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

**Internal Assessment Test 1 – Jun. 2024**

| Sub: | **Data Science and Visualization (Professional Elective)** | | | | | Sub Code: | **21CS644** | Branch: | **CSE** |
|---|---|---|---|---|---|---|---|---|---|
| Date: | 06/06/2024 | Duration: | 90 mins | Max Marks: | 50 | Sem / Sec: | 6/A,B,C | | OBE |

| | Answer any FIVE FULL Questions | MARKS | CO | RBT |
|---|---|---|---|---|
| **1** | (a) Define Data Science. Explain the Drew Conway's venn diagram of data science. | [2+3] | CO1 | L2 |
| | (b) What is Datafication ? Explain with examples. | [2+3] | CO1 | L2 |
| **2** | a) Explain the following concepts with examples.<br>i) Statistical inference<br>ii) Population and Samples | [5] | CO1 | L2 |
| | b) Scenario: A real estate company wants to build a model to predict the prices of houses based on various features.<br>List down the features involved the above scenario. | [5] | CO2 | L2 |
| **3** | a) Explain Linear regression in detail.<br>b) Explain the difference between simple linear regression and multiple linear regression. | [5+5] | CO2 | L2 |
| **4** | Explain exploratory data analysis with example. | [10] | CO2 | L2 |
| **5** | Briefly explain data science process with a neat diagram | [10] | CO2 | L2 |
| **6** | Explain K-nearest neighbour algorithm and solve the following, let k =3 :<br>**<SepalLength=6.2, SepalWidth=2.5, Species=?>** | [5+5] | CO2 | L3 |

| **Sepal Length** | **Sepal Width** | **Species** |
|---|---|---|
| 5.3 | 3.7 | Setosa |
| 5.1 | 3.8 | Setosa |
| 7.2 | 3.0 | Virginica |
| 5.4 | 3.4 | Setosa |
| 5.1 | 3.3 | Setosa |
| 5.4 | 3.9 | Setosa |
| 7.4 | 2.8 | Virginica |
| 6.1 | 2.8 | Verscicolor |
| 7.3 | 2.9 | Virginica |
| 6.3 | 2.3 | Verscicolor |
| 5.1 | 2.5 | Verscicolor |
| 6.3 | 2.5 | Verscicolor |
| 5.5 | 2.4 | Verscicolor |

All the best

**CI** **CCI** **HOD**

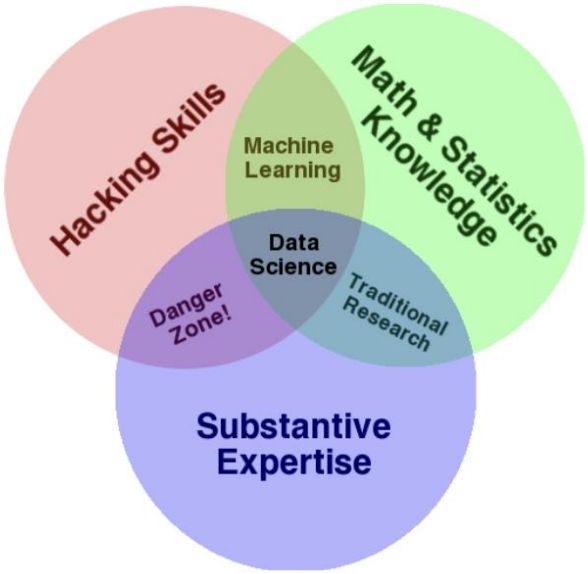
**CO PO Mapping**

| Course Outcomes | | Blooms Level | Modules covered | PO1 | PO2 | PO3 | PO4 | PO5 | PO6 | PO7 | PO8 | PO9 | PO10 | PO11 | PO12 | PSO1 | PSO2 | PSO3 | PSO4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CO1 | Understand the data in different forms | L2 | 1 | 3 | 3 | 3 | 3 | 2 | 2 | - | - | - | - | - | - | - | - | 2 | 3 |
| CO2 | Apply different techniques to Explore Data Analysis and the Data Science Process | L3 | 2 | 3 | 3 | 3 | 3 | 2 | 2 | - | - | - | - | - | - | - | - | 2 | 3 |
| CO3 | Analyze feature selection algorithms & design a recommender system. | L3 | 3 | 3 | 3 | 3 | 3 | 2 | 2 | - | - | - | - | - | - | - | - | 2 | 3 |
| CO4 | Evaluate data visualization tools and libraries and plot graphs. | L3 | 4 | 3 | 3 | 3 | 3 | 2 | 2 | - | - | - | - | - | - | - | - | 2 | 3 |
| CO5 | Develop different charts and include mathematical expressions. | L3 | 5 | 3 | 3 | 3 | 3 | 2 | 2 | - | - | - | - | - | - | - | - | 2 | 3 |

