Diplomawallah.in

Date
Page

SANGAM

Diplomawallah.in

Artificial Intelligence

Artificial Intelligence is a branch of computer science that enables machine to perform tasks that normally require human intelligence.

AI allow machines to learn from experience, understand language, recognize patterns and make decisions.

It is designed to solve problems and automate tasks without human intervention.

AI system can analyze data, adapt to new situation and improve performance overtime.

It is used voice assistant, self driving cars, medical diagnosis, and recommendation systems.

In short, AI makes think, learn and act like humans.

* AI works.

i) Data collection.

AI start with collecting data from various sources such as;

- Text, images, video, and audio
- Sensor data from devices.
- Historical records or online databases.

Ex - Self-driving cars collect data from cameras, sensors & GPS.

Diplomawallah.in

Date
Page

2. Data processing
The collected data is cleaned and organized to remove errors or irrelevant information.
This ensure AI system works with accurate and usable data.

3. pattern Recognition
AI uses algorithm to find patterns and relationship in the data.
- ML helps the system learn from examples.
- Deep learning use neural networks to handle complex data like image or speech.

4. Decision Making
AI makes decision by analying patterns and applying logic or rules.

5. learning & Improvement
AI system Improve over time using feedback from their action: correct decisions are reinforced.

6. Output/Action
finally, AI perform tasks or provide results:
- Generating recommendation.
- Controlling robots, predicting outcomes.

Diplomawallah.in

Purpose of AI

→ The main purpose of AI is to make Machines Smart and capable of performing tasks that normally require human intelligence

Its main Aim

→ Reduce human efforts by automating repetitive tasks.
→ Solve complex problems faster and efficiently.
→ Make accurate decisions using data analysis.
→ Process large amounts of data quickly.
→ Mimic human-like intelligence for learning reasoning and decision making.

Types of Artificial Intelligence:

→ Artificial Narrow Intelligence (ANI) - also called weak AI
→ performs Specific tasks only.
→ Example: Google Search, Siri, Alexa.

→ Artificial General Intelligence (AGI) - Strong AI
- Can perform any intellectual task like a human.
- Example: Still theoretical, not fully developed.

→ Artificial Super Intelligence (ASI)
→ AI Surpass human intelligence in all areas.
- Eg - Future concept purely theoritical.

Diplomawallah.in

Date
Page

Application AI

→ Healthcare. → Disease diagonsis, medical imagiy durig discovery.

→ Finance — Fopaud detection, algorithmic tyading, Coredit scoring.

→ Education — personalized learning, grading automyation.

→ Security — Face recloggition, Surveillance System.

→ Robotics — Industrial automyation, dyones, warehouse robots.

Q. why does ethics of AI Impoutant — and g gi explain with the real life example.

AI Ethics o is a set of moral principles and guidlines that ensure AI System are used responsibility, safely and faiuly. It deals with the question like: —

→ IS the AI making faiu decision?
— IS it protecting puivacy?
— could it harm humani oi society unintentionally?

Diplomawallah.in

Ethics of AI important because: —

1. Ensure fairness - prevent bias in AI decisions.
   Ex- An AI recruitment system should not favour one gender over another.

2. protect privacy
   → Safeguard personal data from misuse.
   Ex - AI in social media should not expose private user data.

3. Increase safety - Avoids harm to human or the environment.
   Ex - Self driving cars must follow strict safety rules to prevent accidents.

4. Build Trust - Ethical AI ensures users can trust the technology.

Example
· Recruitment and Hiring AI
Issue - AI algorithm has given instruction to favour certain gender only.
Example - Amazon scrapped its AI hiring tool because it discriminated against female candidates, as it was trained on historical male dominated data.
Ethical Concern - fairness, avoiding bias in decision-making.

Diplomawallah.in

Q. Describe any two types of cloud service model.

Cloud Computing provides computing resource over the Internet. These resources are delivered in different service models depending on the level of control and responsibility for users.

1. Infrastructure as a service. (IaaS)
IaaS provides computing resource over the internet, such as servers, storage, and networking.
→ user can manage and install their own operating system, application and data.
→ The cloud provider manages the physical infrastructure.

Ex- Amazon web service, microsoft Azure virtual Machines.

2. Software as a Service (SaaS)
SaaS provider ready to use software application over the internet.
→ user access software via a web browser or app.
→ The cloud provider manages everything hardware, OS, and application updates.
Ex- Google workspace (Gmail, docs, drive), Zoom

Diplomawallah.in

AI SDLC

Date
Page

Q. Explain the role of Bigdata in AI/ML.

Bigdata refers to extremely large and complex datasets that are difficult to process using traditional methods.
→ Includes structured data (databases), semi-structured data (logs, XML) and unstructured data (images, videos, text, Social media pages).

It is Characterized by 3Vs -
1. Volume - Huge amount of data.
2. Velocity - data generated rapidly in real-time.
3. Variety - different types of data.

Types
1. Structured data - Data is organized in a fixed format, usually in rows and columns.
Ex- Customer, name, phone number, & email address.

2. Semi-structured data.
→ data that does not follow a strict structure like databases but has some organization.
(tags, key metadata). Ex- JSON, email, XML.
3. unstructured data.
→ data that has no predefined format or structure. It is complex and difficult to process using traditional tools.
Ex- Text documents, PDFs, Images, audio, Social media etc.

Diplomawallah.in

Bigdata is important in AI and ML :—

1. High quality Training data
   → Machine learning algorithm improve accuracy with more data.
2. Better pattern Recognization
   → Bigdata helps AI driven hidden patterns in complex datasets.
3. Real-Time decision making
   → Bigdata allow AI to process live streaming data and make instant decisions.

4. Scalability and continuous learning
   - Large datasets helps AI system adapt & improve over time.

5. Supporting predictive and prescriptive Analytics.
   - AI use Big data to forecast trends and suggest actions.
   Ex - Predicting stock market trends or demand for products in retail.

Ex - Healthcare, E-commerce, Finance, Social media etc.

Diplomawallah.in

Date
Page

SDLC of AI/ML

The AI/ML SDLC describes the complete process of building, training, deploying and maintaining an AI System.

Unlike traditional Software, AI/ML projects are data-drive. So the lifecycle several data-related steps.

1. Problem defination
   In this stage we define :-
   → what problems that AI system will solve.
   → The goal & requirement
   → The expected output & performance. Metrics.
   Ex- predicting house prices, detecting fraud etc.

2. Data Collection
   AI depends on data. In this phase we gather all the necessary datasets from:
   • Databases, Sensors, APIs, Web Scraping
   O/P → Raw datasets.

3. Data Cleaning & Preprocessing,
   The collected data is cleaned & prepared so the model can learn accurately.
   Includes :-
   • Handling missing values, Remove duplicate, Encoding categorical data, Features Engineering.
   O/P → High Quality processed data.

Diplomawallah.in

4. Model Selection

Choose the appropriate machine learning or deep learning algorithm based on:

- problem type (Classification, regression, clustering etc.)
- Dataset size and complexity.
- Required accuracy & speed.

Ex-

Linear Regression, Random Forest, Neural Networks etc.

5. Model Training

This is the phase where the model actually learns using the training data.

The algorithm defines the pattern and builds a predictive model.

Tool used :-

Python, TensorFlow, Scikit-learn

Opp → A trained model.

6. Model Evaluation

The trained model is tested on unseen (test) data to check :

- Accuracy, F1 Score, Confusion Matrix etc.

of steps

Diplomawallah.in

7. Deployment —

The final model is deployed into a real-world environment where users can interact with it (Deployment options(s))

→ AWS, Mobile apps, Edge devices, FlaskAPI

8. Monitoring & Maintenance

Once deployed the model is continuously monitored to ensure :-

- Stable accuracy
- No performance degradation.

9. Continuous Improvement —

New data is collected and the model is retrained or updated for better and reliability.

Diplomawallah.in

Date
Page

Machine learning

The machine learning workflow explain how an AI/ML system is built from start to finish.
It mainly includes three major stages:—

(I) Data Engineering
(II) Model Training
(III) Model Deployment—

These three steps work together to create complete AI solution.

1. Data Engineering
Data Engineering the first & most important step.
In this step we collect and prepare the data so the machine can learn properly.
Activity
→ Collecting data from different sources.
→ Cleaning the data.
→ Converting raw data into useful form.

It provide clean and high-quality data for model.

Diplomawallah.in

2. Model Training
After data ready the next step is to train the machine learning model.
Activities:-
→ Choosing the correct algorithm
→ splitting data into training & test sets.
→ Testing the Model on new data.
→ Adjusting parameters to improve accuracy.

Ex- Training a face Recognition Model.

3. Model Deployment
Once the model performs well, it is deployed so real users and system can use it:-
Activities
→ Converting the model into an API or web service
→ Deploying it on cloud platform.
→ Monitoring the Model's performance.
purpose:-
We use trained Model in real-life applications like apps, websites and business system.

- Data Engineering prepare the data.
- Model Training use the data to build intelligence.
- Model deployment makes that intelligence available to users.



Diplomawallah.in

A Company wants to predict sales using ML what kind of data can they use & why?

To predict the sales using machine learning the company needs data such as :—

i) Historical Sales data
ii) production Data
iii) customer Data
iv) & location & store data
v) Seasonal and Time data.

To predict sales a company can use historical Sales data, product prices, customer details, marketing activities, Seasonal & location data.
Machine learning use data to learn pattern.

Q. describe

Diplomawallah.in

Describe the process of simple AI model on a public cloud platform including basic steps.

A public cloud platform (like AWS, Google Cloud, or Azure) provides tools to easily build and deploy AI model.

1. upload data.
2. prepare and clean data.
3. Select and Algorithm/model type
4. Train the model.
5. Evaluate the model.
6. Deploy the model.
7. Monitor & improve.

Diplomawallah.in

Date
Page

Traditional Software development VS AI Software development

1. Approach
- Traditional SD:- Works on rule-based logic. developers writes exact instructions (if, else, loops).
- AI SD:- Works on data based learning. The model learns patterns from large datasets rather than fixed rules.

2. Input & output
→ Traditional: Input + predefined Rules → output (System follows what developer coded)
→ AI:- Input + Trained Model → output (Model predicts or decides based on learning).

3. Development process
Traditional:
- Requirement analysis
- design
- Coding
- Testing
- Deployment (fixed structure, predictable behaviour)

Diplomawallah.in

AI :-

- Data Collection
- Data cleaning
- Feature engineering
- Model training
- deployment (depends heavily on quality of data).

4. Skill Requirement →

a Traditional

Knowledge of programming languages, algorithm databases, debugging.

- AI

Knowledge of python, ML algorithm, statistics deep learning, data engineering.

5. Behaviour

→ Traditional result is always consistent because rules don't change.

→ AI result vary because model prediction depend on training data & probability.

6. Example

→ Traditional used in banking system, billing system, calculators.

→ AI used in Chatbots, recommendation systems, face recognition, self driving cars.

Diplomawallah.in

Week.02

Machine learning

Q. Why is groupby important for Segregating and Summarizing data in pandas? Explain with example.

In data analysis, large dataset often contains multiple categorical or groups. To understand the data better, we need to segregate it into groups and calculate summary statistics like mean, sum, or count.

Importance of groupby:-

1. Segregation of data: It divides data into groups based on a column.
2. Aggregation & Summarization: It allow computing statistics (mean, sum, min, max, count) for each groups.
3. Finding patterns & insights: Helps to detect trends, performance differences or anomilies in different groups.
4. Efficiency: Eliminate the need for manual loops over groups.

Diplomawallah.in

5

```
import pandas as pd

data = {
    'Student': ['Sagar', 'Sangam', 'Sagar', 'Sangam',
                'Gautam', 'Gautam'],
    'Subject': ['math', 'Math', 'physics',
                'physics', 'Math', 'physics'],
    'Marks': [85, 90, 95, 80, 70, 75]
}

df = pd.DataFrame(data)

grouped = df.groupby('Student')['Marks'].mean()

print(grouped)
```

df.groupby('Student') → Data is divided by each Student

['Marks'].mean() → Computes the average marks for each Student.

Op →

Student	
Sagar	90.0
Sangam	85.0
Gautam	72.5

Name: Marks, dtype: float64

The op shows the mean marks of each Student Summarizing the dataset.

Diplomawallah.in

Q. Why are pivot and melt functions used in pandas? Explain with an example.

In data analysis, dataset can have wide or long formats:
- wide format: many columns for different variables.
- long format: fewer columns, more rows.

pandas provides pivot and melt functions to reshape datasets:
- pivot → converts long format data into wide format.
- Melt → converts wide format data into long format.

Reshaping is important for data analysis visualization
Importance of pivot and melt function:

1. pivot:
- Helps to reorganize data by turning unique values of a column into new columns.

Diplomawallah.in

- useful for summarizing and comparing groups.

2. Melt:
- Helps to unpivot wide data into a long format.
- useful for plotting or applying functions that require long-format data.

Eg.

```
import pandas as pd

data = {
    'Date': ['2026-01-01', '2026-01-01',
             '2026-01-02'],
    'Item': ['Apple', 'Banana', 'Apple'],
    'Quantity': [5, 3, 6]
}
df = pd.DataFrame(data)
# pivot
pivot_df = df.pivot(index='Date', column='Item',
                    values='Quantity')
print("pivot Table: \n", pivot_df)
# melt
melt_df = pivot_df.reset_index().melt(id_vars=
    'Date', value_name='Quantity')
print("\nMelted Dataframe: \n", melt_df)
```

Diplomawallah.in

Date
Page

Opp →

Pivot Table:

Item	Apple	Banana
Date		
2026-01-01	5	3
2026-01-02	6	NaN

Melted Dataframe:

	Date	Item	Quantity
0	2026-01-01	Apple	5.0
1	2026-01-02	Apple	6.0
2	2026-01-01	Banana	3.0
3	2026-01-02	Banana	NaN

- pivot converts the Item column values into separate columns (Apple, Banana) for each date.
- Melt converts back into a long format with Date, Item and Quantity columns.

- pivot → wide format summary.
- Melt is used for → long-format analysis or visualization.

Diplomawallah.in

Q3. What is the role and importance of handling missing values in data processing? Discuss its advantages and disadvantages.

In real-world datasets, it is very common to find missing values due to reasons such as data entry errors, incomplete surveys, sensor failures, or data collection issues.

These missing values can negatively affect data analysis and machine learning models.

Role & Importance of Handling Missing Values:

1. Improves Data Quality:
   Handling missing values ensures that the dataset is clean, consistent and reliable for analysis.
2. Accurate Analysis:
   Missing values can distort statistical calculations like mean, median, and variance.
3. Better Model performance:
   Machine learning algorithms usually do not work properly with missing data.
4. Avoids Errors.
5. Prevent Bias.

Diplomawallah.in

Date
Page

Advantage
→ Reliable and meaningful results.
→ ML Models accuracy
→ Runtime and calculation errors.
→ Suitable for visualization and analysis.

Disadvantage of Handling Missing values:

1. Data loss may occur if rows or columns are removed.
2. Incorrect filling can introduce bias.
3. Imputation methods may distort original data patterns.

Eg

```
import pandas as pd

data = {'Marks': [80, 90, None, 70, None]}
df = pd.DataFrame(data)

df['Marks'] = df['Marks'].fillna(df['Marks'].mean())

print(df)
```

None ki mean ki value aa jayegi.

$$\frac{80+90+70}{3} = \frac{240}{3} = 80$$

	Marks
0	80.0
1	90.0
2	80.0
3	70.0
4	80.0

Diplomawallah.in

Q. Explain how data is processed and split into training and testing data in ML.

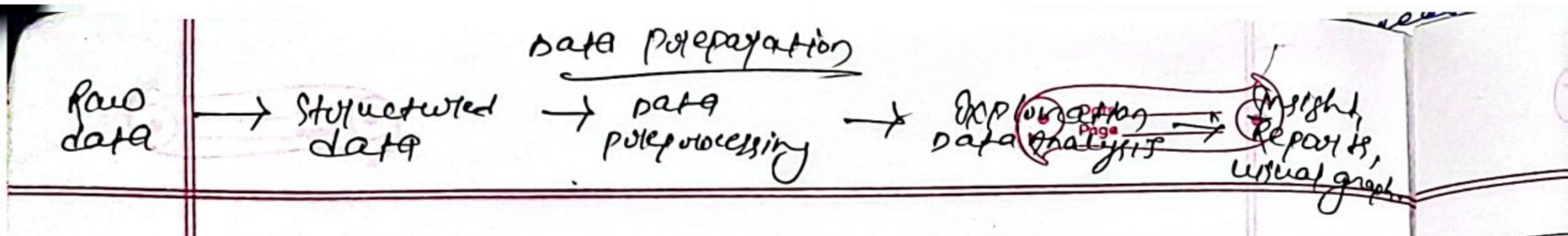
In ML, raw data cannot be directly used to build a model. Therefore, data must be processed and then divided into training and testing datasets to ensure accurate and reliable model performance.

