में हावी होने लगी है। इसकी सबसे बड़ी वजह मशीन का मानव की तुलना में अधिक प्रभावी ढंग से कार्य करना है। वह दिन दूर नहीं जब किसी हॉलीवुड मूवीज की तरह रोबोट्स का दबदबा हमारी दुनिया पर होगा। AI या जिसे हम machine learning भी कहते हैं, इसे निम्न भागों में बाँटा जाता है—

(i) **कमजोर बुद्धिमत्ता** Week AI कमजोर बुद्धिमत्ता जिसे नैरो ए०आई० के नाम से भी जाना जाता है। यह पूरी तरह से नैरो टास्क के कार्यों पर केंद्रित है। Weak AI किसी Strong AI या जनरल आर्टिफिशियल इंटेलिजेंस के विपरीत किसी विशेष समस्या को पूरा करने के लिए होती है। यह मशीन अपना काम करने में बहुत स्मार्ट नहीं होती है परंतु उन्हें ऐसा बनाया जाता है कि वे स्मार्ट लगे। उदाहरण के लिये Ludo Game में जब आप कम्प्यूटर मोड खेलते हैं, तो एक तरफ से टोकन्स खुद ब खुद बढ़ती जाती है। उसके ऐसा करने के लिए सारे रूल्स व नियम पहले से ही सॉफ्टवेयर में फीड कर दिए जाते हैं।

(ii) **मजबूत बुद्धिमत्ता** Strong AI मजबूत बुद्धिमत्ता जिसका उपयोग AI डेवलपमेंट के एक निश्चित माइंडसेट का वर्णन करने के लिए किया जाता है। इस का लक्ष्य उस बिंदु पर Artificial intelligence विकसित करना है, जहाँ मशीनों की बौद्धिक क्षमता कार्यात्मक रूप से इंसानों के बराबर हो। स्ट्राँग AI ऐसी मशीनें बनाता है जो वास्तव में इंसान की तरह सोच और कार्य कर सकती हैं। अभी इसके कोई उचित मौजूद उदाहरण नहीं हैं लेकिन कुछ इंडस्ट्री एक स्ट्राँग AI बिल्ड करने के काफी नजदीक पहुँच चूँकि हैं।

(iii) **प्रतिक्रियाशील मशीनें** Reactive Machines यह मशीन बहुत बेसिक होती है क्योंकि यह मेमोरी स्टोर नहीं करती है और भविष्य में किसी कार्य को करने के लिए पुराने अनुभवों का उपयोग भी नहीं कर पाती है। प्रतिक्रियाशील मशीनें केवल देखकर उस पर रिएक्ट करती हैं। IBM का डीप ब्लू जिसने शतरंज के ग्रांड मास्टर कास्परोव (Kasparov) को हराया इसका एक अच्छा उदाहरण है।

(iv) **आत्म जागरूकता** Self-Awareness यह एक ऐसी Atificial intelligence है, जिसके पास अपनी खुद की चेतना, सेल्फ अवेयरनेस और सुपर इंटेलिजेंस होती हैं। सरल शब्दों में आप, इसे एक तरह का ह्यूमन भी कह सकते हैं लेकिन अभी तक इस तरह का बॉट उपलब्ध नहीं है। अगर भविष्य में यह मुमकिन हो सका तो AI के लिए यह बड़ी उपलब्धि होगी।

(v) **सीमित स्मृति** Limited Memory यह ऐसे AI सिस्टम होते हैं, जो फ्यूचर डिसिजन को इन्फॉर्म करने के लिए पिछले अनुभवों का उपयोग कर सकते हैं। सेल्फ ड्राइविंग कार में कुछ डिसिजन मेकिंग फंक्शन्स को डिजाइन किया गया है।

(vi) **मस्तिष्क का सिद्धांत** Theory of Mind इस प्रकार की AI मशीन को लोगों की भावना विश्वास, विचार, उम्मीद और समाजिक रूप से बातचीत करने में सक्षम बनाया जाता है। हालाँकि इस क्षेत्र में काफी प्रयोग हुए हैं लेकिन अभी ऐसी कोई चीज निकलकर सामने नहीं आई जिससे यह सम्भव हो सके।

प्रश्न 6. AI के अनुप्रयोग बताइए।

उत्तर AI महत्त्वपूर्ण है क्योंकि यह विभिन्न उद्योगों; जैसे—मनोरंजन, शिक्षा, स्वास्थ्य, वाणिज्य, परिवहन और उपयोगिताओं में कठिन मुद्दों को हल करने में मदद कर सकता है। AI एप्लीकेशन को पाँच श्रेणियों में बाँटा जा सकता है—

1. **ज्ञान** दुनिया के बारे में जानकारी प्रस्तुत करने की क्षमता; जैसे—वित्तीय बाजार व्यापार, खरीद भविष्यवाणी, धोखाखड़ी की रोकथाम, दवा निर्माण, चिकित्सा निदान, मीडिया की सिफारिश इत्यादि।
2. **विचार** तर्क कटौती के माध्यम से समस्याओं का हल करने की क्षमता; जैसे—वित्तीय परिसम्पत्ति प्रबन्धन, कानूनी मूल्यांकन, वित्तीय अनुप्रयोग प्रसंस्करण, स्वायत्त प्रणाली, खेल इत्यादि।
3. **संचार** बोले जाने वाली और लिखित भाषा को समझने की क्षमता; जैसे—बोले जाने वाली और लिखित भाषाओं का वास्तविक समय अनुवाद, वास्तविक समय प्रतिलेखन, बुद्धिमान सहायक, आवाज नियंत्रण इत्यादि।
4. **योजना** लक्ष्य निर्धारित करने और प्राप्त करने की क्षमता; जैसे—सूची प्रबंधन, भाग पूर्वानुमान, भविष्य कहने वाला रख-रखाव, भौतिक और डिजिटल नेटवर्क अनुकूलन, नेविगेशन इत्यादि।
5. **अनुभूति** ध्वनियों, चित्रों और अन्य संवेदी आदानों के माध्यम से दुनिया के बारे में चीजों का अनुमान लगाने की क्षमता; जैसे—चिकित्सा निदान, स्वायत्त वाहन, निगरानी इत्यादि।

प्रश्न 7. आर्टिफिशियल इंटेलिजेंस और ह्यूमन इंटेलिजेंस के बीच अंतर बताइए।

उत्तर AI तथा Human intelligence के बीच डिफरेंस या मतभेद को समझने के लिए पहले हमें जानना होगा कि इंटेलिजेंस क्या है? परिभाषानुसार, बुद्धि या इंटेलिजेंस में किसी इनफॉर्मेशन को प्राप्त करने की क्षमता, अनुभवों से सीखने की क्षमता, देखकर समझने और ढंग से विचार करने की क्षमता होती है। अपने नेचुरल व्यवहार के कारण यह बुद्धि संज्ञानात्मक कार्यों; जैसे—अनुभूति, मेमोरी, लैंग्वेज और प्लानिंग को एकत्रित करती है। Artificial intelligence और Human intelligence के बीच अंतर निम्न है—

(i) Human intelligence को दिमाग की गुणवत्ता के रूप में परिभाषित किया जाता है। इसके अंदर पिछले अनुभवों से अनुभव लेने, परिस्थिति के अनुकूल प्रतिक्रिया करने की ताकत, विचारों से निपटने और प्राप्त जानकारी का उपयोग करके स्वयं को परिस्थिति से बाहर निकालने की क्षमता होती है। Human intelligence की ऊर्जा दक्षता लगभग 25 watts होती है। मानव सैकड़ों स्किल्स को मैनेज करना अपनी जिंदगी से सीखता है। मानव में अनुभवी परिदृश्य से फैसला लेने की क्षमता होती है। ह्यूमन ब्रेन एनालॉग होती है।

(ii) AI का काम उन मशीन को डिजाइन करना है, जो ह्यूमन व्यवहार की नकल कर सके। रोबोट वैज्ञानिक द्वारा डिजाइन किये गए निर्देशों का उपयोग करते हैं। AI को इंटेलीजेंस एजेंट द्वारा स्टडी और डिजाइन किया जाता है। AI रिसर्च कई क्षेत्रों के टूल्स और इनसाइट्स का प्रयोग करता है। यह रोबोटिक्स सेंट्रल सिस्टम जैसे कार्यों के लिए भी ओवरलैप करता है। AI की ऊर्जा दक्षता एक मॉडर्न मशीन या लर्निंग मशीन में 2 वाट्स होती है। प्रत्येक जिम्मेदारी पर सिस्टम को सिखाने के लिए समय काफी अधिक लगता है।

प्रश्न 8. आर्टिफिशियल इंटेलिजेंस के फायदे और नुकसान बताइए।

उत्तर आर्टिफिशियल इंटेलिजेंस के कुछ मुख्य लाभ निम्नलिखित हैं—

(i) **कम Errors के साथ** High Accuracy AI मशीन या सिस्टम कम errors और उच्च सटीकता के लिए प्रवीण है क्योंकि, यह पूर्व-अनुभव या जानकारी के अनुसार निर्णय लेता है।

(ii) **हाई-स्पीड** AI सिस्टम बहुत उच्च गति और तेजी से निर्णय ले सकता है, क्योंकि AI सिस्टम शतरंज के खेल में एक शतरंज चैंपियन को हरा सकते हैं।

(iii) **उच्च विश्वसनीयता** AI मशीनें अत्यधिक विश्वसनीय हैं और उच्च सटीकता के साथ एक ही क्रिया को कई बार कर सकती हैं।

(iv) **जोखिम भरे क्षेत्रों के लिए उपयोगी** AI मशीनें बम को defuse करके, समुद्र तल की खोज करने, जहाँ मानव को रोजगार देना जोखिम भरा हो सकता है, जैसी स्थितियों में मददगार हो सकती हैं।

(v) **Digital Assistant** AI उपयोगकर्ताओं के लिए डिजिटल सहायक प्रदान करने के लिए बहुत उपयोगी हो सकता है, जैसेकि वर्तमान में विभिन्न E-commerce वेबसाइटों द्वारा ग्राहकों की आवश्यकता के अनुसार उत्पादों को दिखाने के लिए AI तकनीक का उपयोग किया जाता है।

आर्टिफिशियल इंटेलिजेंस के नुकसान हर तकनीक के कुछ नुकसान होते हैं। इतनी लाभप्रद तकनीक होने के कारण अभी भी इसके कुछ नुकसान हैं जिन्हें हमें AI सिस्टम बनाते समय अपने दिमाग में रखना चाहिए। AI के नुकसान निम्नलिखित हैं—

(i) **High cost** AI की हार्डवेयर और सॉफ्टवेयर की आवश्यकता बहुत महँगी है क्योंकि, इसे वर्तमान विश्व आवश्यकताओं को पूरा करने के लिए बहुत सारे रख-रखाव की आवश्यकता होती है।

(ii) **Can't think out of the box** यहाँ तक कि हम AI के साथ smart मशीनें बना रहें हैं, लेकिन फिर भी वे box से बाहर कार्य नहीं कर सकते हैं, क्योंकि रोबोट केवल उसी कार्य को करेगा जिसके लिए वे प्रशिक्षित हैं, या प्रोग्राम किए गए हैं।

(iii) **No feelings and emotions** AI मशीनें एक outstanding performer हो सकती हैं, लेकिन फिर भी इसमें भावना नहीं होती है इसलिए यह मानव के साथ किसी भी प्रकार का भावनात्मक लगाव नहीं कर सकती है और उचित देखभाल न होने पर कभी-कभी उपयोगकर्ताओं के लिए हानिकारक हो सकता है।

(iv) **Increase dependency on machines** Technology के बढ़ने से लोग उपकरणों पर अधिक निर्भर हो रहे हैं और इसलिए वे अपनी मानसिक क्षमताओं को खो रहे हैं।

(v) **No original creativity** जैसाकि मनुष्य बहुत रचनात्मक है और कुछ नए विचारों की कल्पना कर सकता है, लेकिन फिर भी AI मशीनें मानव बुद्धि की इस शक्ति को हरा नहीं सकती हैं और रचनात्मक और कल्पनाशील नहीं हो सकती हैं।

प्रश्न 9. डेटा साइंस के फायदे और नुकसान बताइए।

उत्तर **डेटा साइंस के फायदे** Benefits of data science डेटा साइंस बिजनेस के निर्णय लेने में काफी काम आता है। यह डेटा को बड़े ही सही तरीके से इस्तेमाल करता है और उसे उपयोगी बनाता है जिससे ही हम उसे इस्तेमाल कर सकें।

डेटा से जो हम निर्णय लेते हैं वह हमें काफी लाभ देता है और कार्य करने की क्षमता को भी बढ़ा देता है। डेटा साइंस लोगों की भर्ती में भी काफी काम आता है जैसेकि लोगों की आंतरिक कार्यों में जो लोग आगे की स्टेज के लिए चुने गए हैं तो उनको भी डेटा साइंस को इस्तेमाल करके इसी तरीके से छाँटा जाता है।

डेटा से एप्टिट्यूड टेस्ट लेना और गेम्स, कोडिंग आदि ह्यूमन रिसोर्स के लोगों के लिए काफी उपयोगी होते हैं, क्योंकि इससे वे लोगों को कंपनी में लेते हैं।

प्रश्न 10. डेटा साइंस के उपयोग बताइए।

उत्तर **डेटा साइंस के उपयोग** Uses of data science डेटा साइंस के फायदे कंपनी के लक्ष्य और संसाधनों पर भी निर्भर करते हैं कि कंपनी किस तरह का काम करती है और किस तरह से संसाधनों को इस्तेमाल करती है। सेल्स और मार्केटिंग डिपार्टमेंट पर भी कंपनी का फायदा निर्भर करता है। उदाहरण के तौर पर हम यह देख सकते हैं कि कुछ कंपनी उपयोगकर्ताओं के डेटा को खरीदती हैं और फिर उसका विश्लेषण करती हैं।

डेटा को सही तरीके से समझा जाता है और उसके बाद उसकी उचित रिपोर्ट बनायी जाती है और फिर कंपनी में इसका पूरा विचार विमर्श होता है, जिससे इस डेटा को प्रभावी बनाया जा सके। यह कैम्पेन करने में भी काफी उपयोगी होता है।

नेट्फ़्लिक्स में भी डेटा पर निर्भर करने वाली एल्गॉरिथम इस्तेमाल होता है जोकि उपयोगकर्ता का इतिहास बताती है कि उसने पहले क्या क्या देखा था नेट्फ़्लिक्स में। मशीन लर्निंग की चीजें भी डेटा साइन्स में उपयोग होती हैं जैसेकि इमेज रेकोगनिशन और स्पीच रेकोगनिशन।

प्रश्न 11. हमें डेटा साइंस की आवश्यकता (need) क्यों है?

उत्तर हमें डेटा साइंस की आवश्यकता क्यों है, परंपरागत रूप से, हमारे पास जो डेटा था, वह अधिकतर स्ट्रक्चर्ड और आकार में छोटा था, जिसे सरल BI टूल का उपयोग करके विश्लेषण किया जा सकता था। परंपरागत प्रणालियों में डेटा के विपरीत जो ज्यादातर स्ट्रक्चर्ड था, आज अधिकांश डेटा अनस्ट्रक्चर्ड या सेमी-सट्रक्चर्ड है।

यह डेटा वित्तीय स्रोतों, टेक्स्ट फाइल, मल्टीमीडिया फॉर्म, सेंसर और उपकरणों जैसे विभिन्न स्रोतों से उत्पन्न होता है। अगर आप अपने ग्राहकों के पिछले डेटा जैसे ब्राउजिंग हिस्ट्री, खरीद इतिहास, आयु और आय से उनकी सटीक आवश्यकताओं को समझ सकते हैं तो क्या होगा। इसमें कोई संदेह नहीं है कि आपके पास पहले भी यह सब डेटा था, लेकिन अब विशाल मात्रा और डेटा के साथ, आप मॉडल को अधिक प्रभावी ढंग से प्रशिक्षित कर सकते हैं और अपने ग्राहकों को प्रॉडक्ट अधिक सटीकता के साथ रेकमेंड कर सकते हैं।

सेल्फ-ड्राइविंग कारें अपने आस-पास के मैप बनाने के लिए रडार, कैमरे और लेजर समेत सेंसर से लाइव डेटा एकत्र करती हैं। इस डेटा के आधार पर, यह निर्णय लेती है कि कब स्पीड तेज हो और कब स्पीड कम होनी चाहिए, कब आगे बढ़ना है, और कब टर्न लेना हैं, आदि। इसके लिए वे एडवांस मशीन लर्निंग एल्गोरिथ्म का उपयोग करती है।

प्रश्न 12. डेटा साइंस क्या है?

उत्तर डेटा साइंस एक ऐसा अध्ययन है जो व्यावसायिक उद्देश्यों के लिए उपयोग किए जाने वाले डेटा स्रोतों के सार्थक जानकारी की पहचान, प्रतिनिधित्व और डेटा विज्ञान निष्कर्षण से संबंधित है।

प्रत्येक मिनट में उत्पन्न होने वाले फैक्ट की भी मात्रा के साथ, उपयोगी अंतर्दृष्टि एक्सट्रैक्ट करने की आवश्यकता है, ताकि बिज़नेस भीड़ से बाहर खड़ा हो सके। डेटा इंजीनियर्स, डेटा माइनिंग, डेटा मंगिंग और अन्य प्रोसेस को सुविधाजनक बनाने के लिए डेटाबेस और डेटा स्टोरेज सेट अपने करते हैं। हर दूसरा ऑर्गनाइजेशन, प्रॉफिट के पीछे भाग रहा है, लेकिन ताजा और उपयोगी अंतर्दृष्टि के आधार पर कुशल रणनीतियों को तैयार करन वाली कंपनियाँ हमेशा लंबे समय तक खेल जीतती है।

प्रश्न 13. डेटा माइनिंग क्या है? इसके लक्ष्यों को भी परिभाषित कीजिए।

उत्तर Data mining को data या knowledge discovery भी कहते हैं। data mining, बहुत बड़े डेटा के समूह में से small डेटा को search करने की प्रक्रिया है। इस प्रक्रिया में परम्परागत statistics, artificial intelligence तथा computer graphics का प्रयोग किया जाता है।

Data mining एक उपयोगी तकनीक है जिसका प्रयोग करके कंपनियाँ बहुत बड़े data के समूह में से महत्त्वपूर्ण information को निकालती है।''

डेटा माइनिंग का प्रयोग करके hidden patterns और useful data को खोजा जाता है और फिर इन pattern और data के आधार पर decision making की जाती है और डेटा माइनिंग की प्रक्रिया का प्रयोग करके organizations बिज़नेस में आने वाली problem को solve करते हैं। data mining में डेटा को analyze करने के लिए data mining tools का प्रयोग किया जाता है। ये tool बहुत ही powerful होते हैं।

Data mining के निम्नलिखित goals होते हैं—

1. **Expanatory** इसमें देखी गयी घटना या परिस्थिति को explain किया जाता है।
2. **Confirmatory** इसमें संभावनाओं से मुक्त परिकल्पनाओं की confirmation की जाती है।
3. **Analyzatry** इसमें नए डेटा को analyze किया जाता है जिससे positive feedback दी जा सके।

प्रश्न 14. डेटा माइनिंग के लाभ और हानि बताइए।

उत्तर **डेटा माइनिंग के लाभ** Advantage of Data Mining इसके लाभ निम्नलिखित हैं—

(i) Data mining की तकनीक के द्वारा कंपनी knowledge पर आधारित information को प्राप्त करती है।

(ii) इसके द्वारा organizations अपने production और operation को बेहतर करते हैं।
(iii) डेटा माइनिंग दूसरे statistical data applications की तुलना में cost effective है अर्थात् इससे cost (मूल्य) की बचत होती है।
(iv) इसके द्वारा आसानी से decisions को लिया जा सकता है।
(v) इसे नए systems में implement करना बहुत ही सरल होता है।
(vi) इसकी speed बहुत ही तेज होती है जिससे बड़े data को कम समय में analyze कर लिया जाता है।
(vii) यह profitable customers को आसानी से खोज लेता है जिससे product को sale करना आसान हो जाता है और customer से relationship भी बेहतर होती है।

इसकी हानियाँ Disadvantage of Data Mining इसकी हानि निम्नलिखित हैं—
(i) इसका बड़ा नुकसान यह है कि इसमें data की security और privacy नहीं होती है, इसमें सभी data को स्टोर किया जाता है जैसेकि social media के message, photos आदि इससे लोगों की privacy खत्म होती है।
(ii) डेटा माइनिंग के द्वारा collect किया गया data ज्यादातर incomplee (अधूरा) होता है।
(iii) इसमें irrelevant (बेकार) data को भी एकत्र किया जाता है।

प्रश्न 15. डेटा माइनिंग की विशेषताएँ क्या हैं?

उत्तर **डेटा माइनिंग की विशेषताएँ** Characteristics of Data Mining इसकी विशेषताएँ निम्नलिखित हैं—
(i) यह future prediction करता है। इसका अर्थ है कि यह भविष्य में होने वाली घटनाओं को predict करता है।
(ii) यह बड़े datasets और database को focus करता है।
(iii) इसमें pattern की prediction, automatic होती है और यह behavior analysis पर आधारित होती है।
(iv) यह काम में आने वाली information को create करता है।

प्रश्न 16. Data mining analysis कितने प्रकार का होता है?

उत्तर **डेटा माइनिंग के प्रकार** Data mining analysis के दो प्रकार होते हैं, जो कि निम्नलिखित है—

1. **Predictive Data Mining Analysis** यह भविष्य में होने वाली घटनाओं को predict करता है, यह निम्न चार प्रकार का होता है—
   (i) Classificiation Analysis
   (ii) Regression Analysis
   (iii) Time Serious Analysis
   (iv) Prediction Analysis
2. **Descriptive Data Mining Analysis** इसका प्रयोग data को उपयोगी information में बदलने के लिए किया जाता है, इसके भी निम्न चार प्रकार होते हैं—
   (i) Clustering Analysis
   (ii) Summarization Analysis
   (iii) Association Rule Analysis
   (iv) Sequence Discovery Analysis

प्रश्न 17. डेटा माइनिंग के उदाहरण क्या हैं?

उत्तर **Data Mining के उदाहरण** एक credit card data mining का इस्तेमाल कर उनके members की buying habits को समझती है। वही cardholders के purchases को analyze कर company उनके shopping habits को study कर सकती है, वहीं वह ये भी जान सकती है कि कैसे अलग-अलग जगहों के लोग किस प्रकार की खरीदारी ज्यादा करते हैं।

वही उस individuals को कुछ specific promotions offer करने में। ये information काफी महत्त्वपूर्ण हो सकती है वहीं ये समान data से उनकी खरीदारी के pattern को भी समझा जा सकता है, फिर चाहे वह किसे भी देश के हो या किसी भी राज्य से क्यूँ न हो।

ये information काफी valuable होती है उन companies के लिए जोकि advertise करना चाहते हैं या कोई नया businesses आरम्भ करना चाहती हैं।

वहीं Online services, जैसेकि Google और Facebook, बहुत मात्रा की data को mine करते हैं जिससे कि वे targeted content और advertisements offer users को कर पायें।

वहीं Google भी, ऐसे ही search queries को analyze करता है वहीं ऐसे popular searches को खोजता है। कुछ specific areas में और उन्हें अपने autocomplete list में डाल देता है (ये वो suggestions होती हैं जोकि दिखाई पड़ती है जैसे ही आप कुछ type करें तब)।

User Activity data को mine कर Facebook पर बहुत से अलग-अलग topics की जानकारी हासिल कर लेता है, वहीं उसी हिसाब से वह ads target करता है जोकि उसी information के ऊपर आधारित होता है।

जहाँ data mining को मुख्य रूप से marketing purposes के लिए इस्तेमाल किया जाता है, वहीं इसके बहुत-से दूसरे uses भी होते हैं। उदाहरण के लिए, healthcare companies इस data mining का इस्तेमाल कर उन links को खोज सकती है जोकि कुछ विशेष genes और diseases के विषय में होती है।

मौसम विभाग भी इन data को mine कर मौसम के pattern को खोज सकता है और साथ में इसकी मदद से आगे की meteorologic events के विषय में पूर्व अनुमान लगा सकता है।

वहीं Traffic management भी इन automotive data को mine पर इस चीज़ का पूर्व अनुमान लगा सकती है कि भविष्य में किस प्रकार की traffic levels होने वाली है और उस हिसाब से highways और streets के लिए सही plans बना सकते हैं।

प्रश्न 18. Data Mining का प्रयोग कैसे किया जाता है?

उत्तर Use of Data Mining **डेटा माइनिंग** में बिल्कुल Raw Data की जाँच की जाती है, जिसके बाद उस Raw Data में से अपनी जरूरत की जानकारियाँ इकट्ठा की जाती है, जिसका प्रयोग विश्लेषण के लिए किया जाता है।

अगर हम एक बिज़नेस में डेटा माइनिंग तकनीक के इस्तेमाल की बात करे तो डेटा माइनिंग द्वारा अपने Customers से जुड़ी जानकारी Collect की जाती है; जैसे—ग्राहकों की पसंद, जरूरतें, माँग इत्यादि और इसी के आधार पर Business की मार्केटिंग रणनीतियाँ बनाई जाती हैं और Sale को बढ़ाया जाता है।

उदाहरण के तौर पर जैसे आप इंटरनेट पर कोई प्रोडक्ट सर्च करते हैं, तो आपके द्वारा सर्च की गयी जानकारी का एक रिकॉर्ड स्थापित हो जाता है, जो एक डेटा के रूप में इंटरनेट पर मौजूद Server's पर Save हो जाता है, और जब कभी आप फिर से इंटरनेट का इस्तेमाल करते हैं, तो आपको पिछले सर्च किए गए प्रोडक्ट से जुड़ी दूसरी ओर नई जानकारियाँ दिखाई देने लगती हैं, तो यह सब Data Mining द्वारा ही होता है, जिसमें Data Mining Process द्वारा ऐसे ही Raw Data को फिल्टर करके इनफार्मेशन जुटाई जाती है और बिज़नेस डेवलपमेनट में उसका इस्तेमाल किया जाता है।

प्रश्न 19. डेटा माइनिंग के अनुप्रयोग क्या हैं?

उत्तर **डेटा माइनिंग के अनुप्रयोग** Applications of Data Mining इसका प्रयोग बहुत सारी जगहों पर किया जाता है। इसके अनुप्रयोग निम्नलिखित हैं—

1. **Healthcare (स्वास्थ्य) के क्षेत्र में** इसका प्रयोग करके मरीज के रोग के बारे में पता लगाया जाता है। यह ऐसे hospitals के बारे में जानकारी देता है जहाँ मरीज का इलाज कम पैसों और कम समय में हो जाए।
2. **Market के क्षेत्र में** डेटा माइनिंग के द्वारा customer के behavior का पता लगाया जाता है। इसमें यह देखा जाता है कि अगर customer ने कुछ समान खरीदा है तो वह इसके साथ दूसरा कौन-सा सामान खरीदेगा।

3. **Education (शिक्षा) के क्षेत्र में** डेटा माइनिंग का प्रयोग करके student के result को predict किया जाता है। यह ये भी बताता है कि कैसे किसी student को teach करे और क्या teach करें।

4. **Fraud को detect करने में** आजकल बहुत सारे frauds हो रहे हैं जिससे लाखों लोगों का पैसा बर्बाद हो जाता है। इससे बचने में डेटा माइनिंग मदद करता है।

प्रश्न 20. Data mining तैयार के चरण लिखिए।

उत्तर **Data Processing के Basic Stages** Basic stages में data processing cycle के मुख्य रूप से तीन चरण होते हैं—

(i) **Input** इस चरण में, input data को processing के लिए एक convenient form में prepare किया जाता है। ये form procesing machine के ऊपर निर्भर करती है। उदाहरण के लिए, जब electronic computers का इस्तेमाल किया जाता है, तब input data को किसी एक मौजूद medium में store किया जाता जैसेकि magnetic disks, tapes या और कुछ।

(ii) **Processing** इस चरण में, input data को produce data में बदला जाता है जोकि ज्यादा useful form होता है। उदाहरण के लिए, किसी company में sales की summary calculate करने के लिए sales orders को देखा जाता है।

(iii) **Output** इस चरण में, इसके पूर्व के processing step के result को collect किया जाता है। Output data का कोई particular form इसके ऊपर निर्भर करता है कि उस data को किस तरह से इस्तेमाल किया जाता है। उदाहरण के लिए, output data में कोई employees के pay-checks भी हो सकते हैं।

प्रश्न 21. मशीन लर्निंग से आप क्या समझते हैं?

उत्तर **मशीन लर्निंग** Machine Learning एक प्रकार की learning है जिसमें machine बिना उसे explicitly programmed किये खुद अपने आप ही learn करती है।

ये AI का एक प्रकार का application है जोकि system को वह ability प्रदान करता है जिससे वह automatically अपने experience से learn और improve कर सकें यहाँ पर हम एक ऐसा program generate कर सकते हैं जो कि उसी program के input और output को integrate कर बनाया गया हो।

प्रश्न 22. Artificial Intelligence और Machine learning में क्या अंतर है?

उत्तर **Artificial Intelligence और Machine learning में अंतर**

Artificial Intelligence	Machine Learning
AI का full form होता है, Artificial intelligence जहाँ पर intelligence को define किया जाता है। एक ऐसी ability जहाँ पर knowledge को acquire और apply किया जाता है।	ML का Full form होता है, Machine Learning जिसे define किया जाता है यह एक प्रकार का feature है जिससे experience से knowledge और skill को acquire किया जाता है।
इसका लक्ष्य मॉडल को सफल बनाना होता है।	इसका लक्ष्य मॉडल की सटीकता को बढ़ाना है।
ये computer program के जैसे कार्य करती है जोकि smart work होता है।	वहीं ये एक simple concept machine होती है जोकि data ग्रहण करती हैं और उसी से learn करती है।
इसका main goal है, natural intelligence को simulate करना जिससे ये complex problem solve कर सके।	इसका main goal है किसी certain task से data learn करना जिससे ये machine के performance को maximize कर सके उसी specific task के लिए।
AI खुद ही decision making होता है।	ML allows करता है system को जिससे वो data से नयी चीजें learn कर सके।

ये एक ऐसा system develop करता है जोकि इंसानों को mimic कर सके जिससे ये किसी circumstances में ठीक तरीके से respond कर सके।	इसमें ये ज्यादा involve रहता है self learning algorithms create करने में।
AI हमेशा किसी problem का optimal solution ढूंढ़ने में विश्वास रखता है।	वहीं ML किसी problem का कोई भी solution चाहे वो optimal हो या न हो ढूंढ़ने में विश्वास रखता है।
AI आखिर में intelligence और wisdom की ओर lead करता है।	वहीं ML (Machine Learning) Knowledge की ओर lead करता है।

प्रश्न 23. मशीन लर्निंग कैसे कार्य करता है?

उत्तर Online shopping में, जहाँ ecommerce websites में रोजाना करोड़ों लोग आते हैं और अपने पसंदीदा चीज़ खरीदते हैं, क्यूँकि यहाँ पर उन्हें चुनने के लिए unlimited range के brands, colors, price range और बहुत कुछ दिखाई पड़ते हैं। Online shopping में, हम ऐसे ही अपने चीज़ नहीं खरीद लेते हैं बल्कि हम बहुत-सी चीजों को पहले देखते हैं और सही का चुनाव करते हैं, ऐसे देखने के लिए हमें बहुत सारे items को खोलना पड़ता है।

इन्हीं Items को बहुत-से advertising platform target कर लेते हैं जिससे हमें recommended list में ऐसे items दिखाई पड़ते हैं जिन्हें हम पहले खोज चुके होते हैं। इसमें आपको आश्चर्यचकित होने की जरूरत नहीं है क्योंकि ये कोई इन्सान नहीं कर रहा है बल्कि इस task को कुछ ऐसे program कर दिया गया है जिससे ये हमारी गतिविधियों को रिकॉर्ड कर सके।

इस चीज़ के लिए Machine Learning हमारे बहुत काम आती है क्योंकि वह हमारे behaviour को पढ़ लेती है और उसी हिसाब से अपने experience से स्वयं को program कर लेता है इसलिए जितनी अच्छी data मिलेगी उतने ही अच्छे से learning models बनकर तैयार होंगे और customers को भी उस हिसाब से लाभ होगा।

अगर हम Tradition Advertisement की बात करें तब उसमें newspaper, magazines, radio प्रमुख थे, लेकिन अब technology बदल रही है और ये smart भी बनती जा रही है जोकि Targeted advertisement (Online ad system) से कर रही है।

ये बहुत ही कारगर विधि है जोकि सिर्फ targeted audience पर ही अपने advertisement show करते हैं जिससे कि conversion rate ज्यादा होता है।

बात सिर्फ Online shopping तक ही नहीं रह गयी है, बल्कि Health Care industries में भी Machine Learning से बहुत-से कार्य किये जाते हैं।

Researchers और Scientists अब ऐसे models prepare किये हुए हैं जोकि cancer जैसे बड़े रोग को पहचानने के लिए उपयोगी है। machines को train करते हैं इसके लिए वह इन machines में cancer cell images feed कर दिए हैं जोकि actual में cancel cells के अलग-अलग variations हैं।

जिससे मरीज के tests के दौरान इन ML System का इस्तेमाल किया जाता है। Cancer cells को detect करने के लिए, जोकि इन्सानों के लिए करना बहुत time taking था, इससे बहुत ही कम समय में बहुत मात्रा के मरीजों के cancer test हो पाते हैं।

इसके अलावा Machine learning का इस्तेमाल IMDB ratings, Google Photos, Google Lens के लिए किया जाता है। ये केवल आप पर depend करता है कि आप Machine learning का इस्तेमाल कहाँ पर और कैसे करना चाहते हैं।

Machine Learning में सही model बनाने के लिए computers को सही मात्रा में data की जरूरत होती है; जैसेकि text, image, audio. इसमें जितनी अच्छी और बेहतर quality की data होती है उतनी की अच्छी model learning होगी। इसके लिए algorithms को कुछ इस ढंग से design किया जाता है जिससे past experience से machine future actions को कर पाता है।

प्रश्न 24. मशीन लर्निंग के लाभ और दोष समझाइए।

उत्तर **मशीन लर्निंग के लाभ** Advantages of Machine Learning

(i) Machine learning के बहुत सारे wide applications है जैसेकि banking और financial sector, healthcare, retail, publishing इत्यादि industries में।

(ii) Google और Facebook machine learning के इस्तेमाल से relevant advertisements push कर पाते हैं। ये सभी advertisements users के past search behaviour पर ही आधारित होते हैं इसलिए इसे targeted ads भी कहा जाता है।

(iii) Machine learning का प्रयोग multi-dimensional और multi-variety data को handle करने के लिए किया जाता है वह भी dynamic environments में।

(iv) Machine learning के प्रयोग से time cycle reduction होता है और resources का efficient utilization भी किया जा सकता है।

(v) अगर कोई चाहता है कि continuous quality, large और complex process environments प्रदान करने के लिए तब भी इसमें machine learning के कारण, ऐसे कुछ tools मौजूद हैं।

(vi) वैसे तो Machine Learning के benefits के अन्तर्गत बहुत सारी चीज़ आती हैं जो कि practically हमारे बहुत काम आ सकते हैं, जैसेकि autonomous computers का development, software programs इत्यादि। साथ ही ऐसे process भी जोकि बाद में automation of tasks हो पाते हैं।

मशीन लर्निंग के दोष Disadvantages of Machine Learning मशीन लर्निंग के दोष निम्नलिखित हैं—

(i) Machine learning की एक major challenge होती है, Acquisition जिसमें, different algorithms पर based होकर data को process किया जाता है और इसे processed किया जाता है किसी respective algorithms के input के हिसाब से इस्तेमाल करने से पहले। इसलिए इसका significant impact होता है results के ऊपर जो कि achieved या obtained किया जाता है।

(ii) एक और शब्द होता है, interpretation. जिसका मतलब है कि results भी एक बहुत major challenge है। इससे ये determine करना होता है कि machine learning algorithms की effectiveness कितनी है।

(iii) Machine algorithm के use limited होते हैं। साथ में ये भी surety नहीं होती है कि algorithms हमेशा सभी imaginable cases में भी काम करेगी।

(iv) Deep learning algorithm की तरह ही machine learning में भी बहुत-से training data की जरूरत होती है।

(v) एक machine learning बहुत ही notable limitation की यह है कि यह errors के प्रति ज्यादा susceptible होते हैं। Brynjolfsson और McAfee इसके actual problem के विषय में ये बताया है कि जब वे कोई error करते हैं, तब उन्हें diagnose और correct करना बहुत ही कठिन होता है। ऐसा इसलिए होता है क्योंकि इसे underlying complexities के नीचे से गुजरना होता है।

(vi) इसमें machine learning system के साथ immediate predictions करने के बहुत ही कम possibilities होती है साथ में ये न भूलें की ये historical data से ही ज्यादातर learn करते हैं। इसलिए जितनी ज्यादा बड़ी data होगी और जितनी ज्यादा देर तक ML को expose किया जाये, इससे ये और भी बेहतर perform कर सकता है।

(vii) ज्यादा Variability का न होना भी machine learning का एक दूसरी limitation है।

प्रश्न 25. मशीन लर्निंग के प्रकार क्या हैं?