| COGNITIVE LEVEL | REVISED BLOOMS TAXONOMY KEYWORDS |
|---|---|
| L1 | List, define, tell, describe, identify, show, label, collect, examine, tabulate, quote, name, who, when, where, etc. |
| L2 | summarize, describe, interpret, contrast, predict, associate, distinguish, estimate, differentiate, discuss, extend |
| L3 | Apply, demonstrate, calculate, complete, illustrate, show, solve, examine, modify, relate, change, classify, experiment, discover. |
| L4 | Analyze, separate, order, explain, connect, classify, arrange, divide, compare, select, explain, infer. |
| L5 | Assess, decide, rank, grade, test, measure, recommend, convince, select, judge, explain, discriminate, support, conclude, compare, summarize. |

| PROGRAM OUTCOMES (PO), PROGRAM SPECIFIC OUTCOMES (PSO) | | | | CORRELATION LEVELS | |
|---|---|---|---|---|---|
| PO1 | Engineering knowledge | PO7 | Environment and sustainability | 0 | No Correlation |
| PO2 | Problem analysis | PO8 | Ethics | 1 | Slight/Low |
| PO3 | Design/development of solutions | PO9 | Individual and team work | 2 | Moderate/ Medium |
| PO4 | Conduct investigations of complex problems | PO10 | Communication | 3 | Substantial/ High |
| PO5 | Modern tool usage | PO11 | Project management and finance | | |
| PO6 | The Engineer and society | PO12 | Life-long learning | | |
| PSO1 | Develop applications using different stacks of web and programming technologies | | | | |
| PSO2 | Design and develop secure, parallel, distributed, networked, and digital systems | | | | |
| PSO3 | Apply software engineering methods to design, develop, test and manage software systems. | | | | |
| PSO4 | Develop intelligent applications for business and industry | | | | |

Answer 1: Data science is the study of data to extract meaningful insights for business. It is a multidisciplinary approach that combines principles and practices from the fields of mathematics, statistics, artificial intelligence, and computer engineering to analyze large amounts of data




Answer 2: Datafication as a process of “taking all aspects of life and turning them into data.” As examples, they mention that “Google’s augmented-reality glasses datafy the gaze. Twitter datafies stray thoughts. LinkedIn datafies professional networks.”

Example- sms, tweet, email

Answer 2: **Statistical Process:** This overall process of going from the world to the data, and then from the data back to the world, is the field of statistical inference.
More precisely, statistical inference is the discipline that concerns itself with the development of procedures, methods, and theorems that allow us to extract meaning and information from data that has been generated by stochastic (random) processes.

Populations and Samples

A population refers to the entire group of individuals, items, or data points that are of interest to a researcher.
It represents the complete set of observations that share certain characteristics and is the target of analysis..

A sample is a subset of the population selected for study. It's impractical or impossible to collect data from

an entire population, so researchers often take a smaller, manageable group from the population to represent
It.

Answer 2-b **1. Property Features**

- **Square Footage**: Total living area in square feet.
- **Number of Bedrooms**: Total number of bedrooms.
- **Number of Bathrooms**: Total number of bathrooms (full and half).
- **Lot Size**: Total area of the lot/property in square feet or acres.
- **Age of the Property**: Time since the property was built or last renovated.
- **Type of Property**: e.g., single-family home, condo, townhouse, etc.
- **Number of Floors**: Number of levels in the property.
- **Parking**: Availability and type (garage, driveway, none).
- **Basement**: Presence and type (finished, unfinished).
- **Attic**: Presence and usability.
- **Construction Quality**: Materials and build quality.
- **Heating System**: Type and efficiency (e.g., gas, electric, central heating).
- **Cooling System**: Type and efficiency (e.g., central air conditioning).
- **Energy Efficiency**: Features like insulation, solar panels.
- **Fireplace**: Presence and number of fireplaces.
- **Flooring**: Types of flooring (hardwood, carpet, tile, etc.).

## 2. Location Features

- **Neighborhood**: General neighborhood rating or type.
- **Proximity to Amenities**: Distance to schools, shopping centers, parks, hospitals, etc.
- **School District Quality**: Ratings of local schools.
- **Crime Rate**: Crime statistics for the area.
- **Public Transportation**: Availability and proximity of public transit options.
- **Walkability Score**: Ease of walking to nearby amenities.
- **Zoning Laws**: Local zoning regulations affecting property use.
- **View**: Scenic views, ocean views, cityscape, etc.
- **Flood Zone**: Proximity to flood zones or other natural hazards.
- **Noise Level**: Proximity to airports, highways, or other noise sources.

