**Q1. Define Data science and big data?**

**Project on DS and Big Data;**  **Name: Yash**

**Course: B.tech{CSE}, Data Science.**

**Roll no: 241302121**

Ans-> Data science is an interdisciplinary field that extracts insights and knowledge from structured and unstructured data using various techniques, tools, and methods.

Big data refers to large, complex datasets that exceed traditional data processing capabilities.

**Q2. Elaborate the points based on data science and big data respectively ?**

Ans->

(a.) Use of technology:- { of data science;}

-> Programming Languages: Python, R, Julia.

-> Data analysis libraries: Pandas, NumPy, Matplotlib.

-> Machine learning frameworks: scikit-learn, TensorFlow, PyTorch.

-> Database management: MySQL, PostgreSQL, MongoDB.

(b.) Use of technology:- { of big data;}

-> Distributed computing: Hadoop, Spark, Apache Flink.

-> NoSQL database: MongoDB, Cassandra, Couchbase.

-> Data warehousing: Amazon Redshift, Google big Query.

-> Cloud storage: AWS S3, Azure Blob storage.

**Q3. Define types of data ?**

[1.] Structured data-> Structured data refers to highly organized and formatted data that is easily searchable, machine-readable, and understandable by humans. It follows a predefined schema or format, making it easily accessible and manageable.

[2.] Semi-structured data-> It is a type of data that has some level of organization and formatting, but doesn’t conform to a rigid or fixed schema.

[3.] Unstructured data-> Unstructured data refers to a data that lacks a predefined format, organization, or schema, making it difficult for computers to search, analyze, and understand.

[4.] Graph based data-> Graph based data represents information as a collection of interconnected nodes or vertices, edges, and properties.

[5.] Machined-generated data-> Machined-generated data refers to data that is automatically produced by machines, device, or algorithms, without human intervention.

[6.] Audio data-> Audio data refers to sound information stored in digital or analog formats.

[7.] Video data-> Video data refers to visual information stored in digital or analog formats.

[8.] Image data-> Image data refers to visual information stored in digital or analog formats.

[9.] Streaming data-> Streaming data refers to a continuous flow of data that is generated and transmitted in real time, often from various sources.

[10.] Network data-> Network data refers to the information that is transmitted, received, or stored within a network.

[11.] Natural language-> Natural language refers to the human language used in everyday communication.

**Q4. Discuss data science process?**

Ans-> The data science process is a systematic approach to extracting insight and knowledge from data. It involves several stages:

A. Problem Formulation:

~ Define business or research problem.

~ Identify key stakeholders and their goals.

~ Determine data requirements.

B. Data Collection:

~ Gather relevant data from various sources.

~ Ensure data quality and integrity.

~ Store data in a suitable format.

C. Data Cleaning:

~ Handle missing values and outliers.

~ Remove duplicates and irrelevant data.

~ Transform data into suitable format.

D. Data Exploration:

~ Visualize data to understand distributions and relationships.

~ Perform statistical analysis and summary metrices.

~ Identify pattern and trends.

E. Feature Engineering:

~ Select relevant feature for modeling.

~ Create new features through transformations.

~ Encode categorical variables.

F. Model Selection:

~ Choose suitable algorithms based on problems type.

~ Consider model complexity and interpretability.

~ Evaluate model performance metrics.

G. Model Training:

~ Train model using selected algorithm.

~ Tune hyperparameters for optimal performance.

~ Monitor model convergence.

H. Model Evaluation:

~ Assess model performance using metrics(accuracy, precision, recall)

~ Compare model using cross-validation.

~ Identify potential biases.

I. Deployment:

~ Integrate model into production environment.

~ Monitor model performance and retrain as needed.

~ Communicate result to stakeholders.

J. Maintenance:

~ Update model with new data.

~ Refine model based on feedback.

~ Ensure model scalability and reliability.

**Q5. Introduce :-**

**1.) AI** -> Artificial Intelligence(AI) refers to the simulation of human intelligence in machines, enabling the to:

**Think:** Reason, Learn, and problem solve.

**Learn:** Adapt to new data and experience.

**Perceive:** Interpret and understand data from sensors and inputs.

**Act:** Make decision and take actions.

**2.) ML** -> Machine learning(ML) is a subset of artificial intelligence(AI) that enables machines to :

**Learn:** From data and experiences.

**Improve:** Performance on tasks overtime.

**Make decision:** Based on predictions and patterns.

**3.) DS** -> Data science(DS) is an interdisciplinary field that extract insights and knowledge from data using key components like: ~ Data Engineering.

~ Machine Learning (ML).

~ Statistics.

~ Visualization.

~ Domain Expertise.

**4.) Applications of DS** ->

[a.] Industrial Applications:- ~ Healthcare: Predictive analytics, disease diagnosis, patient outcomes.

~ Marketing: Personalisation, campaign optimization, social media

Analysis.

[b.] Business Applications:- ~ Customer Relationship Management (CRM).

~ Supply Chain Management (SCM).

[c.] Functional Applications:- ~ Text Analysis: Sentiment analysis, topic modeling, text classification.

~ Time Series Analysis: Forecasting, anomaly detection, seasonal

decomposition.

[d.] Emerging Applications:- ~ Artificial Intelligence (AI).

~ Internet of Things (IOT).

**5.) History of AI** -> Early Years (1950s-1960)

~ Alan Turing’s 1950 paper, “Computing Machinery and Intelligence,”

proposed that Turing Test.

~ 1951: First AI program, Logical Theorist, developed by Allen Newell

and Herbert Simon.

~ 1956: Dartmouth Summer Research Project on Artificial Intelligence

(AI) founded.

~ 1958: Computer Learning Machine developed by Arthur Samuel.