उत्तर **Machine Learning के प्रकार** मुख्यत: तीन प्रकार के मशीन लर्निंग एल्गोरिथम हैं—

1. **Supervised Learning** इस learning में इनपुट के रूप में Labelled Data जिनमें Example तथा Answer शामिल है और फिर एल्गोरिथ्म इन labelled data के आधार पर सही Result का अनुमान लगाता है। Supervised learning अग्र दो प्रकार की होती है—

(i) Regression (ii) Classification

2. **Unsupervised Learning** इनमें इनपुट के रूप में labelled data की ओर answer नहीं दिया जाता। इसमें एल्गोरिथम को डेटा के आधार पर अनुमान लगाना होता है। Unsupervised learning दो प्रकार की होती है—

(i) Clustering (ii) Association

3. **Reinforcement Learning** इस learning में एल्गोरिथम स्वयं के Reward और Feedback को इनपुट के रूप में use करता है।

प्रश्न 26. मशीन लर्निंग के अनुप्रयोग (applications) क्या हैं?

उत्तर **मशीन लर्निंग के अनुप्रयोग** Applications of Machine Learning हमारे दैनिक जीवन में Machine Learning की कई Applications's है जिनका इस्तेमाल हम सभी करते हैं जिनमें से कुछ निम्न प्रकार हैं—

(i) Facebook दुनियाभर में फेसबुक का इस्तेमाल काफी बड़ी मात्रा में किया जाता है और हम सभी इसका इस्तेमाल करते हैं और मशीन लर्निंग का इस्तेमाल फेसबुक में Automatic Friend Tagging Suggestion में किया जाता है जिसमें Face Detection और Image Recognition के आधार पर फेसबुक अपने डेटाबेस में चेक करता है और किसी फोटो या इमेज को पहचान लेता है।

(ii) Shopping Websites आप अगर ऑनलाइन खरीदारी करते हैं तो आपने देखा होगा कि आपके सर्च किए गए प्रोडक्ट से जुड़ी जानकारियाँ आपको हर जगह दिखाई देने लगती हैं, जैसे आपने Amazon पर कुछ सर्च किया और कुछ देर बाद जब आप फेसबुक या यूट्यूब खोलेंगे तो वहाँ भी आपको उसी प्रोडक्ट से जुड़े विज्ञापन दिखाने लगते हैं तो यह सब Machine Learning का कमाल है जिसमें गूगल आपकी हर गतिविधि का ध्यान रखता है और आपको उसी अनुरूप विज्ञापन दिखाता है।

(iii) E-Mail Spam Filter E-Mail इस्तेमाल करते समय आपने देखा होगा कैसे सिर्फ हमारी जरूरत की Mails ही इनबॉक्स में आती हैं और अधिकतर Spam Malls Spam नाम से बने फोल्डर में चली जाती हैं तो इसके पीछे भी Machine Learning इस्तेमाल हो रही होती है जिसमें Machine Learning द्वारा Automatically किसी ईमेल का Content और Source Detect कर लिया जाता है और कुछ गलत पाए जाने पर ईमेल को spam कर दिया जाता है।

(iv) Uber अगर आप यातायात के लिए Uber का इस्तेमाल करते हैं तो आपने देखा होगा किस तरह से Uber खुद कस्टमर की Location का पता लगा लेता है, Real Time में गाड़ी की Actual Location भी दिखती रहती है, ड्राइवर को सबसे छोटे और खुले रास्तों के बारे में भी पता चलता रहता है और साथ ही भारी माँग होने पर अपने Charges में फेरबदल भी करता रहता है तो यह सब Machine Learning से ही संभव हो पाता है।

2 डाटा प्री-प्रोसेसिंग, एनालाइसिस एवं विजुलाइजेशन
Data Preprocessing, Analysis and Visualization

बहुविकल्पीय प्रश्न (MCQ)

प्रश्न 1. Noisy व In cosistent डाटा को हटाने के लिए कौन-सी प्रक्रिया की जाती है?

(a) Data cleaning
(b) Data Reduction
(c) Data integration
(d) Data transformation

उत्तर (a) Data cleaning

प्रश्न 2. निम्न में से किस प्रक्रिया में बहुत-से data sources को एकसाथ combine किया जाता है?

(a) Data Reduction
(b) Data Integration
(c) Data Transformation
(d) Data Cleaning

उत्तर (b) Data Integration

प्रश्न 3. निम्न में से कौन-सा data क्लस्टर के बाहर हो जाता है?

(a) Inconsistent data
(b) Missing data
(c) Noisy data
(d) इनमें से कोई नहीं

उत्तर (c) Noisy data

प्रश्न 4. निम्नलिखित में से कौन-सी डाटा सामान्यकरण (data normalization) तकनीक है?

(a) Numerosity Reduction
(b) Data Reduction
(c) Clustring
(d) Decimal Scaling

उत्तर (d) Decimal Scaling

प्रश्न 5. किस प्रक्रिया में किसी attribute के डाटा को स्केल किया जाता है ताकि यह एक छोटी श्रेणी में आ जाए जैसे कि –1.0 से 1.0 या 0.0 से 0.1 ?

(a) Aggregation
(b) Normalization
(c) Binarization
(d) Clustering

उत्तर (b) Normalization

प्रश्न 6. डाटा एनालिसिस प्रक्रिया में क्या होता है?

(a) डाटा का निरिक्षण
(b) डाटा की प्रोसेसिंग
(c) डाटा की क्लीनिंग
(d) ये सभी

उत्तर (d) ये सभी

प्रश्न 7. कौन-सी डाटा एनालाइसिस तकनीक पिछले डाटा रिकॉर्ड को संक्षिप्त रूप में प्रस्तुत करती है?

(a) डायग्लेस्टिक एनालिसिस
(b) प्रीडिक्टिव एनालिसिस
(c) डीस्क्रिप्टिव एनालिसिस
(d) प्रीस्क्रिप्टिव एनालिसिस

उत्तर (c) डीस्क्रिप्टिव एनालिसिस

प्रश्न 8. निम्न में से किसका उपयोग करके डाटा विजुलाइजेशन किया जा सकता है?

(a) Charts
(b) Maps
(c) Graphs
(d) इन सभी का

उत्तर (d) इन सभी का

प्रश्न 9. कौन-सी डाटा सामान्यकरण (data normalization) तकनीक में attributes की values को decimal points में change किया जाता है?

(a) Z-score normalization
(b) Decimal scaling
(c) Min-max normalization
(d) इनमें से कोई नहीं

उत्तर (b) Decimal scaling

खण्ड 'अ' : अतिलघु उत्तरीय प्रश्न

प्रश्न 1. डेटा प्री-प्रोसेसिंग तकनीक क्या है?

उत्तर Mean Removal, Scaling, Normalization, Binarization, One Hot Encoding and Lable Encoding जैसी कई तकनीकों का उपयोग करके डेटा को प्री-प्रोसेस किया जा सकता है।

प्रश्न 2. माध्य निकालने (Mean Removal) से आप क्या समझते हैं?

उत्तर मशीन लर्निंग में मीन रिमूवल एक प्रकार की डेटा प्री-प्रोससिंग तकनीक है, जिसका उपयोग हर सुविधा से एक माध्य निकालने के लिए किया जाता है ताकि यह शून्य पर केंद्रित हो सके। यह सुविधा से पूर्वाग्रह को हटाने में भी मदद करता है।

प्रश्न 3. क्लस्टरिंग क्या है?

उत्तर इसके द्वारा समान प्रकार के डाटा को एक क्लस्टर में रखा जाता है और जो noisy data होता है वह क्लस्टर के बाहर हो जाता है।

प्रश्न 4. Regression क्या है?

उत्तर इस विधि में regression function का प्रयोग किया जाता है। Regression दो प्रकार का होता है—linear और multiple.

प्रश्न 5. Noisy data से आप क्या समझते हैं?

उत्तर Noisy data useless (बेकार) डाटा होता है तथा इसे machine के द्वारा interpret नहीं किया जा सकता है, दोषपूर्ण (faulty) डाटा को collect करने से एवं data entry में errors आने आदि से noisy data उत्पन्न हो जाता है।

प्रश्न 6. Data integration से आपका क्या तात्पर्य है?

उत्तर इस step में बहुत-से data sources को एक साथ combine किया जाता है। इसको data migration tools और data synchronization tools का प्रयोग करके किया जाता है।

खण्ड 'ब' : लघु एवं दीर्घ उत्तरीय प्रश्न

प्रश्न 1. डाटा प्री-प्रोसेसिंग क्या है?

उत्तर Data preprocessing एक data mining तकनीक है जिसका प्रयोग raw data को महत्त्वपूर्ण और प्रभावी format (रूप) में बदलने के लिए किया जाता है।

Real world में जो data होता है वह अकसर incomplete (अधूरा), noisy और inconsistent होता है।

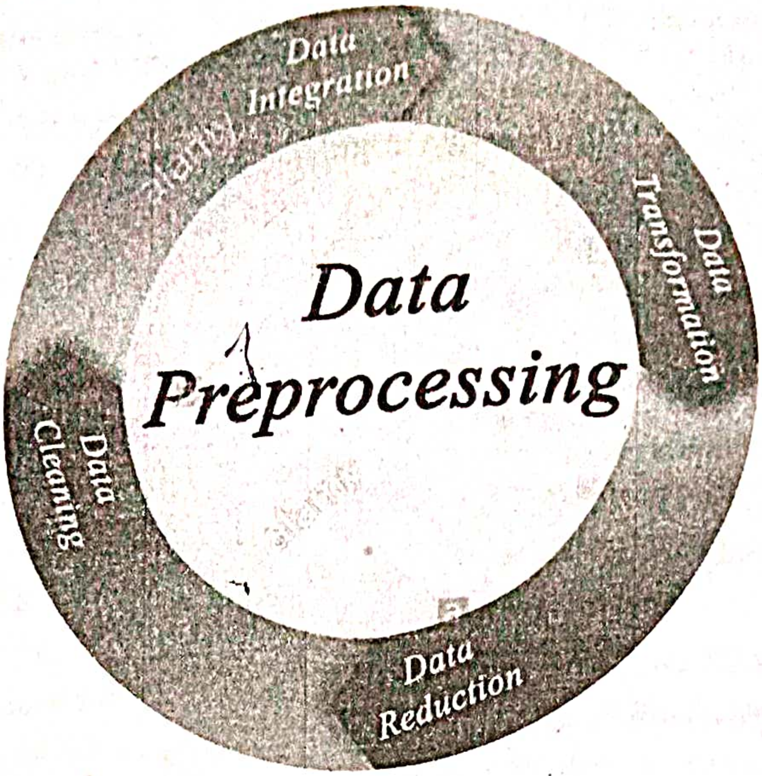
Incomplete का मतलब है कि उसमें attributes की कमी होती है। noisy का मतलब है कि इसमें erros होती है।

Inconsistent का अर्थ है कि डाटा में विसंगतियाँ और डाटा duplicate होता है।

Data preprocessing steps Data preprocessing में निम्नलिखित चरण होते हैं—

1. **Data cleaning** डाटा वह irrelevant (असंगत) हो सकता है और इसके कुछ parts (हिस्से) missing हो सकते हैं, इसके लिए data cleaning की आवश्यकता पड़ती है। इसके अंतर्गत missing data, noisy data आदि को handle किया जाता है।

   (a) **Missing data** यह स्थिति तब उत्पन्न होती है जब डाटा में से कुछ डाटा missing होता है। इसको हम अग्रलिखित प्रकार से handle कर सकते हैं—



चित्र 2.1

(i) **tuples को ignore करना** यह approach तभी उपयुक्त होती है जब हमारे पास बहुत बड़ी मात्रा में dataset होता है और एक tuple के अंदर बहुत सारी values missing रहती है।

(ii) **missing values को fill करना** इसको fill करने के बहुत-से तरीके होते हैं। आप इसे manually भी fill कर सकते हैं।

(b) **Noisy data** Noisy data जो है useless (बेकार) डाटा होता है तथा इसे machine के द्वारा interpret नहीं किया जा सकता है, दोषपूर्ण (faulty) डाटा को collect करने से एवं data entry में errors आने आदि से noisy data उत्पन्न हो जाता है। इसे निम्नलिखित तरीकों से handle किया जा सकता है—

(i) **Binning method** इस विधि का प्रयोग sorted data पर किया जाता है। इसमें पूरे data को एकसमान size के segments में विभाजित कर लिया जाता है और विभिन्न methods का प्रयोग task को पूरा करने के लिए किया जाता है। प्रत्येक segment को अलग-अलग handle किया जाता है।

(ii) **Regression** इस विधि में regression function का प्रयोग किया जाता है। Regression दो प्रकार का होता है—linear और multiple.

(iii) **Clustering** इसके द्वारा समान प्रकार के data को एक cluster में रखा जाता है और जो noisy data होता है वह cluster के बाहर हो जाता है।

2. **Data transformation** इस step के द्वारा, data को data mining की प्रक्रिया के लिए उपयोगी form में बदला जाता है। इसके निम्नलिखित तरीके होते हैं—

(i) **Normalization** Data values को एक विशिष्ट range में मापने के लिए इसका प्रयोग किया जाता है। यह range है—(–1.0 से 1.0 या 0.0 से 1.0).

(ii) **Attribute section** इस तरीके में नए attributes को दिये गये attributes के set से निर्मित किया जाता है।

(iii) **Discretization** इसका प्रयोग numeric attributes की raw values को replace करने के लिए किया जाता है।

(iv) **Hierarchy generation** इसमें low-level के attributes को high level attributes में बदल दिया जाता है; जैसे—attribute "city" को attribute "country" में बदल दिया जाता है।

3. **Data Reduction** Data mining एक ऐसी प्रक्रिया है जिसका प्रयोग बहुत बड़ी मात्रा के data को handle करने के लिए किया जाता है। बड़ी मात्रा के data के साथ काम करने के कारण कभी-कभी analysis करना बहुत कठिन हो जाता है। इस परेशनी को दूर करने के लिए हम data reduction technique का प्रयोग करते हैं। इस technique का मुख्य उद्देश्य storage क्षमता को बढ़ाना और analysis costs को कम करना होता है।

Data reduction के steps निम्नलिखित हैं—

(i) **Data Cube Aggregation** Data cube को निर्मित करने के लिए aggregation opeation को data पर apply किया जाता है।

(ii) **Attributes Subset Selection** इसमें उचित attributes का प्रयोग किया जाता है और शेष attributes को discard (रद्द) कर दिया जाता है।

(iii) **Numerosity Reduction** इसके द्वारा पूरे data को स्टोर करने की बजाय हम केवल data के model को स्टोर करते हैं।

(iv) **Dimensionality Reduction** Encolding विधियों के द्वारा यह data के size को कार्य कर देता है। यह lossy या lossless दोनों में से कोई भी हो सकता है। Dimensionality Reduction के दो प्रभावी methods है—wavelet transforms और PCA (principal component analysis).

प्रश्न 2. Normalization से आपका क्या मतलब है?

उत्तर Normalization का उपयोग किसी attribute के डाटा को स्केल करने के लिए किया जाता है ताकि यह एक छोटी श्रेणी में आ जाये; जैसे—1.0 से 1.0 या 0.0 से 0.1। यह वर्गीकरण एल्गोरिथ्म के लिए आमतौर पर उपयोगी है।

Methods of Data Normalization

(i) Decimal Scaling

(ii) Min-Max Normalization

(iii) z-Score Normalization (zero-mean Normalization)

प्रश्न 3. Decimal scaling से आपका क्या अभिप्राय है?

उत्तर Decimal scaling एक डाटा सामान्यकरण (data normalization) तकनीक है। इस तकनीक में हम attributes की values decimal points में change करते हैं। Decimal points की यह movement पूरी तरह से सभी attributes की values के बीच अधिकतम मूल्य पर निर्भर करती है।

Decimal Scaling Formula

A value v of attribute A is can be normalized by the following formula

Normalized value of attribute = **(vi/10j)**

Example of Decimal scaling

OGPA	Formula	CGPA Normalized after Decimal scaling
2	2/10	0.2
3	3/10	0.3

प्रश्न 4. Min max normalization से आपका क्या तात्पर्य है?

उत्तर Min max normalization डाटा को सामान्य करने के लिए सबसे सामान्य तरीकों में से एक है। उस सुविधा का न्यूनतम मान 0 में रुपांतरित हो जाता है, अधिकतम मान 1 में बदल जाता है और प्रत्येक अन्य मान 0 से 1 बीच एक दशमलव में परिवर्तित हो जाता है।

Min Max normalization Example

Marks
8
10
15
20

The minimum value of the given attribute. Here Min is **8.**

Max:

The maximum value of the given attribute. Here Max is **20.**

V: V is the respective value of the attribute. For example here V1=8, V2=10, V3=15 and V4=20

newMax:

1

newMin:

0

$$v' = \frac{v - \min}{\max_A - \min_A}(new - \max_A - new \min_A) + new - \min_A$$

For Marks as 8 :

$$\text{Min max} = \frac{(V - \min marks)}{Max\ marks - Min\ marks}(newMax - newMin) + newMin$$

$$Min\ Max = \frac{(8-8)}{20-8} * (1-0) + 0$$

$$Min\ Max = \frac{(0)}{12} * 1$$

$$MinMax = \mathbf{0}$$

For Marks as 10:

$$MinMax = \frac{(10-8)}{20-8} * (1-0) + 0$$

$$MinMax = \frac{(2)}{12} * 1$$

$$MinMax = \mathbf{0.16}$$

For Marks as 15:

$$MinMax = \frac{(15-8)}{20-8} * (1-0) + 0$$

$$MinMax = \frac{(7)}{12} * 1$$

$$MinMax = \mathbf{0.58}$$

For Marks as 20:

$$MinMax = \frac{(20-8)}{20-8} * (1-0) + 0$$

$$MinMax = \frac{(12)}{12} * 1$$

$$MinMax = \mathbf{1}$$

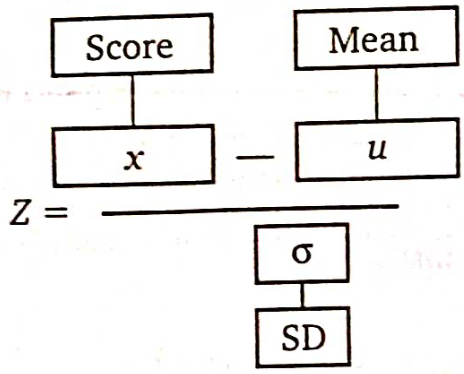
Min Max normalization formula

Marks	Marks after Min-Max normalization
8	0
10	0.16
15	0.58
20	1

प्रश्न 5. Z-Score Normalization से आप क्या समझते हैं?

उत्तर जेड-स्कोर डाटा के सामान्यीकरण में मदद करता है। यदि हम z स्कोर सामान्यीकरण की सहायता से डाटा को सरल रूप में सामान्य करते हैं।

Z-Score formula



How to calculate Z-Score of the following data?

Marks
8
10
15
20

$$S = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}$$

$$\text{Standard deviation} = \sqrt{\frac{\Sigma\,(\textit{every individual value of marks} - \textit{mean of marks})^2}{n}}$$

Mean of marks = 8 + 10 + 15 + 20/4 = 13.25

$$= \sqrt{\frac{(8-13.25)^2 + (10-13.25)^2 + (15-13.25)^2 + (20-13.25)^2}{4}}$$

$$= \sqrt{\frac{(-5.25)^2 + (-3.25)^2 + (1.75)^2 + (6.75)^2}{4}}$$

$$= \sqrt{\frac{27.56 + 10.56 + 3.06 + 45.56}{4}} = \sqrt{\frac{86.74}{4}} = \sqrt{21.6} = 4.6$$

Mean = 13.25

Standard deviation = 4.6

$$\text{ZScore} = \frac{x-\mu}{\sigma} = \frac{8-13.25}{4.6} = -\mathbf{1.14}$$

$$ZScore = \frac{x-\mu}{\sigma} = \frac{10-13.25}{4.6} = -\mathbf{0.7}$$

$$Zscore = \frac{x-\mu}{\sigma} = \frac{15-13.25}{4.6} = \mathbf{0.3}$$

$$ZScore = \frac{x-\mu}{\sigma} = \frac{20-13.25}{4.6} = \mathbf{1.4}$$

Marks	**Marks after Min-Max normalization**
8	–1.14
10	–0.7
15	0.3
20	1.4

प्रश्न 6. Binarization क्या है?

उत्तर Binarization में थ्रेसहोल्ड से ऊपर के सभी मान एक में तब्दील हो जाते हैं और थ्रेसोल्ड के बराबर या उसके नीचे जीरो में तब्दील हो जाते हैं। एक सरल उदाहरण में image ग्रेस्केल 0 – 255 स्पेक्ट्रम से 0 – 1 सेक्टर में बदलना binarization है। Binarization continuous attributes और discrete attributes को binary attributes परिवर्तित करने की प्रक्रिया है।

Simple techniques is :

Assigning numerical value.

Finding number of binary attribute required.

Conversion in to binary.

Say there is an categorical attribute with 'm' number of values.

Assigning numerical values :

Numbers assigned will be between [0, m –1]

For ordinal attribute → assignment follows order.

Finding number of binary attribute required :

Say n be the number of binary atttibutes.

$$n[\log_2 m]$$

Conversion into binary :

Number assigned is converted to its respective binary value.

Ex: if number of binary atttibute is three in numbers, then

2 – 010

if number of binary attribute is four in numbers, then

2 – 0010

Lets us consider an example to learn the process in detail.

{awful, poor, ok, good, great}

	Attribute Values	**Integer Value**
Assigning numerical value	awful	0
	poor	1
	ok	2
	good	3
	great	4

Identifying number of binary attributes : $n = [\log_2 m]$ i.e., $n = [\log_2 5] \approx 3$	Attribute Values	Integer Value	x_1	x_2	x_3
	awful	0			
	poor	1			
	ok	2			
	good	3			
	great	4			

Binary Conversion	Attribute Values	Integer Value	x_1	x_2	x_3
	awful	0	0	0	0
	poor	1	0	0	1
	ok	2	0	1	0
	good	3	0	1	1
	great	4	1	0	0

Complication	Attribute Values	Integer Value	x_1	x_2	x_3	
	awful	0	0	0	0	
	poor	1	0	0	1	
	ok	2	0	1	0	
	good	3	0	1	1	→ **Relationship exists between two attribute**
	great	4	1	0	0	

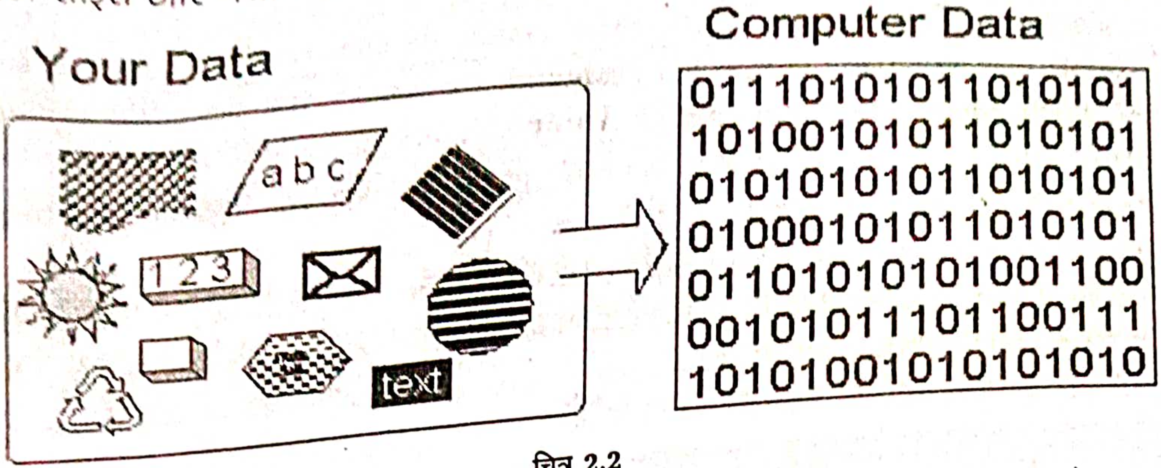
Overcoming the issue :

Number of binary attributes = Number of values

Attribute Values	Integer Value	X_1	X_2	X_3	X_4	X_5
awful	0	1	0	0	0	0
poor	1	0	1	0	0	0
ok	2	0	0	1	0	0
good	3	0	0	0	1	0
great	4	0	0	0	0	1

प्रश्न 7. श्रेणीबद्ध एन्कोडिंग (Categorical Encoding) क्या है?

उत्तर आमतौर पर, किसी भी संरचित डाटासेट में कई कॉलम शामिल होते हैं, संख्यात्मक के साथ-साथ श्रेणीबद्ध चर (categorical variables)। एक मशीन केवल संख्याओं को समझ सकती है। यह text को समझ नहीं सकती है। यह अनिवार्य रूप से मशीन लर्निंग एल्गोरिथ्म के साथ भी है।



चित्र 2.2

यह मुख्य कारण है कि हमें Categorical column को Numerical column में बदलने की आवश्यकता है ताकि एक मशीन लर्निंग एल्गोरिथम इसे समझ सके। इस प्रक्रिया को Categorical एन्कोडिंग कहा जाता है।

Different Approaches to Categorical Encoding Categorical variables को handle करने के निम्न तरीके हैं—

(i) Label Encoding (ii) One-Hot Encoding

प्रश्न 8. Label Encoding से आप क्या समझते हैं?

उत्तर लेबल शब्द या संख्या हो सकते हैं। आमतौर पर, प्रशिक्षण डाटा को पठनीय बनाने के लिए शब्दों के साथ लेबल किया जाता है। लेबल एन्कोडिंग शब्द लेबल को संख्याओं में परिवर्तित करता है।

Example

A → 0
B → 1
C → 2
D → 3
E → 4

Country	Age	Salary
India	44	72000
US	34	65000
Japan	46	98000
US	35	45000
Japan	23	34000

Country	Age	Salary
0	44	72000
2	34	65000
1	46	98000
2	35	45000
1	23	34000

As you can see here, label encoding uses alphabetical ordering. Hence, India has been encoded with 0, the US with 2, and Japan with 1.

Challenges with Label Encoding In the above scenario, the Country names do not have an order or rank. But, when label encoding is performed, the country that the model captures the relationship between countries such as India < Japan < the US.

This is someting that we do not want! So how can we overcome this obstacle? Here comes the concept of **One-Hot Encoding.**

वन-हॉट एन्कोडिंग श्रेणीबद्ध चर के इलाज के लिए एक और लोकप्रिय तकनीक है। यह बस श्रेणीबद्ध विशेषता में अद्वितीय मूल्यों की संख्या के आधार पर अतिरिक्त सुविधाएँ बनाता है। प्रत्येक श्रेणी में अद्वितीय मूल्य को एक सुविधा के रूप में जोड़ा जाएगा।

''वन-हॉट एन्कोडिंग डमी चर बनाने की प्रक्रिया है।''

0	1	2	Age	Salary
1	0	0	44	72000
0	0	1	34	65000
0	1	0	46	98000
0	0	1	35	45000
0	1	0	23	34000

Here, 3 new features are added as the country contains 3 unique values – India, Japan, and the US. In this technique, we solved the problem of ranking as each category is represented by a binary vector.

प्रश्न 9. डाटा एनालिसिस क्या है?

उत्तर **डाटा एनालिसिस** डाटा एनालिसिस एक ऐसी प्रक्रिया है, जिसके द्वारा Raw और Unstructured डाटा में उपयोगी जानकारियाँ निकाली जाती हैं, ताकि निकाली गई इन जानकारियों के आधार पर प्रभावी निर्णय लिए जा सकें।

Data Analysis प्रक्रिया में डाटा का निरीक्षण उसकी प्रोसेसिंग, क्लीनिंग, ट्रांसफॉर्मिंग और मॉडलिंग शामिल होती है। इन सभी प्रक्रियाओं का इस्तेमाल कर Raw डाटा में अपने काम की जानकारी निकाल ली जाती है, ताकि इस जानकारी के आधार पर बेहतर निर्णय लिए जा सकें।

प्रश्न 10. डाटा एनालिसिस क्यों किया जाता है?

उत्तर डाटा एनालिसिस करने के पीछे का कारण सीधे तौर पर उपयोगी जानकारियाँ जुटाना है, ताकि जुटाई गई जानकारी के अनुसार आगे की प्रभावी रणनीति तैयार की जा सके और उपयुक्त कदम उठाए जा सके।

डाटा एनालिसिस आज किसी भी बिजनेस का एक प्रमुख हिस्सा है, जहाँ पर बिजनेस के पिछले सारे डाटा का डाटा एनालिसिस या डाटा एनालिस्ट के द्वारा विश्लेषण किया जाता है और एनालिसिस की एक तय प्रक्रिया के द्वारा डाटा को प्रोसेस किया जाता है, जिसके बाद बिजनेस से जुड़ी उपयोगी जानकारियाँ प्राप्त होती हैं, जिन्हें Processed Data भी कहा जाता है।

इन उपयोगी जानकारियों से बिजनेस की पिछली पूरी रिपोर्ट मिल जाती है। उदाहरण के तौर पर जैसे बिजेनस को डेवलप करने के लिए क्या कदम उठाए गए थे, उनसे बिजनेस को क्या फायदा या नुकसान हुआ इत्यादि और फिर इन्हीं आँकड़ों के अनुसार आगे के लिए उचित निर्णय लिए जाते हैं।

प्रश्न 11. Data Analysis की प्रक्रिया क्या है?