## 3. Economic Features

- **Property Taxes**: Annual property tax amount.

- **Homeowners Association (HOA) Fees**: Monthly or annual fees for community upkeep.
- **Local Market Trends**: Recent trends in property prices in the area.
- **Interest Rates**: Current mortgage interest rates.

## 4. Other Features

- **Renovations and Upgrades**: Recent improvements or updates to the property.
- **Historical Sales Data**: Previous sale prices of the property.
- **Occupancy Status**: Whether the property is currently occupied or vacant.
- **Future Developments**: Planned infrastructure or commercial developments nearby.

Answer 3: **a) Linear Regression: Detailed Explanation**

Linear regression is a statistical technique used to model and analyze the relationship between one dependent variable (outcome) and one or more independent variables (predictors). The main goal is to find the best-fitting linear relationship that explains how changes in the predictors affect the outcome.

**Key Concepts**

1. **Linear Relationship**: Assumes the relationship between the variables can be represented by a straight line. The equation of this line is:

   $y=\beta0+\beta1x1+\beta2x2+\cdots+\betapxp+\epsilon$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilony=β0 +β1 x1 +β2 x2 +⋯+βp xp +ϵ

   where:
   - yyy is the dependent variable (outcome).
   - β0\beta_0β0 is the intercept (the value of yyy when all xix_ixi are zero).
   - βi\beta_iβi are the coefficients (the change in yyy for a one-unit change in xix_ixi ).
   - xix_ixi are the independent variables (predictors).
   - ϵ\epsilonϵ is the error term (captures the difference between the observed and predicted yyy).
2. **Ordinary Least Squares (OLS)**: The most common method for estimating the coefficients. OLS minimizes the sum of the squared differences between the observed and predicted values:

$\min_{\beta} \sum_{i=1}^n (y_i - \hat{y_i})^2$

where $\hat{y_i}$ is the predicted value of $y$ for the $i$-th observation.

3. **Assumptions**:
    - **Linearity**: The relationship between $y$ and $x$ is linear.
    - **Independence**: Observations are independent of each other.
    - **Homoscedasticity**: The variance of residuals is constant across all levels of the independent variables.
    - **Normality**: The residuals (errors) are normally distributed.
4. **Coefficient Interpretation**: Each coefficient $\beta_i$ represents the change in the dependent variable for a one-unit change in the corresponding independent variable, holding all other variables constant.
5. **Goodness-of-Fit**:
    - **R-squared ($R^2$)**: Indicates the proportion of the variance in the dependent variable explained by the independent variables.
    - **Adjusted R-squared**: Adjusts $R^2$ for the number of predictors, providing a more accurate measure when multiple predictors are used.
6. **Diagnostics**: Include checking for multicollinearity, autocorrelation, and the presence of outliers, among others.

#### Example

Consider predicting house prices ($y$) based on square footage ($x$).

$y = \beta_0 + \beta_1 x + \epsilon$

If we estimate $\beta_0 = 50{,}000$ and $\beta_1 = 200$, the model suggests that each additional square foot adds $200 to the house price, and a house with 0 square feet would cost $50,000 (the intercept, which may not be meaningful in practice but is part of the model).

## b) Difference Between Simple Linear Regression and Multiple Linear Regression

#### Simple Linear Regression

- **Definition**: Models the relationship between one dependent variable and a single independent variable.
- **Equation**:
  $y = \beta_0 + \beta_1 x + \epsilon$

- **Usage**: Best used when the relationship is straightforward and influenced predominantly by one predictor.
- **Example**: Predicting house prices solely based on square footage.

#### Multiple Linear Regression

- **Definition**: Models the relationship between one dependent variable and two or more independent variables.
- **Equation**:

  $y=\beta0+\beta1x1+\beta2x2+\cdots+\betapxp+\epsilony = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilony=\beta0 +\beta1 x1 +\beta2 x2 +\cdots+\betap xp +\epsilon$
- **Usage**: Used when the outcome is influenced by several factors and can provide a more comprehensive understanding of the dependent variable.
- **Example**: Predicting house prices based on square footage, number of bedrooms, and location.

