Internal Assessment Test 1 – June 2024

Sub:	Data Science and Visualization					Sub Code:	21CS644	Branch:	ISE
Date:	06/6/2024	Duration:	90 min's	Max Marks:	50	Sem/Sec:	VI / A, B		OBE

	Answer any FIVE FULL Questions	MARKS	CO	RBT
1 a)	What is Datafication? Discuss with examples.	6M	CO1	L1
1 b)	Clarify the work of the Data Scientist in academia and industry.	4M	CO1	L2
2	Discuss the following concepts with an example: i. Statistical inference ii. Population iii. Samples iv. Types of Data.	10M	CO1	L2
3	What is data science? Is it new, or is it just statistics or analytics rebranded? Is it real, or is it pure hype? And if it's new and if it's real, what does that mean?	10M	CO1	L2
4	Implement the Brooklyn housing data.	10M	CO2	L5
5	Briefly explain Data Science Process with a neat diagram.	10 M	CO2	L2
6 a)	Taking k=3, compute the class of a new data entry using KNN given brightness=20, Saturation=35, given the set of data points (40,20, Red), (50,50, Blue),(60,90, Blue), (10,25, Red), (70,70, Blue), (60,10, Red) and (25,80, Blue) where the first attribute is brightness and second is saturation.	5 M	CO2	L3
6 b)	Apply K(=2) – Means algorithm over the data (185,72), (170,56), (168,60), (179,68),(182,72),(188,77) up to two iterations and show the clusters. Initially choose first two objects as initial centroids.	5 M	CO2	L6

USN | | | | | | | | | | |

Faculty Signature CCI Signature HOD Signature

SOLUTION

Ans-1(a) Datafication refers to the process of transforming various aspects of human life, activities, and interactions into data that can be stored, analyzed, and utilized for various purposes. This concept has gained prominence with the proliferation of digital technologies and the increasing amount of data generated in today's world. Here's a discussion with examples:

1. **Personal Datafication**: Personal activities and behaviors are increasingly being tracked and recorded through digital devices and platforms. For example, fitness trackers like Fitbit collect data on users' physical activities, sleep patterns, and heart rates. Social media platforms gather vast amounts of data on users' preferences, interactions, and behaviors to tailor content and advertisements.
2. **Smart Cities**: Cities are becoming increasingly "smart" through datafication. Sensors embedded in urban infrastructure collect data on traffic patterns, energy consumption, air quality, and more. For instance, traffic management systems use data from cameras and sensors to optimize traffic flow and reduce congestion in real-time.
3. **Healthcare Datafication**: Electronic health records (EHRs) digitize patient information, making it easier to store, access, and analyze medical data. Wearable devices and health monitoring apps track individuals' health metrics, allowing for personalized healthcare interventions and early detection of health issues.
4. **Business Datafication**: Companies collect and analyze vast amounts of data to gain insights into customer behavior, market trends, and operational efficiency. For example, e-commerce platforms analyze customer browsing and purchasing behavior to recommend products and personalize marketing campaigns. Supply chain management systems use data analytics to optimize inventory levels and streamline logistics.
5. **Education Datafication**: Educational institutions are increasingly adopting data-driven approaches to improve learning outcomes and personalize instruction. Learning management systems (LMS) track students' progress and engagement, enabling educators to identify areas for improvement and tailor teaching methods accordingly.
6. **Financial Datafication**: Financial institutions leverage data analytics to assess credit risk, detect fraudulent activities, and personalize financial products and services. For instance, credit scoring models analyze individuals' financial histories and behaviors to determine their creditworthiness and set interest rates.

Overall, datafication is transforming various aspects of society and business, offering opportunities for innovation, efficiency improvements, and better decision-making.

Ans-1 (b) **Data Scientist in Academia:**

1. **Research and Analysis**
2. Data Collection and Curation:
3. **Model Development:**
4. Publication and Communication:
5. Grant Writing:

Data Scientist in Industry:

- **Business Insights:** Data scientists in industry focus on extracting actionable insights from data to drive business decision-making and strategy. They work closely with stakeholders from different departments, such as marketing, finance, operations, and product development.

- **Data Cleaning and Preparation:** A significant portion of a data scientist's time in industry is spent on data cleaning, preprocessing, and integration. They ensure that data from various sources are cleansed, standardized, and prepared for analysis.

- **Model Building and Deployment:** Industry data scientists develop predictive models, recommendation systems, or optimization algorithms to solve specific business problems. They leverage machine learning, deep learning, and other analytical techniques to extract patterns and insights from large datasets.