उत्तर **डाटा एनालिसिस की प्रक्रिया** Method of Data Analysis किसी Raw डाटा में से उपयोगी जानकारी निकालने के लिए Data Analysis Process का इस्तेमाल किया जाता है। यह एक प्रक्रिया है जिसका पालन करने के बाद ही एनालिसिस पूरी हो पाती है और उपयोगी जानकारी सामने निकल कर आती है।

डाटा एनालिसिस प्रोसस के अंतर्गत अग्रलिखित चरण शामिल है—

Data Requirement यह डाटा एनालिसिस का सबसे पहला और मुख्य चरण है, जिसमें आपको अपनी जरुरत को समझना होता है, अर्थात् किस प्रकार का डाटा एनालिसिस आप चाहते हैं और उससे क्या परिणाम की इच्छा रखते हैं। इस चरण का उद्देश्य आपकी डाटा एनालिसिस की जरुरत को समझना होता है, जैसे क्या, कैसे और क्यों ताकि स्पष्ट रहे।

Data Collection पहले चरण के बाद आपके सामने स्पष्टता आ जाएगी और इसका अगला चरण है—डाटा कलेक्शन का। यह एक महत्त्वपूर्ण चरण होता है क्योंकि इसमें सही डाटा श्रोतों के चुनाव पर ही एनालिसिस का परिणाम निर्भर करता है। डाटा कलेक्शन में सबसे पहले Internal Sources से डाटा जुटाया जाता है।

प्रश्न 12. डाटा एनालाइसिस के उपयोग लिखिए।

उत्तर **डाटा एनालाइसिस के उपयोग** Application of Data Analysis वास्तविक जीवन में डेटा विश्लेषण का उपयोग निम्नलिखित क्षेत्रों में किया जाता है—

(i) **Business organization** व्यावसायिक संगठनों में बेहतर व्यावसायिक निर्णय लेने के लिए, मार्केट संबंधी अनुसंधान के लिए, उत्पाद संबंधी अनुसंधान के लिए, ग्राहकों की पसंद एवं नापसंद को समझने के लिए, विशेष समय या मौकों पर पैदा होने वाले माँग का पता लगाने के लिए, employees के प्रदर्शन के इतिहास को देखने के लिए इत्यादि कार्यों में डेटा एनालिसिस का उपयोग किया जाता है।

(ii) **Security** पुलिस तथा विभिन्न प्रकार के सरकारी संगठनों द्वारा डाटा एनालिसिस के तकनीक का उपयोग विभिन्न प्रकार के अपराधिक प्रवृत्ति के लोगों का पता लगाने एवं अपराधियों को ढूँढ़ने के उद्देश्य से किया जाता है।

(iii) **Transportation** Data analytics का उपयोग करके परिवहन के क्षेत्र में क्रांति लाई जा सकती है। आजकल शहरों में ट्रफिक जाम की समस्या बहुत आम हो गई है। लोगों को ट्रैफिक जाम के कारण अपना मूल्यवान समय सड़कों पर व्यर्थ ही गवाना पड़ता है। इससे बचने के लिए डेटा एनालिटिक्स का उपयोग करके एक बेहतर यातायात प्रणाली का उपयोग किया जा सकता है।

(iv) **Fraud and Risk Detection** क्रेडिट कार्ड तथा ऋण देने वाले संगठन जैसेकि बैंक आदि अपने ग्राहकों द्वारा किए जाने वाले धोखाधड़ी से बचने के लिए डाटा एनालिसिस का उपयोग करते हैं। डाटा एनालिसिस का उपयोग करके यह बहुत आसानी से पता लगाया जा सकता है कि कौन-से ग्राहक ऋण लेने के बाद डिफॉल्टर हैं और किस ग्राहक में इतना सामर्थ्य है कि वह ऋण को चुका देगा।

(v) **Cities Planning** शहरों की बढ़ती आबादी के अनुसार आधुनिक ढंग से नए City प्लान करने के लिए भी डाटा एनालिसिस का उपयोग किया जाता है। इसके उपयोग से यातायात की सुविधा, ट्रांसपोर्ट और सुरक्षा संबंधी सुविधाओं को बेहतर तरीके से शहरों में लागू करने में सुविधा होती है।

(vi) **Healthcare** डाटा एनालिसिस ने हेल्थ केयर के क्षेत्र में क्रांति ला दी है। दवा बनाने वाली कम्पनियाँ एवं डॉक्टरों के लिए अपने मरीजों के स्वास्थ्य को ट्रैक करना काफी आसान हो गया है, जिसके कारण मरीजों का ध्यान बेहतर तरीके से रखा जा रहा है तथा दवा बनाने वाले संगठन बेहतर दवाइयों का निर्माण कर पा रहे हैं।

(vii) **Searching** सर्च इंजन जैसेकि Google, Bing, Yahoo आदि उपयोगकर्ता द्वारा खोजी जा रही जानकारियों के उत्तर में बेहतर से बेहतर वेबसाइटों के प्रदर्शित करने के लिए डाटा एनालिटिक्स की तकनीक का उपयोग करते हैं।

(viii) **Digital Advertisement** डिजिटल एडवर्टाइजमेंट एवं मार्केटिंग के क्षेत्र में डाटा एनालिटिक्स का बहुत अधिक उपयोग होता है। मार्केटिंग में विभिन्न प्रकार के ऐड कैंपेन के रिजल्ट को मॉनिटर करने के लिए डाटा एनालिटिक्स एक क्रांतिकारी टूल की तरह मदद करता है।

इन सबके अलावा भी डाटा एनालिटिक्स के कुछ और व्यापारिक अनुप्रयोग हैं; जैसेकि—Airline Route Planning, Price Comparison, Gaming, Speech Recognition इत्यादि।

प्रश्न 13. डाटा एनालाइसिस के प्रकारों का विस्तार से उल्लेख कीजिए।

उत्तर **डाटा एनालाइसिस के प्रकार** Types of Data Analysis डाटा एनालिसिस के कुछ प्रमुख Techniques और Methods निम्नलिखित रूप से हैं—

1. **डीस्क्रिप्टिव एनालिसिस** Descriptive Analysis यह डाटा एनालिसिस का सबसे साधारण रूप है, जिसका उपयोग सबसे अधिक होता है। इसमें पिछले डाटा रिकॉर्ड को संक्षिप्त रूप में प्रस्तुत करता है, जिसके कारण डाटा को आसानी से समझा जा सकता है। साधारण शब्दों में कहे तो डीस्क्रिप्टिव एनालिसिस यह बताता है कि डाटा में क्या हुआ है (what happened)?

   आमतौर पर descriptive analysis का उपयोग Key Performance Indicators (KPI) निकालने के लिए किया जाता है। KPI की मदद से यह पता चलता है कि किसी चुने हुए बेंचमार्क के आधार पर कोई व्यवसाय कैसा प्रदर्शन कर रहा है।

2. **डायग्नोस्टिक एनालिसिस** Diagnostice Analysis इसका उपयोग यह समझने के लिए किया जाता है कि ऐसा क्यों हुआ? अर्थात् "Why did it happen?" यहाँ से डाटा का विश्लेषण प्रारंभ होता है। यह डाटा के व्यवहार पैटर्न की पहचान करने में मदद करता है।

3. **प्रीडिक्टिव एनालिसिस** Predictive Analysis यह उपलब्ध डाटा के आधार पर यह बताने का प्रयास करता है कि "क्या होने की संभावना है?" (what is likely to happen?) साधारण शब्दों में कहे तो इसमें पिछले डाटा के आधार पर भविष्य के परिणामों के बारे में भविष्यवाणी की जाती है। यह समझना भी महत्वपूर्ण है कि इसके द्वारा किये गए भविष्यवाणी की सटीकता और गुणवत्ता पूरी तरह से उपलब्ध डाटा पर निर्भर करती है।

4. **प्रीस्क्रिप्टिव एनालिसिस** Prescriptive Analysis यह काफी उपयोगी डाटा एनालिसिस तकनीक है लेकिन इसे करने के लिए काफी अच्छे quality के हार्डवेयर और सॉफ्टवेयर की जरूरत होती है। इसकी मदद से यह पता लगाया जा सकता है कि वर्तमान परिस्थितियों में किसी समस्या के समाधान के लिए संगठन का कौन-सा निर्णय लेना चाहिए।

प्रश्न 14. डाटा एनालाइसिस के लाभ क्या हैं?

उत्तर **डाटा एनालाइसिस के लाभ** Advantages of Data Analysis डाटा एनालिसिस के प्रमुख लाभ निम्नलिखित रूप से हैं—

(i) **Improved Decision Making** यह किसी भी संगठन से सम्बन्धित बेहतर और कारगर फैसले आसानी से लेने में मदद करता है। आँकड़ों एवं तथ्यों के आधार पर लिए गए नियमों के गलत होने की संभावना बहुत कम होती है।

(ii) **More Effective Marketing** यह ग्राहक के रुझान को बेहतर तरीके से समझने में मदद करता है जिससे कंपनियाँ बेहतर Ads Campaign बनाकर अपने ग्राहकों के सामने बेहतर तरीके से जानकारियों को प्रस्तुत कर सकती हैं।

(iii) **Better Customer Service** डाटा एनालिटिक्स का उपयोग करके ग्राहकों को उनकी आवश्यकताओं के अनुरूप सेवा प्रदान किया जा सकता है। यह उनकी संतुष्टि के स्तर को बढ़ाता है और संगठन के प्रति उनके रिश्तों में मजबूती प्रदान करता है।

प्रश्न 15. Data Visulaisation से आप क्या समझते हैं?

उत्तर Data Visulaisation डाटा को Graphical या Chart के रूप में प्रस्तुत करने की तकनीक है जिसे डाटा की Visual Representation कहते हैं।

डाटा के इस Visual Representation से डाटा को समझना, Analyze करना उस पर Research करना और उससे प्रोसेस करना बहुत आसान हो जाता है।

अलग-अलग तरह के डाटा के लिये अलग-अलग तरह की Visualization के तरीके प्रयोग किये जाते हैं। आज डाटा को Graphs या Chart के अलावा और भी कई तरीकों से Present किया जा सकता है; जैसे—Info graphics, dials and gauges, geographic maps, sparklines, heat map, and detailed bar, pie and fever charts आदि।

Data Visulaisation के तीन मुख्य Principle हैं—

1. **Visulaisation** डाटा Visual रूप में प्रस्तुत होना।

2. **Insight** पूरे डाटा को सही तरह से प्रस्तुत करना ताकि इसका कोई गलत अर्थ न निकले।

3. **Sharing** डाटा आसानी से समझ में आ सके और अगर कोई तीसरा व्यक्ति भी इसे देखे तो उसे भी आसानी से समझ सके।

प्रश्न16. डाटा विजुलाइजेशन के प्रकार लिखिए।

उत्तर **डाटा विजुलाइजेशन के प्रकार** डाटा विजुलाइजेशन के प्रकार निम्नलिखित हैं—

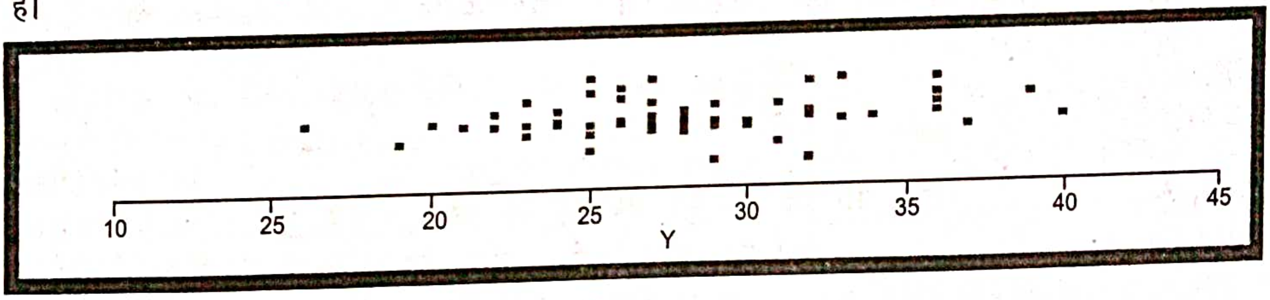
1. Graph
2. Bar Charts
3. Geographic Maps
4. Pie Charts
5. Dials & Gauges
6. Heat Maps
7. Infographic
8. Sparklines
9. Fever Charts

प्रश्न 17. Univariate plot शब्द से क्या समझते हैं?

उत्तर A univariate plot shows the data and summarizes its distribution.

Examples : Dot plot and Box plot

(i) **Dot plot** एक डॉट प्लॉट, जिसे स्ट्रिप प्लॉट के रूप में भी जाना जाता है, individual observations को दर्शाता है।



चित्र 2.3

A dot plot gives an indication of the spread of the data and can highligt clustering or extreme values.

(ii) **Box plot** Box plot का structure काफी साधारण होता है। एक बॉक्स प्लॉट डाटा की पाँच-संख्या का सारांश दिखाता है।

the minimum,
first quartile,
median,
third quartile,
and maximum

Example नीचे दी हुई संख्याओं की length 11 है अर्थात् उनकी length odd है।

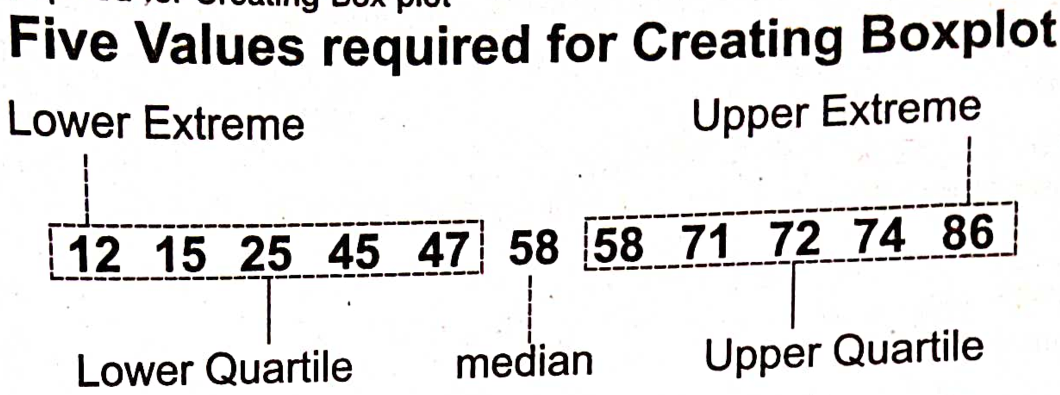
अगर x है, तब c(45,12,47,25,86,15,71, 74,58,72,58)

सबसे पहले इन numbers को ascending order में लगाया जाता है।

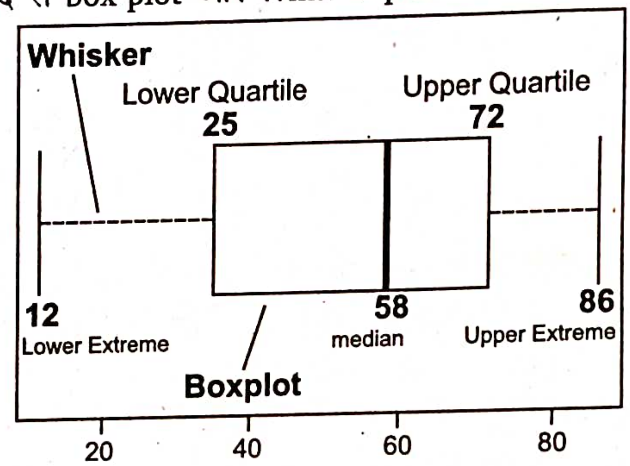
x का ascending Order : 12,15,25,45,47,58,58,71,72,74,86 है।

उसके बाद संख्याओं के Lower Extreme, Lower Quartile, Median, Upper Quartile और Upper Extreme को ढूंढे।

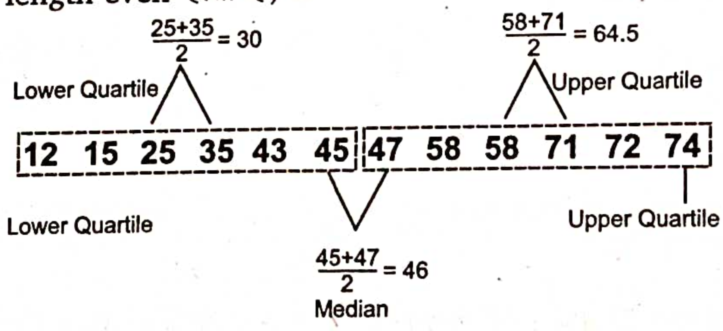
Five Values required for Creating Box plot



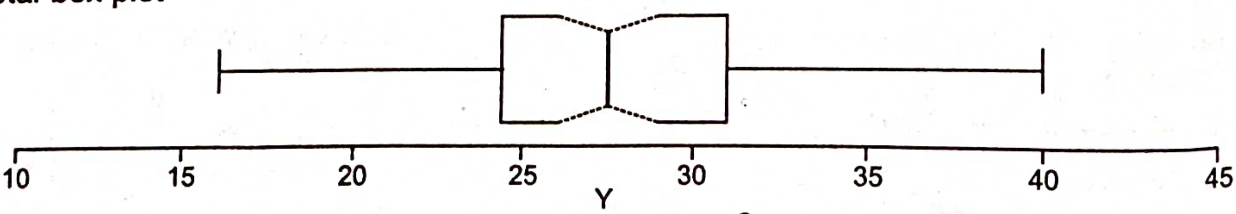
अब इन पाँच values की मदद से box plot और Whisker plot को create किया जाएगा।



अगर दी गई संख्याओं की length even होती है, तो

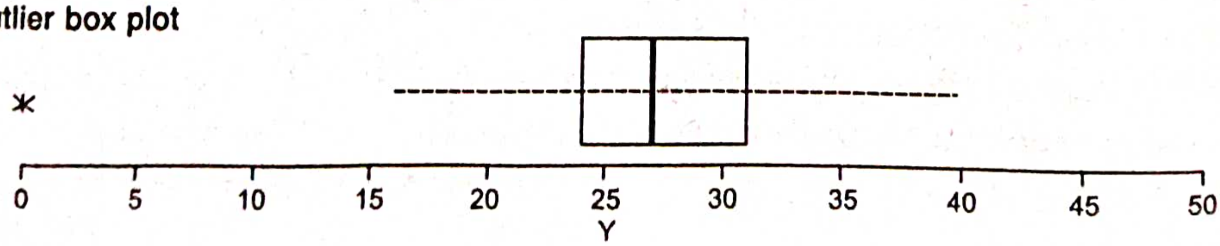


Skeletal box plot



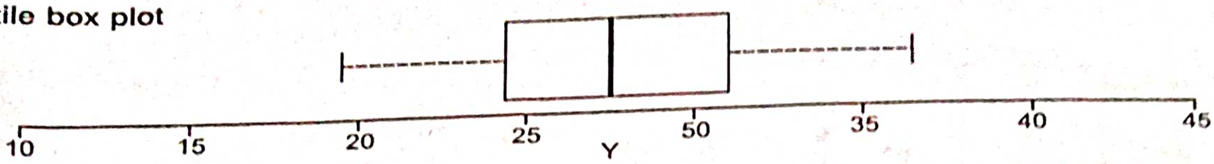
A sketetal box plot shows the median as a line, a box from the 1st to 3rd quartiles and whiskers with end caps extending to the minimum and maximum. Optional notches in the box represent the confidence interval around the median.

Outlier box plot



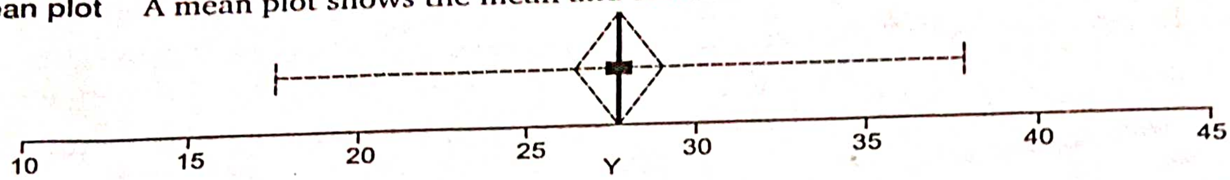
Outlier box plot, skeletal box plot का variation है जो संभावित आउटलेयर्स की पहचान भी करता है।

Quantile box plot



A quantile box plot is a variation on the skeletal box plot and shows the whiskers extending to specific quantiles rather than the minimum and maximum values.

- **Mean plot** A mean plot shows the mean and standard deviation of the data.



A line or dot represent the mean. A standard error or confidence interval measures uncertainty in the mean and is represented as either an error bar or diamond.

An optional error bar or band represents the standard deviation. The standard deviation gives the impression that the data is from a normal distribution centered at the mean value, with most of the data within two standard deviations of the mean. Therefore, the data should be approximately normally distributed. If the distribution is skewed, the plot is likely to mislead.

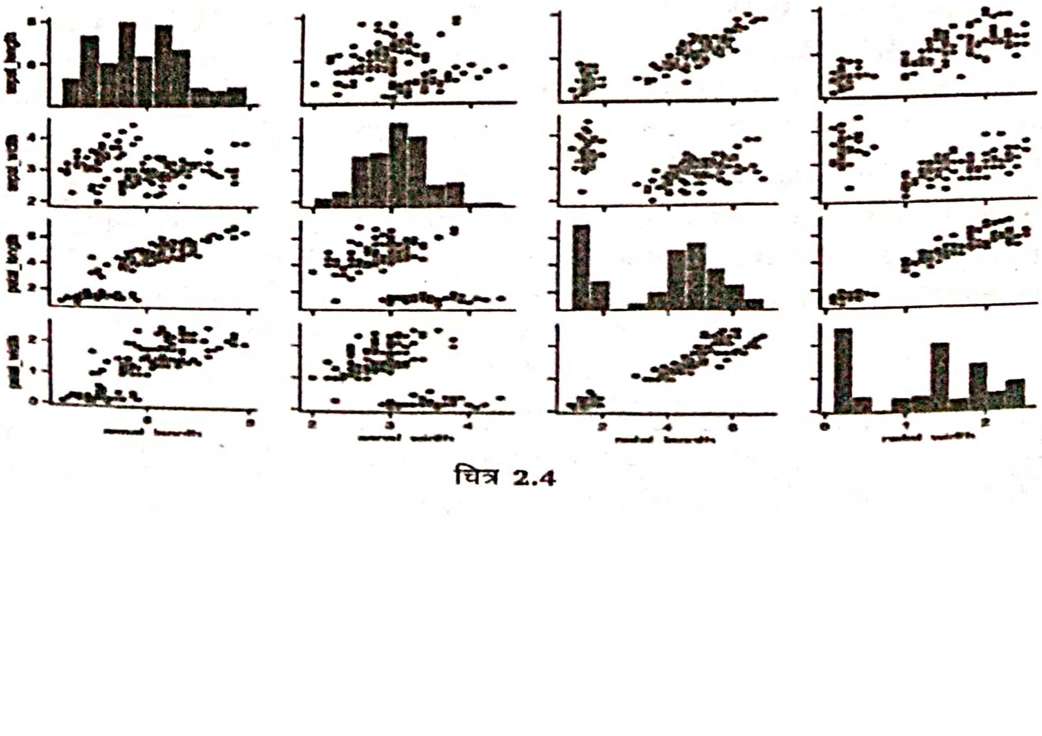
प्रश्न 18. Multivariate plots से क्या समझते हैं?

उत्तर Multivariate descriptive displays or plots are designed to reveal the relationship among several variables simultaneously.

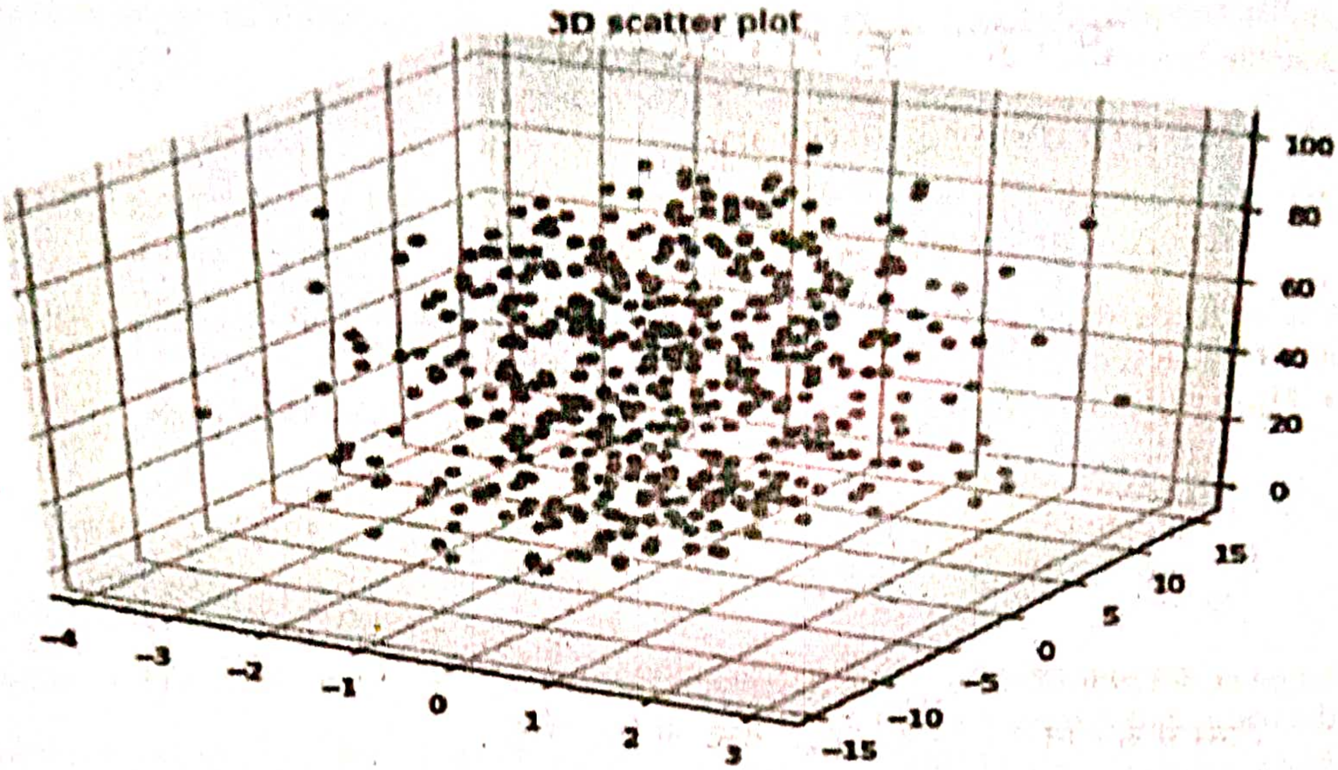
Multivariate analysis is performed to understand interactions between different fields in the dataset (or) finding interactions between variables more than 2. Examples : Pair plot and 3D scatter plot.

Pair plot A pairplot plot a pairwise relationships in a dataset. The pairplot function creates a gird of Axes such that each variable in data will be shared in the y-axis across a single row and in the x-axis across a single column. That creates plots as shown below.

3D scatter plot A 3d Scatter Plot is a mathematical diagram, the most basic varsion of three-dimensional plotting used to display the properties of data as three variables of a dataset using the cartesian coordinates.



चित्र 2.4

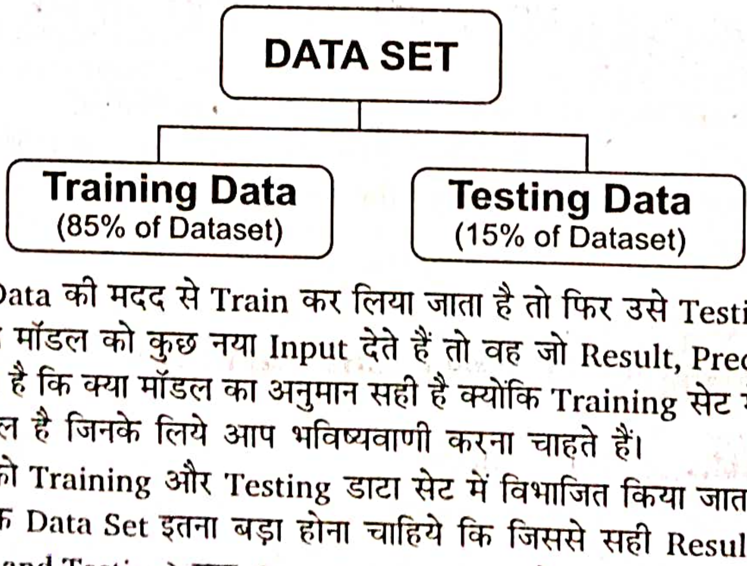


चित्र 2.5

Machine Learning या Data Mining में जब हमें मॉडल को Develop करना होता है तो हमें उसके लिये Data की जरुरत होती है फिर इस Data set को मुख्य रूप से तीन कार्यों के लिये उपयोग किया जाता है—

1. Training Data set 2. Volidation Data set 3. Testing Data set

आमतौर पर, जब किसी डाटा सेट को Training सेट और Testing सेट से अलग करते हैं, तो उसमें से अधिकतर डाटा का उपयोग Training के लिए किया जाता है, और उस डाटा का एक छोटा हिस्सा Testing के लिए उपयोग किया जाता है।



जब मॉडल को Training Data की मदद से Train कर लिया जाता है तो फिर उसे Testing Data की मदद से Test किया जाता है और जब हम मॉडल को कुछ नया Input देते हैं तो वह जो Result, Predict करता है इससे हमें यह निर्धारित करना आसान होता है कि क्या मॉडल का अनुमान सही है क्योंकि Training सेट में मौजूद डाटा में उसके लिए पहले से ही Values शामिल है जिनके लिये आप भविष्यवाणी करना चाहते हैं।

जब किसी एक डाटा सेट को Training और Testing डाटा सेट में विभाजित किया जाता है तो इसके लिए दो चीजों का ध्यान रखना होता है कि Data Set इतना बड़ा होना चाहिये कि जिससे सही Results मिल सके।

दोनों Dataset (Training and Testing) एक Complete Dataset को represent करना चाहिए अर्थात् Testing डाटा के characterstics, Training Dataset से अलग नहीं होना चाहिये।

3 सांख्यिकीय निष्कर्ष
Statistical Inference

बहुविकल्पीय प्रश्न (MCQ)

प्रश्न 1. गणित की कौन-सी शाखा में आँकड़ों का संग्रहण, प्रदर्शन, वर्गीकरण आदि किया जाता है?

(a) अंकगणित (b) बीजगणित

(c) रेखागणित (d) सांख्यिकी

उत्तर (d) सांख्यिकी

प्रश्न 2. सांख्यिकी की कौन-सी श्रेणी को डेटा के संग्रहण अथवा वर्णन के लिए इस्तेमाल किया जाता है?

(a) वर्णात्मक सांख्यिकी (b) अनुमानित सांख्यिकी

(c) (a) व (b) दोनों (d) इनमें से कोई नहीं

उत्तर (a) वर्णात्मक सांख्यिकी

प्रश्न 3. निम्न में से कौन-सा नॉन पैरामीट्रिक टेस्ट का उदाहरण है?

(a) Logistic Regression (b) Native Bayes Model

(c) Decision tree Model (d) इनमें से कोई नहीं

उत्तर (c) Decision tree Model

प्रश्न 4. मशीन लर्निंग में डिस्टेंस मेट्रिक्स कितने प्रकार की होती है?

(a) दो (b) तीन

(c) चार (d) पाँच

उत्तर (c) चार

प्रश्न 5. Analysis of variance (ANOVA) सांख्यिकीय निष्कर्ष किनके बीच के अन्तर का पता लगाने के लिए उपयोग किया जाता है?

(a) प्रसरण (Variances)

(b) माध्यम (Means)

(c) अनुपात (Proportions)

(d) Only two parameters

उत्तर (c) अनुपात (Proportions)

प्रश्न 6. निम्न में से कौन-सा सांख्यिकीय निष्कर्ष (statistical inference) का मुख्य घटक है?

(a) Sample size

(b) Size of the observed differences

(c) Variability in the sample

(d) उपर्युक्त सभी

उत्तर (d) उपर्युक्त सभी

प्रश्न 7. जनसंख्या पैरामीटर के बारे में कौन-से परीक्षण से जानकारी होती है?

(a) पैरामीट्रिक परीक्षण (b) नॉनपैरामीट्रिक परीक्षण

(c) (a) व (b) दोनों (d) इनमें से कोई नहीं

उत्तर (a) पैरामीट्रिक परीक्षण

प्रश्न 8. निम्न में से कौन-सा सांख्यिकीय निष्कर्ष है?

(a) Pearson correlation (b) Chi-square statistics

(c) One sample hypothesis testing (d) ये सभी

उत्तर (d) ये सभी

खण्ड 'अ' : अतिलघु उत्तरीय प्रश्न

प्रश्न 1. Statistical से आप क्या समझते हैं?

उत्तर सांख्यिकी गणित की वह शाखा है जिसमें आँकड़ों का संग्रहण, प्रदर्शन, वर्गीकरण और उसके गुणों के आकलन का अध्ययन किया जाता है।

प्रश्न 2. Statistical inference क्या होता है?

उत्तर किसी आँकड़े का सांख्यिकीय विश्लेषण करके उसमें छिपे हुए 'गुण' को प्रकट करना सांख्यिकीय निष्कर्ष (Statistical inference) कहलाता है।

प्रश्न 3. पैरामीट्रिक और नॉनपैरामीट्रिक टेस्ट क्या है?

उत्तर पैरामीट्रिक परीक्षण वह है जिसमें जनसंख्या पैरामीटर के बारे में जानकारी होती है। दूसरी ओर, नॉनपरमेट्रिक टेस्ट वह है जहाँ शोधकर्ता को जनसंख्या पैरामीटर के बारे में कोई पता नहीं है।

प्रश्न 4. Population को परिभाषित कीजिए।

उत्तर Population एक entire group है जिसके बारे में आप निष्कर्ष (conclusions) निकालना चाहते हैं।

प्रश्न 5. Sample (नमूना) को समझाइए।

उत्तर Sample एक विशिष्ट समूह है जिससे आप डेटा एकत्र करेंगे। Sample का size हमेशा population के कुल size से कम होता है।

प्रश्न 6. Statistical Inference की उपयोग बताइए।

उत्तर Hypothesis testing और confidence intervals, statistical inference के अनुप्रयोग (applications) हैं।

प्रश्न 7. Probability Distribution क्या है?

उत्तर संभाव्यता वितरण (probability distribution) यादृच्छिक प्रयोग (random experiment) के प्रत्येक परिणाम की संभावना देता है। यह विभिन्न संभावित घटनाओं की संभावनाएँ प्रदान करता है।

खण्ड 'ब' : लघु एवं दीर्घ उत्तरीय प्रश्न

प्रश्न 1. सांख्यिकी क्या है? सांख्यिकी के प्रकार भी बताइए।

उत्तर सांख्यिकी एक गणितीय विज्ञान है जिसमें किसी वस्तु/अवयव/तंत्र/समुदाय से सम्बन्धित आँकड़ों का संग्रह, विश्लेषण, व्याख्या या स्पष्टीकरण और प्रस्तुति की जाती है। यह विभिन्न क्षेत्रों में लागू है—अकादमिक अनुशासन (academic disciplines), इससे प्राकृतिक विज्ञान, सामाजिक विज्ञान, मानविकी, सरकार और व्यापार आदि।

सांख्यिकों को दो अलग-अलग श्रेणियों में वर्गीकृत किया जा सकता है—

(i) Descriptive Statistics (ii) Inferential Statistics

सांख्यिकी में, वर्णनात्मक आँकड़े (Descriptive Statistics) डेटा का वर्णन करते हैं, जबकि ह्रासमान आँकड़े (Inferential Statitics) आपको डेटा से भविष्यवाणियाँ करने में मदद करते हैं।

सांख्यिकीय तरीकों को डेटा के संग्रह के संग्रहण अथवा वर्णन के लिए इस्तेमाल किया जा सकता है। इसे **वर्णनात्मक सांख्यिकी** (descriptive statistics) कहा जाता है। इसके अतिरिक्त, डेटा में पैटर्न को इस तरह से मॉडल किया जा सकता है कि वह निष्कर्षों की यादृच्छिकता और अनिश्चितता का कारण बने और फिर इस प्रक्रिया को उस विधि, या

जिस जनसंख्या का अध्ययन किया जा रहा हो, उसके बारे में अनुमान लगाने के लिए किया जाता है। इसे **अनुमानित** सांख्यिकी (inferential statistics) कहा जाता है। वर्णनात्मक तथा अनुमानित सांख्यिकी, दोनों में **व्यावहारिक सांख्यिकी** सम्मिलित है।

प्रश्न 2. सांख्यिकीय अनुमान (Statistical Inference) को परिभाषित कीजिए।

उत्तर **सांख्यिकीय अनुमान** Statistical Inference Statistical inference, परिणाम का विश्लेषण करने और डाटा विषय से याछच्छिक भिन्नता के लिए निष्कर्ष बनाने की प्रक्रिया है। इसे हीन सांख्यिकी (inferential statistics) भी कहा जाता है। सांख्यिकीय निष्कर्ष में, जनसंख्या का एक उपसमुच्चय (एक सांख्यिकीय नमूना) चुना जाता है, जो सांख्यिकीय विश्लेषण में, जनसंख्या का प्रतिनिधित्व करता है। यदि कोई नमूना उचित रूप से चुना गया तो, उस नमूने की तत्संबंधी विशेषताओं से, पूरी जनसंख्या की विशेषताएँ, जिससे वह नमूना निकाला गया हो, अनुमानित की जा सकती है। Hypothesis testing और confidence intervals, statistical inference के अनुप्रयोग (applications) हैं।

प्रश्न 3. Statistical Inference के मुख्य घटक बताइए।

उत्तर The components (घटक) used for making statistical inference (सांख्यिकीय निष्कर्ष) are :

(i) Sample Size
(ii) Variability in the sample
(iii) Size of the observed differences

प्रश्न 4. सांख्यिकीय निष्कर्ष (Statistical Inference) के प्रकार क्या हैं?