1. Data processing
Data processing is the step in which raw data is converted into a suitable form of ML:-
→ Missing values are removed.
→ Duplicate and incorrect data removed.
→ Categorical data changed into numerical.

2. Train-Test Split:
After data processing, the dataset is divided into two parts:
- Training data: used to train the Machine learning model.
- Testing data: used to test the model's performance on unseen data.

Generally 80% of the data is used for training and 20% for testing.

Diplomawallah.in



Q. Why does visualization help in Exploratory Data Analysis (EDA)?

Exploratory Data Analysis is the process of understanding data before Applying machine learning models.
Data visualization play an important role in EDA because it help to analyze data visually and easily.

Role of visualization in EDA:

- visualization helps to understand the distribution of data.
- It makes it easy to identify patterns and trends in data.
- It help in detecting outliners and anomalies.
- Relationship b/w variables can be clearly observed.
- Complex data becomes simple and understable.

Common visualization Techniques:

- Histogram
- Box plot
- Bar chart
- Line graph.
- Scatter plot

Diplomawallah.in

Q. How can matplotlib or seaborn be used to display the distribution of numerical data from a pandas dataframe using a histogram?

In data Analysis, understanding the distribution of numerical data is very important!

A histogram is a graphical representation that shows how data values are distributed over different intervals.

Matplotlib and seaborn are commonly used python libraries to create histogram from pandas dataframes.

Using Histogram for data Distribution:
- A Histogram divides data into bins (ranges).
- It shows the frequency of values in each bin.
- It helps to understand whether data is normal, skewed or has outliers.

Diplomawallah.in

Date
Page

using Matplotlib:
Matplotlib is a basic visualization library used to plot histogram.
Ex.

```
import matplotlib.pyplot as plt
plt.hist(df['column_name'])
plt.xlabel('values')
plt.ylabel('Frequency')
plt.title('Histogram of Numerical Data')
plt.show()
```

using Seaborn.
Seaborn provides a more attractive and statistical visualization

```
import seaborn as sns
sns.histplot(df['column_name'])
```

Advantage of Histogram
→ Easy to understand data distribution
→ Helps detect skewness
→ Identifies outliers
→ Useful in Exploratory Data Analysis (EDA)

Dp
Departm
HR 32
IT 45
Name: Salar

Diplomawallah.in

Q. How do you group data by a columns, Calculate the mean, and display the result in pandas?

In data analysis, grouping data helps to summarize and analyze large datasets. pandas provides the groupby() function to group data based on a column and apply aggregation function such as mean.

Steps

→ Select the data frame
→ use the groupby() function on the required column.
→ Apply the mean() function.
→ Display result.

Eg- import pandas as pd

data = {
'Department': ['IT', 'IT', 'HR', 'HR'],
'Salary': [40000, 50000, 30000, 35000]}

df = pd.DataFrame(data)
result = df.groupby('Department')['Salary'].mean()
Print(result)

Diplomawallah.in

Date
Page

- groupby('Department') groups data based on department.
- ['Salary'] selects the numerical column.
- mean() calculates the average salary for each group.

Conclusion
Using groupby() with mean() in pandas allows efficient grouping and summarization of data, making data analysis easier and more meaningful.

Diplomawallah.in

Q. How can you use pivot and melt function to check and reshape a dataset in pandas?

In data analysis, dataset are often not in the required format. pandas provided pivot and melt functions to reshape data for better analysis and visualization.

1. pivot function:
The pivot() function is used to convert long format data into wide format.

Uses of pivot:
- Restructures data
- Helps in summarizing data
- Makes comparison easier.

Ex
df.pivot(index='Date', column='product', values='Sales')

2

Diplomawallah.in

2. Melt function :

The melt() function is used to convert wide format data into long format.

*Use Of melt :
- Simplifies data structure
- useful for visualization
- Helps in data cleaning.

Ex
pd.melt(df, id_vars=['Date'], value_vars=['Sales', 'Profit'])

Conclusion
pivot and melt functions are powerful tools in pandas that help reshape and clean datasets, making data analysis more flexible & effective.

Diplomawallah.in

Q. How can outliers be detected in a numerical dataset using the IQR method?

Outliers are data values that very different from other observation in a dataset.

Detecting outliners is important because they can affect the accuracy of data analysis and ML models. The Interquartile Range (IQR) method is a commonly used statistical technique

Steps Of IQR method:

1. Arrange the data in ascending order.
2. Calculate the first quartile (Q1).
3. Calculate the third quartile (Q3).
4. Compute IQR = Q3 - Q1
5. Calculate lower and upper bounds:
   - lower bound = Q1 - 1.5 X IQR
   - upper bound = Q3 + 1.5 X IQR
6. Any value below the lower bound or above the upper bound is considered an outlier.

Diplomawallah.in

```
Q1 = df['Column'].quantile(0.25)
Q3 = df['Column'].quantile(0.75)
IQR = Q3 - Q1

outliers = df[(df['Column'] < Q1 - 1.5 * IQR)
              | (df['Column'] > Q3 + 1.5 * IQR)]

print(outliers)
```

Advantage

- Simple and effective
- Not affected by extreme values
- Works well for skewed data.

Diplomawallah.in

Date
Page

* Cloud Computing

Cloud Computing is a technology that provides Computing resources such as servers, storage, databases, networking, software and services over the internet, allowing users to access and use them without owning physical hardware.

In Simple words

cloud computing means using someone else Computer on the internet to store data and run programs instead of using your own computer.

The user does not worry about buying, installing or maintaining hardware & s/f.

* Why Cloud Computing is Needed

Earlier companies had to:

- Buy expensive servers.
- Maintain h/w & s/w.
- Hire technical staff
- Upgrade system regularly.

Cloud computing removed all these problems by providing resources online anytime anywhere.

Diplomawallah.in

Date
Page

* Features

1. On-demand Access — Resources are available whenever required.
2. Internet Based — Services are accessed using the internet.
3. Scalable — Resources can be increased or decreased easily.
4. Cost Effective — pay only for what we use.
5. High Availability: — Services are always available with backup and recovery.

* Advantage

- No need to buy physical servers.
- Low cost & maintenance.
- High Storage capacity
- Fast performance.
- Data backup & recovery.

Ex Google drive, Gmail, Dropbox, AWS, Microsoft Azure, Netflix, etc.

Diplomawallah.in

Date
Page

✓ Cloud Computing deployment Models:

A Cloud deployment Model defines the type of cloud environment used for computing and the way cloud resources are managed, accessed and controlled.

* It tells where the cloud is located and who can use it.

There are 4 Main deployment models :→

1. public Cloud

public cloud is a cloud environment that is owned and operated by third-party providers and available for public use our the internet.

* Characteristics:

- shared by many users, Easily accesible, low cost
- Advantage

- No hardware cost, Easy to use, Highly flexible,

Ex- Amazon AWS, Microsoft Azure, Google cloud.

Diplomawallah.in

2. private cloud
private cloud is a cloud infrastructure that is used by only one organization.
Characteristics
• High security, full control
Advantage:
Strong data security, better privacy, more control over resources.
Ex - used by banks, government organizations, large companies.

3. Hybrid cloud
Hybrid cloud is a combination of public cloud and private cloud, allowing data and application to be shared b/w them.
Characteristics:
• Flexible, cost effective, secure and scalable.
Ex
Enterprises handling sensitive customer data.

4. Community cloud:
Community cloud is shared by multiple organization that have similar requirements such as security, policy or compliance.
Characteristics:
• Shared among a group, more secure than public cloud, cost is shared.

Diplomawallah.in

Ex- Hospitals, banks, research institutions, universities.

✗ Cloud Service Model

Cloud Service Models describe the type of cloud services provided to users and the level of control they have over resources.

There are three Cloud Service Models:-

1. IaaS - Infrastructure as a Service.
2. Paas - Platform as a Service.
3. Saas - Software as a Service.

IaaS	Paas	Saas
Low	medium	Low.
High	User manages	provider
User manages	apps	manage everything
OS & apps	Developers	End users
System admins		
user.		

Diplomawallah.in

Parameter	public cloud	private cloud
cost	low initial cost with a pay as you go model. No need to invest in h/w or infrastructure	High initial cost due to dedicated infrastructure, h/w & maintenance.
Security	Managed by the cloud provider.	offer higher security and privacy as infrastructure is dedicated to a single organization.
Scalability	High scalable with virtually unlimited resources available on demand.	Scalability is limited to the organization own infrastructure capacity.
performance	performance may very due to shared resources.	provides more consistent and predictable performance with dedicated resources.
Control	Limited control over infrastructure & Configuration.	Full control over infrastructure, configuration & policies.
Maintenance	Maintenance, updates, and upgrades are handled by the service Providers.	Organization is responsible for Maintenance and upgrades.

Diplomawallah.in

Date
Page

✱ Serverless Services

Serverless Computing is a cloud computing model in which the cloud provider automatically manages servers, infrastructure and scaling, and the user only focuses on writing and running application code.

Although servers exist, the user does not need to manage them, so it is called serverless.

✱ In traditional computing, developer must:
- Buy servers, configure them, Maintain hardware, Manage scaling.

In Serverless computing, all these responsibilities are handled by the cloud service provider.
Developer simply upload their code and it runs automatically when required.

Diplomawallah.in

Working

1. Developer writes the code.
2. Code is uploaded to the cloud platform.
3. The cloud provider executes the code when an event occurs (request, upload, click etc.)
4. The user is charged only for the time the code runs.

Feature

- No server management, Automatic scaling High availability, pay-per execution model
- Fast deployment.

Advantage

- Reduces operational cost.
- No infrastructure maintenance
- Faster Development
- High performance.

Disadvantage

- Limited control over server environment
- Vendor dependency • Cold start delay in some cases.

Example.

AWS Lambda, Google cloud function, Azure functions, IBM cloud function.

Diplomawallah.in

* Major Cloud Service providers

Cloud Service providers are Companies that offer cloud computing Services such as Storage, Servers, n/w, databases, S/w and development tools over the Internet.

They own and manage large data centers and provide Service to user on a pay as you use basis.

Some major cloud Service providers.

1. Amazon Web Services (AWS)

AWS is the largest and most popular Cloud Service provider in the world.

* Services offered

- EC2 - virtual Server
- S3 → Storage
- RDS → database
- lambda → Serverless computing
- VPC → Networking

Diplomawallah.in

Date
Page

Advantage
• Highly reliable, Global infrastructure, Strong Security.

2. Microsoft Azure
Azure is Microsoft cloud platform mainly used by enterprises.
Services offered:
• Azure virtual machines.
• Azure Storage
• Azure SQL database
• Azure function
• App Services.

Advantage
• Strong integration with windows, Good for enterprise solutions, High Security.

3. Google cloud platform (GCP)

GCP is Google cloud platform, Strong in data analytics and AI.
Services offered
• Compute Engine, Cloud Storage, Big Query, Kubernet Engine, cloud functions.

Diplomawallah.in

4. IBM Cloud
IBM Cloud focuses on enterprise solutions, AI and hybrid cloud.
Services offered
• Virtual servers, Cloud database, Watson AI Kubernets services.

5. Oracle Cloud
Oracle Cloud specialize in database services.
* Services offered
• Oracle database Cloud
• Compute Cloud
• Storage cloud
• Application cloud.

Diplomawallah.in

Date
Page

* Virtualization

Virtualization is a technology that allow a single physical computer (server) to be divided into multiple virtual machines each acting like an independent computer.

* Virtualization means creating virtual version of physical resources such as server, storage, network and operating system.

* Cloud Computing VS Big data

* Cloud Computing

Cloud Computing is the delivery of computing resources such as server, storage, databases networking & s/w over the Internet on pay as you go use basis.

* Big data

Big data refers to extremely large and complex datasets that cannot be processed using traditional data processing techniques.

Diplomawallah.in

It is characterized by 5Vs
1. Volume 2. Velocity 3. Variety
4. Veracity 5. Value.

1. Volume
Volume refers to the huge amount of data generated every seconds,

→ Data comes from social media, sensors transaction, videos, images etc.
→ The size of this data is very large (in TB, PB, ZB).
Ex- facebook, youtube, Amazon generates terabytes of data daily.

2. Velocity
Velocity mean the speed at which data is generated, collected and processed.
→ Bigdata system must handle data in real time or near real time.
Example:- stock market transactions, online payment live streaming data.

Diplomawallah.in

3. Variety

Variety refers to different types of data formats.

Data can be:

- Structured (tables, databases)
- Semi-Structured (XML, JSON)
- Unstructured (images, videos, text, emails)

Ex

photos on Instagram, Videos on youtube
text on Twitter.

4. ~~Variety~~ Veracity

Veracity refers to the accuracy and reliability of data.

→ Not all data is correct. Some data may be incomplete, noisy or misleading.
Bigdata system must handle uncertain & inconsistent data.
eg:- fake reviews, duplicate records, incorrect sensor data.

Diplomawallah.in

5. Value

- Value refers to the usefulness of data.

↑ The main goal of Bigdata is to extract meaningful information that helps in decision making.

Ex

Customer behaviour analysis, business prediction, market trends.

* Bigdata - Sources of data.

Sources of Bigdata are the different places from where huge amount of data are generated continuously in digital form.

- Major sources of Bigdata.

1. Social media

Social media platform generate massive data every second though:

post, comments, likes, photos & videos.

Ex

Facebook, Instagram, Twitter, youtube.

Diplomawallah.in

2. Machine & Sensor Data
Data generated by Machines, Sensors and devices automatically.
Ex
IoT sensor, Smart meters, UPS System, Weather Sensors.

3. Transactional Data.
Data created during business transactions.
Ex
Online Shopping, Bank transactions, ATM usage, Credit/debit card records.

4. Mobile Devices
→ Smartphone generate data through:
- Apps
- Calls
- Messages
- Location tracking

5. Enterprise System
Organization produce data through internal System.
Ex
ERP System, CRM S/W, HR records, Sales database.

Diplomawallah.in

6. Web & log data
Data generated from websites and servers,
Ex -
• website clicks, server logs, search history
cookies.

7. Multimedia data
Includes images, audio, video files,
Ex
CCTV recordings, medical images, online
videos, voice recordings.

Diplomawallah.in

Unit chang.

Date
Page

* python packages for Machine learning and Deep learning:

→ Scientific Computing libraries.

python is the most popular language for machine learing and deep learning because of its powerful scientific computing libraries.

These libraries provide tools for data handling Mathematical computation, model building and visualization

Major Scientific Computing libraries in python

1. Numpy (Numerical python)
   → store and process large numerical datasets.
2. Scipy (Scientific python)
   → used for advanced scientific and Mathematical computations.

* features

Optimization, Integration, Statistics, Signal and image processing.

Diplomawallah.in

Role in ML/DL
→ Support to complex mathematical calculations used in ML models.

3. pandas
→ Used for data manipulation and analysis.
Features
Dataframe and Series, Data cleaning, Handling missing value, Data transformation
Role in ML/DL
Prepares and manage datasets for training models.

4. Matplotlib (Data visualization)
→ Helps understand data patterns and model performance.

5. Seaborn
Create attractive and informative statistical graphics.

6. Scikit-learn
Main library for machine learning.

7. Tensorflow
Used for deep learning.

8. PyTorch
→ popular Deep learning library.

Diplomawallah.in

* Visualization Libraries (python)

Visualization libraries are used to represent data graphically so that the patterns, trends and relationship in data can be easily understood.

Major visualization libraries.
1. matplotlib.
2. Seaborn
3. plotly
4. Bokeh

* Algorithmic libraries

Algorithmic libraries provide ready-made algorithm for building machine learning models.

