DSC IV234

Assignment - 1

1 What are the Key interdisciplinary fields t contribute to Data science? Explain their ro

→ Data science is an interdisciplinary fi that combines various domains to extract insight from data.

i Mathematics and statistics

→ Role:- provides fundamental techniques for data analysis, hypothesis testing, probability and inferential statistics.

→ Key concepts:- Linear algebra, probability distributions, regression analysis, and hypothesis testing

ii Computer science and programming

→ Role:- Enables data collection, processing, storage, and algorithm implementation.

→ Key concepts:- Data structures, algorithms, databases, and software engineering

IU2341231

iii] Machine Learning & Artificial Intelligence

⇒ Role:- Develops predictive models and automate decision-making using data-driven techniques.

⇒ Key concepts:- supervised and unsupervised learning, neural networks, deep learning, and reinforcement learning, neural networks, deep learning, and reinforcement learning.

iv] Domain Expertise

⇒ Role:- Helps in understanding the real-world context of the data and ensures insights are meaningful

⇒ Key concepts:- Industry-specific knowledge, business analytics, and decision-making frameworks

v] Data Engineering

⇒ Role:- Helps Manages large-scale data pipelines storage, and data transformation for efficient analysis

⇒ Key concepts: ETL (Extract, Transform, Load) Big Data technologies (Hadoop, Spark), and cloud computing

IU2341231

2] What are the main responsibilities of a Data scientist? How do they differ from Data Analysts and Data Engineers? What key skills are required to become a Data scientist? Discuss both technical & non-technical skills.

⇒ A Data scientist is responsible for analyzing complex data to uncover insights, build predictive models, and help organizations make data driven decisions.

i Data collection & cleaning: Extracting data from various sources, handling missing values, and ensuring data quality

ii Exploratory Data Analysis (EDA): Understanding data patterns, trends, and distributions using statistical and visualization techniques.

iii Building Machine Learning Models: Developing predictive models using algorithms like regression, classification, clustering, and deep learning.

iv Data Interpretation & Communication: presenting findings using dashboards, reports, and data storytelling to non-technical stakeholders.

v A/B Testing & Experimentation: Conducting Experiments to optimize business strategies.

JV2341231551

vi Big Data Handling: Working with large-scale data using cloud computing, Spark, or Hadoo

vii Deploying Models & Automation: Integrating ML models into production environments for real-t decision-making.

viii Ensuring Data Ethics & Compliance: Handling data responsibly while following regulations like GDPR.

Responsibilities	Key skills
• Role:- Data scientist	
→ Builds ML models, performs predictive analysis, and finds deep insights	ML, AI, statistics, python, SQL
• Role:- Data Analyst	
→ Analyzes past data trends, creates reports & dashboards for business decisions	SQL Excel, Tableau, power BI
• Role:- Data Engineer	
→ Builds data piplines, optimizes storage & ensures data availability	Big Data, SQL, ETL, Cloud computing

IU2341231551

→ Data Scientists works on predictions and automation.

→ Data Analysts focus on descriptive analysis and reporting

→ Data Engineers handle infrastructure and data pipelines.

• Key skills Required to Become a Data scientist

○ Technical skills

1. Programming: Python, R, SQL
2. Machine Learning & AI: supervised/unsupervised learning, deep learning, NLP
3. Statistics & Mathematics: Probability, hypothesis testing, regression analysis
4. Data Manipulation: Pandas, Numpy, SQL
5. Big Data Technologies: Hadoop, Spark
6. Data Visualization: Matplotlib, seaborn, Tableau
7. cloud computing: Aws, Google cloud, Azure
8. Model Deployment: Flask, FastAPI, Docker

IU2341231555

2 Non-Technical skills

1 Critical Thinking: Ability to frame problems & interpret results

2 Communication: Explaining complex data insights to non-technical audiences.

3 Business Acumen: Understanding business needs & aligning solutions.

4 Teamwork & collaboration: Working with cross functional teams.

5 Curiosity & problem solving: Eagerness to explore new data trends and technologies

IV2341231551

3] Explain the relationship between Data science and Data. What are the key characteristics (7Vs) of Big Data? Explain in detail.