- **Performance Monitoring and Optimization:** After deploying models into production, data scientists monitor their performance, evaluate their effectiveness, and fine-tune parameters as needed. They may also develop monitoring systems to detect anomalies or drifts in data over time.

- **Collaboration and Cross-functional Teams:** Industry data scientists collaborate with cross-functional teams, including software engineers, business analysts, and domain experts, to translate data-driven insights into actionable strategies and solutions. They communicate complex technical concepts to non-technical stakeholders effectively.

- **Product Development:** In some industries, data scientists are directly involved in the development of data-driven products or services. They contribute to product ideation, prototype development, and iterative improvement based on user feedback and market demand.

Ans-2 (a) Discuss with example:

i. Statistical inference
ii. Population
iii. Samples
iv. Types of Data

Statistical inference- Statistical inference is the process of drawing conclusions or making predictions about a population based on sample data from that population. It involves using statistical methods to analyze the characteristics of a sample and then generalize the findings to the larger population from which the sample was drawn.

Example- Suppose a pharmaceutical company develops a new drug intended to lower blood pressure in patients with hypertension. The company wants to determine whether the new drug is more effective than the existing standard treatment.

Here's how statistical inference can be applied in this scenario:

1. **Null Hypothesis (H0):** The null hypothesis states that there is no difference in the effectiveness of the new drug compared to the standard treatment. Symbolically, it can be represented as H0: $\mu_{new} = \mu_{standard}$, where μ_{new} is the population mean blood pressure reduction with the new drug, and $\mu_{standard}$ is the population mean blood pressure reduction with the standard treatment.
2. **Alternative Hypothesis (H1):** The alternative hypothesis contradicts the null hypothesis and states that the new drug is more effective than the standard treatment. Symbolically, it can be represented as H1: $\mu_{new} < \mu_{standard}$, indicating that the population mean blood pressure reduction with the new drug is less than that with the standard treatment.
3. **Sample Collection:** The pharmaceutical company conducts a clinical trial where a sample of patients with hypertension is randomly assigned to receive either the new drug or the standard treatment. Blood pressure measurements are taken before and after the treatment period for each patient.
4. **Data Analysis:** After collecting the data, the company calculates the sample means and standard deviations for both groups (new drug and standard treatment). They then perform a statistical test, such as a t-test or z-test, to determine whether the difference in mean blood pressure reduction between the two groups is statistically significant.
5. **Inference:** Based on the results of the statistical test, the company can either reject or fail to reject the null hypothesis. If the p-value is less than a predetermined significance level (e.g., 0.05), they reject the null hypothesis and conclude that there is sufficient evidence to suggest that the new drug is more effective than the standard treatment. If the p-value is greater than the significance level, they fail to reject the null hypothesis, indicating that there is not enough evidence to conclude that the new drug is more effective.
6. **Conclusion:** If the null hypothesis is rejected, the pharmaceutical company may proceed with further development and testing of the new drug, potentially leading to its approval for widespread use. If the null hypothesis is not rejected, the company may need to reconsider the efficacy of the new drug or explore alternative approaches.

ii) Population- The population is the entire collection of units (people, animals, objects, etc.) that the researcher wants to generalize their findings to. It represents the larger group to which the study's results will be applied.

Example- Suppose a researcher is interested in studying the average income of all households in a particular city. In this case:

Population: The population would consist of all households in that city. This includes every household, regardless of size, income level, or any other characteristics.

iii) Sample- A sample refers to a subset of individuals, items, or observations selected from a larger group or population. This subset is chosen to represent the characteristics of the larger population accurately. Samples are often used in research and data analysis because it is usually impractical or impossible to collect data from every member of the population of interest.

Example: Survey on Smartphone Usage

Suppose a market research firm wants to conduct a survey to understand smartphone usage habits among college students in a particular city. Here's how they might go about it:

1. **Population:** The population consists of all college students in the city who own smartphones. This includes students from different colleges, majors, ages, and backgrounds.
2. **Sampling Frame:** The sampling frame is a list of all colleges and universities in the city. The market research firm obtains this list from local educational institutions or databases.

iv) Types of data- In statistics and data analysis, data can be classified into various types based on their nature, characteristics, and measurement scales. The main types of data include:

1. **Nominal Data:** Nominal data, also known as categorical data, represent categories or labels with no inherent order or numerical value. Examples include gender (male, female), marital status (single, married, divorced), and type of car (sedan, SUV, truck). Nominal data can be represented using numbers, but the numbers have no mathematical significance.
2. **Ordinal Data:** Ordinal data represent categories with a natural order or ranking. While the differences between categories are not necessarily equal, there is a meaningful sequence. Examples include education level (high school, bachelor's degree, master's degree), customer satisfaction ratings (poor, fair, good, excellent), and Likert scale responses (strongly disagree, disagree, neutral, agree, strongly agree).
3. **Interval Data:** Interval data represent values measured on a scale where the differences between values are consistent and meaningful, but there is no true zero point. Examples include temperature measured in Celsius or Fahrenheit, calendar dates, and IQ scores. In interval data, arithmetic operations such as addition and subtraction are meaningful, but multiplication and division are not.
4. **Ratio Data:** Ratio data are similar to interval data but have a true zero point, indicating the absence of the measured quantity. In ratio data, all arithmetic operations are meaningful. Examples include height, weight, age, income, and number of items purchased. Ratio data allow for the calculation of meaningful ratios, such as the ratio of one value to another.

Q-3 What is data science? Is it new, or is it just statistics or analyticsrebranded? Is it real, or is it pure hype? And if it’s new and if it’s real, what does that mean?

Ans- Data science is an interdisciplinary field that involves the use of scientific methods, algorithms, and systems to extract insights and knowledge from structured and unstructured data. It combines elements from various disciplines such as statistics, computer science, machine learning, domain expertise, and data visualization to analyze and interpret complex datasets.

While data science shares some similarities with statistics and analytics, it also incorporates additional components, such as advanced computing techniques and big data technologies, to handle large volumes of data efficiently. Data science often involves the entire data lifecycle, including data collection, cleaning, preprocessing, analysis, modeling, interpretation, and visualization.

Here are some key points to consider regarding data science:

New Discipline: While the underlying concepts of data analysis and statistical modeling have been around for centuries, the term "data science" gained popularity in recent years as advancements in technology and the proliferation of big data created new opportunities and challenges for extracting insights from data.

Interdisciplinary Nature: Data science integrates techniques and methodologies from multiple disciplines, including statistics, computer science, information theory, and domain-specific knowledge. It combines mathematical and computational approaches with subject matter expertise to solve complex problems across various domains.

Real-world Applications: Data science has real-world applications in diverse fields, including business, healthcare, finance, marketing, social sciences, and more. It helps organizations make data-driven decisions, improve processes, optimize performance, and gain competitive advantages.

Hype vs. Reality: While there has been significant hype surrounding data science, fueled by media attention and industry trends, it is undoubtedly a real and valuable field. However, it's essential to distinguish between the hype and the genuine capabilities of data science. While data science has the potential to generate valuable insights and drive innovation, it is not a panacea and requires careful planning, rigorous methodology, and ethical considerations.

Implications: The emergence of data science has profound implications for society, economy, and technology. It has led to the creation of new job roles, such as data scientists, data engineers, and data analysts, and has transformed industries by enabling data-driven decision-making, personalized experiences, predictive analytics, and automation.

In conclusion, data science represents a significant evolution in the way we analyze, interpret, and leverage data to extract insights and create value. While it builds upon principles from statistics and analytics, it also incorporates advanced computational techniques and domain expertise to address the challenges of big data and complex real-world problems. As a result, data science is both new and real, with far-reaching implications for various aspects of modern society.

Ans- 4 Implementing linear regression on Brooklyn housing data

```
# Import necessary libraries
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score

# Load Brooklyn housing data into a DataFrame
data = pd.read_csv('brooklyn_housing_data.csv')
```

```
# Display the first few rows of the dataset
print(data.head())

# Extract features (X) and target variable (y)
X = data[['sqft', 'bedrooms', 'bathrooms', 'year_built']]
y = data['price']

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Initialize and train the linear regression model
model = LinearRegression()
model.fit(X_train, y_train)

# Make predictions on the testing set
y_pred = model.predict(X_test)

# Evaluate the model performance
mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)
r2 = r2_score(y_test, y_pred)

print(f'Mean Squared Error: {mse}')
print(f'Root Mean Squared Error: {rmse}')
print(f'R-squared: {r2}')

# Display the coefficients of the linear regression model
print('Coefficients:')
for feature, coef in zip(X.columns, model.coef_):
    print(f'{feature}: {coef}')

# Optional: Visualize the predicted vs. actual prices
import matplotlib.pyplot as plt

plt.scatter(y_test, y_pred)
plt.xlabel('Actual Price')
plt.ylabel('Predicted Price')
plt.title('Actual vs. Predicted Prices')
plt.show()
```

Q- 5 Briefly explain Data Science Process with a neat diagram.