#### Comparison Table

| Feature | Simple Linear Regression | Multiple Linear Regression |
|---|---|---|
| **Number of Predictors** | One | Two or more |
| **Equation** | $y=\beta0+\beta1x+\epsilony = \beta_0 + \beta_1 x + \epsilony=\beta0 +\beta1 x+\epsilon$ | $y=\beta0+\beta1x1+\beta2x2+\cdots+\betapxp+\epsilony = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilony=\beta0 +\beta1 x1 +\beta2 x2 +\cdots+\betap xp +\epsilon$ |
| **Complexity** | Simpler, easier to interpret | More complex, requires understanding interactions |

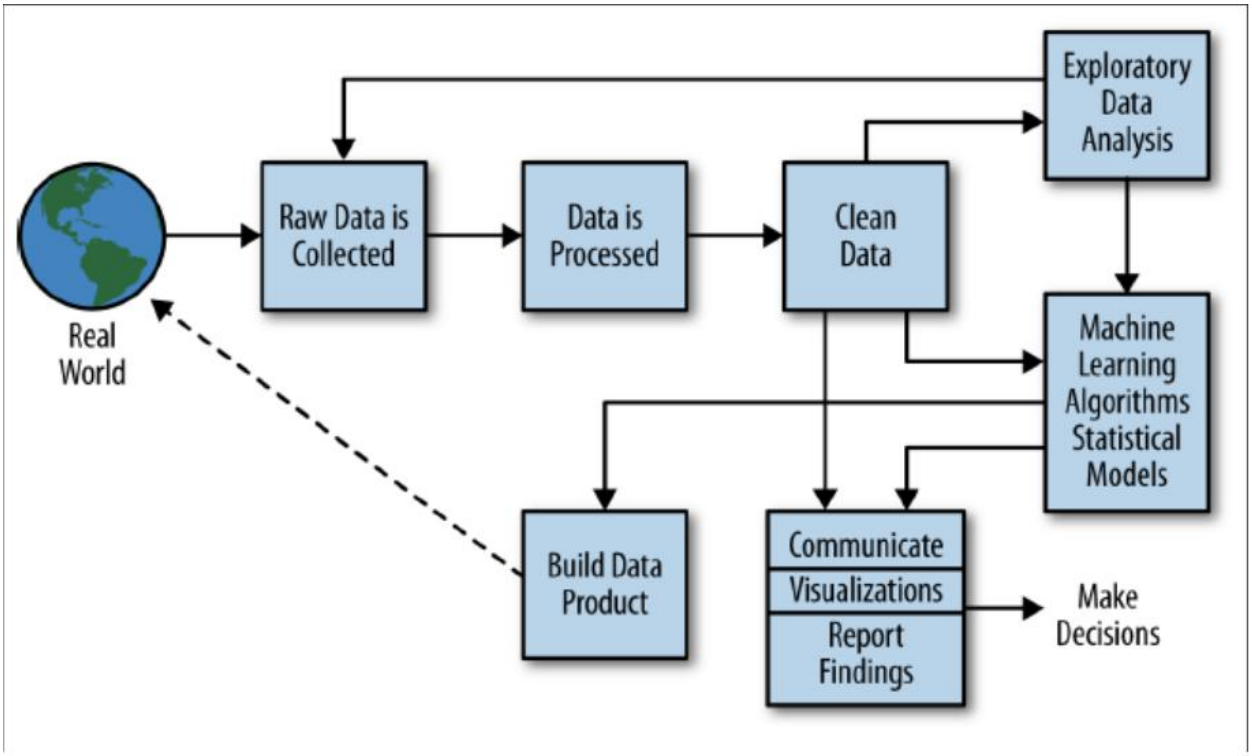
| | | |
|---|---|---|
| **Fit** | Fits a straight line to the data | Fits a plane (or hyperplane) to the data |
| **Assumptions** | Easier to check | More assumptions to validate |
| **Interpretation** | Direct interpretation of one predictor | Interpretation of each predictor adjusted for others |
| **Multicollinearity** | Not an issue | Can be an issue, needs diagnostics |
| **Applications** | Initial exploration of relationships | Comprehensive modeling of complex relationships |

In summary, **simple linear regression** is ideal for exploring the relationship between a single predictor and an outcome, while **multiple linear regression** is better suited for scenarios where the outcome is influenced by multiple predictors.

Answer 4: The basic tools of EDA are plots, graphs and summary statistics.

Generally speaking, it's a method of systematically going through the data, plotting distributions of all variables (using box plots), plotting time series of data, transforming variables, looking at all pairwise relationships between variables using scatterplot matrices, and generating summary statistics for all of them.

Answer 5:




Answer 6: Versicolor is the class for test case

1.Calculate the euclidean distance between the test instance to all the training instance
2. Sort the distance based on smallest to largest
3. Since k=3, select the 3 nearest neighbors.
4.Based on the nearest neighbors, the majority of the label in the training instance can be selected for the final label for the test instance.
5.Label for test instance :Versicolor is the class for test case