उत्तर **सांख्यिकीय निष्कर्ष के प्रकार** Types of Statistical Inference There are different types of statistical inferences that are extensively used for making conclusion. They are :

(i) One sample hypothesis testing
(ii) Confidence Interval
(iii) Pearson Correlation (सहसंबंध)
(iv) Bi-variate regression
(v) Muilti-variate regression
(vi) Chi-square statistics (चाई-वर्ग परीक्षण) and contingency table
(vii) (Analysis of variance) ANOVA or T-test (भिन्नता का विश्लेषण)

प्रश्न 5. Statistical Inference का महत्त्व समझाइए?

उत्तर **सांख्यिकीय अनुमान का महत्व** Importance of Statistical Inference डाटा की सही तरीके से जाँच करने के लिए इंफेरेंशियल स्टेटिस्टिक्स महत्त्वपूर्ण है। एक सटीक निष्कर्ष बनाने के लिए, अनुसंधान परिणामों की व्याख्या करने के लिए उचित डाटा विश्लेषण महत्त्वपूर्ण है।

विभिन्न क्षेत्रों में विभिन्न टिप्पणियों के लिए भविष्य की भविष्यवाणी में इसका प्रमुख रूप से उपयोग किया जाता है। यह हमें डाटा के बारे में अनुमान लगाने में मदद करता है। सांख्यिकीय निष्कर्ष में विभिन्न क्षेत्रों में आवेदन की एक विस्तृत शृंखला है, जैसे—

(i) व्यापार विश्लेषण (Business Analysis)
(ii) कृत्रिम बुद्धिमत्ता (Artificial Intelligence)
(iii) वित्तीय विश्लेषण (Financial Analysis)
(iv) धोखाधड़ी का पता लगाना (Fraud Detection)
(v) मशीन लर्निंग (Machine Learning)
(vi) शेयर बाजार (Share Market)
(vii) फार्मास्युटिकल सेक्टर (Pharmaceutical Sector)

प्रश्न 6. Statistical Inference के उदाहरण दीजिए।

उत्तर An example of statistical inference is given below.

From the shuffled pack of cards, a card is drawn. This trial is repeated for 400 times and the suits are given below :

Suit	Spade	Clubs	Hearts	Diamonds
No. of times drawn	90	100	120	90

While a card is tried at random, then what is the probability of getting a :

1. Diamond cards
2. Black cards
3. Except for spade

Solution :

By statistical inference solution,

Total number of events = 400

i.e., 90 + 100 + 120 + 90 = 400

1. **The probability of getting diamond cards :**
   Number of trials in which diamond card is drawn = 90
   Therefore, P(diamond card) = 90/400 = 0.255
2. **The probability of getting black cards :**
   Number of trials in which black ard showed up = 90 + 100 = 190
   Therefore, P(black card) = 190/400 = 0.475
3. **Except for spade**
   Number of trials other than spade showed up = 90 + 100 + 120 = 310
   Therefore, P(except spade) = 310/400 = 0.775

प्रश्न 7. प्रायिकता वितरण (probability distribution) का एक उदाहरण क्या है?

उत्तर If two coins are tossed, then the probability of getting 0 heads is ¼, 1 head will be ½ and both head will be ¼. So, the probability P(x) for a random experiment or discrete random variable x, is distributed as :

P(0) = ¼

P(1) = ½

P(2) = 1/4

P(0) = ¼ + ½ + ¼ = 1

प्रश्न 8. A coin is tossed twice. X is the random variable of the number of heads obtained. What is probability distribution of x?

उत्तर First write, the value of X = 0, 1 and 2 as the possiblity are there that

No head comes

One head and one tail comes

And head comes in both the coins

Now the probability distribution could be written as ;

P(X=0) = P(Tail+Tail) = ½ * ½ = ¼

P(X=1) = P(Head+Tail) or P(Tail+Head) = $\frac{1}{2} * \frac{1}{2} + \frac{1}{2} * \frac{1}{2} = \frac{1}{2}$

P(X=2) = P(Head+Head) = $\frac{1}{2} * \frac{1}{2} = \frac{1}{4}$

We can put these values in tabular form;

X	0	1	2
P(X)	1/4	1/2	1/4

प्रश्न 9. The weight of a pot of water chosen is a continuous random variable. The following table gives the weight in kg of 100 containers recently filled by the water purifier. It records the observed values of the continuous random variable and their corresponding frequencies. Find the probability of changes for each weight category.

Weight W	Number of Containers
0.900 – 0.925	1
0.925 – 0.950	7
0.950 – 0.975	25
0.975 – 1.000	32
1.000 – 1.025	30
1.025 – 1.050	5
Total	100

उत्तर We first divide the number of containers in each weight category by 100 to give the probabilities.

Weight W	Number of Containers	Probability
0.900 – 0.925	1	0.01
0.925 – 0.950	7	0.07
0.950 – 0.975	25	0.25
0.975 – 1.000	32	0.32
1.000 – 1.025	30	0.30
1.025 – 1.050	5	0.05
Total	100	1.00

प्रश्न 10. पैरामीट्रिक सांख्यिकी क्या है?

उत्तर पैरामीट्रिक आँकड़े वे आँकड़े हैं जिनमें डाटा/नमूनों को सामान्य वितरण से लिया गया माना जाता है। पैरामीट्रिक आँकड़ों की परिभाषा है "आँकड़े जो मानते हैं कि डाटा एक प्रकार की संभाव्यता वितरण से आया है और वितरण के मापदंडों के बारे में अनुमान लगाता है।" अधिकांश ज्ञात प्रारंभिक सांख्यिकीय विधियाँ इसी समूह की हैं। वास्तव में, वे आमतौर पर नहीं किए जा सकते हैं। इसलिए, यह आँकड़े अधिक मान्यताओं पर आधारित हैं। यदि डाटा/नमूने सामान्य रूप से वितरित किए जाते हैं या लगभग-सामान्य से वितरित किए जाते हैं, तो सूत्र सटीक परिणाम और अनुमान लगा सकते हैं। हालाँकि, अगर सामान्य रूप से वितरित होने की धारणा गलत है, तो पैरामीट्रिक आँकड़े काफी भ्रामक हो सकते हैं।

प्रश्न 11. गैर-पैरामीट्रिक सांख्यिकी क्या है?

उत्तर गैर-पैरामीट्रिक आँकड़े वितरण-मुक्त आँकड़ों के रूप में भी जाने जाते हैं। इस सांख्यिकीय प्रकार का लाभ यह है कि इसे एक धारणा बनाने की जरूरत नहीं है जैसाकि पहले पैरामीट्रिक्स के साथ बनाया गया था। गैर-पैरामीट्रिक सांख्यिकीय गणना साधनों की तुलना में ध्यान करने के लिए मध्यस्थों को लेते हैं। इसलिए, यदि एक या दो माध्य मान से

विचलन होता है, तो उनका प्रभाव उपेक्षित होता है। आमतौर पर पैरामीट्रिक आँकड़े इससे अधिक पंसद किए जाते हैं क्योंकि इसमें एक अपरंपरागत विधि की तुलना में झूठी परिकल्पना को अस्वीकार करने की अधिक शक्ति है। सबसे ज्ञात गैर-पैरामीट्रिक परीक्षणों में से एक ची-स्क्वायर परीक्षण है। कुछ पैरामीट्रिक परीक्षणों जैसेकि पैरामेड नमूना टी-टेस्ट के लिए विलकॉक्सन टी टेस्ट, मान-व्हिटनी यू टेस्ट के लिए स्वतंत्र नमूने टी-टेस्ट, स्पीयरमैन के सहसंबंध के लिए पीयरसन के सहसंबंध आदि के लिए गैर-समरूप एनालॉग है। एक नमूना टी-टेस्ट के लिए, कोई भी नहीं है।

प्रश्न 12. पैरामीट्रिक और नॉन-पैरामीट्रिक टेस्ट के बीच अंतर बताइए।

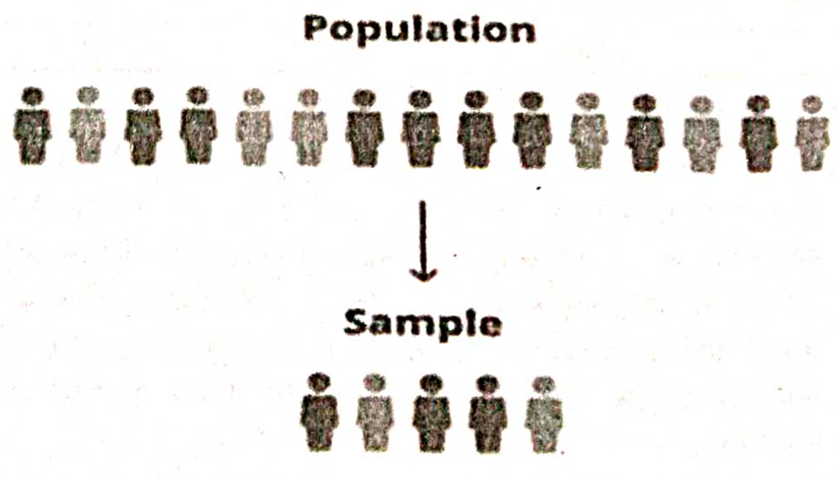
उत्तर पैरामीट्रिक और गैरपारंपरिक परीक्षण के बीच बुनियादी अंतर निम्नलिखित बिंदुओं पर चर्चा करते हैं—

1. एक सांख्यिकी परीक्षण, जिसमें जनसंख्या पैरामीटर के बारे में विशिष्ट धारणा बनाई जाती है, पैरामीट्रिक परीक्षण के इस रूप में जाना जाता है। गैर-मीट्रिक स्वतंत्र चर के मामले में उपयोग किए जाने वाले सांख्यिकीय परीक्षण को गैर-पैमाना परीक्षण कहा जाता है।
2. पैरामीट्रिक परीक्षण में, परीक्षण सांख्यिकीय वितरण पर आधारित है। दूसरी ओर गैर-सममितीय परीक्षण के मामले में परीक्षण आँकड़ा मनमाना है।
3. पैरामीट्रिक परीक्षण में, यह माना जाता है कि ब्याज के चर का मापन अंतराल या अनुपात स्तर किया जाता है। नॉनपैरामीट्रिक परीक्षण के विपरीत, जिसमें नाममात्र या क्रमिक पैमाने पर ब्याज के चर को मापा जाता है।
4. सामान्य तौर पर, पैरामीट्रिक परीक्षण में केंद्रीय प्रवृत्ति का माप माध्य है, जबकि नॉनपैरामीट्रिक टेस्ट के मामले में औसत दर्जे का है।
5. पैरामीट्रिक परीक्षण में, जनसंख्या के बारे में पूरी जानकारी है। इसके विपरीत, नॉनपैरामीट्रिक टेस्ट में, जनसंख्या के बारे में कोई जानकारी नहीं हैं
6. पैरामीट्रिक परीक्षण की प्रयोज्यता केवल चरों के लिए होती है, जबकि अप्रस्तुत परीक्षण चर और विशेषताओं दोनों पर लागू होता है।
7. दो मात्रात्मक चर के बीच संबंध की डिग्री को मापने के लिए, पियर्सन के सहसंबंध के गुणन का उपयोग पैरामीट्रिक परीक्षण में किया जाता है, जबकि स्पीयरमैन के रैंक सहसंबंध का उपयोग नॉनपैरामीट्रिक परीक्षण में किया जाता है।
8. पैरामीटिक परीक्षण के Examples : Logistic Regression, naive Bayes Model, etc.
   नॉनपैरामीट्रिक टेस्ट के Examples : KNN, Decison Tree Model, etc.

प्रश्न 13. Population and Sample की व्याख्या कीजिए।

उत्तर Population एक entire group है जिसके बारे में आप निष्कर्ष (conclusions) निकालना चाहते हैं। आँकड़ों में, जनसंख्या का मतलब जनसंख्या में सभी तत्वों के औसत के रूप में परिभाषित किया गया है। यह समूह की विशेषता का एक मतलब है, जहाँ समूह वस्तुओं, व्यक्तियों आदि जैसे तत्वों के तत्वों को संदर्भित करता है और विशेषता ब्याज की वस्तु है। चूँकि जनसंख्या बहुत बड़ी है और ज्ञात नहीं है, जनसंख्या का मतलब अज्ञात स्थिर है।

Sample एक विशिष्ट समूह है जिससे आप डाटा एकत्र करेंगे। Sample का size हमेशा population के कुल size से कम होता है।



चित्र 3.1

शोध में, a population doesn't always refer to people. It can mean a group containing elements of anything you want to study, जैसे कि objects, events, organizations, countries, species,organisms इत्यादि।

प्रश्न 14. Distance metrics का क्या अर्थ है?

उत्तर डिस्टेंस मेट्रिक्स कई मशीन लर्निंग एल्गोरिथम का एक महत्त्वपूर्ण हिस्सा है। इन डिस्टेंस मेट्रिक्स का उपयोग पर्यवेक्षित (supervised) और अनुपयोगी (unsupervised) शिक्षा दोनों में किया जाता है, आमतौर पर डाटा बिंदुओं के बीच समानता की गणना करने के लिए।

एक effective distance metric हमारे मीशन लर्निंग मॉडल के प्रदर्शन को बेहतर बनाता है, चाहे वह वर्गीकरण कार्यों के लिए हो या क्लस्टरिंग के लिए।

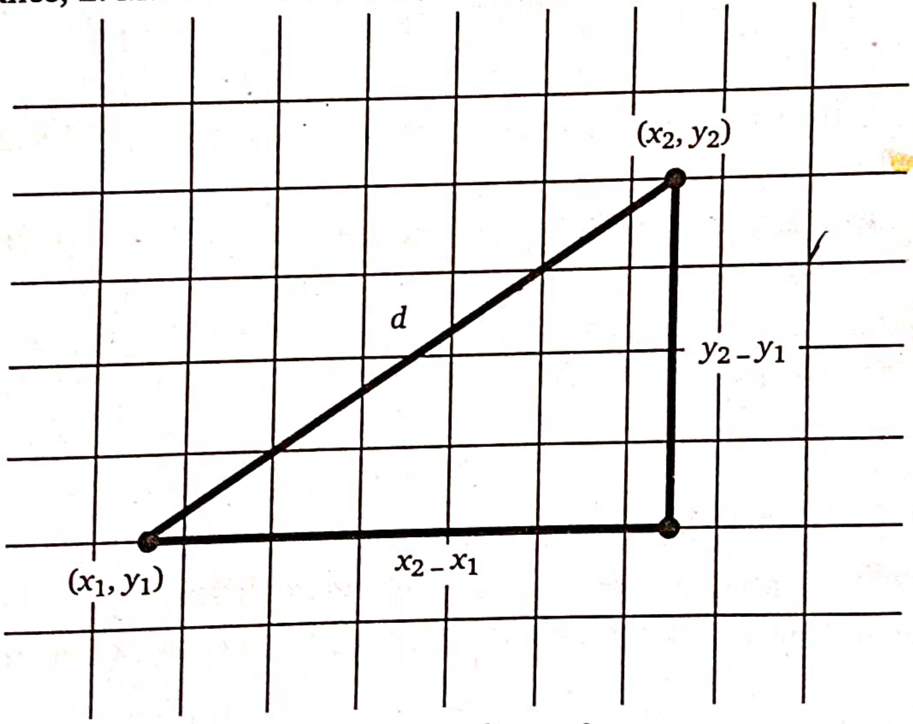
प्रश्न 15. मशीन लर्निंग में डिस्टेंस मेट्रिक्स कितने प्रकार की होती है?

उत्तर **मशीन लर्निंग में डिस्टेंस मेट्रिक्स का मापन** Measures of Distance Metrics in Mechine Learning Clustering एक डाटा analysis टूल है जिसमें डाटा तथा ऑब्जेक्ट्स को इस प्रकार अलग-अलग समूहों (clusters) में divide किया जाता है कि जो समान गुणों वाले ऑब्जेक्ट्स होते हैं उन्हें एक समूह (cluster) में रखा जाता है और भिन्न गुणों वाले ऑब्जेक्ट्स को दूसरे cluster में रखा जाता है। प्रत्येक cluster के ऑब्जेक्ट्स दूसरे cluster के ऑब्जेक्ट्स से भिन्न होते हैं।

एक cluster में जितने भी ऑब्जेक्ट्स होते हैं उन्हें एक समूह के रूप में treat किया जाता है।

अधिकांश clustering approaches objects के pair के बीच समानता या अंतर का आकलन करने के लिए distance metrices का उपयोग करते हैं। most pouular 4 types of distance Metrices in machine Learning निम्न हैं—

1. Euclidean Distance, 2. Manhattan Distance, 3. Minkowski Distance, 4. Hamming Distance



चित्र 3.2 यूक्लिडियन दूरी

प्रश्न 16. डाटा माइनिंग में यूक्लिडियन दूरी (Euclidean Distance) की गणना कीजिए।

उत्तर **यूक्लिडियन दूरी** Euclidean Distance Euclidean distance is considered the traditional metric for problems with geometry. It can be simple expalined as the ordinary distance between two points. यह cluster analysis में सबसे अधिक उपयोग किए जाने वाले एल्गोरिथम में से एक है। Data mining में इस formula का उपयोग करने वाला एल्गोरिथम K-mean है। Mathematically it computes the root of squared differences between the coordinates between two objects.

$$d(p, q) = d(q, p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \ldots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

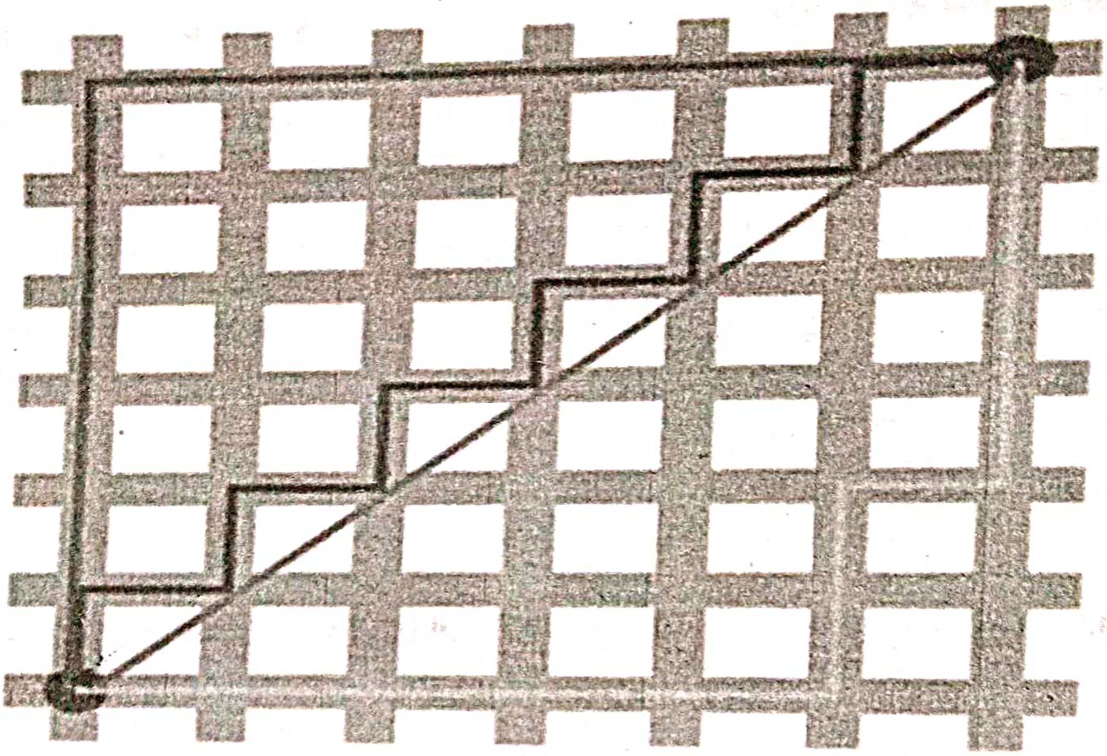
प्रश्न 17. डाटा माइनिंग में मैनहट्टन दूरी Manhattan Distance की गणना कीजिए।

उत्तर **मैनहट्टन दूरी** Manhattan Distance This determines the absolute difference among the pair of the coordinates.

Suppose we have two points P and Q to determine the distance between these points we simply have to calculate the perpendicular distance of the points from X-Axis and Y-Axis.

In a plane with P at coordinate (x_1, y_1) and Q at (x_2, y_2).

Manhattan distance between P and $Q = |x_1 - x_2| + |y_1 - y_2|$.



चित्र 3.3

यहाँ रेड लाइन की कुल दूरी दोनों बिंदुओं के बीच मैनहट्टन की दूरी होती है।

प्रश्न 18. डाटा माइनिंग में Minkowski Distance की गणना कीजिए।

उत्तर Minkowski distance यह यूक्लिडियन और मैनहट्टन Distance Measure का generalized form है।

In an N-dimensional space, a point is represented as, $(x_1, x_2, .., x_N)$

Consider two points P_1 and P_2.

$P_1 : (X_1, X_2, ..., X_N)$

$P_2 : (Y_1, Y_2, ..., Y_N)$

Then, the Minkowski distance between P_1 and P_2 is given as :

$$\sqrt[P]{(x_1 - y_1)^P + (x_2 - y_2)^P + \ldots + (x_N - y_N)^P}$$

- When $p = 2$, Minkowski distance is same as the Euclidean distance.
- When $p = 1$, Minkowski distance is same as the Manhattan distance.

4 खोजपूर्ण डाटा विश्लेषण और डाटा विज्ञान प्रक्रिया

Exploratory Data Analysis and the Data Science Process

बहुविकल्पीय प्रश्न (MCQ)

प्रश्न 1. EDA की full form क्या है?

(a) Edit Data Analysis
(b) Entry Data Application
(c) Extra Data Analysis
(d) Exploratory Data Analysis

उत्तर (d) Exploratory Data Analysis

प्रश्न 2. EDA की process के दौरान कौन-सा method प्रयोग किया जाता है?

(a) Hypothesis testing
(b) Distance matrics
(c) Parametric test
(d) Non-parametric test

उत्तर (a) Hypothesis testing

प्रश्न 3. EDA के प्रचार में किसने योगदान दिया?

(a) जॉर्ज केन्टोर (Georg cantor)
(b) पाइथागोरस (Pythagoras)
(c) मैक्स प्लैंक (Max planck)
(d) जॉन टकी (John tukey)

उत्तर (d) जॉन टकी (John tukey)

प्रश्न 4. खोजपूर्ण डाटा विश्लेषण (Exploratory Data Analysis) क्या है?

(a) डाटा विश्लेषण करने के लिए एक rigid framework
(b) डाटा का अनुभव प्राप्त करने का एक प्रारम्भिक तरीका
(c) डाटा विश्लेषण की विशुद्ध रूप से मात्रात्मक विधि
(d) डेटा सेट का विश्लेषण करने का एक तरीका

उत्तर (d) डेटा सेट का विश्लेषण करने का एक तरीका

प्रश्न 5. डाटा विश्लेषण में Exploratory graphs की क्या भूमिका है?

(a) वे formal representation के लिए बने हैं।
(b) वे आमतौर पर जल्दी बन जाते हैं।
(c) axis, legends व अन्य विवरण साफ और बिल्कुल विस्तृत होते हैं।
(d) इनका उपयोग औपचारिक मॉडलिंग के स्थान पर किया जाता है।

उत्तर (b) वे आमतौर पर जल्दी बन जाते हैं।

प्रश्न 6. EDA निम्न में से किस पर निर्भर करता है?

(a) Visual techniques
(b) Assumptions
(c) Fixed models
(d) Testing for statistical significance

उत्तर (a) Visual techniques

प्रश्न 7. Data Science Process में निम्न में से किसका उपयोग किया जाता है?

(a) Artificial Intelligence का
(b) Machine Learning का
(c) (a) व (b) दोनों का
(d) इनमें से किसी का भी नहीं

उत्तर (c) (a) व (b) दोनों का

खण्ड 'अ' : अतिलघु उत्तरीय प्रश्न

प्रश्न 1. Data Analysis के प्रकार क्या हैं?

उत्तर Data Analysis दो प्रकार के होते हैं—

1. Confirmatory Data Analysis
2. Exploratory Data Analysis

प्रश्न 2. EDA के मुख्य उद्देश्य क्या हैं?

उत्तर EDA के चार उद्देश्य इस प्रकार हैं—

1. Discover Patterns
2. Spot Anomalies
3. Frame Hypothesis
4. Check Assumptions

प्रश्न 3. EDA के methods बताइए।

उत्तर Exploration के लिए दो Methods इस प्रकार हैं—

1. Univariate Analysis
2. Bivariate Analysis

प्रश्न 4. EDA की process के दौरान कौन-से methods प्रयोग किये जाते हैं?

उत्तर EDA की Process के दौरान प्रयोग किये जाने वाले methods निम्न प्रकार हैं—

1. Trends
2. Distribution
3. Mean
4. Median
5. Outlier
6. Spread measurement (SD)
7. Correlations
8. Hypothesis testing
9. Visual Exploration

प्रश्न 5. EDA की full form क्या है?

उत्तर EDA की Full form है— Exploratory Data Analysis.

खण्ड 'ब' : लघु एवं दीर्घ उत्तरीय प्रश्न

प्रश्न 1. Exploratory Data Analysis क्या है?

उत्तर Data Statistics में Exploratory Data Analysis या व्याख्यात्मक डाटा विश्लेषण एक approach है जिसमें Data Set को analyze किया जाता है और उनकी मुख्य विशेषताएँ (गुणों) को visual method के द्वारा प्रस्तुत किया जाता है। इसके लिये सांख्यिकीय मॉडल (statistic model) का उपयोग किया भी जा सकता है और नहीं भी, लेकिन EDA का मुख्य रूप से उपयोग यह देखने के लिए है कि डाटा हमें Formla Modeling या hypothesis testing tasks के परे या उनसे अधिक और क्या बता सकता है।

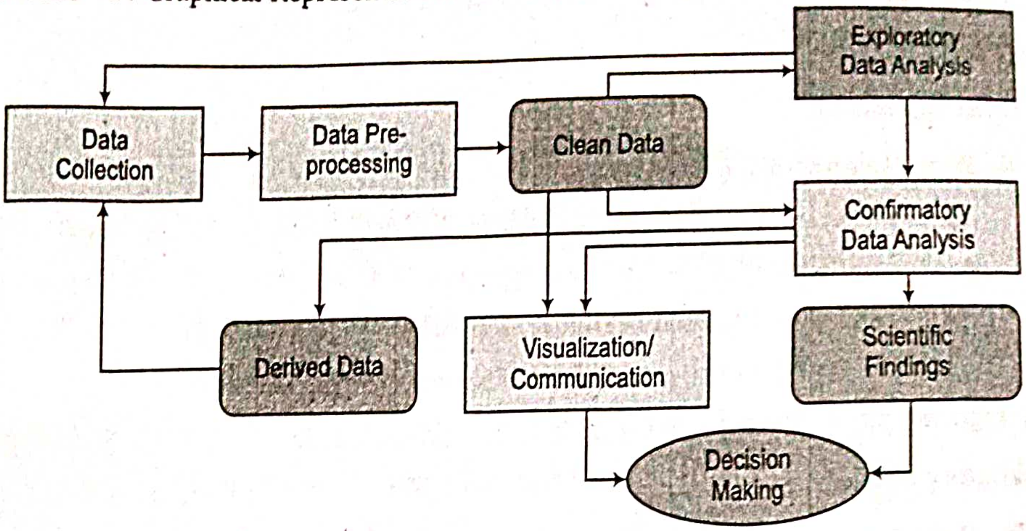
Exploratory Data Analysis डाटा पर प्रारंभिक जाँच करने की प्रक्रिया को बताता है जिसके निम्न चरण है—

(i) Data Set में Data Patern की खोज की जा सके।

(ii) विसंगतियों (Anomalies) का पता लगाया जा सके।

(iii) Hypothesis को test किया जा सके और

(iv) Statistics और Graphical Representations की मदद से Assumptions की जाँच की जा सके।



चित्र 4.1

प्रश्न 2. EDA के उद्देश्य बताइए।

उत्तर EDA के उद्देश्य निम्नलिखित हैं—

(i) Observed Phenomena (घटना) कारणों के बारे में Hypothesis Suggest करना।

(ii) जिन पर Statistic Result आधारित होगा उन अनुमानों का आकलन करना।

(iii) उचित सांख्यिकीय उपकरणों और तकनीकों के चयन में मदद करना।

(iv) आगे के डाटा संग्रह के लिए, सर्वेक्षण या प्रयोगों के माध्यम से एक आधार प्रदान करना।

(v) EDA की कई तकनीकों को Data Mining में और Data Analystics में भी उपयोग किया जाता है।

प्रश्न 3. Exploratory Data Analysis के key concept बताइए।

उत्तर **की-कॉन्सेप्ट एक्सपलोरेटरी डाटा एनालाइसिस** Key Concepts of Exploratory Data Analysis एक्सपलोरेटरी डाटा एनालाइसिस के की-कॉन्सेप्टर निम्नलिखित हैं—

Two types of Data Analysis :

1. Confirmatory Data Analysis

2. Exploratory Data Analysis

Four Objectives of EDA :

1. Discover Patterns

2. Spot Anomalies

3. Frame Hypothesis

4. Check Assumptions

Stuff done during EDA

1. Trends

2. Distribution

3. Mean
4. Median
5. Outlier
6. Spread measurement (SD)
7. Correlations
8. Hypothesis testing
9. Visual Exploration

प्रश्न 4. Data Science क्या है?

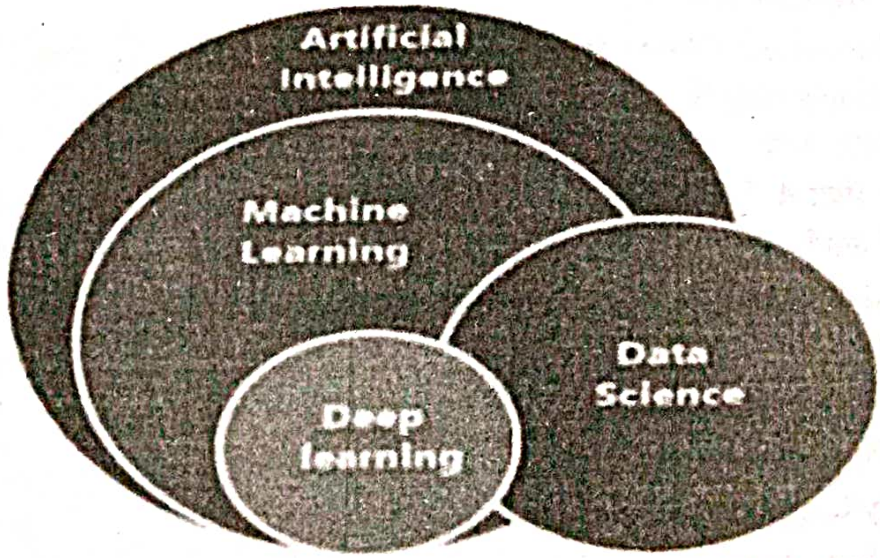
उत्तर Data Science भी डाटा पर आधारित एक विज्ञान की तरह है जिसमें हम डाटा कि खोज से लेकर उस पर Analysis, उसके प्रभावपूर्ण उपयोग तथा उसके Management पर भी कार्य करते हैं और जब हम केवल डाटा साइंस की बात करते हैं तो यहाँ पर Machine Learning, डाटा साइंस का एक Subset है।

Data Science = Data + Machine Learning + Statistics

डाटा साइंस वह ज्ञान है जिसके द्वारा हम Raw डाटा को Information के रूप में बदलते हैं। डाटा साइंस में हम Exploratory Data Analysis के साथ-साथ उस पर मशीन लर्निंग के एल्गोरिथम भी apply करते हैं।

प्रश्न 5. Data Science और Machine Learning में क्या है।

उत्तर डाटा साइंस और मशीन लर्निंग दोनों एक-दूसरे से कुछ हद तक जुड़े हुए हैं और समान है लेकिन दोनों अपने आप में बहुत विस्तृत क्षेत्र को दर्शाते हैं। दोनों के पास अपनी-अपनी विशेषताएँ हैं।



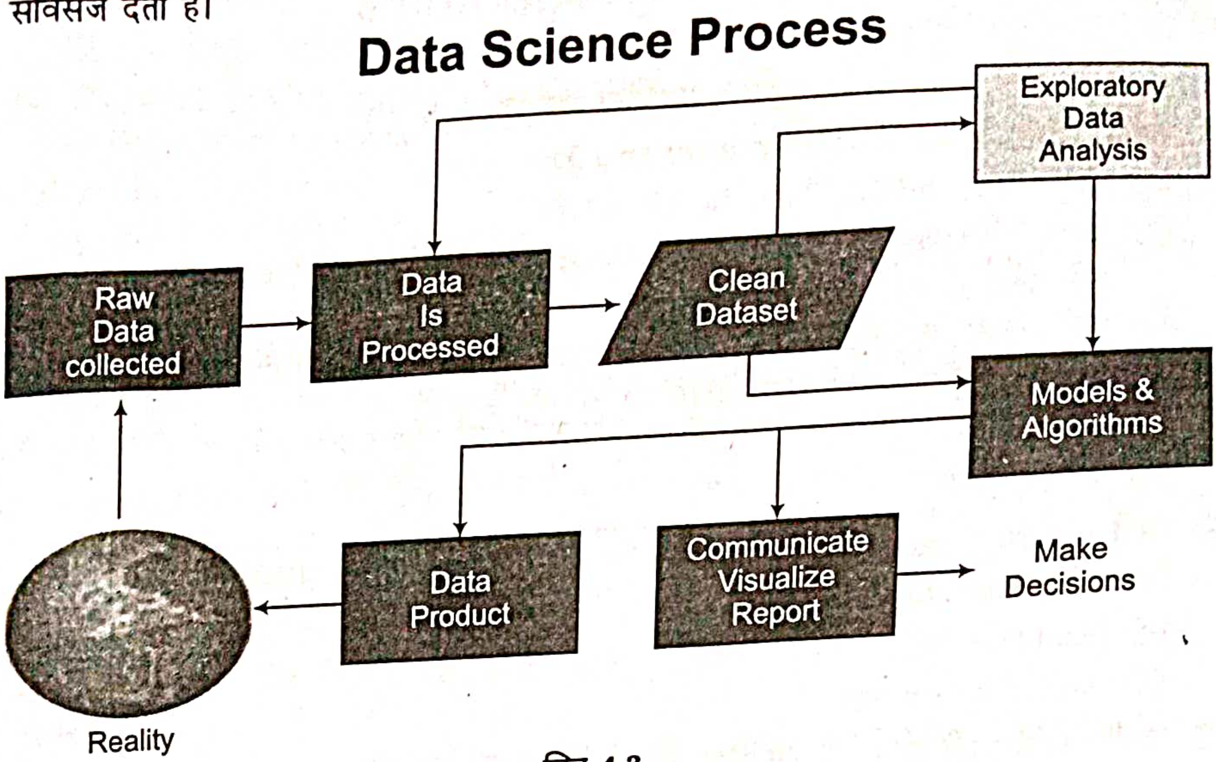
चित्र 4.2

इसका overview दर्शाया गया जिसमें Data Science और Machine Learning में क्या संबंध है इसके बारे में दिखाया गया है।

प्रश्न 6. Data Science Process से आप क्या समझते हैं?