Eg Major Algorithmic libraries

1. S.CIKit-learn

Most popular ML library in python.
* provides Algorithms.
→ Linear Regression
→ logistic Regression
→ decision Tree

Diplomawallah.in

→ Random Forest
→ K-means
→ SVM
→ Naive Bayes.

2. TensorFlow
used for deep learning algorithms, provides
- Neural Networks,
- CNN
- RNN
- LSTM
- Deep learning models.

3. PyTorch
Another powerful deep learning library.
*Features
- Flexible neural network building
- Easy experimentation.
- High performance.

4. Keras
High-level deep learning library built on Tensorflow.
used for fast Development of deep learning Models.

Diplomawallah.in

Date
Page

1

* Environment Setup: Install Required package & Explore package.

1. Environment setup for ML & DL.

Environment Setup means preparing the computer system with all necessary sfw and libraries required to develop Machine learning and deep learning application.

Steps to set up python Environment

Step 1: Install python
python --version
download & install python form official website.

Step 2: Install package Manager (pip)
pip is used to install in python packages.
pip --version.

Step 3: Install Required Packages
pip install numpy
" " pandas
" " Matplotlib
" " Scipy

Diplomawallah.in

Step 4: To explore Installed packages.

1. Numpy

```
import numpy as np
print(np.array([1,2,3]))
```

2. pandas

```
import pandas as pd
data = pd.Series([10,20,30,40])
print(data)
```

3. Matplotlib

```
import matplotlib.pyplot as plt
plt.plot([1,2,3],[4,5,6])
plt.show()
```

4. Scikit-learn
Used for ML algorithm:

```
from sklearn.linear_model import
LinearRegression
model = LinearRegression()
```

5. Tensorflow

```
import tensorflow as tf
print(tf.__version__)
```

6. PyTorch

```
import torch
print(torch.__version__)
```

Diplomawallah.in

Date
Page

* Data visualization with python

Data visualization is the graphical representation of data using charts, graphs and plots.
It help to understand large and complex datasets easily & quickly.

Python provides powerful visualization libraries such as Matplotlib, seaborn, plotly and Bokeh for effective data visualization.

* Visualization fundamentals.

visualization fundamentals are the basic principles of presenting data visually.

1. Types of visualization
   - Line chart, Barchart, pie chart, Histogram, Scatter plot, Box plot, Heat Map.
2. Components of a Good Visualization
   - Title, x-axis and y-axis labels.
   - legends, colours and scale
   - proper layout.

Diplomawallah.in

3. purpose of visualization

- To represent data clearly.
- To identify trends and patterns.
- To detect outliers.
- To compare data values.

* Why visualization is important?

1. Easy to understanding : Human understen images faster than numbers. visualization convert complex data into easy - to read graphics.

2. Detect patterns trends.
Trends, correlations and pattern become visible through graphs.

3. Better decision making
Managers and organization use visual reports for fast and accurate decisions.

4. Error detection.
data errors and outliers can be easily identified.

5. Effective communication
visualization helps communicate data insights clearly to others.

Diplomawallah.in

Date
Page

* 'Coordinate System and Axes'.
A coordinate system is used to locate points on a graphs using two axis.

- x-axis — horizontal
- y-axis — vertical.

~~Purpose~~

* Directory of visualization

Data type	visualization Type
→ Amount	Bar chart, column chart
→ distribution	~~His~~ Histogram, Box plot
→ Proportion	pie chart
→ x-y Relationship	line chart, Scatter plot
→ uncertainty	Error Bars, Confidence Intervals.

* Basics of python visualization with ~~matlot~~ matplotlib

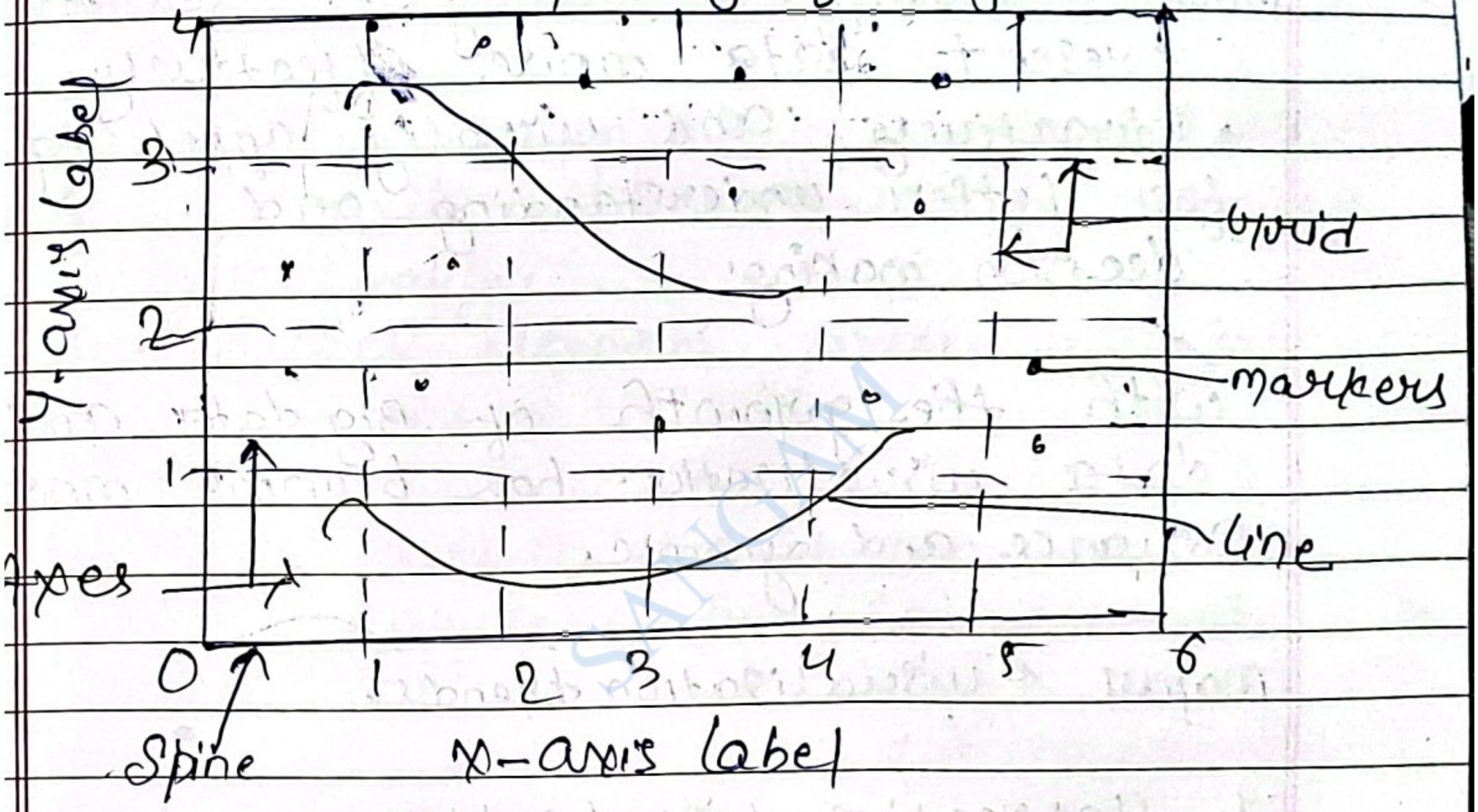
* Anatomy of a figure
A matplotlib figure contains:-
- Figure → whole image
- Axes — plotting area.

Diplomawallah.in

Date
Page

p. Title
• legend
• x & y labels.

Anatomy of a figure



— Blue Signal
— Black Signal

4 legends.

Diplomawallah.in

Visualization Trends.

Visualization trends refers to the modern techniques and styles used to present data more effectively, interactively and visually appealing for better understanding and decision making.

With the growth of Big data and AI, data visualization has become more advance and dynamic.

Major Visualization trends.

1. Interactive Visualization
User can interact with data by Zooming, filtering, clicking and hovering.
-Ex
Dashboards, web based charts (plotly, Power BI, Tableau)

2. AI - Driven Visualization
AI helps in automatically detecting patterns, trends and insights.
Ex - Auto generated reports, predictive charts.

Diplomawallah.in

3. Real-Time visualization
Data is displayed instantly as it is generated.
Ex - Stock market dashboards, Live traffic monitoring, Social media analytics.

4. Data Storytelling
Combining data with narrative to communicate insights clearly.
Ex - Business reports, presentation, journalism graphics.

5. Mobile friendly visualization
charts optimized for smartphones and tables.

6. Minimalist design
Simple and clean design focusing on important data.

7. Augumented & virtual Reality visualization
using AR/VR for immersive data exploration.

Diplomawallah.in

* Visualizing Categorical Data

Categorical data represents groups or categories such as gender, City product type departments etc.
Visualization helps compare value across different categories easily.

1. Visualize proportions.
used when we want to show how much each category contributes to the whole.
Common charts
• pie chart • Bar chart

2. Visualize data on Multi-plot Grid
- Used to compare multiple graphs in one figure.

3. Composite views for Informative Data Summary
Composite view combine multiple charts to present view in one visual.

Diplomawallah.in

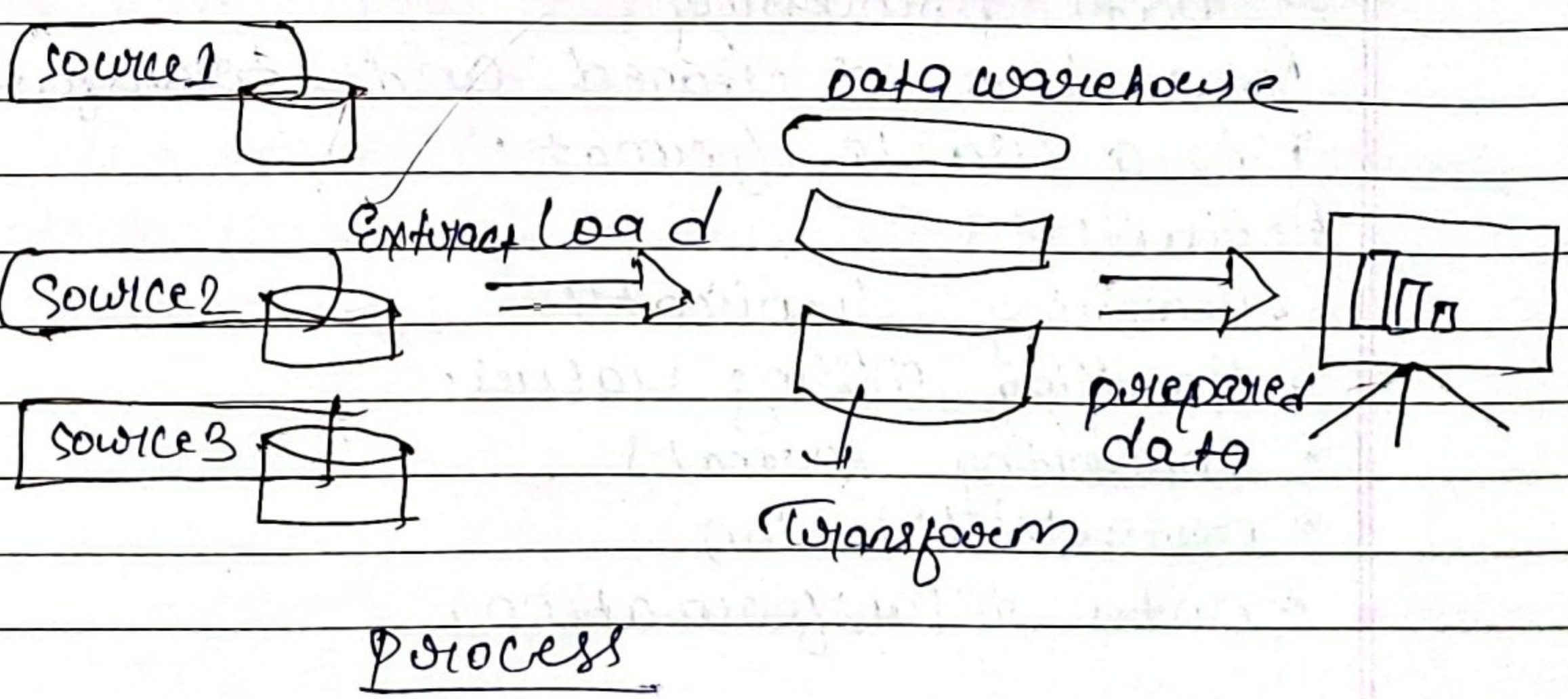
Date
Page

purpose:
- Show relationship, compare patterns
- Summarize data effectively.

Data Engineering pipeline

A data Engg. pipeline is a systematic process that collect raw data from different sources, process it, store it and prepare it for analysis, reporting and machine learning.

It ensure that data flow ~~automatically~~ smoothly from its origin to the final user in a clean, reliable and usable form.



Diplomawallah.in

1. Data collection
→ Data gathered from various sources.
Sources like
→ Database, files (CSV, Excel)
• Sensors & IoT device, Web application
Social media, APIs.

2. Data Ingestion
Data ingestion is the process of transferring collected data into the system.
It can be :-
• Batch ingestion — data collected at fixed intervals.
• Real-time ingestion — data collected continuously.

3. Data processing
Raw data is cleaned and transformed into a usable format.
Activities :
• Removing duplicate.
• Handling missing values.
• Converting formats.
• Data validation
• Data transformation.

Diplomawallah.in

4. Data Storage
• Processed data is stored securely for future use.
Storage Systems :-
Databases, data warehouse, data lakes, cloud storage.

5. Data Analysis
Data is Analyzed to extract insights.
Techniques :-
Statistical analysis, Machine learning, data mining.

6. Data visualization & Reporting
Final insights are presented using charts, dashboards and reports.

* Data Engg. pipeline is the backbone of modern data-driven systems.
It transform raw data into valuable information that driver intelligent decisions and innovations.

Diplomawallah.in

Date
Page

Types of data

Data can be classified different ways :-

* Based on collection method
a) Cross sectional data
Data collected at a single point of time.
Ex- marks of students in 2025.
b) Time series data
Data collected over a period of time.
Ex- Daily temperature of patna for one month.

* Based on number of variables.
a) univariate data : data involving one variable.
Ex - Height of student
(b) ~~Multivalue~~ multivariate data
data involving two or more variable.
Ex - Height, weight and age of students.

2. Variable Types
1. Categorical data - Data represented by categories.
Ex - Gender, blood groups department.
2. ordinal data
- Categorical data with order or ranking.

Diplomawallah.in

Date
Page

* Probability

In AI & ML probability is used to handle uncertainty, make prediction and take decisions when outcomes are not completely known.
Almost every ML model works by estimating the probability of different possible outcomes.

* In AI & ML probability is a numerical measure of how likely a particular output or event is, given the available data.

P(Y|X)

Which means probability of output Y given input features X.

* ML models do not give only decisions they give confidence in decisions.

Ex - A Spam classifier predicts, Email is Spam with probability 0.98.

Diplomawallah.in

Random Experiment →

* In ML a random experiment is the process of training a testing model on data where the outcome cannot be predicted with complete certainty.

Ex - Running a classifier on new data & observing the predicted class.

* Conditional and joint probability.

In AI & ML models learn the relationship b/w input features (X) and output labels (Y) using probability.

Two very important concepts for this learning probability

- Conditional probability
- Joint probability

* Conditional probability

Conditional probability is the probability of an event occuring given that another event has already occured.

In ML terms, it represents the probability of output Y when input X is known.

Diplomawallah.in

$P(Y|X)$

This is the core idea of supervised learning predict the output Y based on feature X.

Ex Disease prediction model:

$P(\text{Disease} | \text{Symptoms})$

Spam filter:

$P(\text{Spam} | \text{email content})$

Mathematical formula:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

2. Joint probability

Joint probability is the probability that two events occur together.

In ML, it measure how often a particular input X and output Y appear together.

$P(X, Y)$

* Joint probability is used to understand the relationship b/w features and labels.

Ex

$P(\text{Email contains "free"}, \text{Email is spam})$

* Relationship Between Conditional and Joint probability

$P(X, Y) = P(Y|X) \times P(X)$

Diplomawallah.in

* Bayes' Theorem

Bayes Theorem is a mathematical formula used to calculate the conditional probability of an event by reversing the given condition.

In AI & ML Bayes Theorem helps models update their beliefs when new data becomes available.