Ans- The data science process typically involves several stages or steps, each aimed at extracting insights and value from data. Here's a brief explanation of the data science process along with a simplified diagram:

1. **Problem Definition:** The process starts with clearly defining the problem or question to be addressed. This involves understanding the objectives, defining success criteria, and identifying the relevant data sources.
2. **Data Collection:** Next, data is collected from various sources, including databases, APIs, files, sensors, and web scraping. This may involve gathering structured data from databases and spreadsheets, as well as unstructured data from text documents, images, and videos.
3. **Data Cleaning and Preprocessing:** Raw data often contains errors, missing values, inconsistencies, and noise. In this stage, data is cleaned, transformed, and prepared for analysis. This includes tasks such as removing duplicates, handling missing values, standardizing formats, and feature engineering.
4. **Exploratory Data Analysis (EDA):** EDA involves visualizing and exploring the data to gain insights, identify patterns, and detect relationships. Descriptive statistics, data visualization techniques, and exploratory data analysis tools are used to understand the structure and characteristics of the data.
5. **Feature Engineering:** Feature engineering involves creating new features or variables from the existing data to improve model performance. This may include transformations, scaling, dimensionality reduction, and creating interaction terms.
6. **Model Selection and Training:** In this stage, various machine learning algorithms are evaluated and selected based on the problem type, data characteristics, and performance metrics. Models are trained on the training data using techniques such as supervised learning, unsupervised learning, or reinforcement learning.
7. **Model Evaluation:** Trained models are evaluated using validation datasets or cross-validation techniques to assess their performance and generalization ability. Performance metrics such as accuracy, precision, recall, F1 score, and ROC curves are used to evaluate model performance.
8. **Model Deployment:** Once a satisfactory model is identified, it is deployed into production or operational systems for real-world use. This involves integrating the model into existing workflows, building APIs for inference, and monitoring model performance in production.
9. **Monitoring and Maintenance:** After deployment, the model performance is monitored continuously to ensure it remains effective and reliable over time. This may involve tracking performance metrics, detecting drift, retraining models periodically, and updating model parameters.

Here's a simplified diagram illustrating the data science process:

Problem Definition --> Data Collection --> Data Cleaning & Preprocessing --> EDA --> Feature Engineering --> Model Selection & Training --> Model Evaluation --> Model Deployment --> Monitoring & Maintenance

Ans 6(a)

1. Distance from (40, 20, Red):

$$\sqrt{(20-40)^2+(35-20)^2} = \sqrt{(-20)^2+(15)^2} = \sqrt{400+225} = \sqrt{625} = 25$$

2. Distance from (50, 50, Blue):

$$\sqrt{(20-50)^2+(35-50)^2} = \sqrt{(-30)^2+(-15)^2} = \sqrt{900+225} = \sqrt{1125} \approx 33.54$$

3. Distance from (60, 90, Blue):

$$\sqrt{(20-60)^2+(35-90)^2} = \sqrt{(-40)^2+(-55)^2} = \sqrt{1600+3025} = \sqrt{4625} \approx 68.0$$

4. Distance from (10, 25, Red):

$$\sqrt{(20-10)^2+(35-25)^2} = \sqrt{(10)^2+(10)^2} = \sqrt{100+100} = \sqrt{200} \approx 14.14$$

5. Distance from (70, 70, Blue):

$$\sqrt{(20-70)^2+(35-70)^2} = \sqrt{(-50)^2+(-35)^2} = \sqrt{2500+1225} = \sqrt{3725} \approx 61.0$$

6. Distance from (60, 10, Red):

$$\sqrt{(20-60)^2+(35-10)^2} = \sqrt{(-40)^2+(25)^2} = \sqrt{1600+625} = \sqrt{2225} \approx 47.17$$

7. Distance from (25, 80, Blue):

$$\sqrt{(20-25)^2+(35-80)^2} = \sqrt{(-5)^2+(-45)^2} = \sqrt{25+2025} = \sqrt{2050} \approx 45.31$$

Let's select the k=3k = 3k=3 nearest data points, which are (10, 25, Red), (40, 20, Red), and (25, 80, Blue). Among these points, two are classified as Red and one as Blue. Therefore, the majority class is Red.

So, the class of the new data entry (brightness=20,saturation=35) using KNN with k=3 is Red.