उत्तर Data Science Process किसी भी Database से काम का data निकालने की प्रोसेस को ही data science कहा जाता है। अब इसके लिए आप artificial intelligence का इस्तेमाल करें या फिर machine learning का दोनों से ही डाटा निकालना सरल है। डाटा साइंस की मदद से किसी भी डाटाबेस से महत्वपूर्ण इनफार्मेशन निकाली जाती है जिससे किसी कंपनी को उसका लाभ हो सके।

Data Science किसी भी कंपनी के लिए बहुत ही लाभदायक सिद्ध होती है। खासतौर पर सोशल मीडिया जैसी कंपनियों के लिए क्योंकि इसके इस्तेमाल से ये कंपनियाँ अपने कस्टूमर को अच्छी तरह से समझ पाती हैं और उसके हिसाब से ये आपको सर्विसेज देती हैं।



चित्र 4.3

प्रश्न 7. Exploratory Data Analysis के Tools बताइए।

उत्तर **Techniques and Tools** There are a number of tools that are useful for EDA, but EDA is characterized more by the attitude taken than by particular techniques.

Typical graphical techniques used in EDA are :

(i) Box plot
(ii) Histogram
(iii) Multi-vari chart
(iv) Scatter plot
(v) Graphs and
(vi) Summary Statistics

प्रश्न 8. Data Exploration and Preprocessing चरण लिखिए।

उत्तर **Stpes in Data Exploration and Preprocessing**

1. Identification of variables and data types
2. Analyzing the basic metrics
3. Non-Graphical Univariate Analysis
4. Graphical Univariate Analysis
5. Bivariate Analysis
6. Variable transformations
7. Missing value treatment
8. Outlier treatement
9. Correlation Analysis
10. Dimensionality Reduction

5 मशीन लर्निंग एल्गोरिथम
Machine Learning Algorithms

बहुविकल्पीय प्रश्न (MCQ)

प्रश्न 1. मशीन लर्निंग के बारे में कौन-सा कथन सत्य है?

(a) मशीन लर्निंग कम्प्यूटर विज्ञान का एक क्षेत्र है।

(b) मशीन लर्निंग एक प्रकार की artificial intelligence है जो एल्गोरिथम या विधि का प्रयोग करके raw data से Patterns निकालती है।

(c) मशीन लर्निंग का मुख्य फोकस कम्प्यूटर सिस्टम को स्पष्ट रूप से प्रोग्राम किए बिना या मानवीय हस्तक्षेप के बिना अनुभव से सीखने की अनुमति देना है।

(d) उपरोक्त सभी

उत्तर (d) उपरोक्त सभी

प्रश्न 2. मशीन लर्निंग (machine learning) निम्न में से किसका एक सबसेट है?

(a) Deep learning (b) Artificial learning

(c) Data learning (d) इनमें से कोई नहीं

उत्तर (b) Artificial learning

प्रश्न 3. मशीन लर्निंग की कौन-सी तकनीक डेटा में आउटलेयर्स का पता लगाने में मदद करती है?

(a) एनोयाली डिटेक्शन (Anamoly Detection) (b) वर्गीकरण (Classification)

(c) क्लस्टरिंग (Clustring) (d) ये सभी

उत्तर (d) ये सभी

प्रश्न 4. मशीन लर्निंग के जनक कौन है?

(a) Geoffrey Hill (b) Geoffrey Everest Hinton

(c) Geoffrey chaucer (d) इनमें से कोई नहीं

उत्तर (b) Geoffrey Everest Hinton

प्रश्न 5. निम्न में से कौन-सा supervised learning एल्गोरिथम नहीं है?

(a) PCA (b) Naive B ayes (c) Linear Regression (d) Decision Trees

उत्तर (a) PCA

प्रश्न 6. KNN एल्गोरिथम का कहाँ पर उपयोग किया जाता है?

(a) Agriculture मे (b) Finance sector में (c) Medical sector में (d) इन सभी में

उत्तर (d) इन सभी में

प्रश्न 7. Narual Network के अन्दर कितनी लेयर होती हैं?

(a) दो (b) तीन (c) चार (d) पाँच

उत्तर (b) तीन

प्रश्न 8. Neural Network को किस algorithm की सहायता से सीखा जाता है?

(a) Supervised Learning से (b) Unsupervised Learning से

(c) Reenforcement Learning से (d) इन सभी से

उत्तर (d) इन सभी से

प्रश्न 9. निम्न में से कौन-सा Unsupervised Machine Learning का एल्गोरिथम है?

(a) KNN Algorithm (b) Linear Regression

(c) Decision trees (d) Clustering Algorithm

उत्तर (d) Clustering Algorithm

प्रश्न 10. Support Vector Machine Algorithm का प्रयोग कहाँ किया जाता है?

(a) Image को classify करने के लिए (b) Handwriting को recognize करने के लिए

(c) Text को Categorize करने के लिए (d) इनमें से कोई नहीं

उत्तर (d) इनमें से कोई नहीं

खण्ड 'अ' : अतिलघु उत्तरीय प्रश्न

प्रश्न 1. Machine Learning क्या है?

उत्तर मशीन लर्निंग एक ऐसी प्रोसेस है जिसमें कम्प्यूटर डाटा से सीखते हैं तथा मनुष्य की भाँति act तथा predict करते हैं तथा समय के साथ-साथ सीखने की क्षमता को बेहतर बनाते हैं।

प्रश्न 2. Regression से आपका क्या तात्पर्य है?

उत्तर यह एक प्रकार का supervised problem होता है, एक case जहाँ की outputs continuous होती है discrete के बदलिए।

प्रश्न 3. Decision Tree क्या है?

उत्तर Decision Tree एक बहुत ही पॉपुलर Machine Learning algorithm है जो supervised learning के अंतर प्रॉब्लम को classify करने के लिए उपयोग में लाई जाती है।

प्रश्न 4. न्यूरल नेटवर्क की शुरुआत कब हुई थी?

उत्तर पहला न्यूरल नेटवर्क 1943 में न्यूरोफिजियोलॉजिस्ट वारेन मैककलोच और लॉजिशियन वाल्टर पिट्स द्वारा उत्पादित किया गया था लेकिन उस समय उपलब्ध प्रौद्योगिकी में उन्हें बहुत कुछ करने की अनुमति नहीं दी थी।

प्रश्न 5. Applications of Support Vector Machine Algorithm बताइए।

उत्तर 1. Image classification 2. Hadwritting Recognizing

3. Hypertext and Text classification 4. Bioligical sciences classification

खण्ड 'ब' : लघु एवं दीर्घ उत्तरीय प्रश्न

प्रश्न 1. मशीन लर्निंग है?

उत्तर Machine learning का भविष्य सच में बहुत ही उज्ज्वल है। ये उन technology में से है जिनकी limit हम जैसे इन्सान ही तय करते हैं। कहने का अर्थ यह है कि हमारा जितना बड़ा imagination होगा उतना ही हम Machine learning का भी इस्तेमाल अपने कार्यों के लिए कर सकते हैं।

बहुत-सी चीज़ें जिन्हें हमारे older generation impossible सोचते थे वह अब हमारा वर्तमान हो चुका है। साथ ही समय के साथ-साथ हम भी ऐसे चीज़ों का experience कर रहे हैं जोकि कभी एक सपना हुआ करता था।

Personally, ये सोचता हूँ कि machine learning एक catalyst की तरह हो सकता है जोकि हमारे future को बदलने के लिए हमारा सहायक होने वाला है। हम machine learning पर अब इतने ज्यादा dependent हो गए हैं कि उनके बिना जीवन imagination के बाहर सा मालूम पड़ता है।

उदाहरण के लिए, जब हम Taxi बुक करते हैं Ola या Uber में तब ये हमें trip का cost, कितना distance, कौन-सा route जैसे information हमें पहले ही दिखा देता है। इसलिए हम कह सकते हैं कि Machine Learning का future सच में बहुत ही अनोखा होने वाला है।

प्रश्न 2. मशीन लर्निंग के प्रकार बताइए।

उत्तर **मशीन लर्निंग के प्रकार** Types of machine learning मशीन लर्निंग के प्रकार निम्नलिखित हैं—

1. Supervised learning, 2. Unsupervised learning

1. **Supervised learning** इस learning में इनपुट के तौर पर विभिन्न प्रकार के labeled example तथा answer दिए जाते हैं, जिससे algorithm इन उदाहरणों से सीखती है। इन इनपुटों के आधार पर सही परिणाम predict करती है।

   उदाहरणार्थ ईमेल में स्पैम फिल्टर, अर्थात् email में स्पैम फिल्टर होता है जिससे कि स्पैम messages स्पैम फोल्डर में चले जाते हैं।

   supervised learning के दो कॉमन प्रकार classification तथा regression हैं।

2. **Unsupervised learning** यह थोड़ी-सी कठिन लर्निंग है क्योंकि इसमें correct answer तथा label कुछ नहीं दिया जाता है। इसमें जो algorithm है वह डाटा में से patterns को analyze करती है।

   उदाहरणार्थ google news.

   unsupervised learning के दो कॉमन प्रकार dimension reduction और clustering हैं।

प्रश्न 3. सुपरवाइजड लर्निंग के प्रकार कौन-से हैं?

उत्तर **सुपरवाइजड लर्निंग तकनीकी के प्रकार** Types of Supervised Learning Techniques सुपरवाइजड लर्निंग के दो Algorithms हैं, अर्थात् सुपरवाइजड लर्निंग में डाटा का एनालिसिस करने के लिए दो तकनीकों का उपयोग किया जाता है, Regression तथा Classification और इन दोनों ही अल्गोरिथम का उपयोग Prediction के लिए किया जाता है।

Classification अल्गोरिथम द्वारा Datasets को श्रेणियों में बाँटा जाता है, यहाँ पर अलग-अलग मापदंड़ों के आधार पर Datasets के वर्गों में बाँट दिया जाता है, उदाहरण के तौर पर जैसे Gmail अपने मेल डाटा को श्रेणियों में बाँट देता है, (Email, SPAM, Advertisement, Promotion) इत्यादि, तो यदि आपको डाटा का Output श्रेणियों में चाहिए तो वहाँ पर Classification अल्गोरिथम का उपयोग किया जाएगा।

Regression अल्गोरिथम का उपयोग Continuous Value को Predict करने के लिए किया जाता है। यह एक प्रकार की तकनीक है, जिसमें पिछले अनुभव अर्थात् डाटा के आधार पर आगे की भविष्यवाणी की जाती है। उदाहरण के तौर पर यदि किसी कंपनी को अपने अगले 5 साल बाद के Profit का अन्दाजा लगाना हो, तो यहाँ पर Regression Algorithm का उपयोग किया जाएगा जिसमें अल्गोरिथम कंपनी के पिछले 5 से 7 साल के प्रॉफिट का Analysis करेगा और उसी अनुसार अपनी भविष्यवाणी देगा।

प्रश्न 4. सुपरवाइजड मशीन लर्निंग के चरण क्या हैं?

उत्तर **सुपरवाइजड मशीन लर्निंग के चरण** Steps of Supervised Machine Learning किसी भी Supervised Learning की समस्या को Solve करने के लिये हमें निम्न Steps को Follow करना होगा—

1. **Determine the type of Training Examples** सबसे पहले यह पता लगाना कि किस प्रकार का डाटा Training Set में इस्तेमाल होगा।
2. **Gather a training set** फिर इनका Training Set तैयार करना, इसमें Input Objects के साथ में उनके Output Objects भी शामिल किये जाते हैं।
3. **Determine the input Feature Representation of the Learned Function** फिर Function जिसे learn करना है उसके Input Feature Represention को determine करना अर्थात् Input Function में इतनी पर्याप्त information होनी चाहिए कि उससे सही Output प्राप्त कर सके। उसमें न तो ज्यादा और न ही कम information होनी चाहिये क्योंकि अगर ऐसा होता है तो Model या तो सही तरह से learn नहीं कर पायेगा या बहुत ज्यादा ही सीख लेगा जिससे कि Error होने कि सम्भावना होती है।
4. **Determine the Structure of the Learned Function and Corresponding Learning Algorithm** फिर इस Function का Structure पता करना मतलब कि वह Support Vector Machine (SVM) को use करेगा या Decision Trees को।
5. **Complete the Design** फिर इस पूरी design को complete करना।

6. **Run the Learning Algorithm on the Gathered Training Set** तो अब हमें इस एल्गोरिथम को हमारे Training Set के लिये Run करना है।
7. **Evaluate the Accuracy of the Learned Function** लर्निंग के बाद Resultant Function कि performances को measure करना।

प्रश्न 5. मशीन लर्निंग के एल्गोरिथम बताइए।

उत्तर **मशीन लर्निंग के एल्गोरिथम** Algorithms of Machine Learning सबसे ज्यादा प्रयोग होने वाले कुछ एल्गोरिथम निम्नलिखित हैं—

(i) Support Vector Machines (SVM)
(ii) Liner Regression
(iii) Logistic Regression
(iv) Naive Bayes
(v) Linear Discrimant Analysis
(vi) Decision Trees
(vii) K-Nearest Neighbor Algoithm (KNN)
(viii) Neural Networks (Multilayer perceptron)
(ix) Similarity

Engineers के द्वारा कई तरह के Algorithm; उपयोग किए जाते हैं जिन्हें दो Groups में विभाजित किया जाता है। इन एल्गोरिथम को ही Supervised Learning के methods कहते हैं। ये methods हैं—Regression और Classification। दोनों का main goal होता है कि input data में Relationship या Structure को determine करे जो हमें correct output data को प्राप्त करवा सके और दोनों का Goal होता है कि वह attribute variable से dependent variable की value पता कर सके। केवल दोनों में अंतर ये होता है कि Regression के लिये dependent attribute; numerical होता है और Classification के लिये dependent variable; Categorical होता है।

प्रश्न 6. Unsupervised Machine Learning के Algorithm कौन-कौन से हैं?

उत्तर Unsupervised Learning; Supervised Learning से ज्यादा कठिन या Complex है क्योंकि इसमें डाटा labelled नहीं होता जिससे इन्हें प्रोसेस करना थोड़ा कठिन हो जाता है और वह इसके लिये Clostering Algorithm का उपयोग करता है। इसके अलावा वह दूसरे प्रकार के algirithm भी प्रयोग करता है। जो निम्नलिखित हैं—

Algorithms

(i) Custering Algorithms
    (a) Hierarchical clustering, (b) K-means, (c) Mixture, (d) DBSCAN
(ii) OPTICS algorithm
    (a) Anomaly detection
    (b) Local Outlier Factor
(iii) Neural Networks Autoencoders
    (a) Deep Belief Nets
    (b) Hebbian Learning
    (c) Generative Adversarial Networks
    (d) Self-organizing map

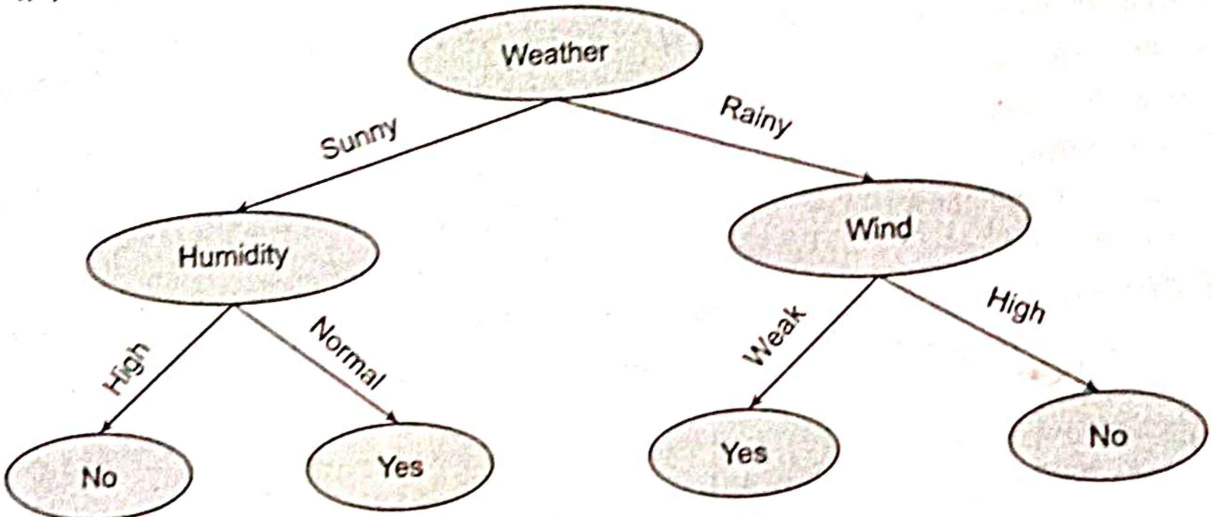
Unsuprvised learning algorithms,; supervised की तुलना में कई ज्यादा Complex Task पर्फोम कर सकता है।

प्रश्न 7/ Decision tree की संक्षेप में व्याख्या कीजिए।

उत्तर **Decision tree** Decision tree एक फ्लो-चार्ट की तरह का स्ट्रक्चर होता है; जिस प्रकार tree में पत्तियाँ, जड़ तथा शाखाएँ होती हैं उसी प्रकार इसमें leaf नोड तथा branches होती हैं।

Decision tree में सबसे ऊपर की नोड को root नोड कहते हैं। इसमें प्रत्येक leaf नोड एक class को प्रदर्शित करती है। Decision tree का प्रयोग Decision making के लिए किया जाता है। ज्यादातर Decision tree बाइनरी होते हैं अर्थात् एक नोड के केवल दो चाइल्ड होते हैं।

उदाहरणार्थ—



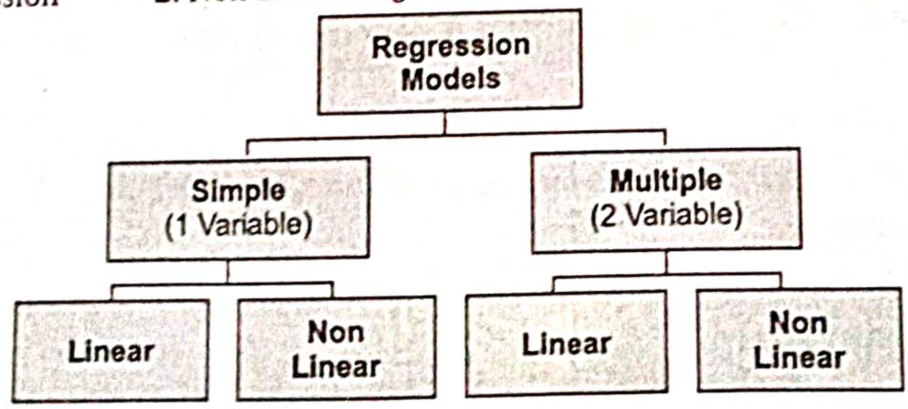
चित्र 5.1

प्रश्न 8. रिग्रेशन (Regression) से आपका क्या अर्थ है?

उत्तर **रिग्रेशन** Regression रिग्रेशन में हम ऐसे Continuous Output का अनुमान लगाते हैं। जिसमें Output Variables; Continous Values होते हैं जैसेकि Amount, Volume या Size. जैसे किसी Student के उसके tests के आधार पर Final Exam के marks का अनुमान लगाना जोकि 0 से 100 के बीच हो सकते हैं तो Regression; continous quantity को predict करने का Task है।

Regression Problem को भी दो तरह से बताया जा सकता है—

1. Linear Regression 2. Non-Linear Regression



चित्र 5.2

प्रश्न 9. प्रतिगमन (Regression) के प्रकार से आप क्या समझते हैं?

उत्तर **प्रतिगमन के प्रकार** Types of Regression वैसे तो regression कई प्रकार के होते हैं लेकिन हम इनमें से कुछ मुख्य प्रकार के बारे में बात करेंगे जो macine learning के अंदर ज्यादातर उपयोग में आते हैं—

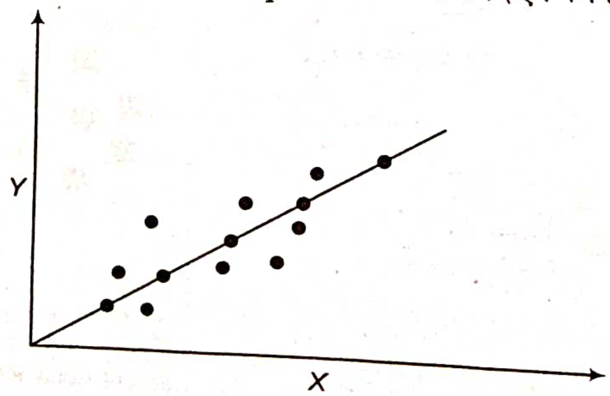
1. Linear Regression, 2. Polynomial Regression, 3. Logistic Regression, 4. Non-Linear Regression

1. **Linear Regression** Linear Regression के अंदर दूसरे independent variable की मदद से एक dependent variable का रिजल्ट pedict करते हैं।
Linear Regression भी दो प्रकार के होते हैं—
(i) Simple Linear Regression
(ii) Multiple Linear Regression
(i) **Simple Linear Regression** Simple Linear Regression में केवल एक independent variable का उपयोग किया जाता है दूसरे dependent variable के रिजल्ट का पता (predict) करने के लिए। Simple Linear Regression की mathematical equation को चित्र 5.3 (image) में देख सकते हैं।
(ii) **Multiple Linear Regression** Multiple Linear Regression के अंदर दो या दो से अधिक independent variable होते हैं जिनका इस्तेमाल dependent variable के रिजल्ट को पता करने के लिए किया जाता है। उदाहरण के लिए मान लो अगर किसी घर की भविष्य में क्या कीमत होगी इसका पता लगाना है तो Multiple Linear Regression के अंदर हम दो या दो से अधिक independent variable (जैसे—घर की location, घर कितना साल पुराना है, घर में कितने फ्लोर है, घर की size आदि) का उपयोग करेंगे। Multiple Linear Regression की mathematical equation को चित्र 5.3 (image) में देख सकते हैं।

2. **Polynomial Regression** Polynomial Regression के अंदर independent variable की power एक से ज्यादा होती है। Polynimial Regression की mathematial equaiton को चित्र 5.3 (image) में देख सकते हैं। यहाँ independent variable की power एक से ज्यादा है और जैसे-जैसे ये पावर बढ़ती जाएगी hypothesis और भी ज्यादा complex होती चली जाएगी।

3. **Logistic Regression** Logistic Regression का उपयोग किसी भी event के होने न होने की संभावना को बताता है। इसके अंदर dependent variable की वेल्यू Binary (जैसे—True/False, Yes/No, 0/1) होती है। उदाहरण के लिए जैसे कल बारिश होगी या नहीं होगी, इंडिया मैच जीतेगी या नहीं जीतेगी आदि जैसी कंडीशन को Logistic Regression के द्वारा हल कर सकते हैं।

4. **Non-Linear Regression** Non-Linear Regression के अंदर parameter की power भी एक ज्यादा हो सकती है। इसके अंदर parameter की power independent variable की power के अनुसार बदलती रहती है। इसके अंदर Linear Regression और Polynomial Regression की समस्या को भी हल किया जा सकता है। Non-Linear Regression की mathematical equation को नीचे दिए चित्र 5.3 (image) में देख सकते हैं।



चित्र 5.3

1. Simple Linear Regression :

$$Y = \theta_0 + \theta_2 X$$

यहाँ पर, Y = dependent variable
X = Independent variable
θ_0 and θ_1 Parameter हैं।

2. Multiple Linear Regression :

$$Y = \theta_0 + \theta_2 X_1 + \theta_2 \ X_2$$

यहाँ पर, एक से ज्यादा Independent variable है।

3. Polynomial Regression :

$$Y = \theta_0 + \theta_1 X^2 + \theta_2 X^3$$

4. Non-Linear Regression :

$$Y = \theta_0^1 + \theta_1^2 x^2 + \theta_2^2 \ x^3$$

प्रश्न 10. K-Nearest Neighbor (KNN) Algorithm क्या है?

उत्तर KNN Algorithm एक बहुत ही Simple, Machine Learning Algorithm है, जो Supervised Machine Learning पर आधारित है जो डाटा को classify करने का काम करती है। KNN की मदद से किसी भी नए input डाटा की कैटेगरी को बताया जा सकता है, कि वह किस केटेगरी से belong करेगी।

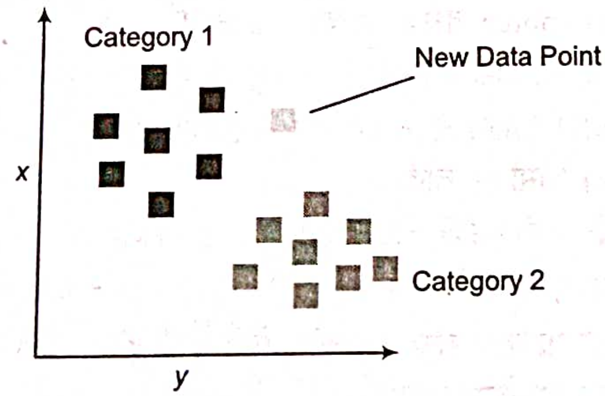
इसको हम Regression के लिए भी उपयोग में ला सकते हैं परन्तु इसका ज्यादातर इस्तेमाल classification problem को solve करने के लिए ही किया जाता है।

K-Nearest Neighbor (KNN) Algorithm एक Supervised Machine Learning Algorithm है, इसलिए इसमें मॉडल को तैयार करने के लिए labeled data का इस्तेमाल किया जाता है, और फिर जब भी कोई नया unlabeled data इस KNN मॉडल को दिया जाता है, तो वह उसे ट्रैनिंग के दौरान दिए गए labeled data की मदद से नए unlabeled data को आसानी से classify कर पाता है। KNN Algorithm को Lazy Learning Algorithm भी कहा जाता है क्योंकि इसमें algorithms को सिखाने के लिए पहले से डाटा की जरूरत पड़ती है उसके बाद ही algorithm कोई डिसीजन ले पाता है।

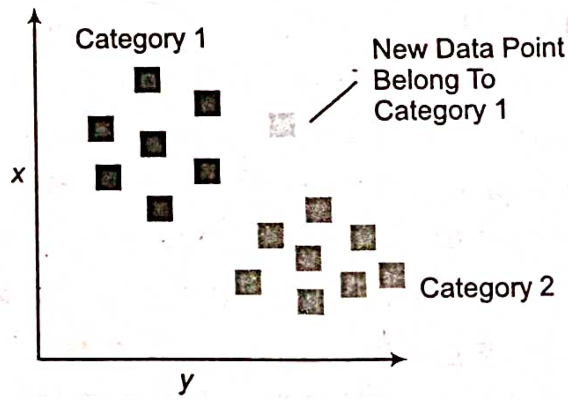
प्रश्न 11. K-Nearest Neighbour (KNN) एल्गोरिथम की क्या आवश्यकता है?

उत्तर Need of K-Nearest Neighbour (KNN) Algorithm

KNN Algorithm की जरूरत को समझने के लिए हम एक example का सहारा लेंगे जिसमें हमारे पास दो प्रकार के डाटा होते हैं, category-1 डाटा तथा category-2 डाटा। अब जब भी कोई नया डाटा आता है और हमें यह पता करना होता है कि वह डाटा किस केटेगरी से बिलॉन्ग करेगा तो उसके लिए कैनन (KNN) का इस्तेमाल कर सकते हैं।



Before Apply KNN Algorithms



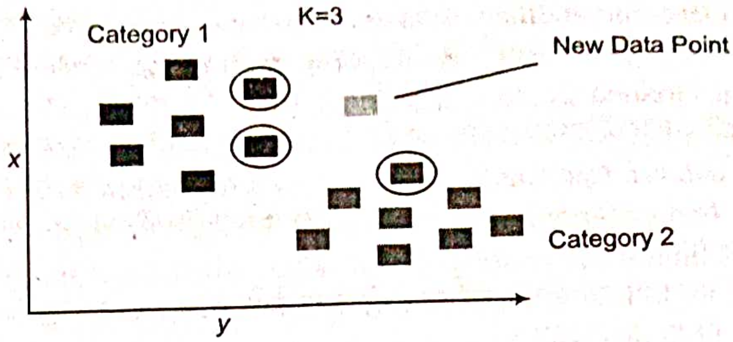
Before Apply KNN Algorithms

चित्र 5.4

प्रश्न 12. KNN Algorithm की कार्य-विधि समझाइए।

उत्तर KNN Algorithm की कार्य-विधि को हम निम्नलिखित चरणों के द्वारा समझ सकते हैं—

1. **KNN Algorithm** के अंदर हम सबसे पहले एक वेरिएबल K की कोई एक वैल्यू कंसीडर कर लेते हैं यहाँ पर K की वैल्यू नियरेस्ट नेवर की संख्या बताएगा। यहाँ पर हम कोई ओल्ड वैल्यू को K की वैल्यू कंसीडर करेंगे जिससे एल्गोरिथम को डिसीजन लेने में आसानी हो।



चित्र 5.6

2. दूसरे step में हम new data point से Nearest Neighbor की डिस्टेंस को कैलकुलेट करेंगे, इसके लिए हम निम्नलिखित Euclidean Distance के सूत्र का प्रयोग करेंगे—

Euclidean Distance Formula

$$D = \sqrt{(X_0 - X_1)^2 + (X_0 - X_2)^2}$$

जहाँ, D = Distance

O = Observed value

A = Actual alue

3. तीसरी step में KNN Algorithm, new data point से नंबर ऑफ Nearest Neighbor point निकालती है और यह भी पता करती है कि कितने Nearest Neighbor point, new data point से बिलॉन्ग करते हैं।
4. चौथी स्टेप में KNN Algorithm, new data point की कैटेगरी को डिसाइड कर लेती है। New data point जिस category के ज्यादा Nearest Neighbor point से belong करेगा उसका उस category का मान लिया जाएगा।

प्रश्न 13. KNN K-Nearest Neighbor (KNN) एल्गोरिथम के फायदे और नुकसान पर चर्चा कीजिए।

उत्तर Advantages of KNN K-Nearest Neighbor (KNN) Algorithm इसके फायदे निम्नलिखित हैं—

1. KNN Algorithm को इंप्लीमेंट करना बहुत ही आसान होता है।
2. KNN Algorithm को classification, Regression तथा searching के लिए भी प्रयोग किया जा सकता है।
3. KNN Algorithm, Noisely data के लिए बहुत ही robust तरीके से कार्य करते हैं।
4. KNN Algorithm लार्ज अमाउंट ऑफ डाटा के लिए भी इस्तेमाल किया जा सकता है।

Disadvantages of KNN K-Nearest Neighbor (KNN) Algorithm इसके दोष निम्नलिखित हैं—

1. Example डाटा ज्यादा होने पर इसकी डाटा प्रोसेसिंग धीमी हो जाती है।
2. हमेशा k की वैल्यू को डिटरमाइंड करना पड़ता है जो कभी-कभी थोड़ा complex हो जाता है।
3. KNN में कैलकुलेशन cost ज्यादा होती है क्योंकि सभी डाटा प्वाइंट्स के बीच की डिस्टेंस को निकालना होता है।
4. KNN एक Lazy Learning Algorithm है क्योंकि इसके लिए पहले से डाटा available होना जरूरी होता है क्योंकि पहले यह उस डाटा को स्टोर करता है और उसके बाद आगे की प्रोसेस करता है।

प्रश्न 14. K-Nearest Neighbor Algorithm की applications बताइए।

उत्तर **Application of K-Nearest Neighbor(KNN) Algorithm** KNN Algorithm को real life में बहुत सारी जगह उपयोग में लाया जा रहा है और लाया जा सकता है जो निम्नलिखित है—

1. **KNN Use in Medicine** KNN Algorithm का उपयोग हार्ट अटैक, डायबिटीज, ब्लड प्रेशर आदि को predict करने में किया जा सकता है। ब्रेस्ट कैंसर को भी KNN की मदद से predict किया जा सकता है।

2. **KNN Use For Recommendation System** YouTube, Netflix तथा बहुत सारे search engine कंटेंट का सर्व करने तथा अपने यूजर एक्सपीरियंस को बढ़ाने के लिए KNN एल्गोरिथम का इस्तेमाल करते हैं।
3. **KNN Use For Finance Sector** फाइनेंशियल रिस्क को समझने, लोन मैनेजमेंट, स्टॉक मैनेजमेंट, बैंकिंग सेक्टर, मनी लॉन्ड्रिंग एनालिसिस आदि जगहों में KNN का इस्तेमाल किया जाता है।
4. **KNN can be use for Text Categorization** KNN Algorithm का इस्तेमाल टेक्स्ट कैटिगराइजेशन में भी किया जाता है। **Text Categorization** के लिए यह एक बहुत ही पॉपुलर एल्गोरिथम है।
5. **Agriculture** Climate forecasting में भी KNN Algorithm का इस्तेमाल किया जा सकता है जो Agriculture में बहुत ही फायदेमंद साबित हो सकता है। K-Nearest Neighbor की मदद से फसलों की गुणवत्ता को भी predict किया जा सकता है।

प्रश्न 15. विभिन्न प्रकार के distance measures क्या हैं?