Statement of Baye's Theorem

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

* ML models use Bayes Theorem to revise prediction based on new information.

Ex - Disease prediction

We want to calculate:

P(Disease | Symptoms)

Using Bayes Theorem:

$$P(\text{Disease} | \text{Symptoms}) = \frac{P(\text{Symptoms} | \text{Disease}) \times P(\text{Disease})}{P(\text{Symptoms})}$$

Diplomawallah.in

This allow model to make more accurate prediction.

Q: In a spam detection system:
- 40% emails are spam
- 60% emails are not spam
- 80% of spam emails contain the word "free"
- 10% of non-spam emails contain the word "free"

If an email contains "free", what is the probability that it is spam?

Step1: define event
S = email spam, N = email not spam
F = email contains "free"
P(S) = 0.4, P(N) = 0.6
P(F|S) = 0.8; P(F|N) = 0.1

Step2: Use Bayes' Theorem

$$P(S|F) = \frac{P(F|S) \times P(S)}{P(F|S) \times P(S) + P(F|N) \times P(N)}$$

Step3: Substitute

$$P(S|F) = \frac{0.8 \times 0.4}{(0.8 \times 0.4) + (0.1 \times 0.6)} = \frac{0.32}{0.32 + 0.06} = \frac{0.32}{0.38}$$

Answer

$P(\text{Spam} | \text{"free"}) = 0.842 = 84.2\%$

Diplomawallah.in

Date
Page

* probability distribution

A probability distribution describes how the value of a random variables are distributed and how likely each value is;

In AI/ML probability distributions help models:.

- Handling uncertainity
- Make prediction
- understand data behaviour.

* Central limit theorem

It states that, for a sufficiently large sample size, the distribution of the sample mean becomes approximately normal, regardless of the original population distribution.

Importance in ML.

- use normal distribution for analysis
- make predictions
- Estimate confidence intervals.

Diplomawallah.in

Date
Page

* univariate Data Analysis

univariate data analysis is the analysis of a single variable to understand its distribution, pattern and behaviour.

It is the first step of data analysis before building ML models, characterizing data with descriptive statistics

Descriptive statistics summarize data numerically.

Main Measure	Meaning
Measure	
Mean	Average value
Median	Middle value
Mode	Most frequent value
Range	Max - Min
Variance	Spread of data
Standard deviation	Dispersion of data
Minimum & maximum	Smallest and largest values.

* univariate distribution plots.

- Histogram — shows frequency
- Density plot — Smooth version of histogram
- Box plot — shows median.

Diplomawallah.in

Date
Page

* Hypothesis

A hypothesis is a statement or assumption about a population parameter that can be tested using sample data.

* A hypothesis is a claim that we want to prove or disprove using data.

Types

1. Null hypothesis (H_0)
It states that there is no effect, no difference or no relationship.

Ex - H_0 : There is no difference b/w the average marks of two classes.

2. Alternative Hypothesis (H_1 or H_a)
The alternative hypothesis states that there is an effect, a difference, or a relationship.

Ex H_1 : There is a difference b/w the average marks of two classes.

Diplomawallah.in

Purpose of Hypothesis Testing
→ Make decision using data.
→ Validate assumptions.
→ Draw conclusion with confidence.

* Steps of Hypothesis Testing
1. formulate hypothesis (Ho and H1)
2. Choose Significance level (α)
3. Select test statistic
4. Calculate P. value
5. Make decision (accept or reject Ho)

Example
Suppose we introduce a new ML model and we want to test if it improves accuracy.
Ho : New model doesnt improve accuracy.
H1 : New model improves accuracy.

* P-Value and Significance level.
P-value and Significance level (α) are used to decide whether the null hypothesis should be accepted or rejected.

* P-value is the probability of obtaining the observed result assuming that the null hypothesis is true.
In simple: * P-value tells how the evidence is against the null hypothesis.

Diplomawallah.in

Date
Page

Significance level

Significance level is the maximum probability of rejecting the null hypothesis when it is actually true.
denoted by α.

Common value

0.05 (5%), 0.01 (1%), 0.10 (10%)

Decision Rule

Condition	Decision
P-value $\leq \alpha$	Reject H_0
P-value $> \alpha$	fail to Reject H_0

Ex If

P-value = 0.03

α = 0.05

Since 0.03 < 0.05 — Reject H_0

Errors, Test statistics, Types and Interpreting Test statistics :-

1. Error in hypothesis Testing

When making a decision about hypothesis two type of error can occur:

Diplomawallah.in

Type 1 error (α error)
Rejection the null hypothesis when it is actually true.

Ex - Type 2 error (β error)
Failing to reject the null hypothesis when H is actually false.

* Test Statistics
Test statistics is a numerical value calculated from sample data is used to decide whether to reject the null hypothesis.

Ex z-statistics, t-statistics, Chi-Square statistics, F-statistics (ANOVA)

Diplomawallah.in

Unit-5

Date
Page

Data preprocessing
Data preprocessing is the process of cleaning, transforming, and organizing raw data into a suitable format before it is used for analysis or for building machine learning models.

Preprocessing improves the Quality of Data.

Importance of Data preprocessing
→ Improves Data Quality
→ Increase Model Accuracy
→ Reduce Overfitting
→ Save time and Resources.
→ Ensure Consistency.
→ Handles Missing and Noisy data.

* Steps of pre processing Data

- Assessing data Quality
- Accuracy - Is data correct?
- Completness - Are values missing?
- Consistency - Is data uniform?
- Uniqueness - Any duplicate records?
- Validity - values in correct format?

Diplomawallah.in

Date
Page

* Data Anomalies.

→ Anomalies are unusual or incorrect data values that differ from expected pattern.

Types

- outliers
- Duplicate records.
- Inconsistent entries.
- Invalid values.

* Detect missing values with pandas.

using

df.info()

using

df.isnull().sum()

(Shows count of missing values per coloumn).

* Data Integration

Data integration is the process of combining data from multiple sources into a single, unified view.

It helps create consistent, accurate and complete datasets for analysis and ML.

Data is usually collected from databases, files, sensors, web sources.

Diplomawallah.in

Date
Page

* Data Reduction

Data Reduction is the process of reducing the volume of data while maintaining its important information.
It makes data analysis faster, easier and more efficient.

* Data Transformation

Data transformation is the process of converting data into a suitable format or structure for analysis and machine learning models.

* Normalization

Normalization scales data values b/w 0 and 1.

$$X_{new} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

used when features have different ranges.

* Standardization

Standardization converts data to have:

- mean = 0
- Standard deviation = 1

$$X_{new} = \frac{X - \mu}{\sigma}$$

used when data follows normal distribution.

Diplomawallah.in

Unit 6

Data Spliting

Data Splitting is the process of dividing a dataset into different parts so that ML models can be trained, tuned and tested properly.

Types

1. Training set
   → used to train the ML model.
   usually 60-70% of the data.
2. Validate set
   used to tune model parameter and select best model.
   usually 10-20% of the data.
3. Testing set
   → used to evaluate final model performance.
   usually 20-30% of the data.

70% Training, 15% validation 15% Testing

* underfitting and overfitting

In ML the main goal is to build a model that learns the true pattern of the data and perform well on unseen data!

Two common problems during model Training are underfitting and overfitting.

Diplomawallah.in

Date
Page

* Underfitting
→ underfitting occurs when a model is too simple and fails to capture the underlying pattern of the data.

* overfitting
overfitting occurs when a model learn the training data too well, including noise, and performs poorly on new unseen data.

* Machine learning pipelines,

Model Training

Supervised learning
- It is a machine learning technique where the model from labelled data.
Eg- Study Hours → Marks
Rainfall → Crop yield.

Diplomawallah.in

*Regression

Regression is a supervised machine learning technique used to predict a continuous numerical value based on one or more input variables.

* Regression finds the relationship b/w input (X) and output (Y) and uses it to make predictions.

Ex - predicting Student marks from Study hours.
- predicting house price from size.
- predicting crop yield from rainfall.

Types of Regression with examples.

1. Linear Regression
predicts outputs using a straight line relationship.
Equation. $y = mx + c$
Ex - predict marks based on hours studied.

2. Multiple Linear Regression
uses multiple input variables.
Ex House price prediction using size, location, rooms.

3. polynomial Regression
Handles non-linear relationship by introducing power of input

Diplomawallah.in

Ex
predicting growth of crops over time.

4. Ridge Regression (L2 Regularization)
Reduces overfitting by shrinkage coefficients.
Used when: Data has multicollinearity.

different Regression types are selected based on data size, complexity and problem requirements.

* Regularization in ML
Regularization is a technique used to reduce overfitting by adding a penalty to the model loss function.
It keeps the model simple and stable.

* Real time Application
1. Education Sector (predicting exam marks)
2. Agriculture
3. Healthcare
4. Business & Sales
5. Real Estate
6. Finance & Banking
7. Weather forecasting.

Diplomawallah.in

Date
Page

* Linear Regression

Linear regression is a supervised ml algorithm used to predict a continuous output variable based on one or more input variables by establishing a linear relationship b/w them.

It finds the best fitting straight line through the data points.

$$y = mx + c$$

where

y = predicted output

x = input feature

m = slope

c = intercept

Types

1. Simple linear Regression

Involve one independent variable one dependent variable.

Ex- predicting marks based on hours studied.

2. Multiple linear Regression

use two or more independent variables.

Ex- predicting house price using area, location and number of rooms.

3. polynomial linear Regression

models non-linear relationship using polynomial terms.

Ex- predicting crop growth over time.

Diplomawallah.in

Application of Linear Regression
- student performance prediction
- Sales forecasting
- Crop yield estimating
- Stock price analysis
- weather prediction.

Model Evaluation & Testing

After training a regression model, it must be evaluated and tested to measure how well it predicts unseen data. Model Evaluation ensures that the model is accurate, reliable and generalize well.

Evaluation metrics for Regression.

1. Coefficient of determination (R^2 Score)

R^2 measure how well the independent variables explain the variation in the dependent variable.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Range : 0 to 1

R^2 close to 1 → Good model

R^2 close to 0 → poor model.

Diplomawallah.in

Interpretion

$R^2 = 0.85$ means the model explains 85% of the variation in the o/p.

2. Root mean Squared Error (RMSE)

RMSE measured the average prediction error of the model.

$$RMSE = \sqrt{\frac{1}{n}\sum (y_{actual} - y_{prediction})^2}$$

Lower RMSE ⇒ Better model performance.

· Gradient Descent

Gradient descent is an optimization algorithm used to minimize the loss function by updating model parameters.

Working

1. Initialize parameters
2. Calculate loss
3. Compute gradient
4. update parameters
5. Repeat until convergence.

Update Rule

$$\theta = \theta - \alpha \frac{\partial loss}{\partial \theta}$$

Diplomawallah.in

Date
Page

* Multiple Regression

Multiple Regression is a supervised learning technique used to predict a continuous dependent variable using two or more independent variables.

It models the relationship b/w input features and o/p as a linear equation.

* Normal equation

The normal equation provides a direct mathematical solution to find optimal regression coefficients.

$$\theta = (X^T X)^{-1} X^T y$$

It is used instead of gradient descent for small dataset.

Application

- House price prediction
- Sales forecasting
- Salary estimation
- Crop yield prediction
- Stock market analysis.

Diplomawallah.in

Date
Page

* Cross Validation

Cross-Validation is a model evaluation technique used to test how well a ml model will perform on unseen data.

Instead of using only one train-test split, the dataset is divided into several parts and the model is trained and tested multiple times.

How cross validation works (K-fold (CV)

1. Dataset is divided into K equal parts
2. Model is trained on K-1 folds and tested on the remaining fold.
3. This process is repeated K times each time using a different as the test set.
4. Final performance = Average of all test results.

Why we need?

1. Preventing overfitting

A model may perform well on one train test split but fail on new data. Cross-validation test the model on different data portions, reducing overfitting.

Diplomawallah.in

Sagar Sangam

Date
Page

2. Gives Reliable performance Estimate
Performance from a single split may be misleading.
Cross validation gives more accurate and stable evaluation.
3. Uses Data Efficiently
All data is both for training & testing which is important when dataset small.
4. Helps in Model Selection.
5. Improves Generalization
↳ Ensure that the model works well on unseen real world data.

* Cross Validation Techniques

1. Hold out Method
In this method, the dataset is split into two parts:
- Training Set
- Testing Set

Common split ratio : 70% Training - 30% testing

2. leave one out Cross Validation (LOOCV)
- If dataset has N samples, the model is trained N times.

Diplomawallah.in

• Each time, N-1 samples are used for training and 1 sample for testing.

K- Fold Cross validation.

Dataset is divided into K equal parts (folds)

process:

1. Train Model on K-1 folds
2. Test on remaining fold
3. Repeat for all K folds.
4. Final score = average of all results.

Common values: K = 5 or 10.

Comparison	Accuracy	Speed	Data usage
Hold-out	Medium	fast	moderate
LOOCV	High	Slow	Maximum
K-fold	High	Moderate	High.

Diplomawallah.in

Unit-07

Date
Page

Some Other Regression Algorithms

1. Random forest Regressor
Ensemble learning method using many decision trees to improve accuracy and reduces overfitting.

2. Support Vector Regression (SVR)
Uses Support Vector machines to predict continuous values, very powerful for complex non-linear data.

3. Lasso Regression.
Linear Regression with L1 regularization, automatically perform feature selection.

Evaluating

Model	Strength
Linear Regression	Simple, fast
Random forest	High accuracy, robust
SVR	Excellent for complex patterns
Lasso	Feature selection + regularization.

Diplomawallah.in

* Supervised learning

- Classification is a supervised learning technique used to predict a category or class label for given input data.

In classification the output is Discrete (not continuous)

Ex
- Email → spam/not spam
- Disease - Yes/NO
- Image - cat/Dog/Bird

Types of Classification

1. Binary Classification
Classification with two possible classes.
Ex spam/not spam, Pass/fail, Yes/no

2. Multi class Classification
Classification with more than two classes.
Ex digits 0-9
Animal type: Dog/cat/Bird

3. Multi-level Classification.
Each data point can belong to multiple classes simultaneously.
Ex. music genre tagging
Photo with labels: person, car, Tree

Diplomawallah.in

4. Imbalanced Classification
occurs when some classes have much more than others.
Ex Fraud detection (very few fraud cases, many normal transactions.)

* Common Classification Models:
- Logistic Regression
- K-Nearest Neighbours
- Naive Bayers
- Decision Tree
- Random Forest
- Neural networks.

* Application
- Spam email filtering
- Disease diagnostic
- Face recognition
- Credit fraud detection.

* Decision Tree
Decision tree is supervised learning algorithm used for classification and regression.
It represent decision in the form of a tree structure, where:

Diplomawallah.in

- Root node → feature selection
- Internal node - decision rules
- leaf node - find output

It mimics human decision making.

* Entropy

Entropy measures the impurity or randomness in data.

