DATA SCIENCE

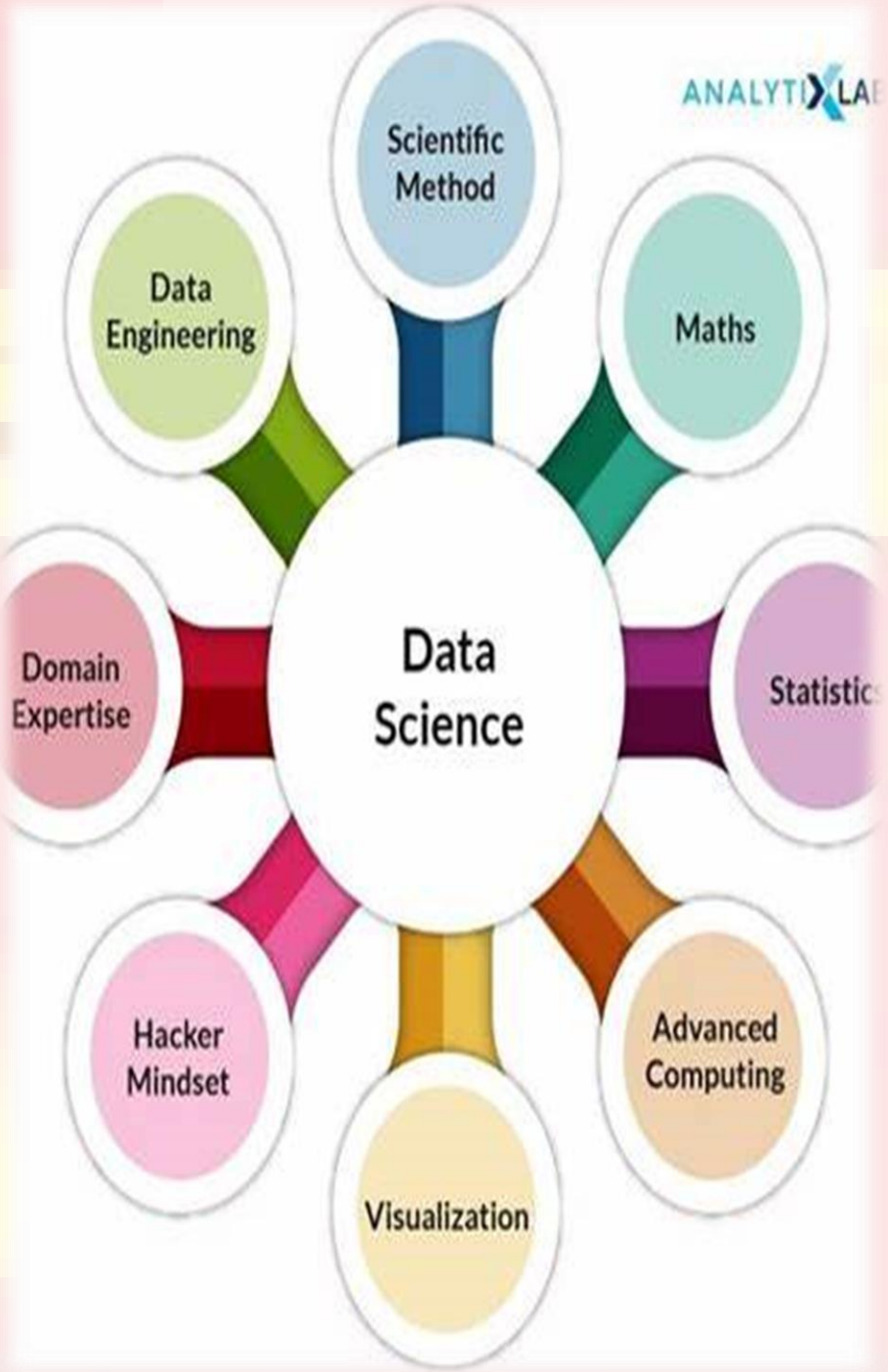
CBSE
AI - 417

DATA SCIENCE
INTRODUCTION

Data Science is a field that uses

- Scientific (mathematical & statistical) methods
- processes algorithms
- systems to extract knowledge
- insights from huge volumes of structural and unstructured data to apply in AI applications.

etc.