→ Big Data refers to large, complex datasets that traditional data processing tools cannot handle efficiently. It focuses on collecting, storing, and managing massive amounts of data.

→ Data science is the field that extracts valuable insights from data using statistics, machine learning, and AI.

• Relationship

1. Big Data as a source for Data science:
Data science relies on Big Data to train models and uncover patterns.

2. Big Data Technologies Enable Data science:
Tools like Hadoop, Spark, and cloud computing help process large datasets efficiently

3. Data science Extracts Insights from Big Data
Using ML, AI, and analytics, Data science transforms Big Data into actionable insights.

IU2341231551

7Vs of Big Data

1. Volume :- The sheer size of data generated from sources like social media, IOT, and transactions.

2. Velocity :- The speed at which data is generated and processed in real time.

3. Variety :- Different data formats, such as structured (db), semi-structured (JSON, XML), and unstructured (videos, images).

4 Veracity :- The accuracy and reliability of data.

5 value :- The usefulness of data in decision-making

6 Variability :- The inconsistency of data due to time-sensitive fluctuations.

7 Visualization :- The ability to represent complex data using graphs, dashboards, and charts.

IV2341231551

4 What is Exploratory Data Analysis (EDA)? How do visualization techniques help in understanding data?

→ Exploratory Data Analysis (EDA) is a crucial step in Data science that involves summarizing, visualizing and understanding the data before applying machine learning or statistical models.

- Objectives of EDA

1. Identify patterns and trends

→ Helps in understanding relationships between variables.

2. Detect missing or erroneous data :- Find anomalies or outliers in datasets.

3. Understand the distribution of data: Checks skewness, kurtosis and spread of data.

4. Generate insights for feature selection: Determines important variables for predictive modeling.

IV2341231551

- Visualization Techniques Help in Understanding Data

1. Histogram :- shows the frequency distribution of a single variable.

2. Box Plot :- Identifies outliers and data spread using quartiles.

3. Scatter Plot :- Displys relationship between two numerical variables

4. Heatmap :- Visualizes correlations between multiple variables.

5 Pie chart :- Represents proportions within a dataset

i Visualization in EDA simplifies complex data makes raw data more digestible.

ii Visualization in EDA Reveals in hidden patterns hidden problems detects trends, correlations, and anomalies

iii It Enhances decision-making helps businesses and analysts draw meaningful conclusions.

IU2341231551

5] What is the data retrieval phase? What are the common sources of data in Data science

→ The Data retrieval phase is a critical steps in the Data Science workflow where raw data is gathered from various sources for further processing, analysis, and decision-making.

This phase ensures that data is extracted efficiently, cleans properly, and stored in an accessible format for analysis.

- Data retrieval involves multiple steps following

1 Identifying Relevant Data sources

→ Finding structured, semi-structured, or unstructured data that is essential for the given prob.

2. Fetching Data

→ Using different techniques such as SQL queries, APIs, web scraping, or direct file imports from spreadsheets, databases, and cloud storage.

3. Data cleaning and preprocessing

→ Handling missing values, duplicates, inconsistencies, and ensuring data integrity before analysis.

IV2341231551

4. Transforming Data

→ Standardizing formats, normalizing values, and encoding categorical data for compatibility with machine learning models.

5 Storing Data

→ Keeping the retrieved data in secure storage solutions such as db, data warehouses, cloud platforms, or local file systems for further analysis

· Common sources of data

1 Structured Data sources

→ Structured data is highly organized and stored in tabular formats, making it easy to process using relational databases and spreadsheets.

- RDBMS :- MySQL, PostgreSQL, MS SQL server, and Oracle DB
- Data warehouses :- Large-scale storage solutions like Amazon Redshift, Google BigQuery, and Snowflake.
- Spreadsheets :- Ms Excel, Google Sheet and csv files.