Entropy = 0 → pure dataset

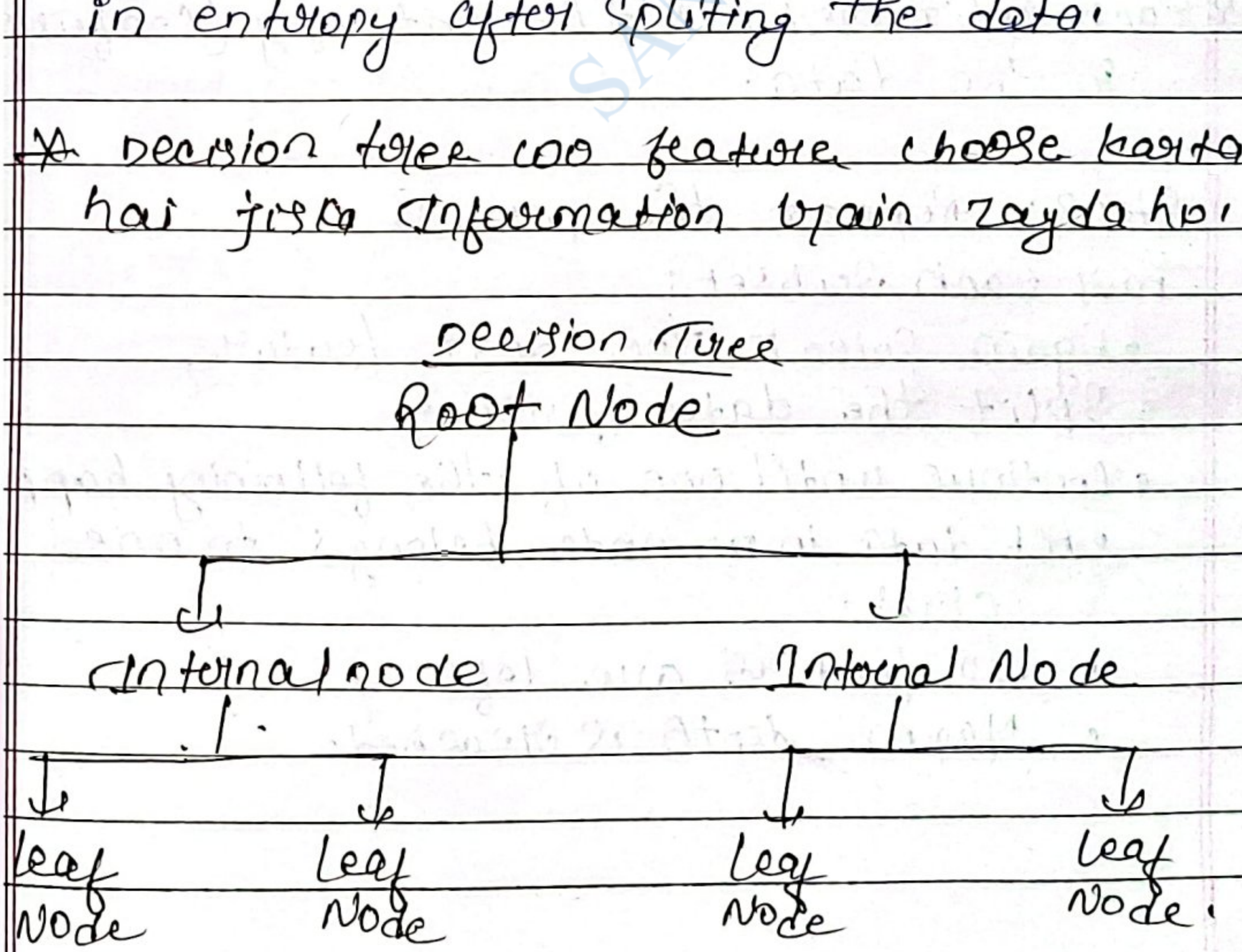
Entropy = 1 → maximum impurity.

Information Gain

Information Gain Measure the reduction in entropy after Spliting the data.

* A decision tree ko feature choose karta hai jiska Information Gain zayda ho.

Decision Tree



Diplomawallah.in

Date
Page

Decision Tree working

Step 1: Select the Best feature
The algorithm checks all features and selects the on that gives the highest information Gain or lowest Gini Index.
This feature becomes the root node of the tree.

Step 2: Split the dataset
The dataset is divided into subset based on the selected feature values.
* Each split reduces the entropy (confusion) in the data.

Step 3: Repeat the process for each subset:
- Again select the best feature
- Split the data further
- Continue until one of the following happens:
  - All data in a node belongs to one class.
  - No features are left
  - Maxm depth is reached.

Diplomawallah.in

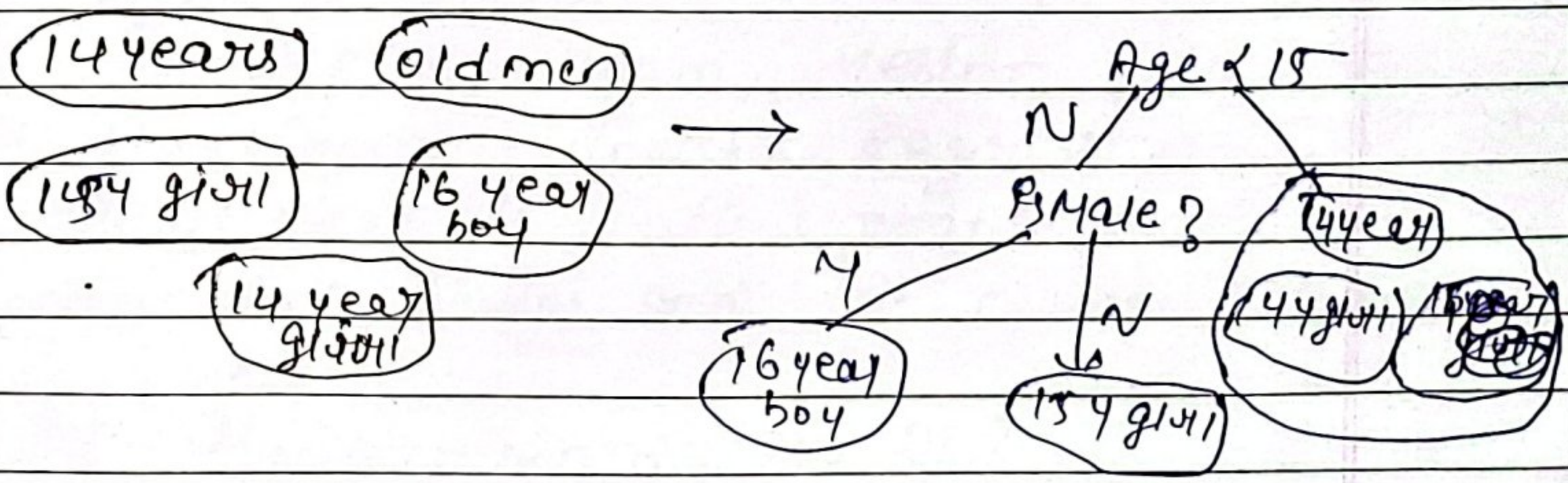
Step 4: Create leaf Nodes
when no further Spliting is possible
, a leaf node is formed.
The leaf node contains the final decision
or class label.

Step 5: prediction
for new data point:
- Start at the root
- follow the decision rules.
- move down the tree
- Reach a leaf node → get prediction.

Ex :

Input Age, Gender
Occupation

Does the person like
Computer games



Diplomawallah.in

Date
Page

* Hyperparameter Tuning for decision Tree

Hyperparameter tuning is the process of finding the best values for model parameters that control learning to improve model performance and reduce overfitting.

Important decision Tree Hyperparameters.

max_depth → Limit tree depth
min_samples_split → Minimum samples to split
min_samples_leaf → minimum samples per leaf.
max_features → features used for split
criterion → split quality measure.

Diplomawallah.in

Unit - 08

Date
Page

Logistic Regression (Category based predict)

Logistic Regression is a supervised learning algorithm used for binary classification problems.

It predicts the probability that an input belongs to a particular class.

Working

Logistic Regression uses the sigmoid function to map predicted values between 0 and 1.

Types

1. Binary logistic Regression
used when the o/p variable has two possible classes.
- Ex - Spam/no spam, yes/no

2. Multinomial logistic Regression
→ used when the o/p variable has more than two classes and the classes are not ordered.
Ex - Handwritten digit - 0-9

3. Ordinal Logistic Regression
used when the o/p variable has more than two classes with a natural order.
Ex - Customer rating :- poor/Average/Good.

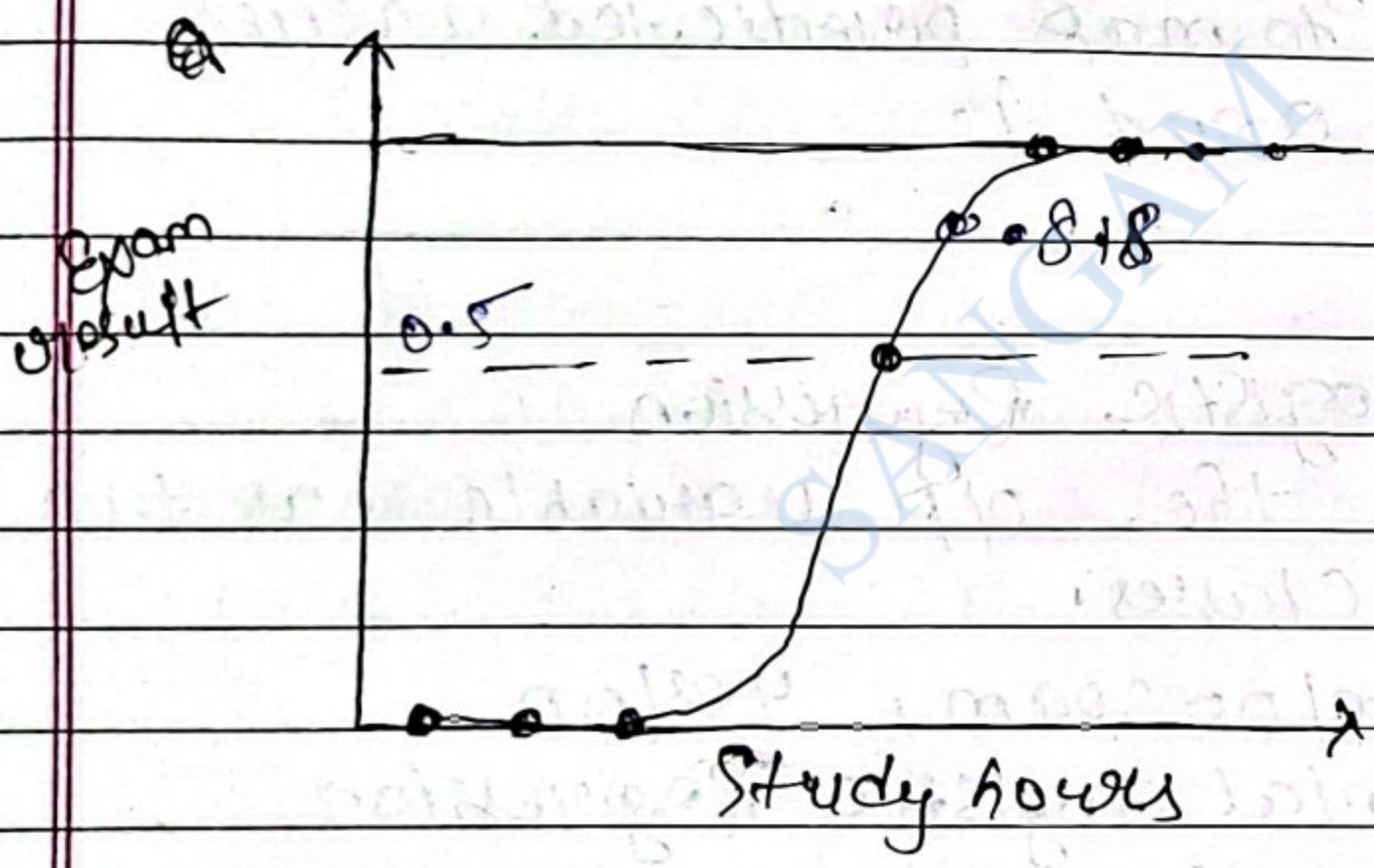
Diplomawallah.in

Sigmoid Changes the value in 1 & 0.

Date Page

* process of working

Study hours	Exam Result
2	0
3	0
4	0
5	1
6	1
7	1
8	1



$$y = \frac{1}{1 + e^{-(a_0 + a_1 x)}}$$

$a_0 = ?$ $a_1 = ?$

Variable. Study hours of Student.

eg- $x = 5$ $= -1.05 + 0.6 \times 5$

$a_0 = -1.05$ $= 1.5$

$a_1 = 0.6$

Diplomawallah.in

Classification Regression

$y = \frac{1}{1+e^{-b5}}$

$= .818$

81% Chances approx.

Linear Regression	Logistic Regression
Supervised regression model	Supervised classification model.
The dependent variable (the variable you want to predict) is continuous.	dependent variable is categorical and binary (0 or 1), making it suitable for calculating the probability of an event.
Equation of linear Regression. $\rightarrow Y = a_0 + a_1 x_1$	Equation of Logistic Regression. Sigmoid function $y = \frac{1}{1+e^{-(a_0+a_1 x)}}$
• we predict the value by an integer number.	we predict the value by 1 or 0. • Category based.

Housesize	Price		
100	1 lakh	yes	0
200	2 lakh	no	1
300	3 lakh	yes	0
400	4 lakh	no	1
↓	↓		
!!	4		

Diplomawallah.in

SVM

Date
Page

* Support Vector machine (SVM)

- Support Vector machines (SVMs) are Supervised machine learning algorithm used for classification and regression tasks.

- Hyperplane: A hyperplane is a decision boundry that separates data points of different classes in the feature Scope.



- Mathematical Representation: $w \cdot x + b = 0$

Ex

If we classify animals into cats (+1) & dogs (-1) based on weight and height, the hyperplane could be: $3 \times \text{weight} + 2 \times \text{height} - 50 = 0$

- Support Vectors:- These are the data points that are closet to the hyperplane. These points influence the orientation and position of the hyperplane.

Diplomawallah.in

Date
Page

• margin: The margin is the distance between the hyperplane and the nearest data of either class. SVM tries to maximize this margin to improve classification confidence.

• Hard margin: Assumes data is perfectly separable by the hyperplane. All points must lie outside the margin.

• Soft margin: Allows some misclassification or margin violations for non-linearly separable data.

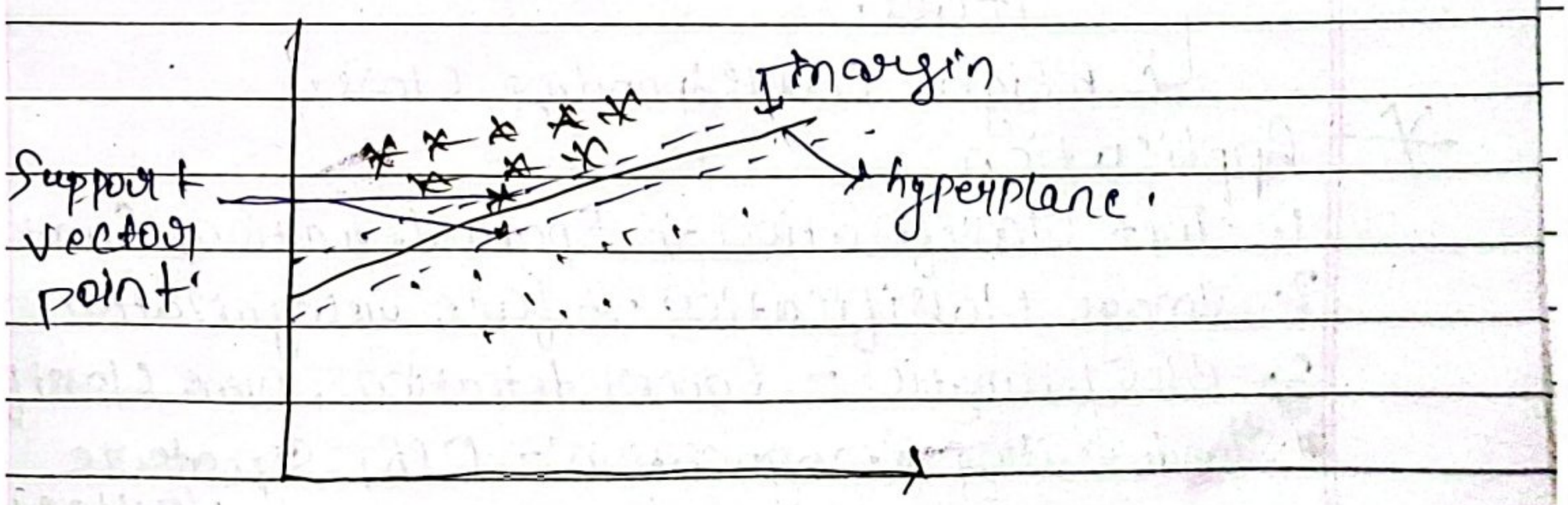
• Kernel function: A kernel function transforms data into a higher-dimensional space to make it linearly separable.

Types:

• Linear Kernel – Suitable for linearly separable data.

• Polynomial Kernel: For curved boundaries.

• Radial Basis function: Captures complex relationship.



Diplomawallah.in

Date
Page

SVM working

(1) Draw possible Hyperplanes
→ SVM checks all possible lines that can separate the data.

(2) Choose maximum margin Hyperplane
Among all possible boundaries, SVM selects the one with the largest margin.

(3) Identify Support vectors.
Only few datapoints closest to the hyperplane influence the decision boundary. i.e
Support vectors.