Application of Data Science

1. Fraud & Risk Detection

Data science plays a critical role in fraud detection and risk management. It helps organizations identify and prevent fraudulent activities, minimize risks, and make informed decisions

2. Healthcare

- ✓ Diagnosis with Medical Image analysis
- ✓ Genetics and Genomics Research
- ✓ Drug Discovery & Development

3. Internet Search

- ✓ Search engines use data science tools and algorithms to provide the best search results for the searched query in a fraction of a second.

  Without the data science, we could not have experienced the Google and other search engines we have today.

Application of Data Science

4. Targeted Advertising

For target advertising, the data science works with data based on browsing habits, past purchases, or any other recent activities and pics trends.

So Two persons simultaneously browsing a shopping site on two different devices will see different advertisements depending upon their past searches and other variables. It is because of the efficient use of data science tools.

The digital ads reach to the intended targeted customers and thus get a much higher CTR(Click through rate) than traditional advertisements.

5. Recommender Systems (RSs)

A Recommender system(RS) refers to a system that is capable of predicting the future preference of a set of items for a user and recommending the top items.

A recommender system is used by many online retail hubs like Amazon, entertainment and digital content sites such as Netflix, Hotstar, Discovery, Travel & Hotel Booking sites like MakeMyTrip, Expedia and so forth.etc.

Application of Data Science

6. Airline Route Planning – Data Science is helping airline companies in crucial decisions like -

- ✓ Popular demand of fliers
- ✓ Planning air routes according to demand
- ✓ Decisions with flight delays
- ✓ Inflight food supply choices
- ✓ Deciding about fares and discounts
- ✓ Loyal customers' benefits

7. Weather Predictions in agriculture Sector

- ✓ The amount and types of cloud
- ✓ Snowfall and precipitation
- ✓ Maximum, minimum, and dew point temperatures
- ✓ Humidity relative
- ✓ The direction and speed of wind
- ✓ Low pressure areas
- ✓ Events like fog, frost, hail, thunderstorms and gales of winds

SUMMARY- Application of Data Science

1. Fraud & Risk Detection
2. Healthcare
3. Internet Search
4. Targeted Advertising
5. Recommender Systems (RSs)
6. Airline Route Planning
7. Weather Predictions in the Agriculture Sector

Let Us Revise

- Data Science is a field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data to apply in AI applications.
- The field of data science combines concepts and methods from various fields of statistics, mathematics, data analysis, machine learning, computer science and so forth.
- Data Science is a broader term that makes use of Data Analytics, which analyses and gathers insights from past data.
- Data science has found applications in all data heavy fields like Finance and Banking for fraud and risk detection, healthcare, Internet search, Recommender systems, air route planning, targeting advertising, weather prediction for agriculture sector and so on.
- Targeted advertising is a type of online advertising that is oriented toward audiences who share certain characteristics, depending on the product or person being promoted.
- A Recommender System (RS) refers to a system that is capable of predicting the future preference of a set of items for a user, and recommend the top items.

Data Collection

Offline and online Data Collection Sources

Offline Data Collection	Online data Collection
Sensors	Open-sourced Government Portals, such as surveyofindia.gov.in
Surveys	Online Survey Sites
Interviews	Reliable websites etc. e.g. Kaggle.com
Focus Groups	Online Forums
Observations	Online polls
Records and documents	Open-sourced statistical website like – dat.gov.in

Data Formats/ Types Used in Data Science

Method	Description
CSV (Comma-separated values)	CSV is a way of storing information in a tabular format in plain text file.
Spreadsheets	Microsoft Excel, OpenOffice Calc etc.
XLSX	Microsoft Excel open XML Format Spreadsheet file
JSON (Javascript Object Notation)	Structured data in text-based format.
SQL (Structured Query Language)	Stored data in DBMS
XML (eXtensible Markup Language)	Structured data both in Human and Machine readable format.

Summary - Revisiting AI Project Cycle (Data Science)

- Like any other AI project, a data science project also undergoes the phases of AI project cycle where each phase is data-centric.
- After collecting data for a data science project, data is first cleaned, standardized and the missing data is handled.
- Then the data is analysed and visualized to know about the trends and key characteristics of data.
- Then for data modelling, the treated dataset is divided into training dataset and testing dataset and the AI-based model is trained using the training dataset using one of the training techniques - supervised/ unsupervised or reinforced learning.
- For model evaluation, the testing dataset is used, and the predicted values are compared with the actual results to determine the efficacy and efficiency of the model developed and trained.
- For collecting data for data science projects, there are many online and offline sources.
- There are many data types and formats used in data science projects such as **CSV, XLSX, spreadsheets, SQL, XML, JSON etc.**

Python for Data Sciences

1. Numpy
2. Pandas
3. Matplotlib
4. Basic Statistics

Python for Data Science

1. **NumPy Arrays** - A NumPy array is simply a grid that contains values of the same/homogenous type.

```
>>>import numpy as np
>>>List=[1,2,3,4]
>>>a1 = np.array(List)
>>>print(a1)
Output– [1,2,3,4]
```

Note –

- NumPy arrays are also referred to as ndarrays (N-dimensional NumPy data arrays).
- NumPy array looks similar to List but unlike Python List, we can't change the size of the array.
- Unlike List Every NumPy array can contain elements of homogenous types.