Ans- 6(b)

Solution:
Given, number of clusters to be created (K) = 2 say c1 and c2,
number of iterations = 2 and The given data points can be represented in tabular form as:

Instance	X	Y
1	185	72
2	170	56
3	168	60
4	179	68
5	182	72
6	188	77

Taking first two objects as initial centroids:
Centroid for first cluster c1 = (185, 72)
Centroid for second cluster c2 = (170, 56)

$d(c1, 3) = \sqrt{(185-168)^2 + (72-60)^2} = \sqrt{(17)^2 + (12)^2} = \sqrt{289 + 144} = \sqrt{433}$

$d(c2, 3) = \sqrt{(170-168)^2 + (56-60)^2} = \sqrt{(2)^2 + (-4)^2} = \sqrt{4 + 16} = \sqrt{20}$

Here, d(c2, 3) < d(c1, 3)

So, data point 3 belongs to c2.

$d(c1, 4) = \sqrt{(185-179)^2 + (72-68)^2} = \sqrt{(6)^2 + (4)^2} = \sqrt{36 + 16} = \sqrt{52}$

$d(c2, 4) = \sqrt{(170-179)^2 + (56-68)^2} = \sqrt{(-9)^2 + (-12)^2} = \sqrt{81 + 144} = \sqrt{225}$

Here, d(c1, 4) < d(c2, 4)

So, data point 4 belongs to c1.

$d(c1, 5) = \sqrt{(185-182)^2 + (72-72)^2} = \sqrt{(3)^2 + (0)^2} = \sqrt{9}$

$d(c2, 5) = \sqrt{(170-182)^2 + (56-72)^2} = \sqrt{(-12)^2 + (-16)^2} = \sqrt{144 + 256} = \sqrt{400}$

Here, d(c1, 5) < d(c2, 5)

So, data point 5 belongs to c1.

$d(c1, 6) = \sqrt{(185-188)^2 + (72-77)^2} = \sqrt{(-3)^2 + (-5)^2} = \sqrt{9 + 25} = \sqrt{34}$

$d(c2, 6) = \sqrt{(170-188)^2 + (56-77)^2} = \sqrt{(-18)^2 + (-21)^2} = \sqrt{324 + 441} = \sqrt{765}$

Here, d(c1, 6) < d(c2, 6)

So, data point 6 belongs to c1.

Instance	X	Y	Distance(C1)	Distance(C2)	Cluster
1	185	72			c1
2	170	56			c2
3	168	60	√433	√20	c2
4	179	68	√52	√225	c1
5	182	72	√9	√400	c1
6	188	77	√34	√765	c1

Iteration 2: Now calculating centroid for each cluster:

$$Centroid\ for\ first\ cluster\ c1 = \left(\frac{185+179+182+188}{4}, \frac{72+68+72+77}{4}\right) = (183.5, 72.25)$$

$$Centroid\ for\ second\ cluster\ c2 = \left(\frac{170+168}{2}, \frac{56+60}{2}\right) = (169, 58)$$

Now, again calculating similarity:

$$d(c1,3) = \sqrt{(183.5-168)^2 + (72.25-60)^2} = 19.7563$$
$$d(c2,3) = \sqrt{(169-168)^2 + (58-60)^2} = 2.2361$$

Here, d (c2, 3) < d (c1, 3)
So, data point 3 belongs to c2.

$$d(c1,4) = \sqrt{(183.5-179)^2 + (72.25-68)^2} = 6.1897$$
$$d(c2,4) = \sqrt{(169-179)^2 + (58-68)^2} = 14.1421$$

Here, d (c1, 4) < d (c2, 4)
So, data point 4 belongs to c1.

$$d(c1,5) = \sqrt{(183.5-182)^2 + (72.25-72)^2} = 1.5207$$
$$d(c2,5) = \sqrt{(169-182)^2 + (58-72)^2} = 19.1050$$

Here, d (c1, 5) < d (c2, 5)
So, data point 5 belongs to c1.

$$d(c1,6) = \sqrt{(183.5-188)^2 + (72.25-77)^2} = 6.5431$$
$$d(c2,6) = \sqrt{(169-188)^2 + (58-77)^2} = 26.8701$$

Here, d (c1, 6) < d (c2, 6)
So, data point 6 belongs to c1.

Data in tabular form:

Instance	X	Y	Distance(C1)	Distance(C2)	Cluster
1	185	72	1.5207	21.2603	c1
2	170	56	21.1261	2.2361	c2
3	168	60	19.7563	2.2361	c2
4	179	68	6.1897	14.1421	c1
5	182	72	1.5207	19.105	c1
6	188	77	6.5431	26.8701	c1