(4) Apply Kernel Trick.
If data is not linearly separable, SVM use kernel function to transform the data into higher dimensions.

(5) Classification of new data
For new data
↳ Check on which side of the hyperplane it lies.
↳ Assign corresponding class.

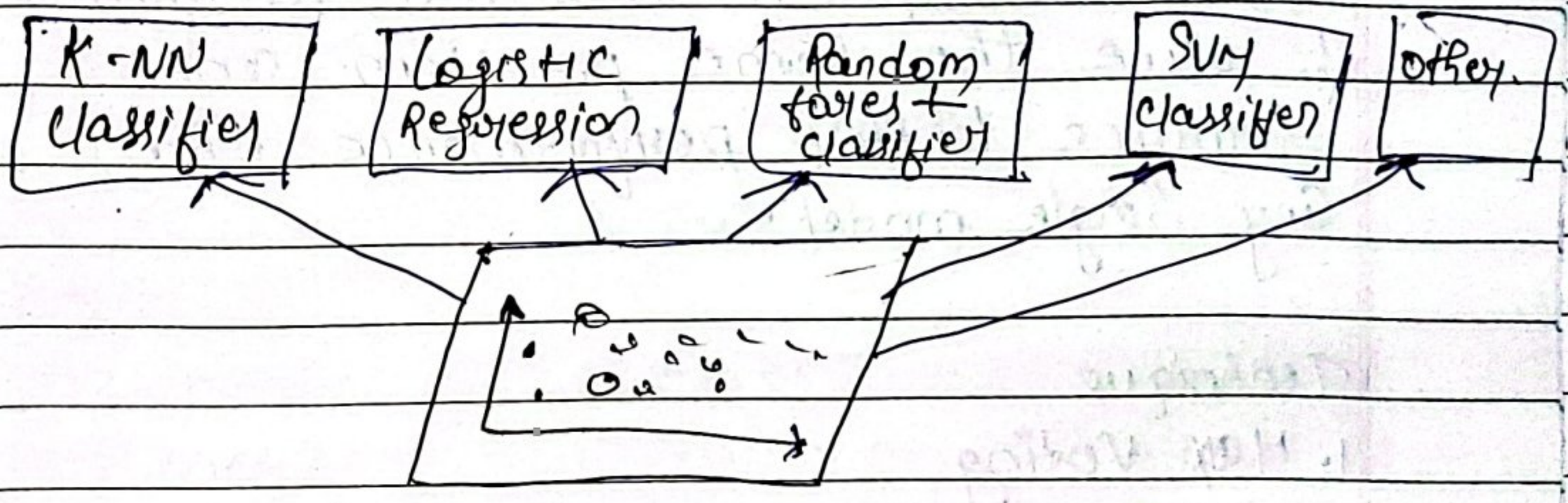
X Application
1. Text Classification → Spam detection, Email filtering
2. Image Classification :- face recognization,
3. Bioinformatics - Cancer detection, Gene Classification.
4. Handwriting Recognization :- OCR, Signature Verification.

Diplomawallah.in

Date
Page

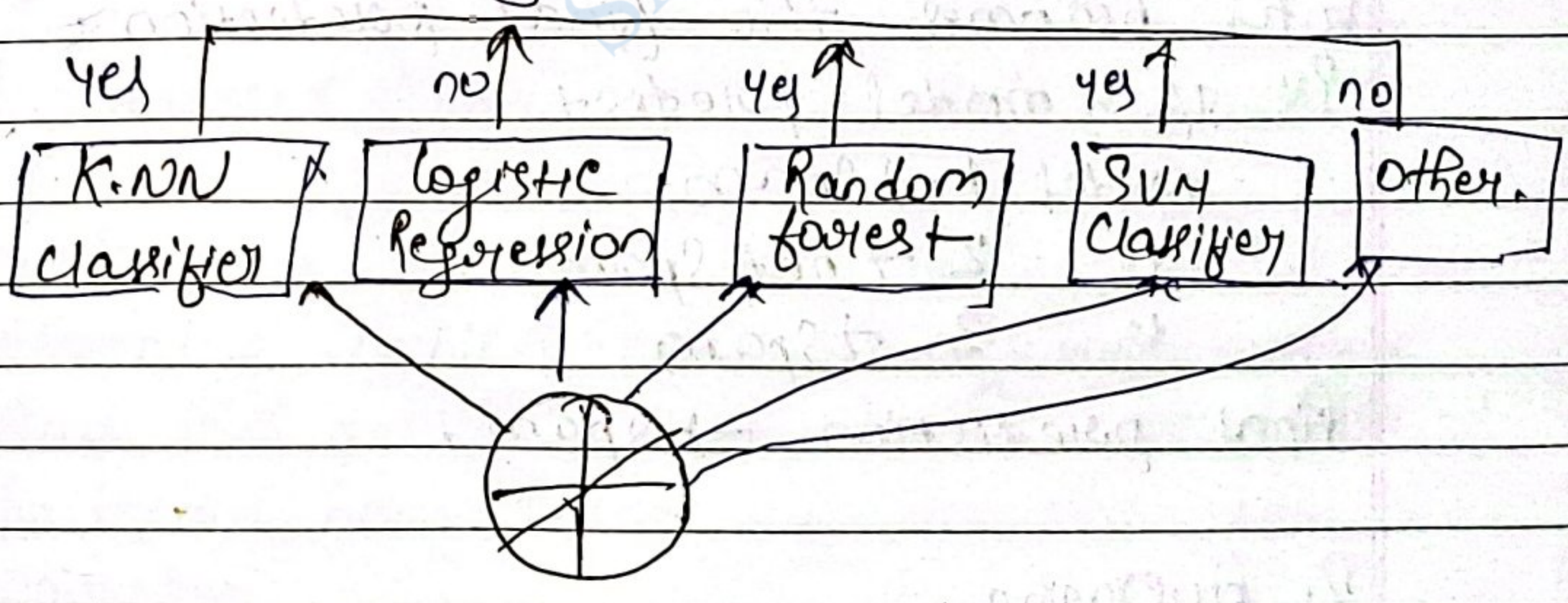
* Ensemble learning

It uses multiple learning algorithm to obtain better predictive performance.



If we here donot ensemble the output taken by of any one model.

But ensemble says collection of data from different learning.



yes - 3 ↘ The yes is in majority.
NO - 2 ↳ prediction high.

Diplomawallah.in

* Some Ensemble Technique

Ensemble learning is a ML technique where multiple models are combined to solve the same problem and produce better performance than any single model.

Technique

1. Hard Voting

Used mainly for classification problems.

Each model gives a prediction and the class with the maximum number of votes becomes the final prediction.

Eg If 3 model predict

Model 1 → Spam

" 2 → not Spam

" 3 → Spam

final prediction → Spam.

2. Averaging

used mainly for regression problems.

* Each model predicts a numerical value and the final o/p is the average of all predictions.

Eg - predictions : 50, 60, 55

final o/p → (50+60+55)/3 = 55

Weak learner - single model
Strong learner - multiple model.

Diplomawallah.in

3. Weighted Average
Some models are more accurate than others, so their predictions are given more important using weights.
formula :-
final = $w_1P_1 + w_2P_2 + w_3P_3$
where $w_1 + w_2 + w_3 = 1$

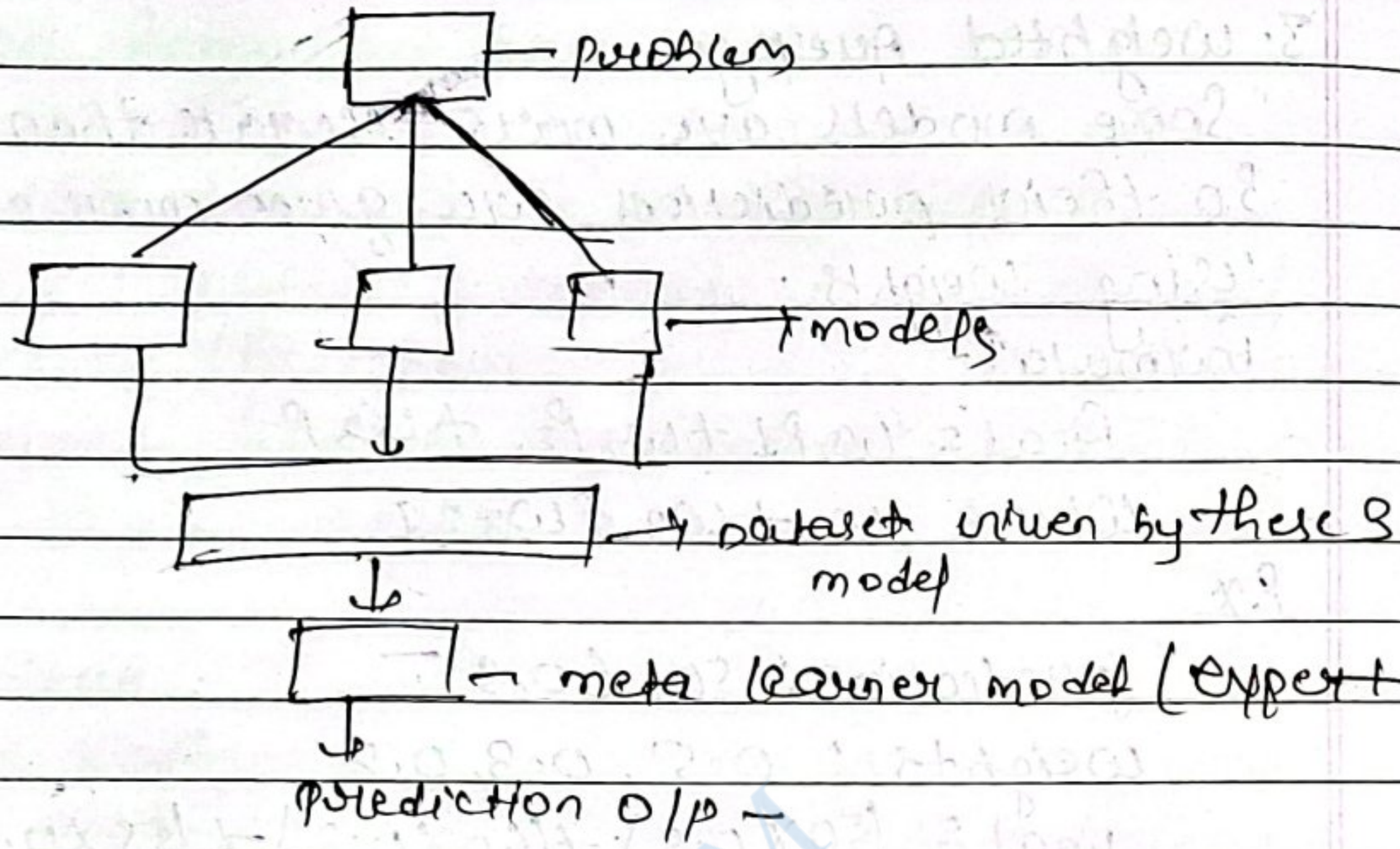
Ex
prediction: 50, 60, 55
weights : 0.5, 0.3, 0.2
final = $(50 \times 0.5) + (60 \times 0.3) + (55 \times 0.2) = 54.5$

Ensemble methods increase accuracy, reduce overfitting and improves reliability.

* Advanced Ensemble Technique

1. Stacking
Stacking combine multiple different models and uses another model (called meta-learner) to learn how to best combine their predictions.
working
1. Train several base models.
2. Collect their predictions.
3. Train a new model on those prediction.
4. Final prediction is made by the meta learner.

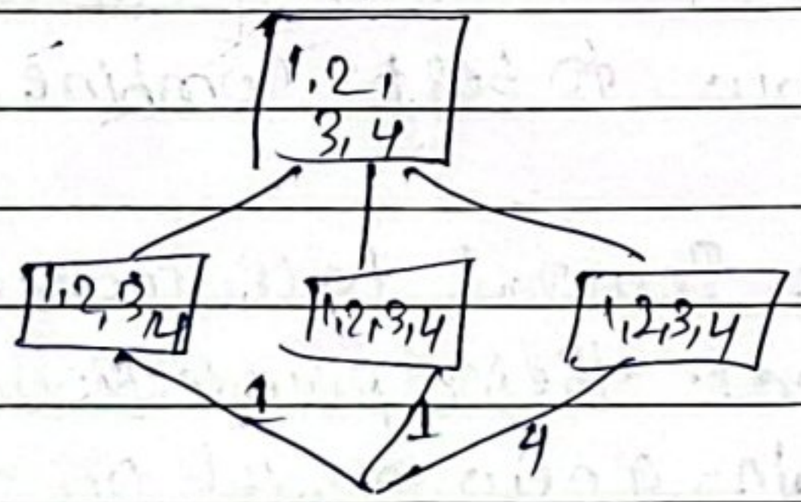
Diplomawallah.in



2* Bagging (Bootstrap Aggregating)
Bagging builds many models on different random samples of the training data & combines their predictions.

Word ① Create many bootstrap samples
② Train a model on each sample
③ Combine prediction by voting or averaging.

Ex- Randomforest is based on bagging.



Diplomawallah.in

Boosting

Boosting builds models sequentially where each new model focuses on correcting errors made by previous models.

Logic ① Train first model

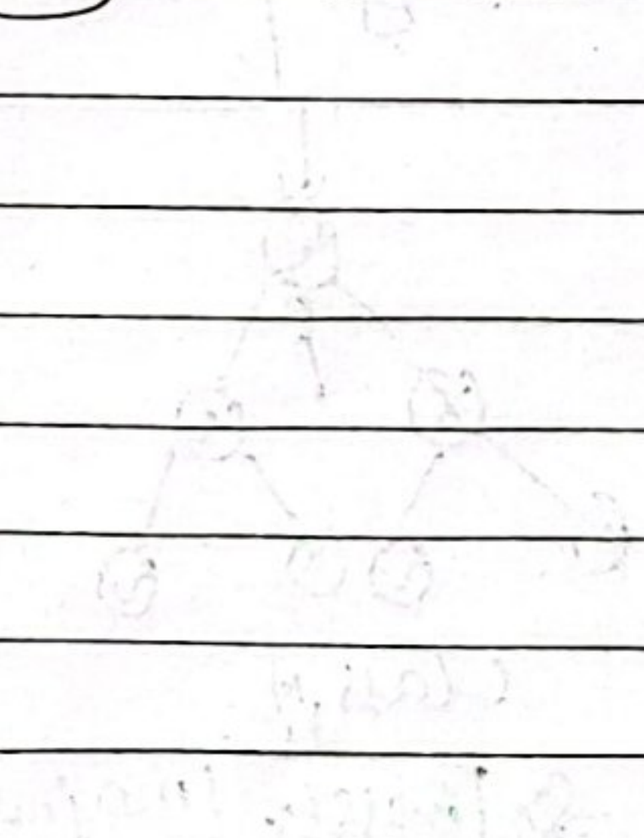
② Increase weight of wrongly predicted samples.

③ Train next model.

④ Repeat

⑤ Combine all model into strong classifier.