Python for Data Science

2. PANDAS- Pandas or Python pandas is Python's library for data analysis.

Note –

- The Pandas library has two primary data structures.
- Series(1-dimensional) and DataFrame (2-dimensional) that can handle the vast majority of typical cases in finance, statistics, social science, and many areas of engineering.
- Well suited for tabular data with heterogeneously typed columns.

Index	Data
1	'A'
2	'B'
3	'C'
4	'D'
5	'E'
1-dimensional data structure	

		Columns		
		A	B	C
INDEX	0	'Hello'	'Column B'	NaN
	1	'NO INFO'	'NO INFO'	'NO INFO'
	2	'A'	'Column B'	NaN
	3	'A'	'Column B'	NaN
	4	'A'	'Column B'	NaN
DataFrame object (2-dimensional Data Structure)				

Python for Data Science

3. MATPLOTLIB- It is a Python library used for Data Visualization

- **Data Visualization**– represents graphical or visual representation of information in the forms of charts /graphs and maps etc.

- For Data Visualization in Python, the Matplotlib Library's Pyplot Interface is used.

- Data Visualization is immensely useful in decision-making.

- Data visualization unveils patterns, trends, outliers, correlations etc.
- Basic Plot Types are
    1. Line chart - A Line chart is created using plot() function
    2. Bar Chart - A bar hart is created using bar() and barh() function
    3. Scatter Plot - A Scatter plot is created using scatter() function
    4. Pie Chart – A pie chart is created using pie() function
    5. Histogram Plot - Histogram is created using hist() function
    6. BoxPlot Chart

LET US REVISE

- NumPy is the core library for scientific computing in Python that can work with high-performance ndarrays and thus, it is widely used in data science projects too.
- The main data structure in NumPy is the ndarray, which is a shorthand name for an N-dimensional array, a container of items of the same type.
- NumPy arrays store homogeneous data, unlike list that store heterogeneous data and are faster than lists and take less memory than lists.
- Pandas is a software library written for the Python programming language for data manipulation and analysis, and is a popular choice for data sciences.
- Matplotlib is a Python library used for Data Visualization. Data Visualization is an essential component of data sciences.
- Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.
- You can create bar plots, scatter plots, histograms, and a lot more visualizations with Matplotlib.

- Mean the mathematical average of a set of two or more numbers Mode the value that appears most often in a set of data values
- Median the middlemost number or centre value in the set
- Outlier an observation that is numerically distant from the rest of the data
- Standard deviation the dispersion of a dataset relative to its mean
- Variance the average squared deviations from the mean

K-Nearest Neighbour Model

KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

Example: Suppose, we have an image of a creature that looks similar to cat and dog, but we want to know either it is a cat or dog. So for this identification, we can use the KNN algorithm, as it works on a similarity measure. Our KNN model will find the similar features of the new data set to the cats and dogs images and based on the most similar features it will put it in either cat or dog category.

KNN (K-Nearest Neighbour) algorithm is a Supervised Learning algorithm

K-Nearest Neighbour Model

KNN (K-Nearest Neighbour) algorithm is a Supervised Learning algorithm that classifies a new data point into the target class, counting on the features of its neighboring data points.

KNN modelling works around four parameters:

1. **Features** – The variables based on which similarity between two points is calculated.
2. **Distance Function** – Distance metric to be used for computing similarity between points.
3. **Neighbourhood (K)** – Number of neighbours to search for.
4. **Scoring function** – The function which finds the majority score and class for the query point.

Let Us Revise

- KNN (K Nearest Neighbour) algorithm classifies a new data point into the target class, counting on the features of its neighbouring data points.
- K Nearest Neighbour algorithm falls under the Supervised Learning category and is used for classification (most commonly) and regression-based problems.
- K refers to the number of neighbours to be considered for a query point.
- The KNN algorithm uses methods and techniques to determine an optimal value for K.