IV23U1231557

2. Unstructured Data Sources

⇒ Unstructured data lacks a predefined format and requires specialized processing techniques, such as NLP and image recognition, to extract useful insights

- Text Data - Emails, chat logs, news articals, product reviews, and social media posts.
- Multimedia Data:- Images, videos, and audio files from platforms like YouTube and CCTV
- Unstructured Documents:- PDFs, scanned documents, handwritten notes.

3 Semi-structured Data sources

⇒ Semi-structured data sources does not conform to traditional relational database structures but still contains some form of organization.

- JSON and XML files:- common formats for data exchange between web services and APIs
- CSV Files:- Flat files that store data in tabular form but lack a formal relational structure.

IU2341231552

6] Discuss some major applications of Data science in healthcare, finance, e-commerce, and social media analytics

⇒ Data science has become a transformative force across various industries, leveraging vast amount of data to drive innovation, optimize processes and enhance decision-making

1. Healthcare

- Predictive Analytics for patient Care

⇒ Data science enables the analysis of historical patient data to forecast disease outbreaks and individual health trajectories, facilitating proactive medical interventions

- Medical Image analysis

⇒ Advanced machine learning algorithms assist in interpreting medical images, aiding in the early detection and diagnosis of conditions such as tumors and neurological disorders.

2. Finance

- Fraud Detection

⇒ Financial institutions utilize data science to identify unusual transaction patterns, enabling the prompt detection and prevention of fraudulent activities

IV2341231551

- Risk Assessment and Management

→ Through predictive modeling, data science evaluates credit histories and market trends to assess risks, guiding lending decisions and in real-time executing trades based on predictive analytics to optimize returns.

3 E-commerce

- Recommendation Systems

→ Ecommerce platforms employ data science to analyze user behavior and preferences, providing personalized product recommendations that enhance the shopping experience.

- Price optimization

→ By examining examining competitor pricing, demand fluctuations, and customer data, data science aids in setting optimal pricing strategies to maximize profits.

1 Social Media Analytics

- Sentiment Analysis

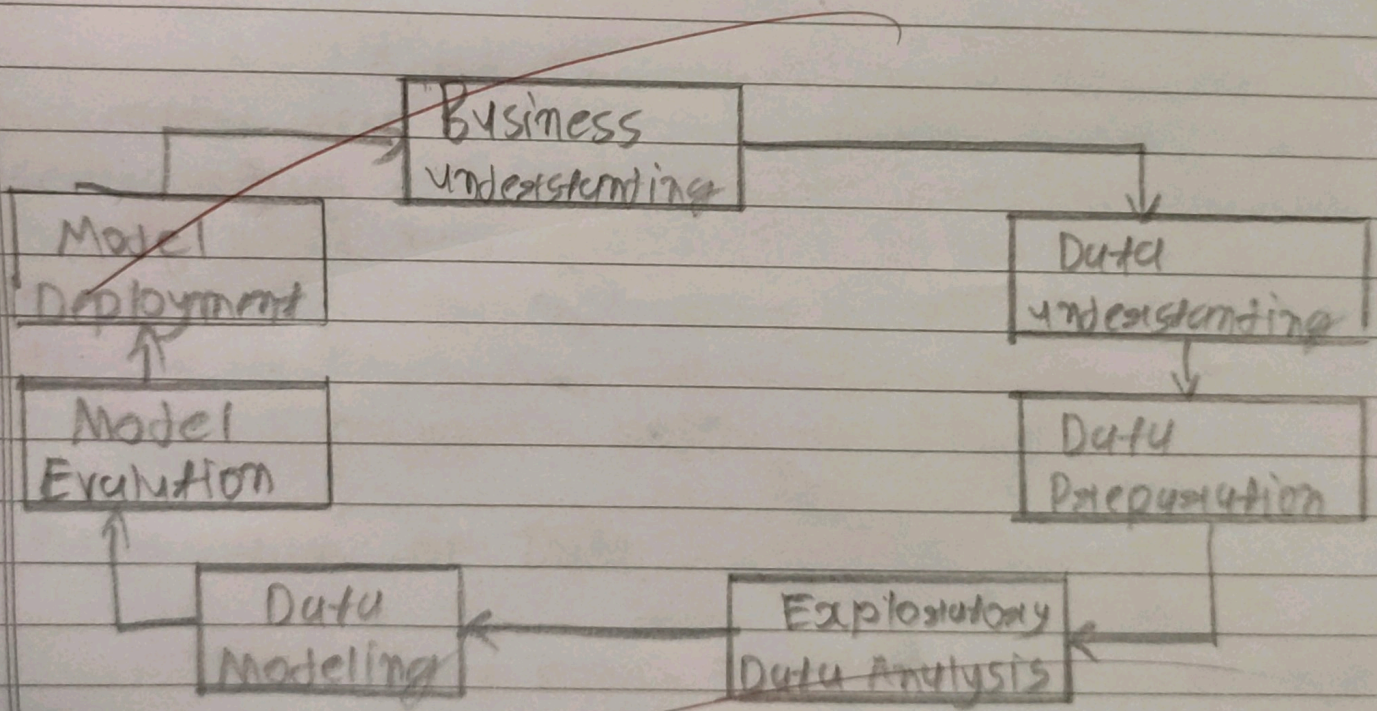
2 Data science techniques analyze user-generated content to gauge public sentiment towards brands, products, or events, informing marketing strategies