Ex - AdaBoost, Gradient Boosting, XGBoost



Bagging

Diplomawallah.in

Date
Page

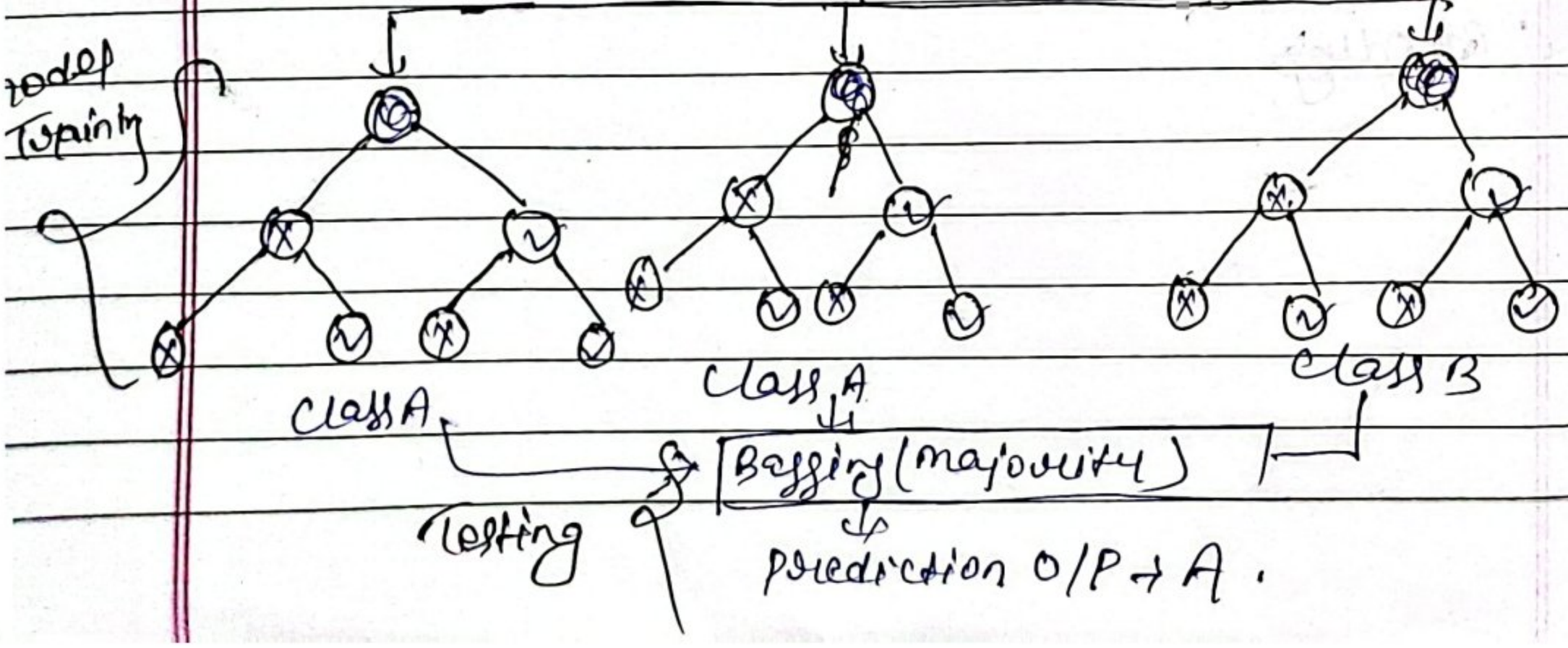
* Random forest Algorithm:

Random forest is an ensemble learning algorithm that combines multiple decision trees to create a more accurate and stable prediction model.

It works on the principle of bagging (Bootstrap Aggregating)

* Working :—

1. Randomly select samples from dataset (with replacement)
2. Builds many decision trees using these samples.
3. Each tree gives its own prediction.
4. Final result is obtained by:
   - voting (for classification)
   - Averaging (for regression)



Diplomawallah.in

Important Hyperimetery

parameters	purpose
n_estimators	Number of trees
max_depth	maxm depth of each tree
min_sample_depth	minimum Sample required to split
min_samples_leaf	minimum samples at leaf
max_features	No. of features Considered.
criterion	Quality of split

Application

- Fraud detection
- medical diagnosis
- stock market analysis
- Credit Scoring
- Image classification.

Diplomawallah.in

unit-09

Date
Page

* unsupervised learning *

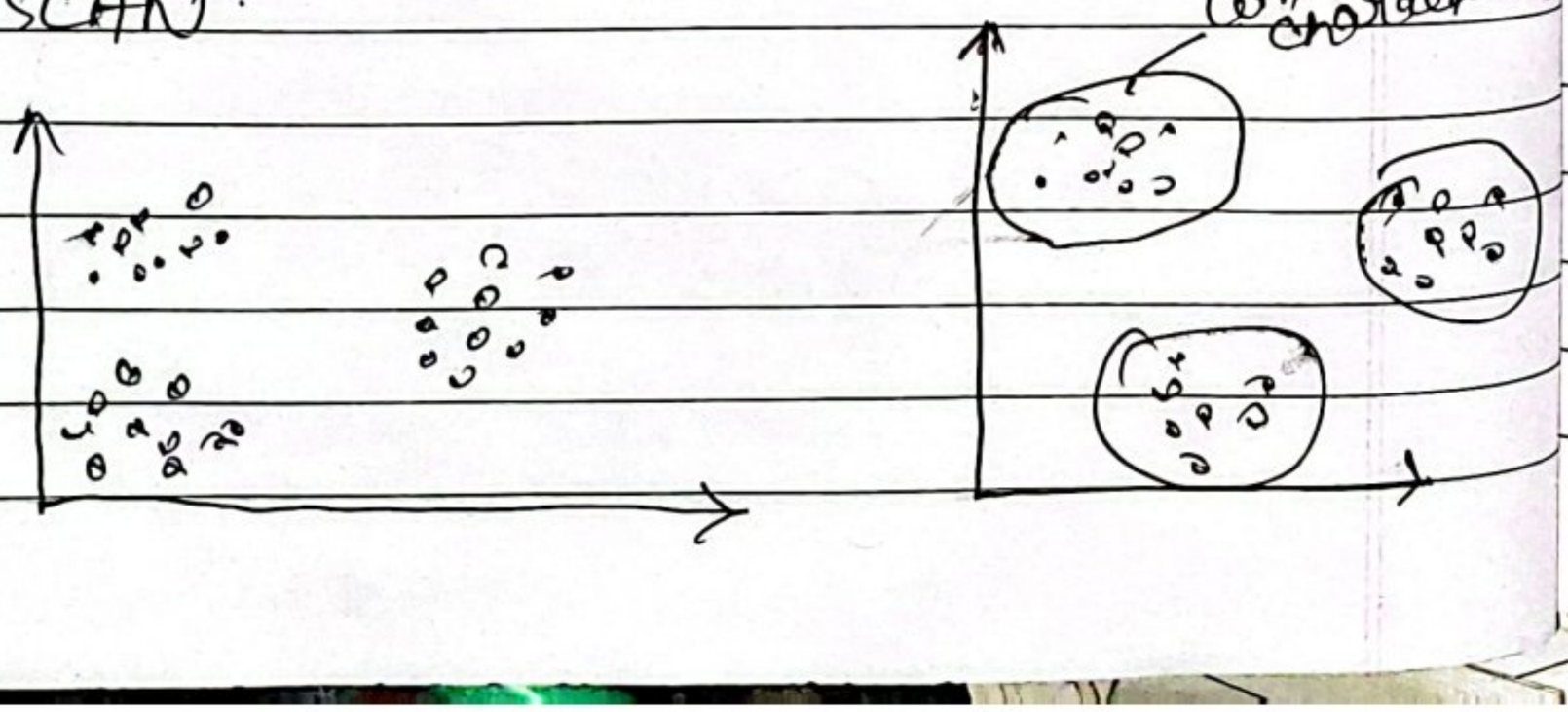
@ unsupervised learning is a type of machine learning where the model is trained on data without any labelled output.
The algorithm tries to discover hidden patterns, structure, or relationship in the data by itself.

* The model learns from data without being told the correct answer.

* Common Approaches in unsupervised learning

1. Clustering
Clustering groups similar data points together based on their characteristics.

Ex
K-means, Hierarchical Clustering, DBSCAN.



Diplomawallah.in

Date
Page

* K means clustering
K-means is an unsupervised learning algorithm that divides data into k clusters, where each data point belongs to the cluster with the nearest mean (centroid).
Use - Customer segmentation, Image compression.

* Hierarchial clustering
Hierarchial cluster builds a hierarchy of clusters using either:
- Agglomerative (bottom-up) approach, or
- Divisive (top-down) approach.

3. DBSCAN (Density Based Spatial Clustering of Application with Noise)
DBSCAN groups data points based on density and can identify clusters of arbitary shape. It also detect noise and outliers.
Use - Anomaly detection, Image processing.

* Challanges in clustering
1. Choosing the Right Number of clusters.
(many algorithm like - K-means requires the number of cluster in advance).
2. High dimensional data (Cluster become less meaningful)
3. Presence of noise and outliers.

Diplomawallah.in

2. Association Rule Learning
finds interesting relationship b/w variables in large datasets.
Ex
- Apriori Algorithm
- FP-Growth Algorithm.

3. Dimensionality Reduction
Reduces number of features while preserving important information.
Ex - PCA (principal Component Analysis)
- LDA

4. Anomaly detection
Identify rare or unusual datapoints.
Ex - Isolation forest, one-class SVM

Diplomawallah.in

Date
Page

* Dimensionality Reduction

Dimensionality Reduction is the process of reducing the number of input features (dimensions) in a dataset while preserving as much important information as possible.

* It removes unnecessary or less important features from data.

Importance of Dimensionality Reduction in ML

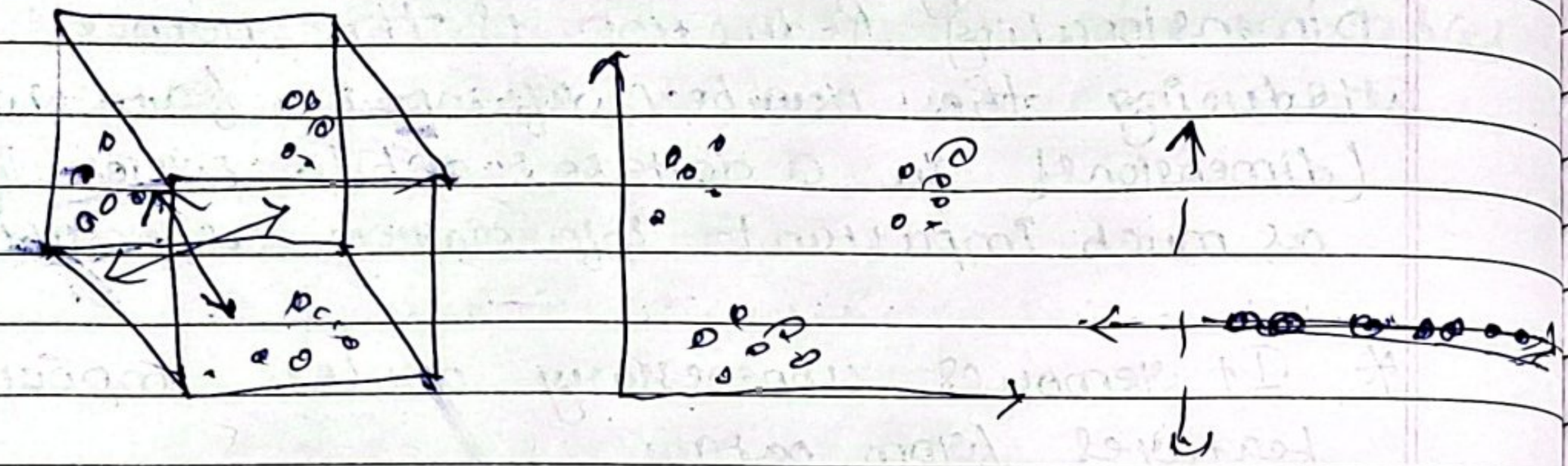
1. Reduces overfitting
→ Too many features can cause models to learn noise. Dimensionality reduction removes irrelevant features and improves generalization.

2. Improves Model performance.
Fewer features → faster training → better efficiency.

3. Reduces Computational cost
Less memory usage and faster computations especially for large datasets.

4. Handles Curse of Dimensionality.
5. Improves visualization.
6. Removes Noise Redudancy.

Diplomawallah.in



* MLOps

MLOps (machine learning operations) is a set of practices that combines ML, DevOps, and Data Engineering to deploy, monitor, manage, and maintain ML models in production efficiently.

* MLops helps take ML models from development to real-world production smoothly.

* Why MLOps?

1. To deploy ML models into production Training a model in a notebook is not enough.
MLops helps in deploying models as scalable services (APIs).

Diplomawallah.in

2. To handle Data Drift
Real-world data changes over time.
MLOps monitors input data and retrains models when performance degrades.

3. To Ensure Model Reliability
MLOps ensures models remain accurate, stable, and reliable in production.

4. To Automate ML pipelines.
Manual process are slow and error-prone.
MLOps automates training, testing, deployment and monitoring.

5. To Improve Collaboration.
MLOps enables smooth collaboration b/w:-
- Data Scientist
- ML Engineers
- DevOps teams.

6. To Enable Continuous Improvement
MLOps supports continuous training, evaluation, and deployment of improved models.

7. To Maintain Model Versioning & Reproducibility
Track version of:
- Data, Models, Experiments.

Diplomawallah.in

Date
Page

Machine learning Pipelines.

Data collection → Feature Engineering → Model training → Evaluation

↓

Maintenance ← Monitoring ← Deployment

An ML pipelines is a structured sequence of steps used to build, train, evaluate, deploy, and maintain a ML model efficiently.

Stages of ML pipeline

1. Data collection
→ Collect data from sources like database, APIs, sensors, files, web, etc.
→ Data can be structured or unstructured.
Ex-
Student marks data, medical records, transaction data.

2. Data preprocessing
- Data cleaning (missing values, noise)
- Data integration
- Data transformation (scaling, encoding)
- Data reduction

Diplomawallah.in

3. Exploratory data Analysis (EDA)
- understand data patterns.
- Identify outliers.
- visualize distribution and relationship.

4. Feature Engineering / selection
- Create new meaningful features.
- Select important features.
- Reduce irrelevant or redudant features.

5. Data Splitting
- Training Set, Validation Set, Testing Set
* Ensure fair evaluation of the model.

6. Model Selection
- Chose appropriate algorithm (Linear Regression, Decision Tree, SVM, KNN etc.)

* Depends on:
  - problem type (Classification / Regression)
  - Data size and nature.

7. Model Training
- Train the model using training data
- Learn parameters and patterns

Diplomawallah.in

8. Model Evaluation
- Use metrices like:
  - Accuracy, precision, Recall (classification)
  - RMSE, R^2 (regression)

Check performance on unseen data.

9. Hyperparameter Tuning
- optimize model parameters.
- Improve accuracy and generalization.

10. Model Deployment
- Deploy model as:
  - API
  - Web service
  - Application.

11. Monitoring & Maintenance
- Monitor performance
- detect data drift
- Retrain model when needed.

Diplomawallah.in

Machine learning & MLOps (Versioning)

→ Versioning is the process of tracking, managing, and maintaining different versions of data, code, models and experiments throughout the ML lifecycle.

* Versioning helps us know what changed, when it changed and why it changed.

Why version is important in ML

ML Systems continuously evolve:

- Data changes.
- Models are retrained.
- parameters are updated.

Without versioning it becomes difficult to:

- Reproduce results.
- Debug issues
- Roll back to a previous working model.

* Model Registry

A Model Registry is a centralized repository used to store, manage, track and control different versions of ML models throughout their lifecycle.

Diplomawallah.in

A model Registry is a catalog of trained ML models with their metadata, versions and deployment states.

model lifecycle in a model Registry.

1. model Training
2. Model Registration
3. Validation & Testing
4. Promotion to Steging
5. Deployment to production
6. Monitoring
7. Archiving or Replacement.

Key features of a model Registry

- Version Control → Track model versions
- Reproducibility — know how model was built.
- Governance → Approval before Deployment
- Rollback — Restore older stable models.
- Collaboration — Shared access for teams

Ex

Model name	version	Accuracy	Stage
Cancer classifier	V1	→ 91%	Archieved
Cancer classifier	V2	94%	production
Cancer classifier	V3	95%	Steging.

Diplomawallah.in

popular Tools for model Registry:

* MLflow Model Registry
* Aws, Sagemaker Model Registry.
* Azure ML Model Registry.
* Kubeflow.

* Model Deployment →

Model deployment is the process of making a trained machine learning model available for real-world use so that it can generate predictions on new data.

* Deployment means using the ML model in production.

* Common deployment method
  1. Batch deployment
     - predictions are made on large data batches.
     - used in reports, analytics.
  2. Real-Time (online) deployment
     - model deployed as an API
     - Instant prediction
     - used in fraud detection, recommendation system.
  3. Edge Deployment
     - model runs on devices (mobile, IoT)

Diplomawallah.in

Date
Page

* Deployment Tools
→ Flask / FastAPI, Docker, Kubernets.
AWS, Azure, GCP.

* Model Training ML
model Training is the continuous process of tracking model performance and behaviour after deployment to ensure it works correctly over time.

Diplomawallah.in

Unit - IV

Date
Page

Deep learning

Deep learning is a subset of ML that uses artificial neural networks with multiple layers (called deep neural networks) to learn complex patterns from large amount of data.