उत्तर मशीन लर्निंग में सबसे अधिक इस्तेमाल होने वाले distance measures के चार प्रकार निम्न प्रकार हैं—

(i) Hamming Distance, (ii) Educlidean Distance, (iii) Manhattan Distance, (iv) Minkowski Distance

प्रश्न 16. Support Vector Machine Algorithm क्या है?

उत्तर Support Vector Machine Algorithm, Supervised Machine Learning के अंतर्गत आती है जिसका मुख्य उद्देश्य डाटा को दो भागों में classify करना होता है। Support Vector Machine, Machine Learning की एक ऐसी Algorithm है जो classification तथा Linear Regression के लिए डाटा को analyze करती है।

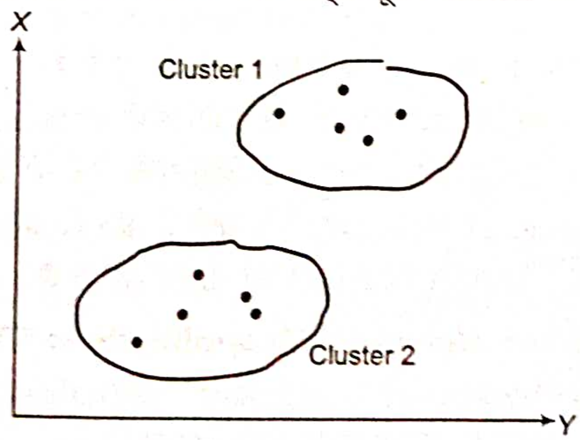
Support Vector Machine Algorithm (SVM) मुख्यत: डाटा को दो भागों में विभाजित करने के लिए प्रयोग की जाती है, जिसके अंदर ट्रेनिंग के तौर पर लेबल डाटा को दिया जाता है उस लेबल डाटा को Support Vector Machine Algorithm एक मार्जिन के बेस पर दो भागों में डिवाइड कर देती है। Support Vector machine (SVM) को Support Vector Network (SVN) के नाम से भी जाना जाता है।

Support Vector Machine (SVM) का प्रयोग Image को classify करने, handwriting को Recognize करने तथा text को categorize करने आदि में किया जाता है।

Support Vector Machine एक प्रकार की non-binary linear classifier है जो लेबल डाटा को दो भागों में डिवाइड कर देती है।

प्रश्न 17. क्लस्टरिंग तकनीक क्या है?

उत्तर **क्लस्टरिंग** Clustering Cluster Analysis को clustering भी कहते हैं। Clustering एक डाटा analysis टूल है जिसमें डाटा तथा ऑब्जेक्ट्स को इस प्रकार अलग-अलग समूहों (clusters) में divide किया जाता है कि जो समान गुणों वाले ऑब्जेक्ट्स होते हैं। उन्हें एक समूह (cluster) में रखा जाता है और भिन्न गुणों वाले ऑब्जेक्ट्स को दूसरे cluster में रखा जाता है। प्रत्येक cluster के ऑब्जेक्ट्स दूसरे cluster के ऑब्जेक्ट्स से भिन्न होते हैं।



चित्र 5.7

हमारी सामान्य जिंदगी के प्रत्येक aspect में भी clustering का role होता है। उदाहरण के लिए किसी restaurant में अलग-अलग प्रकार का food होता है और vehicle showroom में cars, bikes तथा अन्य vehicles होती है। एक cluster में जितने भी ऑब्जेक्ट्स होते है उन्हें एक समूह के रूप में treat किया जाता है।

प्रश्न 18. क्लस्टरिंग तकनीक के प्रकार बताइए।

उत्तर **क्लस्टरिंग तकनीक के प्रकार** Types of Clustering Techniques Clustering निम्नलिखित प्रकार की होती है—

1. **K-means Clustering** K-means clustering को Portioning clustering भी कहते हैं। माना कि हमारे पास 'n' ऑब्जेक्ट्स का डाटाबेस है और हम डाटा के 'k' portion कर देते हैं। प्रत्येक portion एक cluster को प्रदर्शित करती है और [K <_N] K<_N हमें यह प्रदर्शित करता है कि प्रत्येक ऑब्जेक्ट्स एक cluster से संबंधित होता है तथा ये यह भी दर्शाता है कि प्रत्येक Cluster कम-से-कम एक ऑब्जेक्ट्स को contain किये रहता है।
2. **Hierarchical Clustering** Hierarchical Clustering निम्न दो प्रकार की होती है—

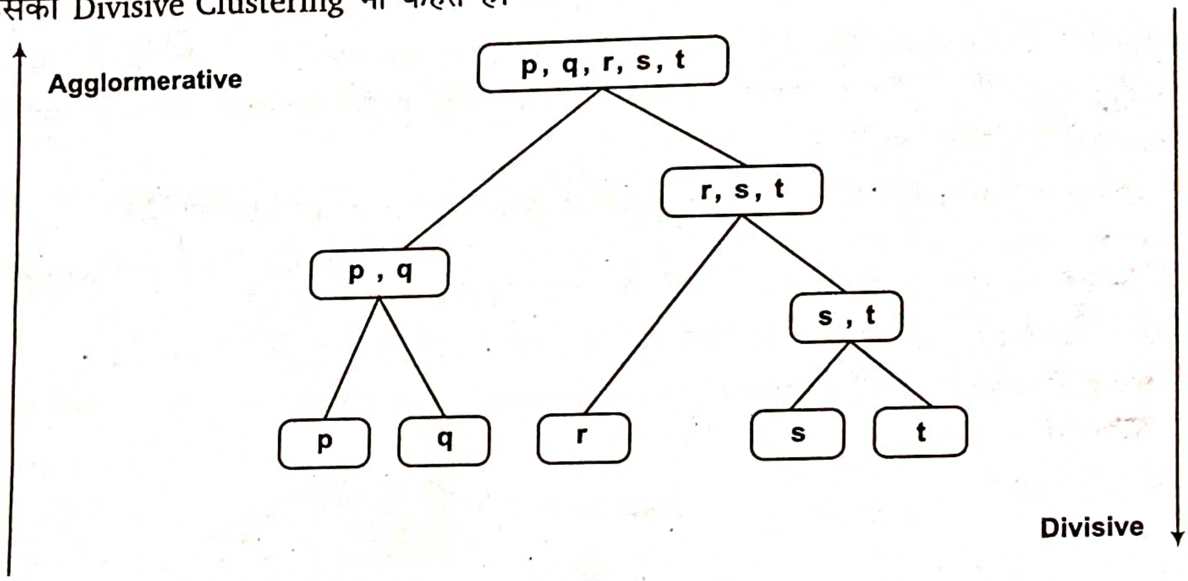
   (a) Bottom-Up (b) Top-Down

   Bottom-Up में प्रत्येक ऑब्जेक्ट्स अलग-अलग समूह में होता है और फिर अगले step में एक ऑब्जेक्ट्स दूसरे ऑब्जेक्ट्स के साथ एक समूह में सम्मिलित होता है और ऐसा तब तक चलता रहता है जब तक कि सभी ऑब्जेक्ट्स एक समूह (Cluster) में नहीं आ जाते हैं।

   Bottom-Up को Agglomerative clustering भी कहते हैं।

   Top-Down में सभी ऑब्जेक्ट्स एक ही समूह (cluster) में होते हैं और ये अगले step में अलग-अलग होते रहते हैं और ऐसा तब तक होता है जब तक कि सभी ऑब्जेक्ट्स अलग-अलग नहीं हो जाते हैं।

   इसको Divisive Clustering भी कहते है।



चित्र 5.8

प्रश्न 19. K-Means Clustring समझाइए।

उत्तर **K-Means Clustring** K-Means एक unsupervised machine learning algorithm है जो clustering की प्रॉब्लम को हल करने में उपयोग में लाई जाती है। इसके अंदर data sets को clusters के अंदर classify कर दिया जाता है। यहाँ पर cluster का मतलब समान प्रकार के data group से है, जो एक ही प्रकार की information को contain करके रखते हैं। यहाँ पर cluster के नंबर को k से represent करते हैं। K-Means algorithms cluster के अंदर के कुछ point को pick करती है उन point को cetroids कहते हैं।

प्रश्न 20. MeanShift Algorithm का परिचय दीजिए।

उत्तर **मीन-शिफ्ट एल्गोरिथ्म का परिचय** MeanShift Algorithm यह एक powerful क्लस्टरिंग एल्गोरिथम है जिसका उपयोग unsupervised learning में किया जाता है। K-means clustering के विपरीत, यह कोई धारणा नहीं बनाता है, इसलिए यह एक गैर-पैरामीट्रिक एल्गोरिथम है।

मीन-शिफ्ट एल्गोरिथम का कार्य (Working of Mean-Shift Algorithm)

निम्नलिखित Steps की मदद से मीन-शिफ्ट क्लस्टरिंग एल्गोरिथम के काम को समझ सकते हैं—

(i) **Step 1 :** First, start with the data points assigned to a cluster of their own.

(ii) **Step 2 :** Next, this algorithm will computer the centroids.

(ii) **Step 3 :** In this step, location of new centroids will be updated.

(iv) **Step 4 :** Now, the process will be iterated and moved to the higher density region.

(v) **Step 5 :** At last, it will be stopped once the centroids reach at position from where it cannot move further.

प्रश्न 21. Dimensionality Reduction Technique पर संक्षिप्त नोट लिखिए।

उत्तर **Dimensionality Redcution** Machine Learning में Dimensionality का मतलब होता है कि आपके Database में कितने Feature या Input Variable हैं।

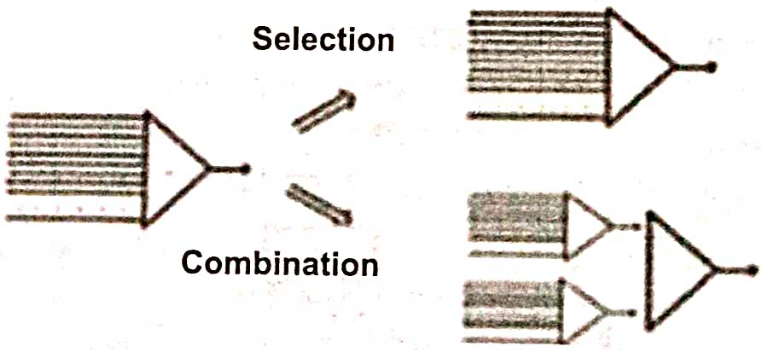
जब आपके database में Features की संख्या, observations की संख्या की तुलना में बहुत ज्यादा होती है तो model को सही तरह से train करने में कठिनाई का सामना करना पड़ता है, इसे "Curse of Dimensionality" कहते हैं। मशीन लर्निंग Classification problem में कई कारक (factors) होते हैं जिनके आधार पर final classification होता है। ये कारक basically; variables होते हैं जिन्हें Features भी कहते हैं। जितने ज्यादा features होते हैं, उन पर कार्य करना उतना ही कठिन होता है। ऐसे में हम Dimensionality Reduction का उपयोग करते हैं।

Dimensionality Reduction एक ऐसा प्रोसेस है जिसके द्वारा Principle Variable के सेट को बनाकर के, Random Variables की संख्या को कम किया जाता है। दूसरे शब्दों में कहें तो High Dimensional Data-set को Low Dimensional Data-set में बदला जाता है।

प्रश्न 22. Dimensionality Reduction Technique के components क्या है?

उत्तर **Dimensionality Reduction के Components** Dimensionality Reduction के मुख्य रूप से 2 Components हैं या दूसरे शब्दों में कहें तो यह दो तरीकों से किया जा सकता है।

1. **Feature Seletion** इसमें एक बड़े डाटा सेट के Subsets बनाये जाते हैं।
2. **Feature Extraction** इसमें Data को high dimensional space से lower dimensional space में बदला जाता है।



चित्र 5.9

प्रश्न 23. Dimensionality Reduction Technique के तरीके क्या हैं?

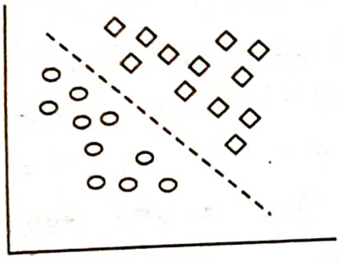
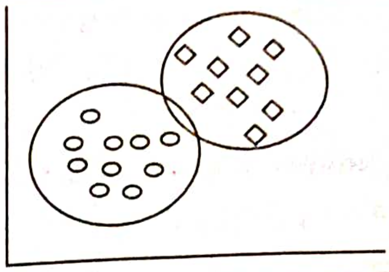
उत्तर **Methods of Dimensionality Reduction** Dimension Reduction को निम्न methods के द्वारा किया जा सकता है—

1. Missing Values
2. Low Variance
3. Decision Tree
4. Random Forest
5. High Correlation
6. Backward Features Elimination

प्रश्न 24. Clustering एवं Classification के बीच अंतर बताइए।

उत्तर

Clustering एवं Classification के बीच अंतर

Classification	Clustering
Process of classifying the data with help of class table	It is similar to classification but there are no predifined class lable
A supervised learning technique	An unsupervised learning technique
Goal of assigning new input to a class	Goal of finding similarities within a given dataset
Works with tabeled data	works with unlabeled data
Known number of classes	Unknown number of classes
	

प्रश्न 25. न्यूरल नेटवर्क क्या है?

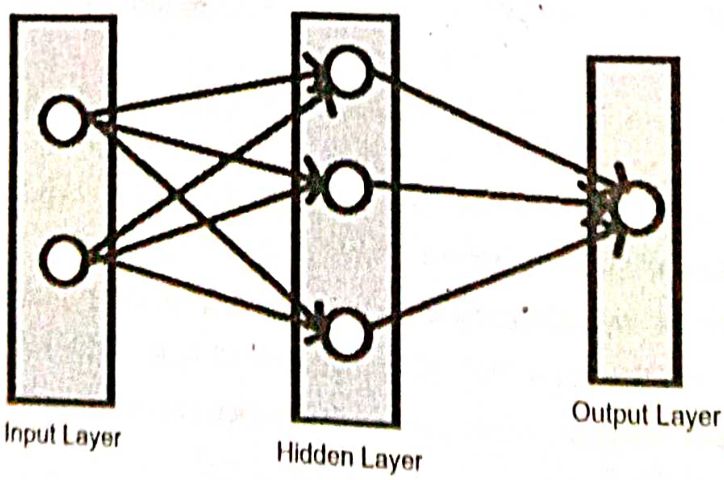
उत्तर न्यूरल नेटवर्क Neural Network, machine learning का ऐसा type (प्रकार) है जो खुद एक human brain (मनुष्य के दिमाग) की तरह अपने मॉडल को तैयार करता है। Neural Network एक artificial neural network बनाता है जो एक algorithm के द्वारा computer को खुद से सीखने की ताकत देता है जब भी कोई नई चीज या डाटा computer देखे तो वो उसे अपने आप ही analyze कर ले, अपने पुराने experience के आधार पर।

जिस प्रकार मनुष्य (human) का nervous system work करता है ठीक उसी तरह artificial neural network भी काम करता है। अगर सीधे शब्दों में कहा जाए तो Neural network, Information को process करने का एक ऐसा type है जो किसी मनुष्य के दिमाग की तरह काम करता है। जिस प्रकार से किसी Human का Brain किसी information को देख कोई निर्णय लेता है, ठीक उसी प्रकार हम इसी चीज को computer या machine से करवाने के लिए neural network का उपयोग करते हैं।

जिस प्रकार Human के brain में neurons आपस में एक-दूसरे से connected रहते हैं, ठीक उसी प्रकार neural network में भी बहुत सारे Nodes का समूह एक-दूसरे से connected रहता है।

प्रश्न 26. Neural Network कैसे कार्य करता है?

उत्तर Neural Network की working process को नीचे दिए चित्र 5.10 के द्वारा समझ सकते हैं—



चित्र 5.10

दिए गए चित्र 5.10 में कुछ nodes हैं जो एक-दूसरे से connected है, यहाँ पर हर node अपनी अगली layer के सारे nodes से connected है। यहाँ पर arrow डाटा processing की direction को बता रहा है, कि डाटा कहाँ से कहाँ flow होगा।

Neural Network के अंदर basically तीन लेयर होती हैं—

1. Input Layer
2. Hidden Layer
3. Output Layer

1. **Input Layer** इसके अंदर डाटा को input के तौर पर दिया जाता है।
2. **Hidden Layer** Hidden Layer के अंदर दिए गए data की processing होती है। Hidden layer एक या एक से ज्यादा भी हो सकती हैं। यह Hidden layer data को कैसे प्रोसेस करती है इसको हम अगले blog में पढ़ेंगे।
3. **Output Layer** यह process हुए data को result के रूप में show करती है। अगर result दिए गए data के अनुसार गलत आता है तो neural network अपनी input value के अंदर change करता है और ऐसी प्रोसेस को फिर से दोहराता है और ये प्रोसेस तब तक चलती है जब तक हमें data के मुताबिक desired output न मिल जाए।

प्रश्न 27. Neural Network कैसे सीखे जाते हैं?

उत्तर Neural Network से किसी task को पूरा करवाने के लिए किसी सामान्य algorithm की तरह programmed नहीं किया जा सकता है, जिस प्रकार किसी छोटे बच्चे को कोई नई चीज सीखाने के लिए उसे उस चीज के बारे में बताना पड़ता है और समझना पड़ता है, ठीक उसी प्रकार neural network को भी Information के तौर पर बार-बार data देकर उसे train करना पड़ता है।

Neural Network को तीन प्रकार से learn कराया जा सकता है—

1. **Supervised Learning** यह बहुत ही आसान strategy है। neural network को learn कराने की क्योंकि इसके अंदर जो डाटा होता है वह labeled data होता है। इस कारण कम्प्यूटर आसानी से समझ जाता है कि यह data किस प्रकार का है और अपने result तक आसानी से पहुँच जाता है।
2. **Unsupervised Learning** इस method को हम वहाँ प्रयोग करते हैं जहाँ हमारे पास labeled data नहीं होता है, इस कारण इसको implement करना थोड़ा कठिन हो जाता है, क्योंकि यहाँ पर डाटा किस प्रकार का है ये पहले से पता नहीं होता है।
3. **Reinforcement Learning** इस algorithm के अंदर **neural network** अपने आपको feedback देता है और उसी feedback से सीखता रहता है।

प्रश्न 28. Artificial Neural Network के अनुप्रयोग क्या हैं?

उत्तर **Applications Of Artificial Neural Network** आज के समय में Artificial Neural Network बहुत सारी जगह प्रयोग हो रहा है उनमें से निम्नलिखित है—

1. **Image Detection** Image Detection के अंदर Artificial Neural Network का उपयोग अभी बहुत ज्यादा मात्रा में किया जा रहा है।
2. **Face Detection** Face को detect करने के लिए भी Artificial Neural Network का उपयोग किया जा रहा है।
3. **Military** इसको हम military में बहुत सारी जगह use कर सकते हैं; जैसे—weapon के orientation में, target को ट्रैक करने में आदि।
4. **Medical** मेडिकल के अंदर कैंसर सेल को analyze करने में भी इसका उपयोग किया जाता है।
5. **Software** pattern, character आदि को recognize करने में भी इसका उपयोग किया जा सकता है।
6. **Credit Card Fraud Detection** Credit Card Fraud Detection के लिए भी Artificial Neural Network का उपयोग किया जा रहा है।

6 माइनिंग सोशल नेटवर्क ग्राफ्स
Mining Social-Network Graphs

बहुविकल्पीय प्रश्न (MCQ)

प्रश्न 1. अकेले यूजर काउंट के आधार पर, निम्न में से कौन-सा सोशल मीडिया प्लेटफॉर्म सबसे अधिक लोकप्रिय बना हुआ है?

(a) Face book
(b) Twitter
(c) Linked in
(d) My space

उत्तर (a) Face book

प्रश्न 2. Community detection के लिए कौन-सी clustering तकनीक प्रयोग करते हैं?

(a) K-means clustering
(b) Hierarchical clustering
(c) Mean shift clustering
(d) इनमें से कोई नहीं

उत्तर (b) Hierarchical clustering

प्रश्न 3. निम्न में से कौन-सी clustering में सभी ऑब्जेक्ट्स एक ही समूह (cluster) में होते हैं?

(a) K-means
(b) Mean shift
(c) Top-Down
(d) Bottom Up

उत्तर (c) Top-Down

प्रश्न 4. सोशल नेटवर्क्स मुख्य रूप से निम्न में से किसको संगठित करते हैं?

(a) ब्रांड्स (brands)
(b) चर्चा (discussions)
(c) मनुष्यों (people)
(d) रुचियाँ (interests)

उत्तर (c) मनुष्यों (people)

प्रश्न 5. निम्न में से कौन दो नोड्स को जोड़ता है?

(a) Edge
(b) Degree
(c) Neighborhood
(d) इनमें से कोई नहीं

उत्तर (a) Edge

प्रश्न 6. सोशल नेटवर्क ग्राफ में लोगों को किससे प्रदर्शित किया जाता है?

(a) Degree
(b) Modes
(c) Edge
(d) इनमें से कोई नहीं

उत्तर (b) Modes

प्रश्न 7. सोशल नेटवर्किंग के बारे में निम्न में से कौन-सा कथन असत्य है?

(a) सोशल नेटवर्किंग का प्रयोग करने से हम हमेशा व्यस्त रह सकते हैं।
(b) सोशल नेटवर्किंग का प्रयोग करने से हमारे बिजनेस को फायदा पहुँच सकता है।
(c) सोशल नेटवर्किंग का प्रयोग करने से हमारे दोस्त कम हो सकते हैं।
(d) सोशल नेटवर्किंग का प्रयोग करने से हम अकेलेपन का शिकार होने से बचते हैं।

उत्तर (c) सोशल नेटवर्किंग का प्रयोग करने से हमारे दोस्त कम हो सकते हैं।

प्रश्न 8. निम्न में से कौन-सी सोशल नेटवर्किंग सर्विस (SNS) users को एक mini-communities वाले groups बनाने के लिए allow करती है?

(a) Profile-based SNS
(b) White-label SNS
(c) Content-based SNS
(d) इनमें से कोई नहीं

उत्तर (b) White-label SNS

खण्ड 'अ' : अतिलघु उत्तरीय प्रश्न

प्रश्न 1. सोशल मीडिया पर कितने लोग हैं?

उत्तर अप्रैल 2020 तक, दुनिया भर में कुल 3.81 बिलियन लोग सोशल मीडिया का उपयोग करते हैं, दुनिया भर में सोशल मीडिया की पैठ दर (penetration rate) 49% है।

प्रश्न 2. सबसे लोकप्रिय सोशल मीडिया प्लेटफॉर्म क्या है?

उत्तर अकेले यूजर काउंट के आधार पर, फेसबुक सबसे लोकप्रिय सोशल मीडिया प्लेटफॉर्म बना हुआ है।

प्रश्न 3. Instagram उपयोगकर्ताओं की औसत आयु क्या है?

उत्तर औसतन, अधिकांश इंस्टाग्राम उपयोगकर्ता 18 से 34 वर्ष की आयु के बीच के हैं।

प्रश्न 4. Community detection के लिए कौन सी clustering technique प्रयोग करते हैं?

उत्तर Community detection के लिए Hierarchical clustering प्रयोग कर सकते हैं।

प्रश्न 5. Hierarchical clustering की approaches कौन-सी होती हैं?

उत्तर Hierarachical clustering की two approaches होती हैं: Agglomerative hierarchical clustering और Divisive hierarchical clustering.

प्रश्न 6. Agglomerative hierarchical clustering (Bottom-Up) को परिभाषित कीजिए।

उत्तर Bottom-Up में प्रत्येक ऑब्जेक्ट्स अलग-अलग समूह में होते हैं और फिर अगले step में एक ऑब्जेक्ट्स दूसरे ऑब्जेक्ट्स के साथ एक समूह में सम्मिलित होता है और ऐसा तब तक चलता रहता है जब तक किसी ऑब्जेक्ट्स एक समूह (cluster) में नहीं आ जाते हैं। इसको Agglomerative Clustering भी कहते हैं।

प्रश्न 7. Divisive hierarchical clustering (Top-Down) को परिभाषित कीजिए।

उत्तर Top-Down में सभी ऑब्जेक्ट्स एक ही समूह (cluster) में होते हैं और ये अगले step में अलग-अलग होते रहते हैं और ऐसा तब तक होता है जब तक कि सभी ऑब्जेक्ट्स अलग-अलग नहीं हो जाते है। इसको Divisive Clustering भी कहते हैं।

प्रश्न 8. Social Network की Varieties बताइए।

उत्तर "Friends" नेटवर्क के अलावा social नेटवर्क के कई उदाहरण हैं—Telephone Networks, Email Networks, Collaboration Networks, messengers आदि।

खण्ड 'ब' : लघु एवं दीर्घ उत्तरीय प्रश्न

प्रश्न 1. सोशल नेटवर्किंग क्या है?

उत्तर Social Networking एक "Network of individuals" है जो पास्परिक संबंधों जैसे दोस्त, सम्बन्धी, ग्राहक और सहकर्मी से मिलकर बनता है। उदाहरण के लिए, Social Media एक social networking service है जहाँ आप अपनी प्रोफाइल बनाकर लोगों से जुड़ते हैं और आपस में communicate और information exchange करते हैं।

ये "Social Networks" आपको लोगों से जुड़ने और सामाजिक relationship बनाने में काफी help करते हैं। Facebook, LinkedIn, Twitter और Instagram आज के सबसे popular social sites में से है। ये एक ऐसी Social Networking है, जहाँ आप virtual community बनाते हैं और online आपस में सूचना का आदान-प्रदान करते हैं।

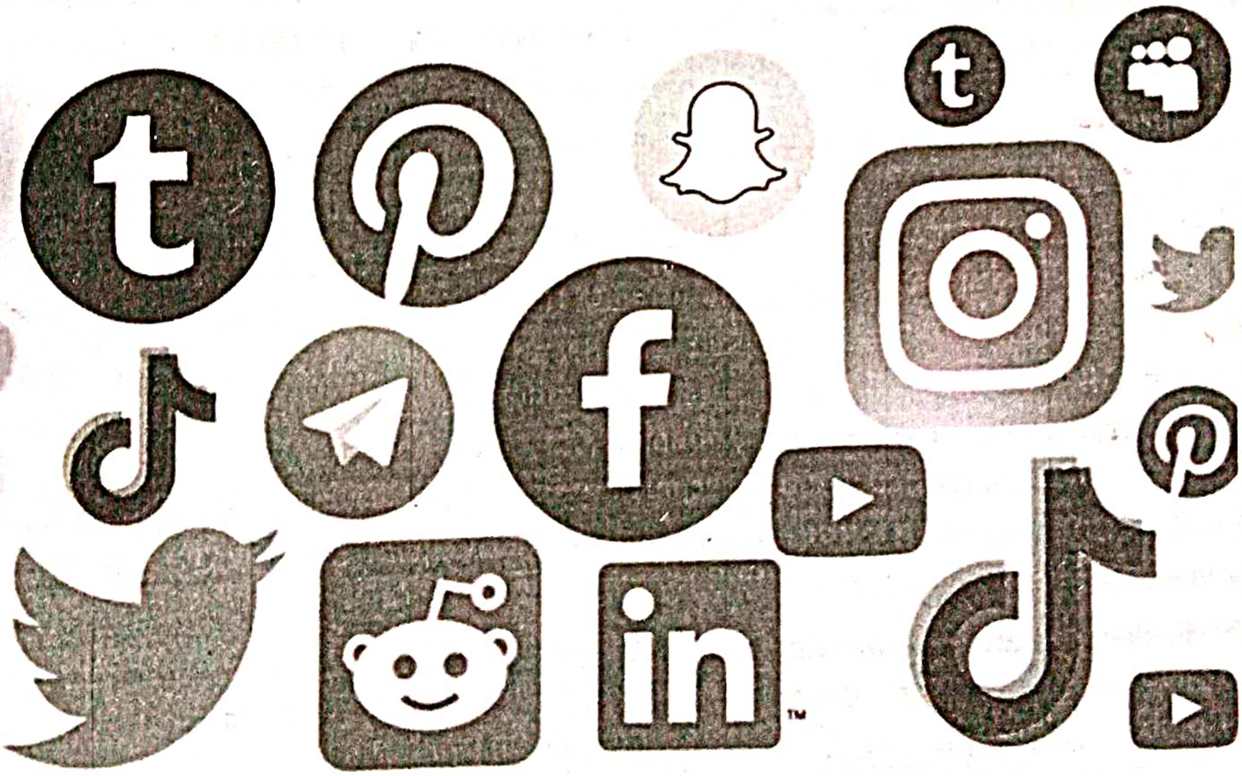
सोशल नेटवर्किंग, दोस्तों, परिवार, सहपाठियों, ग्राहकों और ग्राहकों के साथ संबंध बनाने के लिए इंटरनेट आधारित सोशल मीडिया कार्यक्रमों का उपयोग है।

प्रश्न 2. लोग सामाजिक नेटवर्क का उपयोग क्यों करते हैं?

उत्तर सामाजिक नेटवर्क लोगों को अपने दोस्तों और परिवार के साथ जुड़ा रखने में मदद करते हैं और यह पता लगाने का एक आसान तरीका है कि आपके सामाजिक मंडली में हर दिन क्या हो रहा है। इंटरनेट पर मजेदार और दिलचस्प चीजों को खोजने के लिए सोशल नेटवर्क का इस्तेमाल भी किया जा सकता है क्योंकि, अक्सर आपके मित्र और परिवार आपके जैसे ही कई हितों को साझा करेंगे।

प्रश्न 3. आज सबसे लोकप्रिय सोशल नेटवर्क क्या है?

उत्तर फेसबुक अभी भी एक अरब से अधिक उपयोगकर्ताओं के साथ सबसे बड़ा और सबसे लोकप्रिय सोशल नेटवर्क है। सोशल नेटवर्किंग साइट आज दुनिया भर में एक-दूसरे से जुड़ने का सबसे अच्छा साधन है। आज सिर्फ युवा ही नहीं बल्कि बूढ़े और बच्चे भी सोशल नेटवर्किंग साइट्स के दीवाने हैं लेकिन सोशल नेटवर्किंग साइट्स को प्रयोग करने वाले लोगों को ये नहीं भूलना चाहिए कि आपकी एक छोटी-सी गलती आपको बहुत नुकसान पहुँचा सकती है। आज सोशल नेटवर्किंग साइट्स पर धमकियों, ब्लैकमेलिंग और किडनैपिंग जैसे कई मामले सामने आ चुके हैं लेकिन इसका मतलब ये बिल्कुल नहीं है कि इन सोल नेटवर्किंग साइट्स को बंद या ब्लॉक कर दे। अगर यदि आप इन सोशन साइट्स का सुरक्षित तरीके से प्रयोग करेंगे तो ये साइट्स बिल्कुल सुरक्षित है और बहुत काम की भी।



चित्र 6.1

प्रश्न 4. सोशल नेटवर्किंग के लाभ बताइए।

उत्तर सोशल नेटवर्किंग के लाभ निम्नलिखित हैं—

(i) सोशल नेटवर्किंग के माध्यम से आप बस एक फ्रेंड रिक्वेस्ट को एक्सेप्ट करते ही एक नयी दोस्ती का दौर शुरू हो जाता है।

(ii) सोशल का प्रयोग करने से आप अकेलेपन का शिकार होने से बचते हैं।

(iii) कई बार सोशल नेटवर्किंग साइट्स आपके बिज़नेस को भी फायदा पहुँचा सकता है।

(iv) सोशल नेटवर्किंग साइट्स से आप हमेशा व्यस्त रह सकते हैं।

(v) सोशल नेटवर्किंग साइट के माध्यम से आपको ये एहसास होता है कि आपको कोई सुन रहा है।

(vi) इसके प्रयोग से आपको दोस्तों की कमी महसूस नहीं होती है।

प्रश्न 5. सामाजिक नेटवर्क के उदाहरण बताइए।

उत्तर सामाजिक नेटवर्क के उदाहरण निम्नलिखित हैं—

(i) Bebo
(ii) Classmates
(iii) Facebook
(iv) Friendster
(v) Google+
(vi) Instagram
(vii) LinkedIn
(viii) MySpace
(ix) Orkut
(x) Twitter
(xi) YouTube

प्रश्न 6. सोशल नेटवर्किंग के नुकसान क्या हैं?

उत्तर सोशल नेटवर्किंग के नुकसान निम्नलिखित हैं—

(i) सोशल नेटवर्किंग आपको मानसिक और भावनात्मक रूप से परेशान कर सकता है।

(ii) सोशल नेटवर्किंग एक प्रकार की बीमारी है। आप इसके आदि होते चले जाते हैं जोकि आपकी लिए बहुत घातक सिद्ध हो सकता है।

(iii) आपके लिए सोशल नेटवर्किंग साइट्स पर समय बिताना आपकी जिंदगी का हिस्सा बन जाता है। आप इसको इतनी ज्यादा अहमियत देने लगते हैं कि आप अपने जीवन की महत्त्वपूर्ण चीजों को भूल चुके हैं जो आपके जीवन के लिए सही नहीं है।

(iv) कई बार सोशल नेटवर्किंग साइट्स से आपका डाटा और फोटो चोरी कर लिए जाते हैं जिनका प्रयोग गलत कामों के लिए किया जा सकता है।

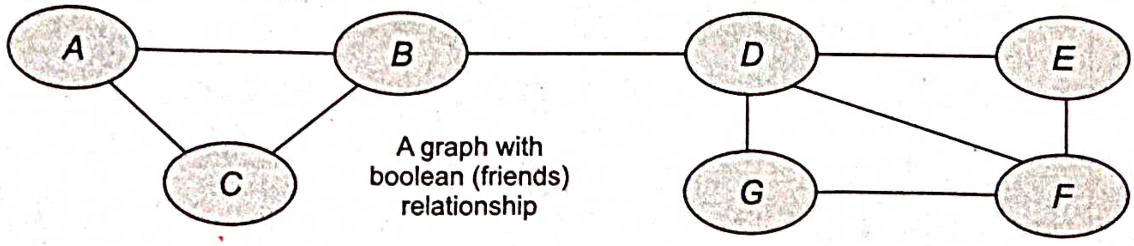
प्रश्न 7. Social Media क्या है?