IU2341231551

• Trend Analysis:

⇒ By monitoring and analyzing social media activity, data science identifies emerging trends, allowing companies to stay ahead in dynamic markets

7] Draw and Explain the Data science lifecycle



1 Business Understanding

⇒ The complete cycle revolves around the enterprise goal. What will you resolve if you do not longer have specific problem? It is extraordinarily essential to apprehend the commercial enterprise goal sincerely - due to the fact that will be your ultimate aim of the analysis.

J023417231551

→ After desirable perception only we c
set the precise aim of evaluation that is in sync
with the enterprise objective. You need to understand
the customer desires to minimize saving loss or if
they prefer to predict the rate of a commodity

2 Data Understanding

→ After enterprise understanding, the
subsequent step is data understanding. This includes a
series of all the reachable data. Here you need to
intently work with the commercial enterprise group
as they are certainly conscious of what information
is present, what facts should be used for this
commercial enterprise problem, and different information

3 Preparation of Data

→ Next comes the data preparation stage.
This consists of steps like choosing the applicable data,
integrating the data by means of merging the data
sets, cleaning it, treating the lacking values through
either eliminating them or imputing them, treating
inaccurate data through eliminating them. Additionally
test for outliers the use of box plots and cope with
them.

IU2341231551

4. Exploratory Data analysis

→ This step includes getting some concep about the answers and elements affecting it, earlier than constructing the real model. Distribution of data inside distinctive variables of a character is explored graphically the usage of bar-graphs, Relations between distinct aspects are captured via graphical representation like scatter plots and warmth maps.

5 Data Modelling

→ Data modeling is the coronary heart of data analysis. A model takes the organized data as input and gives the preferred output. This step consists of selecting the suitable kind of model, whether the problem is a classification problem, or a regression problem or a clustering problem.

6 Model Evaluation

→ Here the model is evaluated for checking if it is geared up to be deployed. The model is examined on an unseen data, evaluated on a cautiously thought out set of assessment metrics. We additionally need to make positive that the model conforms to reality. If we do not acquire a quality and result in the evaluation, we have to re-iterate the complete modelling procedure until the preferred stage of metric is achieved.

IU2341231551

7 Model Deployment

→ The model after a rigorous assessment is at the end deployed in the prefetched structure and channel. This is the last step in the data science life cycle. Each step in the data science life cycle defined above must be laboured upon carefully. If any step is performed improperly, and hence, have an effect on the subsequent step and the complete effort goes to waste.

28/02/25

08/10