उत्तर कई लोगों के मन में social networking v/s Social media में भ्रम है कि ये दोनों सामान हैं, बिल्कुल नहीं ये दोनों सामान नहीं है इनमें थोड़ा-सा अंतर है, Social media (Contents) है जिसे आप अपलोड करते हैं, चाहे वह ब्लॉग, इमेज, वीडियो, स्लाइड शो, पॉडकास्ट, समाचार-पत्र या ईबुक हो और आपने जो share किये हैं वह लोग देख सकते हैं साथ में लोग उन मीडिया को comment, like और share करते हैं।

उदाहरण, सोशल नेटवर्किंग में आप नए स्टूडेंट है और जूनियर 50 Students आपके Social networking website के जरिये friends बन गए हैं और Holiday पर ये सब फ्रेंड्स को एक इमेज भेजना है तो आप क्या करेंगे, कोई एक Social networking website का उपयोग करेंगे जैसे कि Whatsapp, Twitter, Facebook आदि और share करेंगे।

प्रश्न 8. Social Network as a graph से आप क्या समझते हैं?

उत्तर Social Network as a Graph



चित्र 6.2

- Check for the non-randomness criterion
- In a random graph (V, E) of 7 nodes and 9 edges, if XY is an edge, YZ is an edge, what is the probability that XZ is an edge?
  — For a large random graph, it would be close to $|E|/({}^{|V|}C_2) = 9/21 \sim 0.43$

— Small graph : *XY* and *YZ* are already edges, so compute within the rest
— So the probability is $(|E|-2)/(^{|V|}C_2-2)=7/19=0.37$

- Now let's computer what is the probability for this graph in particular
- For each *X*, check possible *YZ* and check if *YZ* is an edge or not
- Example : if *X* = *A*, *YZ* = {*BC*}, it is an edge

X =	**YZ =**	**Yes/Total**
A	BC	1/1
B	AC, AD, CD	1/3
C	AB	1/1
D	BE, BG, BF, EF, Eg, FG	2/6
E	DF	1/1
F	DE, DG, EG	2/3
G	DF	1/1
Total		9/16 ~ 0.56

प्रश्न 9. 2020 में सोशल-नेटवर्क के अन्य प्रकार क्या हैं?

उत्तर 2020 में सोशल-नेटवर्क के अन्य प्रकार निम्नलिखित हैं—

2020 में सोशल-नेटवर्क के अन्य प्रकार Facebook, Twitter, Instagram, linkedIn आदि हैं।

(i) "सामाजिक नेटवर्क" : Facebook, Twitter, Instagram, linkedIn आदि हैं।

(ii) लेकिन कई अन्य प्रकार भी social networks हैं।

(iii) जैसे : टेलीफोन नेटवर्क: नोड फोन नंबर हैं
यदि *A* और *B* ने पिछले एक सप्ताह या महीने, या कभी भी फोन पर बात की तो AB एक Edge है, जितनी बार फोन किया गया, उतनी बार या बातचीत का कुल समय Edges को weighted किया जा सकता है।

(iv) इसी तरह, किसी भी संदेशवाहक नेटवर्क (messanger network) जैसे: **Whatsapp, Email Network :** nodes are email addresses

प्रश्न 10. सामाजिक नेटवर्किंग सेवाओं के प्रकार क्या हैं?

उत्तर **सामाजिक नेटवर्किंग सेवाओं के प्रकार** Types of Social Networking Services सोशल नेटवर्किंग सर्विस (SNS) को आमतौर पर Internet या mobile-based social space के रूप में परिभाषित किया जाता है, जहाँ लोग connect और communicate कर सके। इन services को अपनी विशेषताओं के कारण अलग-अलग वर्गों में रखा जाता है, ये वर्ग निम्नलिखित हैं—

Profile-based SNS प्रोफाइल-आधारित सेवाओं को मुख्य रूप से members के profile pages के आस-पास organised किया जाता है। इनसे किसी individual member की information; जैसे picture और interests के बारे में पता चलता है। Facebook जैसी वेबसाइट इसका एक अच्छा उदाहरण है। अपने विचारों और सामग्री को आप comment या post के माध्यम से share करते हैं।

Content-based SNS ये ऐसे Networks होते हैं, जहाँ आप video और photography content को upload कर share कर सकते हैं और आपको उसमें बदलाव करने की पूर्णतया छूट होती है। Youtube और Flickr इसके अच्छे उदाहरण हैं जहाँ आप अपनी स्वयं की सामग्री लोगों तक पहुँचा सकते हैं। इसके अलावा इन platforms में दूसरे लोगों द्वारा शेयर की गई सामग्री को भी देखा जा सकता है।

White-label SNS अधिकांश Social Networking sites अपने users को group builiding functionality प्रदान करती है अर्थात् यहाँ आपको allow किया जाता है एक mini-communities वाले group बनाने के लिए। Ning

जैसे Social Networks पर आप खुद का एक सोशल नेटवर्क तैयार कर सकते हैं और अपने group में लोगों को invite करते हैं। ये आपके specific interest और activities को support करता है।

Multi-User Virtual Environments Sites जैसे Second life और Word of Warcraft आपको एक online virtual environment प्रदान करते हैं। ये users को एक-दूसरे के avatars (virtual representation) से interact करने की अनुमती देते हैं। मित्र सूची आमतौर पर निजी होती है और सार्वजनिक रूप से सांझा या प्रदर्शित नहीं की जाती है।

Social Search Social search engine एक महत्त्वपूर्ण web development है जिसने social network की popularity को utilize किया है, विभिन्न प्रकार के सामाजिक खोज इंजन है, लेकिन Wink और Spokeo कई Social networking sites सार्वजनिक प्रोफाइल में खोज करके परिणाम उत्पन्न करती है।

प्रश्न 11. सोशल नेटवर्क ग्राफ क्या है?

उत्तर **ग्राफ के रूप में सोशल नेटवर्क** सोशल नेटवर्क को ग्राफ के रूप में modeled किया जाता है जिसे हम कभी-कभी सोशल ग्राफ के रूप में संदर्भित करते हैं एवं entities नोड्स होती है। यदि नोड्स नेटवर्क से संबंधित है तो एक edge दो नोड्स को जोड़ता है। यदि relationship से जुड़ी कोई degree है, तो edges को लेबल करके इस डिग्री का प्रतिनिधित्व किया जाता है। अक्सर, सोशल graph undirtected होते हैं, जैसा कि फेसबुक दोस्तों के ग्राफ के लिए होता है। लेकिन वे directed किए जा सकते हैं, उदाहरण के लिए ट्विटर या Google+ पर followers के graph।

प्रश्न 12. सोशल नेटवर्क ग्राफ में क्लस्टरिंग कैसे कार्य करता है?

उत्तर Clustering of the graph is considered as a way to identify communites. ग्राफ के क्लस्टरिंग में निम्नलिखित चरण शामिल हैं—

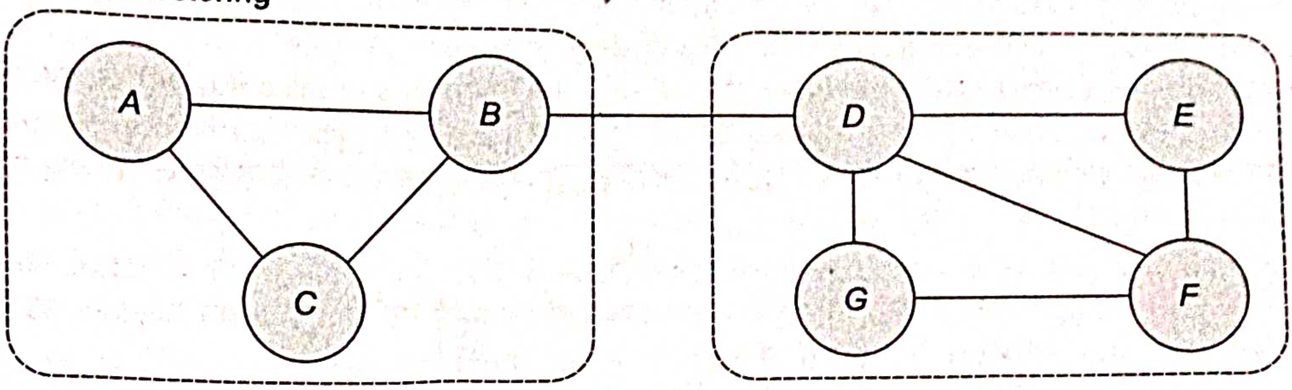
1. Distance Measures for Social-Network Graphs
2. Applying Standard Clustering Methods
   क्लस्टरिंग के लिए दो सामान्य approaches हैं—
   Hierarchical (agglomerative) and point-assignment
3. Betweenness
4. The Girvan-Newman Algorithm

प्रश्न 13. सोशल-नेटवर्क ग्राफ में क्लस्टरिंग से आपका क्या अर्थ है?

उत्तर **सोशल-नेटवर्क ग्राफ में क्लस्टरिंग—**

- Locality property → there are clusters
- Clusters are communities
  - People of the same institute or company
  - People in a photography club
  - Set of people with "Something in common" between them
- Need to define a distance between points (nodes)
- In graphs with weighted edges, different distances exist
- For graphs with "friends" or "not friends" relationship
  - Distance is 0 (friends) or 1 (not friends)
  - Or 1 (friends) and infinity (not friends)
  - Both of these violate the triangle inequality
  - Fix triangle inequality: distance = 1 (friends) and 1.5 or 2 (not friends) or length of shortest path

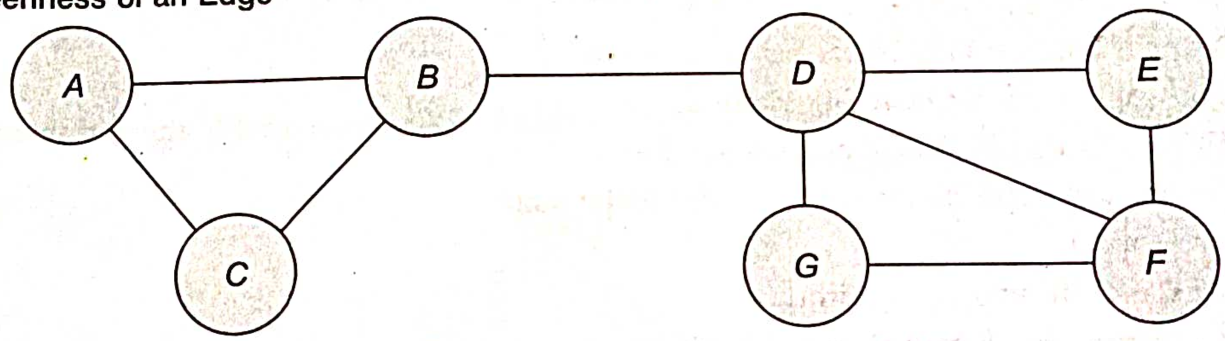
Traditional Clustering



चित्र 6.3

- Intuitively, two communities
- Traditional clustering depends on the distance
  - — Likely to put two nodes with small distance in the same cluster
  - — Social network graphs would have cross-community edges
  - — Severe merging of communities likely
- May join B and D (and hence the two communities) with not so low probability

Betweenness of an Edge



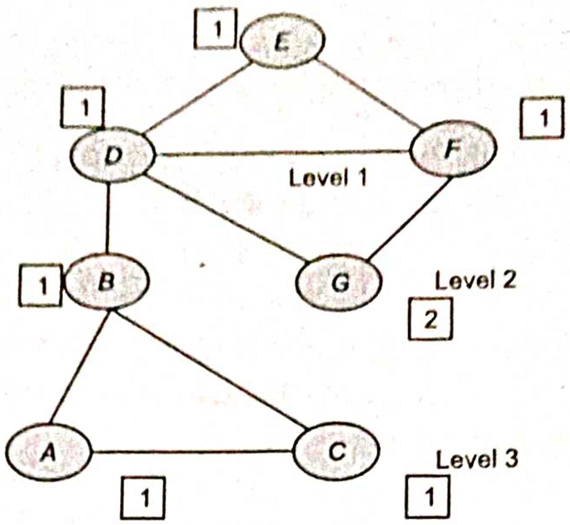
चित्र 6.4

- Betweenness of an edge AB: #of pairs of nodes (X,Y) such that AB lies on the shortest path between X and Y
  - — There can be more than one shortest paths between X and Y
  - — Credit AB the fraction of those paths which include the edge AB
- High score of betweenness means?
  - — The edge runs "between" two communities
- Betweenness gives a better measure
  - — Edges such as BD act a higher score than edges such as AB

प्रश्न 14. गिरवन-न्यून्मैन एल्गोरिथम की व्याख्या कीजिए।

उत्तर **The Girvan-Newman Algorithm :**

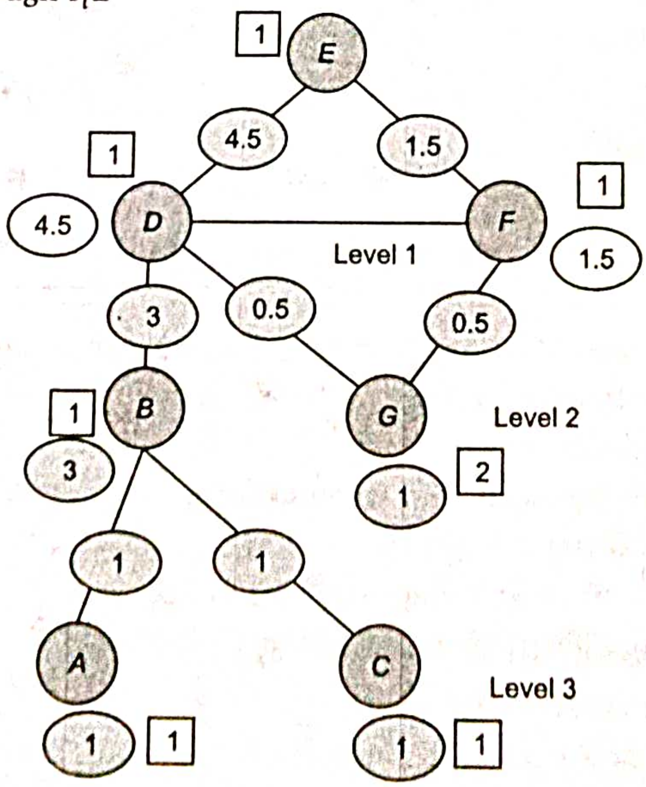
- **Step I** — BFS; Start at a node *X*, perform a BFS with *X* as root
- Observe : level of node *Y* = length of shortest path from *X* to *Y*
- Edges between level are called "DAG" edges
  - — Each DAG edge is part-of at least one shortest path from *X*
- **Step 2** — Labeling : Label each node *Y* by the number of shortest paths from *X* to *Y*



चित्र 6.5

Step 3 — credit sharing :

- Each leaf node gets credit 1
- Each non-leaf node gets 1 + sum (credits of the DAG edges to the level below)
- Credit of DAG edges: Let Y_i $(i = 1, \dots, k)$ be parents of Z, p_i = label (Y_i) credit $(Y_i, Z) = \dfrac{\text{credit }(Z) \times p_i}{(p_i + \cdots p_k)}$
- Intuition: a DAG edge Y_iZ gets the share of credit of Z proportional to the #of shortest paths from X to Z going through Y_iZ



चित्र 6.6

Finally : Repeat Steps 1, 2 and 3 with each node as root. For each edge, betweenness sum credits obtained in all iterations / 2

प्रश्न 15. समुदाय क्या है?

उत्तर समुदाय, ग्राफ के संबंध में, नोड्स के सबसेट के रूप में परिभाषित किया जा सकता है जो एक-दूसरे से घनीभूत रूप से जुड़े होते (densely connected to each other) हैं और समान रूप से एक ही ग्राफ में अन्य समुदायों के नोड्स से जुड़े होते (loosely connected) हैं।

सोशल मीडिया प्लेटफॉर्म; जैसे—फेसबुक, इंस्टाग्राम या ट्विटर, जहाँ हम अन्य लोगों से जुड़ने की कोशिश करते हैं। आखिरकार, कुछ समय बाद, हम अलग-अलग सामाजिक क्षेत्रों से जुड़े लोगों के साथ जुड़े रहते हैं। ये सामाजिक मंडल रिश्तेदारों, स्कूल के साथियों, सहकर्मियों आदि का एक समूह हो सकते हैं।

Graph में communities का पता लगाने के लिए मुख्य रूप से दो प्रकार के तरीके हैं—

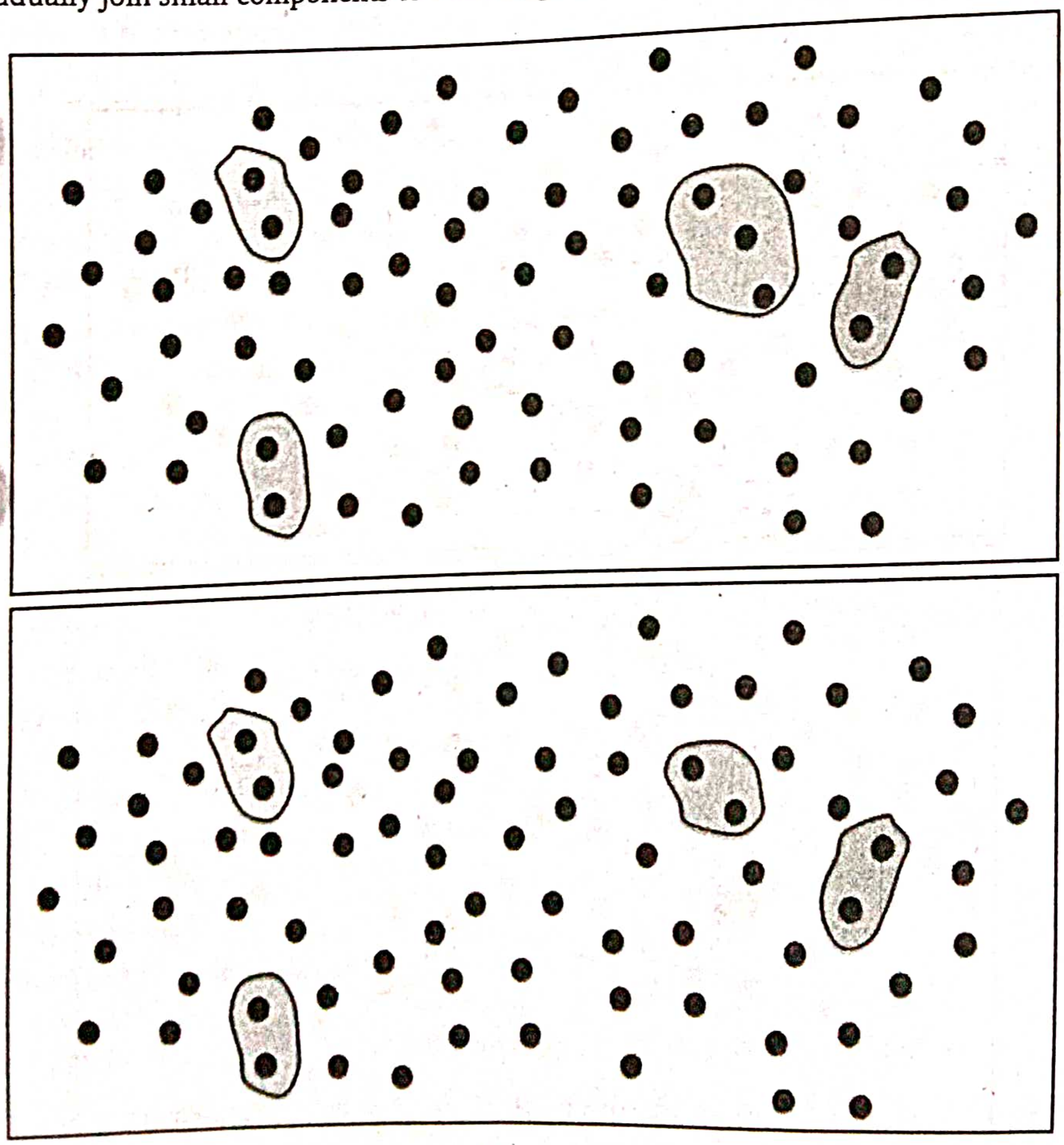
(a) Agglomerative Methods (bottom-up), (b) Divisive Methods (Top-down)

प्रश्न 16. Graph में communites की direct discovery की व्याख्या कीजिए।

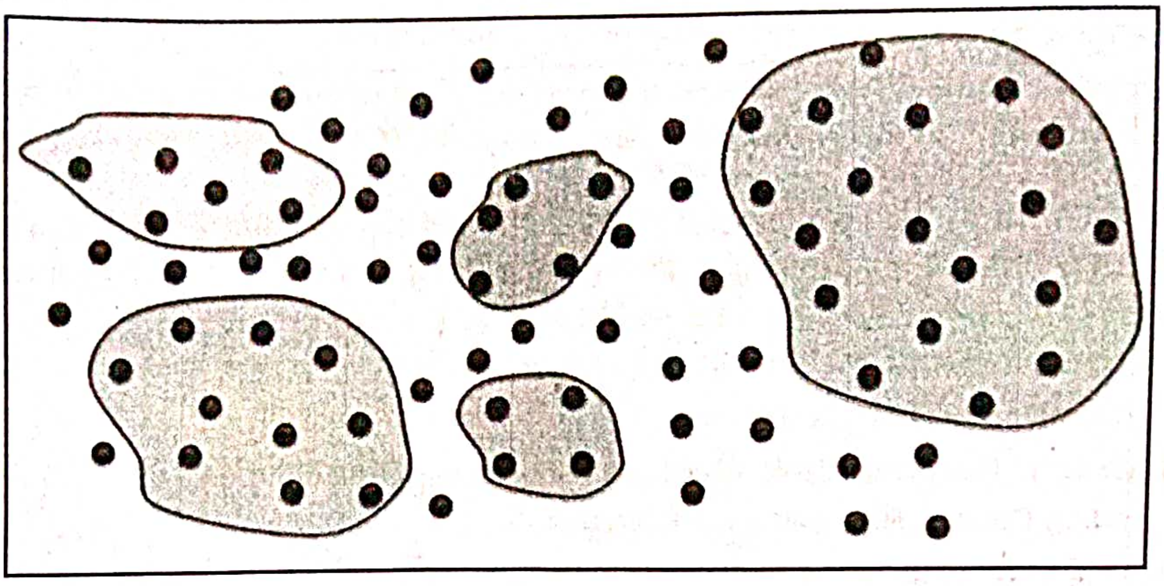
उत्तर Finding Communities using Betweenness :

Method 1: (Bottom-up Approach)

- Keep adding edges (among existing ones) starting from lowest betweenness
- Gradually join small components to build large connected components



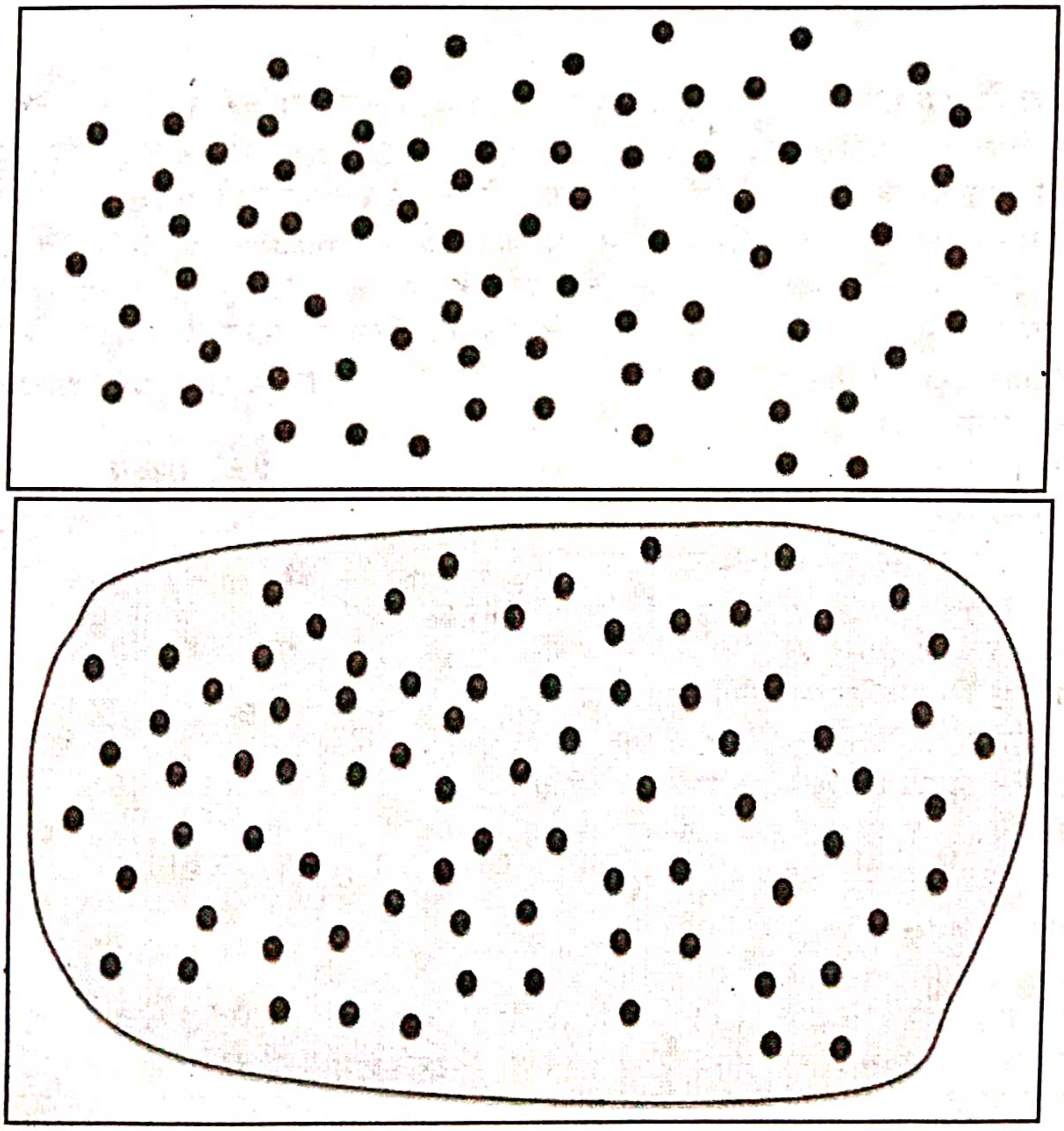
चित्र 6.7



चित्र 6.8

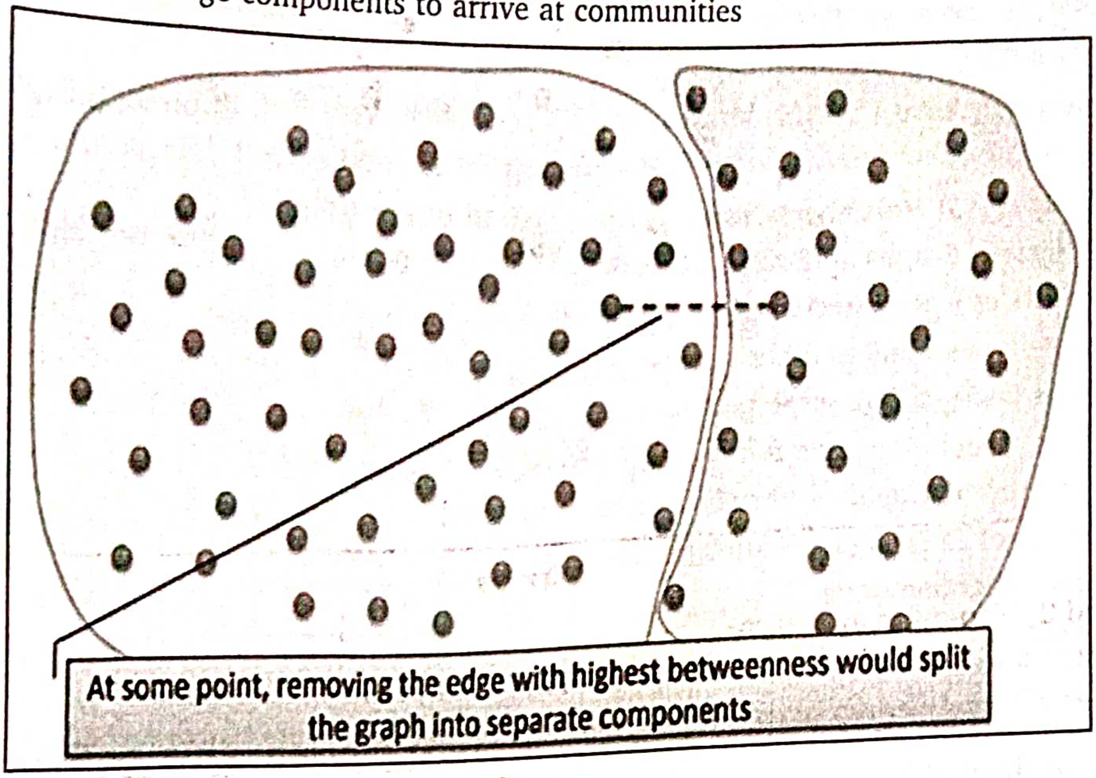
Method 2: (Top-down approach)

- Start trom all existing edges. The graph may look like one big component.
- Keep removing edges starting from highest betweenness



चित्र 6.9

- Gradually split large components to arrive at communities

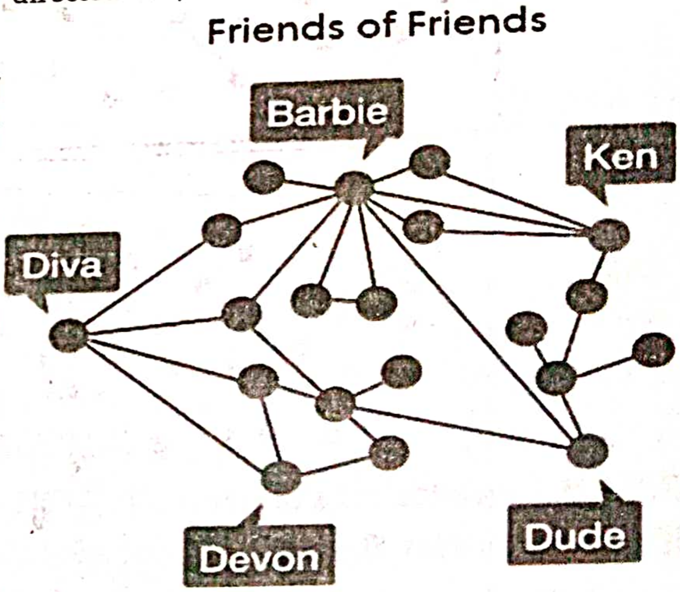


चित्र 6.10

प्रश्न 17. सोशल नेटवर्क ग्राफ क्या है? सोशल नेटवर्क ग्राफ का क्लस्टरिंग कैसे कार्य करता है?

उत्तर **ग्राफ़ के रूप में सोशल नेटवर्क** Social Network as a Graphs सोशल नेटवर्क को ग्राफ के रूप में modeled किया जाता है। जिसे हम कभी-कभी सोशल ग्राफ के रूप में संदर्भित करते हैं। entities नोड्स होती हैं। यदि नोड्स नेटवर्क से संबंधित हैं तो एक edge दो नोड्स को जोड़ता है। यदि relationship से जुड़ी कोई degree है, तो edges को लेबल करके इस डिग्री का प्रतिनिधित्व किया जाता है। अक्सर, सोशल graph undirtected होते हैं, जैसा कि फेसबुक दोस्तों के ग्राफ के लिए होता है। लेकिन वे directed किए जा सकते हैं। उदाहरण के लिए ट्विटर या Google+ पर followers के graph।

एक सामाजिक नेटवर्क ग्राफ ऐसा ग्राफ है जहाँ नोड्स लोगों को प्रतिनिधित्व करते हैं और नोड्स के बीच की रेखाएँ जिन्हें किनारे कहा जाता है, उनके बीच सामाजिक कनेक्शन का प्रतिनिधित्व करते हैं, जैसेकि दोस्ती या एक परियोजना पर एक साथ काम करना। ये रेखांकन अप्रत्यक्ष या निर्देशित हो सकते हैं। उदाहरण के लिए, फेसबुक को एक अप्रत्यक्ष ग्राफ के साथ वर्णित किया जा सकता है; क्योंकि दोस्ती द्विदिश है, ऐलिस और बॉब मित्र हैं, बॉब और ऐलिस मित्र हैं। दूसरी ओर, ट्विटर को एक निर्देशित ग्राफ के साथ वर्णित किया जा सकता है : ऐलिस बॉब के बिना एलिस का अनुसरण कर सकता है।



चित्र 6.11

Example of Social Network Graphs :

यहाँ 20 लोगों के फेसबुक से डेटा और उनके बीच पारस्परिक मित्रता कनेक्शन के साथ निर्मित एक नेटवर्क ग्राफ है। स्पष्ट रूप से बार्बी के पास अपनी पार्टियों (9) में आमंत्रित करने के लिए सबसे अधिक दोस्त हैं, लेकिन अगर निमंत्रण केवल दोस्तों के लिए ही नहीं बल्कि सभी दोस्तों के दोस्तों के लिए भी जाता है, तो किस पार्टी में सबसे अधिक आमंत्रित होंगे?

प्रत्येक बिंदु एक व्यक्ति का प्रतिनिधित्व करता है। प्रत्येक पंक्ति दोनों छोर पर लोगों के बीच पारस्परिक मित्रता का प्रतिनिधित्व करती है।

Dude may only have 3 friends himself, but his friends are very well connected. Dude is either Friends or Friends of Friends with everyone except Diva, and Devon.

प्रश्न 18. Graph में Neighbourhood properties की व्याख्या कीजिए।

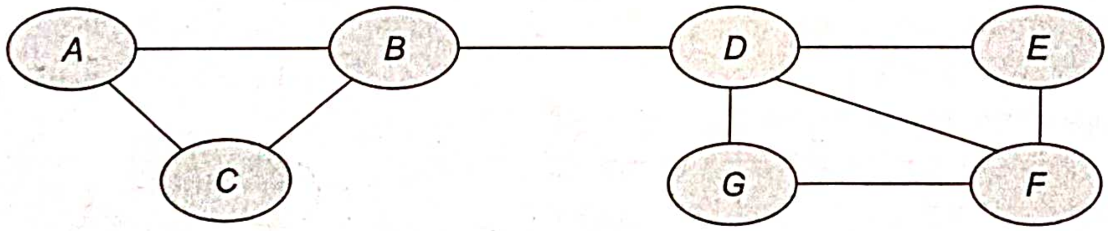
उत्तर Directed Graphs in (Social) Networks

- Set of nodes V and directed edges (arcs) $u \rightarrow v$
- The web: pages link to other pages
- Persons made calls to other persons
- Twitter, Google+: people follow other people
- All undirected graphs can be considered as directed
  — Think of each edge as bidirectional

Paths and Neighborhoods

- Path of length k : a sequence of nodes $v_0, v_1, \ldots, v_k$ from v_0 to v_k so that $v_i \rightarrow v_{i+1}$ is an arc for $i = 0, \ldots, k-1$
- Neighborhood $N(v, d)$ of radius d for a node v : sel of all nodes w such that there is a path from v to w of length $\leq d$
- For a set of nodes V, $N(V, d) := \{w \mid$ there is a path of length $\leq d$ from some v in V to $w\}$
- Neighborhood profile of a node v : sequence of sizes of its neighborhoods of radius $d = 1, 2, \ldots$; that is
  $|N(v, 1)|, |N(v, 2)|, |N(v, 3)|$

Neighborhood Profile :



चित्र 6.12

Neighborhood profile of B

$N(\text{"}B\text{"}, 1) = 4$

$N(\text{"}B\text{"}, 2) = 7$

Neighborhood profile of A

$N(\text{"}A\text{"}, 1) = 3$

$N(\text{"}A\text{"}, 2) = 4$

$N(\text{"}A\text{"}, 3) = 7$

प्रश्न 19. Hierarchical Clustering पर विस्तार से चर्चा कीजिए।

उत्तर Hierarchical Clustering : Hierarchical Clustering निम्न दो प्रकार की होती है—

(a) Bottom-Up

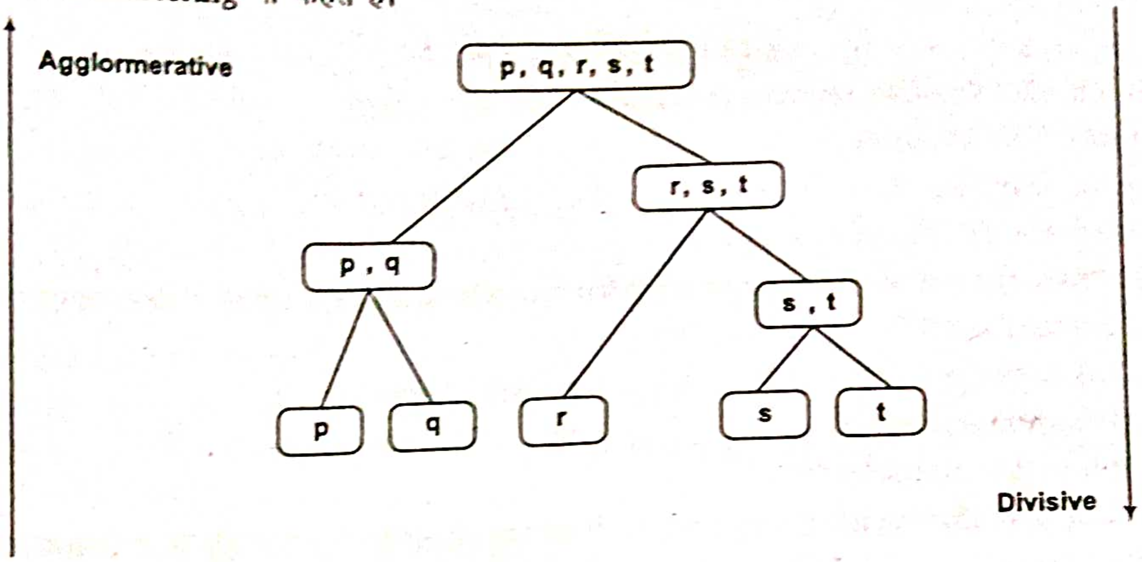
(b) Top-Down

Bottom-Up में प्रत्येक ऑब्जेक्ट्स अलग-अलग समूह में होता है और फिर अगले step में एक ऑब्जेक्ट्स दूसरे ऑब्जेक्ट्स के साथ एक समूह में सम्मिलित होता है और ऐसा तब तक चलते रहता है जब तक कि सभी ऑब्जेक्ट्स एक समूह (cluster) में नहीं आ जाते हैं।

Bottom-Up को Agglomerative clustering भी कहते हैं।

Top-Down में सभी ऑब्जेक्ट्स एक ही समूह (cluster) में होते हैं और ये अगले step में अलग-अलग होते रहते है और ऐसा तब तक होता है जब तक कि सभी ऑब्जेक्ट्स अलग-अलग नहीं हो जाते हैं।

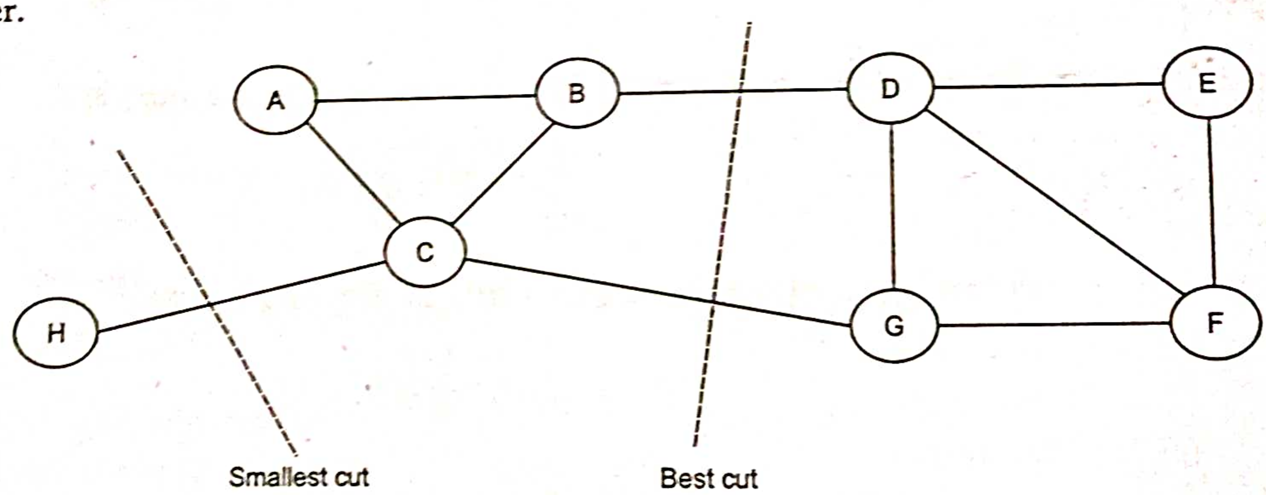
इसको Divisive Clustering भी कहते हैं।



चित्र 6.13

प्रश्न 20. Explain partitioning of a graph in detail.

उत्तर There, it is evident that the best partition put {A, B, C} in one set and {D, E, F, G} in the other.



चित्र 6.14

The smallest cut might not be the best cut

Normalized Cuts A proper definition of a "good" cut must balance the size of the cut itself against the difference in the sizes of the sets that the cut creates.

Suppose we partition the nodes of a graph into two disjoint sets S and T . Let Cut(S, T) be the number of edges that connect a node in S to a node in T . Then the normalized cut value for S and T is

$$\frac{\text{Cut }(S, T)}{\text{Vol }(S)} + \frac{\text{Cut }(S, T)}{\text{V }(T)}$$

7 डाटा साइंस और एथिकल इश्यूज
Data Science and Ethical Issues

बहुविकल्पीय प्रश्न (MCQ)

प्रश्न 1. एक डाटा सांइटिस्ट को निम्न में से किसका ज्ञान होना चाहिए?

(a) मशीन लर्निंग का
(b) डाटा माइनिंग का
(c) एनालिसिस का
(d) इन सभी का

उत्तर (d) इन सभी का

प्रश्न 2. निम्नलिखित में से कौन सोशल नेटवर्किंग प्लेटफॉर्म से संबंधित उचित नैतिक व्यवहार को संदर्भित **(refer)** करता है?

(a) Cyber law
(b) Cyber security
(c) Cyber ethics
(d) Cyber safety

उत्तर (c) Cyber ethics

प्रश्न 3. निम्न में से कौन-सा सॉफ्टेवयर कम्प्यूटर में वायरस को पहचानने व उसे दूर करने में सहायता करता है?

(a) Malware (b) Antivirus (c) Adware (d) इनमें से कोई नहीं

उत्तर (b) Antivirus

प्रश्न 4. निम्नलिखित में से कौन-सा **Antivirus program** है?

(a) Quick Heal (b) Mcafee (c) Norton (d) ये सभी

उत्तर (d) ये सभी

प्रश्न 5. निम्न में से कौन-सी तकनीक का उपयोग करके हम अपना डाटा सुरक्षित रख सकते हैं?

(a) Data Analysis
(b) Data Backup
(c) Data Mining
(d) इनमें से कोई नहीं

उत्तर (b) Data Backup

प्रश्न 6. डाटा को लम्बे समय तक सुरक्षित स्थान पर संग्रहण करने की कौन-सी प्रक्रिया है?

(a) Disposal of Data
(b) Backup
(c) Data Archiving
(d) Archival storage

उत्तर (c) Data Archiving

प्रश्न 7. **Data Security** द्वारा डिजिटल डाटा किस प्रकार बचाया जा सकता है?

(a) किसी भी unauthorized access से
(b) किसी भी प्रकार के modification से
(c) corrupt होने से
(d) उपरोक्त सभी

उत्तर (d) उपरोक्त सभी

प्रश्न 8. निम्न में से किस विषय के लोग डाटा साइंस के क्षेत्र में कार्य करते हैं?

(a) कम्प्यूटर साइंस
(b) भूगोल
(c) हिन्दी
(d) जीवविज्ञान

उत्तर (a) कम्प्यूटर साइंस

खण्ड 'अ' : अतिलघु उत्तरीय प्रश्न

प्रश्न 1. नेटवर्क सुरक्षा क्या है?

उत्तर नेटवर्क सिक्योरिटी एक ऐसा प्रोसेस (Process) होता है जिसके द्वारा किसी नेटवर्क को Unauthorized User Access (बिना इजाजत के उपयोग); जैसे—Phishing, Hacking, Trojan Horse, Spyware, Worm, Malware, आदि से बचाया जाता है। किसी नेटवर्क में नेटवर्क सिक्योरिटी को बढ़ाने के लिए हमें नेटवर्क की Monitoring करनी चाहिए एवं सॉफ्टवेयर और हार्डवेयर Components का प्रयोग करना चाहिए। जैसे—Firewall, Antvirus, आदि।

प्रश्न 2. डाटा सुरक्षा क्या है?

उत्तर किसी डिवाइस या कम्प्यूटर के सभी डाटा को किसी Unauthorized User के Access से बचाने की प्रक्रिया को Data Security कहते हैं। दुनिया भर में डाटा के अनाधिकृत उपयोग और डाटा करप्शन से डाटा को Protect करने की प्रक्रिया को डाटा सिक्योरिटी कहा जाता है।

प्रश्न 3. साइबर सिक्योरिटी क्या है?

उत्तर साइबर सुरक्षा और सुरक्षा फोर्स दोनों ही डाटा की सुरक्षा के लिए रखे जाते हैं जिससे कि किसी भी तरह से डाटा की चोरी न हो और सभी डॉक्युमेंट और फाइल सुरक्षित रहें। बड़े-बड़े कम्प्यूटर विशेषज्ञ और आईटी के प्रशिक्षित लोग इस तरह के कार्य करने में समर्थ होते हैं।

प्रश्न 4. साइबर सिक्योरिटी का दूसरा नाम क्या है?

उत्तर साइबर सिक्योरिटी को इनफार्मेशन टेक्नोलॉजी सिक्योरिटी (Information Technology Security) या इलेक्ट्रॉनिक इंफार्मेशन सिक्योरिटी (Electronic Information Security) के नाम से भी जाना जाता है।

प्रश्न 5. Data privacy important महत्त्वपूर्ण क्यों है?

उत्तर जब ग्राहक अपनी व्यक्तिगत जानकारी कम्पनियों को देते हैं, तो वे उनहें व्यक्तिगत डेटा सौंपते हैं, जिसका उपयोग उनके खिलाफ किया जा सकता है यदि यह गलत हाथों में आता है। यही कारण है कि डेटा गोपनीयता उन ग्राहकों, बल्कि कंपनियों और उनके कर्मचारियों को सुरक्षा उल्लंघनों से बचाने के लिए है।

प्रश्न 6. Data Privacy and Security क्या है?

उत्तर Security is about protecting data from malicious threats, whereas privacy is about using data responsibly.

प्रश्न 7. Data Privacy का क्या अर्थ है?

उत्तर Data privacy or information privacy data security की एक शाखा है, डेटा की उचित हैंडलिंग-सहमति, सूचना और नियामक दायित्वों से संबंधित है। विशेष रूप से, data privacy concern का है कि तीसरे पक्ष के साथ डेटा साझा किया जाता है या नहीं।

खण्ड 'ब' : लघु एवं दीर्घ उत्तरीय प्रश्न

प्रश्न 1. डेटा साइंटिस्ट क्या है?

उत्तर डाटा साइंटिस्ट का कार्य डाटा को कैप्चर करना है। जिसके लिए प्रोग्रामिंग स्किल्स और डाटाबेस स्किल्स की जरूरत होती है। डाटा साइंटिस्ट स्टेट और मैथ्स टूल के जरिए डाटा का विश्लेषण करता है, इसको वह पॉवर प्वाइंट, एक्सेल, गूगल विजुलाइजेशन के जरिए प्रस्तुत करता है।

प्रश्न 2. डाटा साइंस (डाटा विज्ञान) क्या है?

उत्तर डाटा साइंस एक तरह का ज्ञान है जिसमें हम जानकारी को एक साथ इकट्ठा करते हैं जिससे कि हम उसका बिज़नेस और आईटी रणनीतियों के कार्य में ले सकें। हम इस ज्ञान को फिर अच्छे से इकट्ठा करके उससे मूल्यवान संसाधन बनाते हैं।

डाटा साइंस जिनको आता है उनकी आज के जमाने में काफी आवश्यकता होती है क्योंकि डाटा साइंस पर काफी कंपनियाँ निर्भर हैं। ज्यादा मात्रा में डाटा की खोजबीन करने से हमें काफी काम की चीजें मिल जाती हैं और फिर उसमें से हम काम के डाटा को इकट्ठा करके अपने काम के लिए रख लेते हैं।

इससे कंपनी की मुकाबला करने की क्षमता बढ़ती है; हम डाटा साइंस में खोजबीन करते हैं। इससे कंपनी का बिज़नेस भी बढ़ता है।

डाटा साइंस जो क्षेत्र है उसमें गणित, स्टेटिस्टिक्स, और कम्प्यूटर साइंस वाले लोग कार्य करते हैं। यह मशीन लर्निंग, क्लस्टर एनालिसिस, डाटा माइनिंग जैसी तकनीकों का इस्तेमाल करते हैं।

प्रश्न 3. डाटा वैज्ञानिक कौन हैं और उनका क्या कार्य है?

उत्तर **डाटा साइंटिस्ट** Data scientist जैसे ही किसी कंपनी के बिज़नेस में डाटा बढ़ता है वैसे ही डाटा साइंटिस्टों की कंपनियों में जरूरत पड़ने लगती है और उन्हें डाटा को सही से रखने और उसकी सही से रिपोर्ट बनाने के लिए रखा जाता है। जिससे कंपनी इस डाटा को बेच सके और कुछ लाभ कमा सके और कम्पनी की प्रगति हो पाए।

डाटा साइंटिस्ट का प्रमुख कार्य रॉ डाटा को व्यवस्थित करना होता है। सामान्य रूप से डाटा को अव्यवस्थित डाटा में से निकालना होता है और उसे व्यवस्थित करना होता है जिससे की वह डाटा आगे इस्तेमाल हो सके।

इस डाटा की उसके बाद छान बिन होती है और उसमें से काम का डाटा छाँट लिया जाता है। डाटा साइंटिस्ट को मशीन लर्निंग, डाटा माइनिंग, ऐनालिटिकिस आदि का भरपूर ज्ञान होता है और कोडिंग और एल्गोरिथम लिखना भी बखूबी आता है। इसी तरह से डाटा को प्रबंधित और व्याख्या करते हुए डाटा साइंटिस्ट का काम होता है कि वह इस डाटा को इस तरह से बनाए जिससे इसको ग्राफिकली और विडियो, फोटो आदि के रूप में भी दिखा सकें। इस तरीके से डाटा को हम डिजिटली भी रख सकते हैं और बाकी की कंपनियों को बेच सकते हैं जिससे की बिजनेस में काफी इजाफा होता है।

प्रभावी होने के लिए डाटा साइंटिस्ट के अंदर शिक्षा के साथ-साथ भावुक बुद्धि और डाटा एनालिटिक्स का ज्ञान भी भरपूर होना चाहिए। सबसे महत्त्वपूर्ण जो कौशल होता है किसी डाटा साइंटिस्ट में वह यह है कि वह किस तरह से डाटा को रख रहा है और लोगों को समझा पा रहा है और कितने अच्छे तरीके से दर्शा पा रहा है कि इसमें कार्य कैसे होता है।

यह भी जरूरी होता है कि वह अच्छे सॉफ्टवेयर इस्तेमाल कर रहा हो और डाटा का महत्त्व भी बता रहा हो। डाटा साइंटिस्ट डिजिटल जानकारी को चैनल और स्रोतों से बनाते हैं; जैसे—स्मार्ट फोन इंटरनेट ऑफ थिंग्स (IOT) डिवाइस, सोशल मीडिया, सर्वे, इंटरनेट सर्च, खरीददारी। डाटा साइंटिस्ट बहुत सारे डाटा सेट्स में से ऐसे पैटर्न को निकालते हैं जिससे यह डाटा एनालिसिस के द्वारा आसानी से सुलझाया जा सकें। इस प्रक्रिया को हम डाटा माइनिंग भी बोलते हैं।

प्रश्न 4. डाटा साइंस के फायदे बताइए।

उत्तर **डाटा साइंस के फायदे** Benfits of data science डाटा साइंस बिज़नेस के निर्णय लेने में काफी काम आता है। यह डाटा को बड़े ही सही तरीके से इस्तेमाल करता है और उसे उपयोगी बनाता है जिससे की हम उसे इस्तेमाल कर सकें।

डाटा से जो हम निर्णय लेते हैं वह हमें काफी लाभ देता है और कार्य करने की क्षमता को भी बढ़ा देता है। डाटा साइंस लोगों की भर्ती में भी काफी काम आता है जैसाकि जो लोग आगे की स्टेज के लिए चुने गए हैं तो उनको भी डाटा साइंस को इस्तेमाल करके इसी तरीके से छाँटा जाता है।

डाटा से एप्टिट्यूड टेस्ट लेना और गेमस, कोडिंग आदि ह्यूमन रिसोर्स के लोगों के लिए काफी उपयोगी होते हैं क्योंकि इससे वे लोगों को कंपनी में लेते हैं।

प्रश्न 5. डाटा साइंस के उपयोग बताइए।

उत्तर **डाटा साइंस के उपयोग** Application uses of data science डाटा साइंस के फायदे कंपनी के लक्ष्य और संसाधानों पर भी निर्भर करते हैं कि कंपनी किस तरह का काम करती है और किस तरह से संसाधनों को इस्तेमाल करती है। सेल्स और मार्केटिंग डिपार्टमेंट पर भी कंपनी का फायदा निर्भर करता है। उदाहरण के तौर पर हम यह देख सकते हैं कि कुछ कंपनी उपयोगकर्ताओं के डाटा को खरीदती हैं और फिर उसका विश्लेषण करती हैं।

डाटा को सही तरीके से समझा जाता है और उसके बाद उसकी उचित रिपोर्ट बनायी जाती है और फिर कंपनी में इसका पूरा विचार विमर्श होता है, जिससे इस डाटा को प्रभावी बनाया जा सके। यह कैम्पेन करने में भी काफी उपयोगी होता है।

नेट्फ्लिक्स में भी डाटा पर निर्भर करने वाली एल्गोरिथम इस्तेमाल होती है जोकि उपयोगकर्ता का इतिहास बताती है कि उसने पहले नेट्फ्लिक्स में क्या देखा था। डाटा साइंस बहुत ही उभरता हुआ क्षेत्र है और तकनीकी दुनिया में आने वाले समय में यह काफी तरक्की करेगा और हम पूरी तरह इस पर निर्भर होंगे।

मशीन लर्निंग की चीजें भी डाटा साइंस में उपयोग होती हैं जैसे की इमेज रेकोगनिशन और स्पीच रेकोगनिशन।

प्रश्न 6. Cyber Ethics (साइबर एथिक्स) क्या है?

उत्तर Cyber ethics (साइबर नैतिकता) कम्प्यूटर से सम्बन्धित ethics का एक दार्शनिक अध्ययन (study) है, यह users के behavior को कवर करता है तथा कम्प्यूटरों को किस काम को करने के लए program किया गया है और यह किसी समाज या व्यक्ति को कैसे प्रभावित करता है ये सब इसमें शामिल होता है।

सरल शब्दों में कहें तो, "साइबर एथिक्स का मतलब है इन्टरनेट का प्रयोग जिम्मेदारी से करना," वर्षों तक, बहुत सारी सरकारों ने नियम बनाए हैं, जबकि संगठनों ने cyber ethics की policies को explain किया है।

आजकल इन्टरनेट का प्रयोग बहुत सारें बच्चे करने लग गये है, तो इन बच्चों को इसके खतरों के बारें में बताना पहले से बहुत जरुरी हो गया है, teenagers (किशारों) से बात करना बहुत कठिन होता है क्योंकि वे किसी का lecture ज्यादा सुनना पसंद नहीं करते है, उन्हें लगता है कि वे इसे हर तरह से सुलझा लेंगे। यही कारण है कि कम उम्र में उपयुक्त cyber ethics को स्थापित करना महत्त्वपूर्ण है।

1. **Copyrighting or Downloading** कॉपीराइट और डाउनलोडिंग एक बहुत बड़ा issue है क्योंकि ज्यादातर लोग खासकर बच्चे इसके बारें में नहीं जानते हैं, लोग किसी भी चीज जो उन्हें चाहिए उसे सर्च करते है और डाउनलोड कर लेते हैं, वे सोचते हैं कि ऐसा सभी लोग करते हैं, लेकिन यह गलत है और दूसरे व्यक्ति की सामग्री को copy करके इन्टरनेट में पब्लिश करना गलत है।
2. **Hacking (हैकिंग)** Hacking का मतलब है किसी वेबसाइट से बिना अनुमति के किसी private information, या password को चुरा लेना, आजकल hackers की संख्या बहुत बढ़ गयी है, और इन्हें रोकना बहुत जरुरी है। ये वायरस को बनाते हैं और वेबसाइट या सिस्टम को हैक कर लेते हैं, तो हमें पता होना चाहिए कि यह बहुत बड़ा crime है ओर इससे हमें सावधान रहना चाहिए और बचना चाहिए।
3. **Cyberbullying** Cyberbullying का अर्थ है इन्टरनेट में किसी व्यक्ति को धमकी देना या उसे गाली देना या उसे अपशब्द कहना या उसे परेशान करना।

   Cyber bullying किसी भी social media साइट्स जैसे—फेसबुक, या अन्य प्लेटफार्म जैसे : youtube में हो सकती है।

   इसलिए हमें अपने बच्चों को यह बताना चाहिए कि इससे कैसे बचे, क्योंकि यह बहुत dangerous हो सकता है। जब भी किसी के साथ Cyberbullying होती है तो उसे सबसे पहले किसी करीबी को बताना चाहिए और ऐसे लोगों से दूर रहना चाहिए जो Cyberbullying करते हैं।
4. हमें किसी दूसरे का पासवर्ड use नहीं करना चाहिए। इन सब बातों का ध्यान हमें रखना चाहिए जिससे इन्टरनेट की दुनिया अच्छी बन सकें और यहाँ crime कम-से-कम हों।

प्रश्न 7. Data Security क्या है?

उत्तर **Data Security** Data Security से अर्थ डाटा की सुरक्षा से है, यानि Data Security एक ऐसी प्रक्रिया है, जिसमें डिजिटल डाटा यानि कम्प्यूटर द्वारा तैयार डाटा को किसी भी Unauthorized Access, Modification या Corrupt होने से बचाया जा सकता है।

डाटा सिक्योरिटी के अंतर्गत वह सभी जरुरी कदम उठाए जाते हैं, और उन सभी तकनीकों का इस्तेमाल किया जाता है, जिनकी मदद से कम्प्यूटर डाटा को सुरक्षा प्रदान की जा सके।

प्रश्न 8. डाटा सिक्योरिटी क्यों जरुरी है?

उत्तर आज के समय में प्रत्येक छोटा-बड़ा कार्य चाहे वह Personal हो या Professional, कम्प्यूटर और ऑनलाइन माध्यमों द्वारा पूरा किया जाता है, जिसमें सारा कार्य डिजिटल डाटा के रूप में Save रहता है।

अब जब सारा कार्य ही डिजिटल डाटा के रूप में Save रहता है, तो जरुरी है कि उसे सुरक्षित भी रखा जा सके। तब ऐसे में डाटा की सुरक्षा को Data Security Solutions के द्वारा Safe रखा जाता है।

एक Organisation या बिज़नेस के लिए अपना डाटा सबसे महत्त्वपूर्ण होता है, जिसमें बहुत ही महत्त्वपूर्ण जानकारी हो सकती है, जैसे Customer Information, Technology Information या ऐसी किसी भी प्रकार की जानकारी जो एक Organisation कभी भी Open नहीं करना चाहेगी।

इस प्रकार के Data को सुरक्षित रखना काफी महत्त्वपूर्ण हो जाता है, जिसके लिए Data Security से जुड़े तरीकों का इस्तेमाल किया जाता है। ताकि डाटा पर होने वाले किसी भी प्रकार के Cyber Attack या Data Corruption से बचा जा सके।

प्रश्न 9. डाटा सिक्योरिटी के प्रकार कौन-कौन से है?

उत्तर **सिक्योरिटी के प्रकार** जैसाकि आपने ऊपर पढ़ा हर किसी के लिए डाटा कितना अधिक महत्त्वपूर्ण होता है, और डाटा सिक्योरिटी की मदद से डाटा को कैसे सुरक्षित रखा जा सकता है। अब जानते है, Data Security से जुड़ी कुछ Important तकनीक जिन्हें Implement करके डाटा को सुरक्षा प्रदान की जा सकती है।

Data Backup यह डाटा सिक्योरिटी का सबसे शुरुवाती और महत्तवपूर्ण प्रकार है, क्योंकि समय-समय पर डाटा का बैकअप लेना और उसे सुरक्षित रखना बहुत जरुरी है ताकि किसी भी डाटा को Corrupt या डिलीट जैसी स्थिति से बचाया जा सके।

Data Encryption डाटा को एन्क्रिप्ट करने पर वह एक ऐसे कोड में परिवर्तित हो जाता है, जिसे किसी के लिए भी समझ पाना असंभव है और डाटा को फिर से देखने के लिए उसे डिक्रिप्ट करना पड़ता है। इसके लिए आपके पास डिक्रिप्ट करने का कोड होना अनिवार्य होता है। डाटा एनक्रिप्शन करने के लिए कई सॉफ्टवेयर उपलब्ध है, जिनका इस्तेमाल उपलब्ध हैं, जिनका इस्तेमाल किया जा सकता है।

Data Masking डाटा मास्किंग एक ऐसी टेक्निक है जिसमें बिल्कुल असल डाटा की तरह ही एक हूबहू डाटा Structure तैयार किया जाता है, ताकि किसी प्रकार की टेस्टिंग करनी हो तो वह की जा सके। ऐसा करने पर असल डाटा को सुरक्षित रखा जा सकता है और किसी भी ऐसी स्थिति में जहाँ असल डाटा की जरूरत नहीं है वहाँ इसके तरीके को इस्तेमाल किया जा सकता है।

प्रश्न 10. डाटा की गोपनीयता पर चर्चा करें।

उत्तर **Discussion on Data Privacy** सोशल मीडिया और इंटरनेट से लैस डिजिटल दुनिया में कोई भी आपकी निजी जिंदगी और प्राइवेसी तक आसानी से पहुँच सकता है। आज के आधुनिक समय में सभी महत्त्वपूर्ण जानकारियाँ आपके फोन, कम्प्यूटर या लैपटॉप में मौजूद होती हैं। एक तरफ जहाँ ये सब आपके जीवन को गति प्रदान करता है। वहीं दूसरी तरफ एक हल्की सी चूक से ये ऑनलाइन मौजूद डाटा आपके लिए खतरा भी बन जाता है।

व्यक्तिगत जानकारी को सुरक्षित रखने के कुछ उपाय—

1. आपको यह ध्यान रखना होगा कि आप अपने कम्प्यूटर, लैपटॉप, टेबलेट और स्मार्टफोन में उस समय का लेटेस्ट अर्थात् सॉफ्टवेयर का इस्तेमाल करें।
2. अपने एप्लीकेशन को नियमित रूप से सुरक्षा सुधारों के लिए अपडेट करें।
3. सुरक्षा के लिए एंटी-मैलवेयर सॉफ्टवेयर का उपयोग करें और एन्क्रिप्शन के साथ व्यक्तिगत डाटा की सुरक्षा करें।
4. अपने सभी पासवर्ड पर नियंत्रण रखने के लिए एक पासवर्ड मैनेजर का उपयोग करें ताकि आप एक ही पासवर्ड का दो बार उपयोग न कर सके। इसके अलावा लम्बे पासवर्ड की कोशिश करें क्योंकि इन्हें तोड़ना आसान नहीं होता है।

5. यह ध्यान रखें कि आपके फोन में लोकेशन ट्रैकिंग हमेशा ओपन न रहे।
6. अपने जन्मदिन सहित, सोशल मीडिया पर आपके द्वारा डाली गई सभी व्यक्तिगत जानकारी के बारे में सावधान रहे।
7. फोन में टेक्स्ट मैसेज या ईमेल को पढ़ने से पहले या उसका रिप्लाई करने से पहले उसका सोर्स देख लें। इसमें वायरस होने की आशंका होती है।
8. कभी भी असुरक्षित वाईफाई का इस्तेमाल न करें।
9. किसी भी नए डिवाइस (मोबाइल, लैपटॉप, इत्यादि) पर तुरंत अपने प्राइवेसी सेटिंग दर्ज करें।
10. अपने एटीएम, क्रेडिट कार्ड, बैंक स्टेटमेंट का नियमित रूप से जाँच करें, कुछ भी संदिग्ध लगने पर अपने बैंक को तुरंत संपर्क करें।

प्रश्न 11. डाटा वैज्ञानिकों की अगली पीढ़ी कैसी दिखेगी?

उत्तर **Next Generation of Data Scientists** Next-Generation data scientists should be encouraged to become good problem solvers who follow the scientific method, to think deeply about the appropriate use of the data science process, and to use data responsibly and for the common good.

The next generation of data scientist will maintain a breadth of hard technical skills such as mathematics, statistics, probability theory, machine learning, coding, data visualization, and data storytelling. Coding is important, so a good foundation in writing code along with good coding practices like agile software development techniques, code reviews, debugging and version control are particularly valuable.

"The Next Generation Scientists will go beyond just crunching data, creating a model and passing it to the end users."

Few critical trends for next-generation scientists :

- Data Quality
- Automation
- Communication
- Domain Knowledge

प्रश्न 12. साइबर सिक्योरिटी और नेटवर्क सिक्योरिटी में क्या अंतर है?

उत्तर यह एक ऐसा सुरक्षा है जो किसी नेटवर्क से जुड़े कम्प्यूटरों के लिए इस्तेमाल किया जाता है। यह कम्प्यूटर, हार्डवेयर, सॉफ्टवेयर, सूचना और डेटा को साइबर अपराध से बचाने का कार्य करता है। साइबर सिक्योरिटी हैकर्स के Attacks से कम्प्यूटर सिस्टम को सुरक्षा प्रदान करने का कार्य करता है।

आसान शब्दों में कहे तो, साइबर सिक्योरिटी इंटरनेट से जुड़े सभी कम्प्यूटर सिस्टम या नेटवर्क को डिजिटल हमलों से बचाने का कार्य करते हैं।

साइबर सुरक्षा और नेटवर्क सुरक्षा को एक ही सिक्के के दो पहलू माना जाता है। कुछ विशेषज्ञों का कहना है कि नेटवर्क सुरक्षा साइबर सुरक्षा का एक डप डोमेन है। इसमें थोड़ा-सा अलग है जो इस प्रकार है—

नेटवर्क सुरक्षा आंतरिक जानकारी या गतिविधियों कीं सिक्योरिटी के लिए लागू किया जाता है जबकि साइबर सुरक्षा बाहरी खतरों पर ध्यान देता है जैसेकि हैकर्स से सिस्टम को protect करना।

मॉडल प्रश्न-पत्र

डाटा साइंस और मशीन लर्निंग

Data Science and Machine Learning

समय 2.30 घण्टे] **[पूर्णांक : 50**

नोट : (i) सभी प्रश्नों के उत्तर दीजिए।

प्रश्न 1. निम्न में से किन्हीं दो भागों के उत्तर दीजिए— **[2 × 5 = 10]**

(अ) उदाहरण के साथ आर्टिफिशियल इंटेलिजेंस की व्याख्या कीजिए।

(ब) डाटा माइनिंग क्या है? इसके लक्ष्यों को भी परिभाषित कीजिए।

(स) Artificial Intelligence और Machine learning में क्या अन्तर है?

प्रश्न 2. निम्न में से किन्हीं दो भागों के उत्तर दीजिए— **[2 × 5 = 10]**

(अ) Min max normalization से आपका क्या तात्पर्य है?

(ब) Label Encoding से आप क्या समझते हैं?

(स) पैरामीट्रिक सांख्यिकी क्या है?

प्रश्न 3. निम्न में से किन्हीं दो भागों के उत्तर दीजिए— **[2 × 5 = 10]**

(अ) Exploratory Data Analysis के Key Concept बताइए।

(ब) Data Science Process से आप क्या समझते हैं?

(स) Subpervised Machine Learning के चरण क्या है?

प्रश्न 4. निम्न में से किन्हीं दो भागों के उत्तर दीजिए— **[2 × 5 = 10]**

(अ) प्रतिगमन (Regression) के प्रकार से आप क्या समझते हैं?

(ब) K-Nearest Neighbour (KNN) एल्गोरिथम के लाभ और हानि पर चर्चा कीजिए।

(स) क्लस्टरिंग तकनीक के प्रकार बताइए।

प्रश्न 5. निम्न में से किन्हीं दो भागों के उत्तर दीजिए— **[2 × 5 = 10]**

(अ) सोशल नेटवर्किंग के लाभ बताइए।

(ब) गिरवन-न्यूमैन एल्गोरिथम की व्याख्या कीजिए।

(स) Cyber Ethics (साइबर एथिक्स) क्या है?