Deep learning allow machine to learn automatically from data, just like the human brain.

Types of Deep learning Models:

1. Artificial Neural Network (ANN)
2. Convolutional Neural network (CNN) - images, videos.
3. Recurrent Neural Networks (RNN) - Sequence data
4. LSTM/GRU - time-series, text
5. Autoencoders
6. Deep belief Networks (DBN)

Artificial Neural Network (ANN)

ANN is the basic deep learning model inspired by the human brain.

It Consist of:

- Input layer
- one or more hidden layers.
- Output layer.

Diplomawallah.in

2. Convolutional Neural Network
CNN is designed to process image and visual data using convolution and pooling layers.
Uses - Image classification, face recognition, medical image analysis.

3. Recurrent Neural Network
→ RNN is designed for sequential data where previous information matters.
Uses → Speech recognition, Text processing, Time series prediction.

4. Long Short-Term Memory (LSTM)
→ LSTM is an advanced type of RNN that solves the vanishing gradient problem.
Uses → Language translation, Sentiment analysis, Stock price prediction.

5. Gated Recurrent unit (GRU)
GRU is a simplified version of LSTM with fewer gates.
Uses - Sequence modelling, faster training Compared to LSTM.

Diplomawallah.in

6. Autoencoders.
Auto encoders learn compressed representation of data.
Uses - Dimensionality Reduction, Anomaly detection, Data denoising

7. Deep Belief Networks (DBN)
DBN are composed of multiple layers of probabilistic models.
uses - Feature learning, pattern recognition.

Application of Deep learning
1. Computer ~~version~~ vision
2. Speech Recognition
3. NLP
4. Healthcare.
5. Autonomous Vehicles.

* Deep learning Frameworks
- Deep learning frameworks are S/W libraries that provide tools to design, build, train, and deploy deep neural networks easily and efficiently.
They handle:
- Tensor operations
- Automatic differentiation.
- GPU/TPU acceleration.

Diplomawallah.in

popular frameworks

1. Tensor flow developed by Google.
2. Keras - High level API built on top of Tensorflow.
3. pytorch - developed by facebook (meta)
4. MXNet → developed by Apache, used by AWS.
5. Coffe → focused mainly on image processing tasks.

Framework	Strength
Tensorflow	production & scalability
Keras	Simplicity
pytorch	Research & flexibility
MXNet	Cloud deployment
Coffe	computer vision.

Sagar Sangam

Diplomawallah.in

Date
Page

* Neural Network

Neural network is a computational model inspired by the human brain that is used to recognize patterns, learn from data and make decisions.

It consist of interconnected units called neuron that process information using mathematical operations.

In simple words, A neural network learns by adjusting its connection based on experience (data).

Diplomawallah.in

Date
Page

* Biological neuron:

A biological neuron is a nerve cell in the human nervous system that receives, processes and transmits information using electrical and chemical signals.

Neuron are the basic building blocks of the human brain and nervous system.

* Artificial Neuron / perception

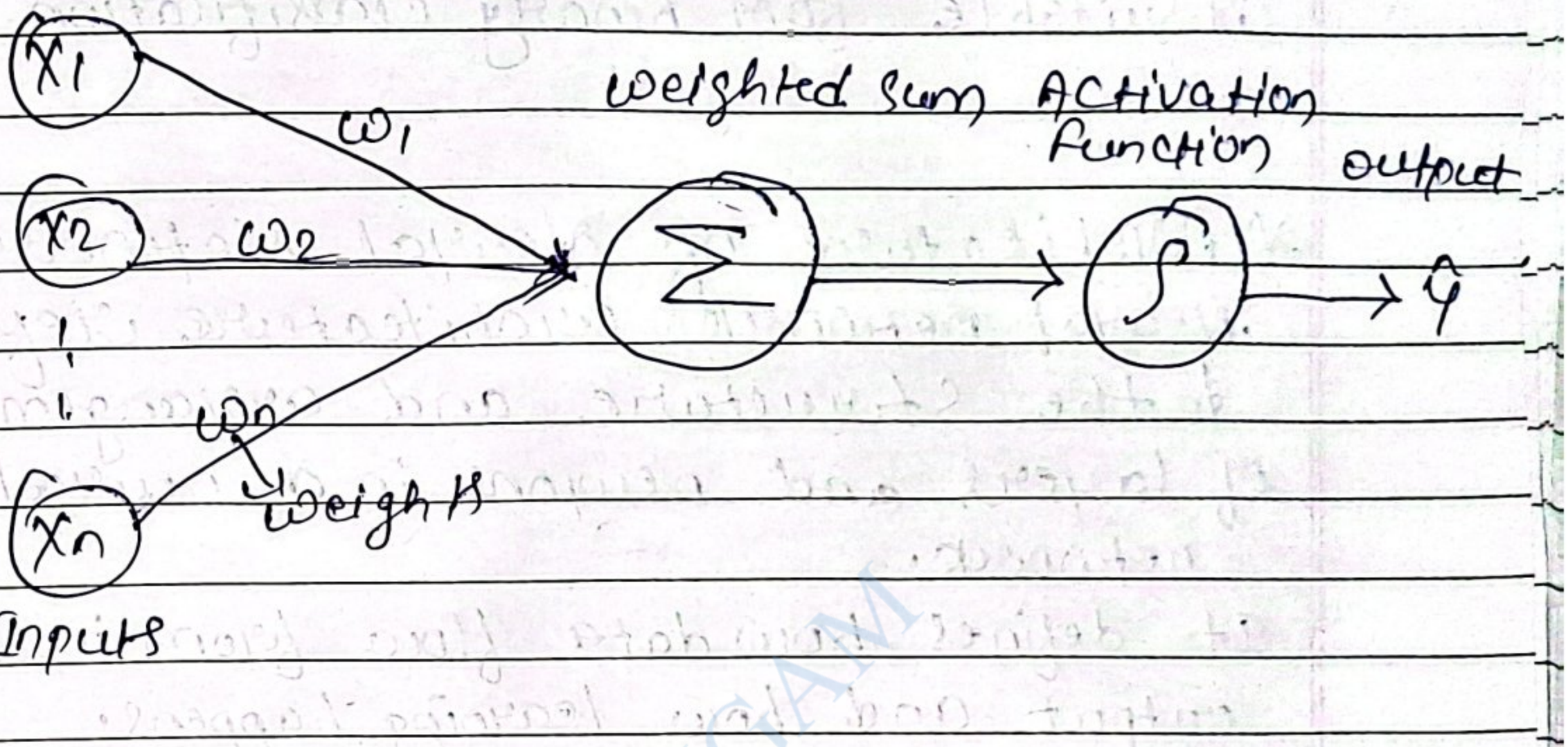
An Artificial neuron, also called a perception is the basic building blocks of a neural network.
It is a mathematical model inspired by a biological neuron that receives inputs, processes them, and produces an output.

* A perception takes input, applies weight, adds bias, passes the result through an activation function, and gives an output.

Diplomawallah.in

Date
Page

Structure of a perceptron.



working process of perceptron

1. Weighted Sum
Each input is multiplied by its weights added with bias:

$$z = w_1x_1 + w_2x_2 + \ldots + w_nx_n + b$$

(input, weight, bias value)

2. Apply Activation function
The weighted sum is passed through an activation function $f(z)$.

common activation (perceptron):

- Step function

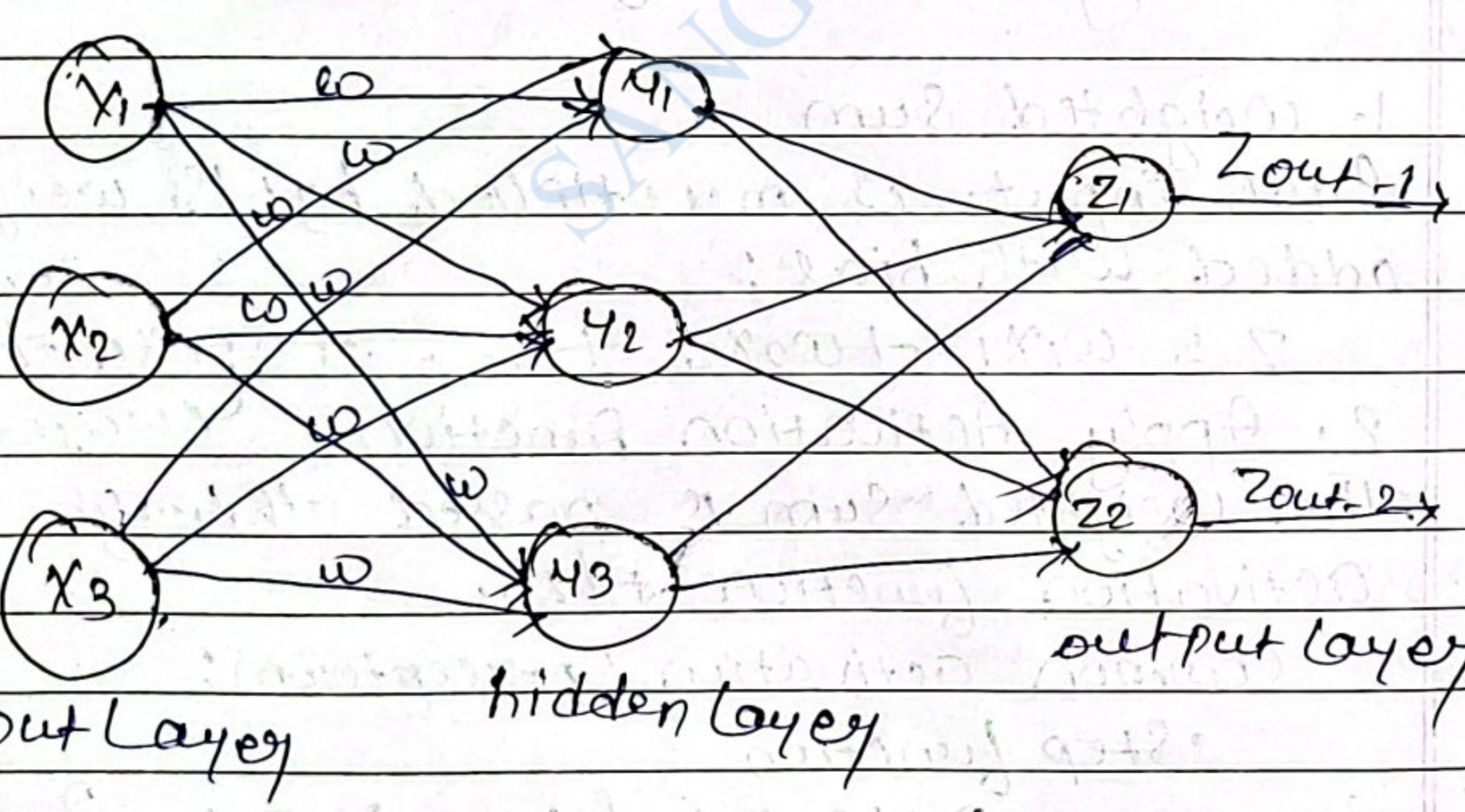
$$\text{output} = \begin{cases} 1, & \text{if } z \geq 0 \\ 0, & \text{if } z < 0 \end{cases}$$

Diplomawallah.in

3. produce output
The perceptron outputs 0 or 1, making it suitable for binary classification.

Architecture of Neural Network

Neural network architecture refers to the structure and arrangement of layers and neurons in a neural network.
It defines how data flow from input to output and how learning happens.



Diplomawallah.in

Date
Page

* Main Component of neural network

1. Input layer
- First layer of the network.
- Received raw input features.
- Number of neuron = number of input features.
Ex - Image pixels, numerical values, text features.

2. Hidden layer(s)
- One or more layers b/w input & output
- perform feature extraction and learning.
- Each neuron applied weights, bias and activation function.
Role :- learn complex & non-linear patterns.

3. Output layer
- final layer of the network.
- Produces prediction or classification result
- Activation depend on problem type;
  - Sigmoid → Binary classification
  - Softmax → multi-class classification
  - linear → Regression.

- other important elements.
* weights :- Strength of connection b/w neuron.
* Bias - Helps model fit data better.
* Activation function - Introduce non-linearity.

Diplomawallah.in

working of NN:

1. Input layer
→ Receives your input features.
→ Each neuron represent one feature.
Ex- Study hours, marks.

2. Forward propagation
→ Step 1: weighted sum
$z = w_1x_1 + w_2x_2 + \dots + w_nx_n + b$
Step 2: Activation function
The weighted sum is passed through an activation function (ReLU, Sigmoid etc.) to introduces non-linearity:
$a = f(z)$
Step 3: pass Through Hidden layer.
→ output of one layer become input of other.
→ This is continue till the output layer.

3. output layer
- produce final prediction.
- Activation depend on problem type:
  - Sigmoid
  - Softmax
  - Linear:

Diplomawallah.in

Date
Page

* Forward propagation
Forward propagation is the process by which input data is passed from the input layer to the output layer of a neural network to generate a prediction.

* Backward propagation
Backward propagation also called back-propagation is the process of updating the weights and biases of a neural network by propagating the error from the o/p layer back to the input layer.

* Back propagation tells the network how wrong it was and how to correct itself.

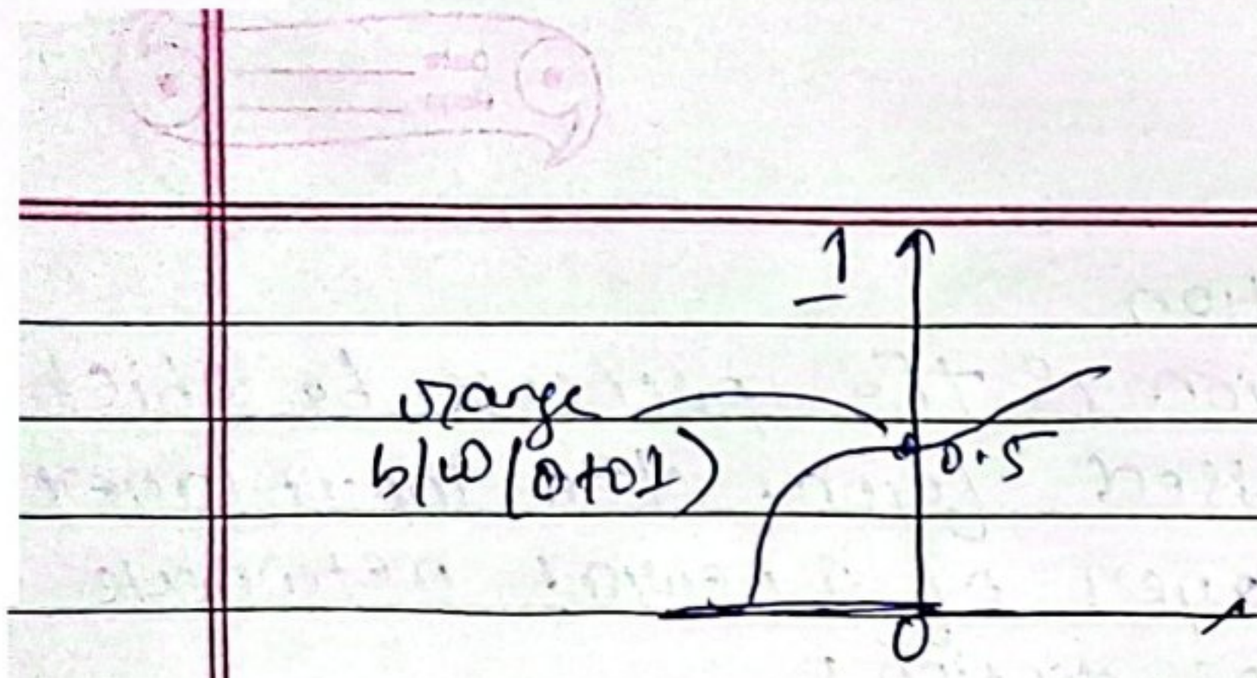
* Activation function (mathematical function)

* Sigmoid
$$F(z) = \sigma(z) = \frac{1}{1+e^{-z}}$$ z = value.

z = weighted sum of input
$z = w_1x_1 + w_2x_2 + \cdots + w_nx_n$
e = Euler's number (≈ 2.718)

- Range of Sigmoid function
  - output range: (0, 1)
  - If z is very large → output ≈ 1
    If z is very small → output ≈ 0

Diplomawallah.in

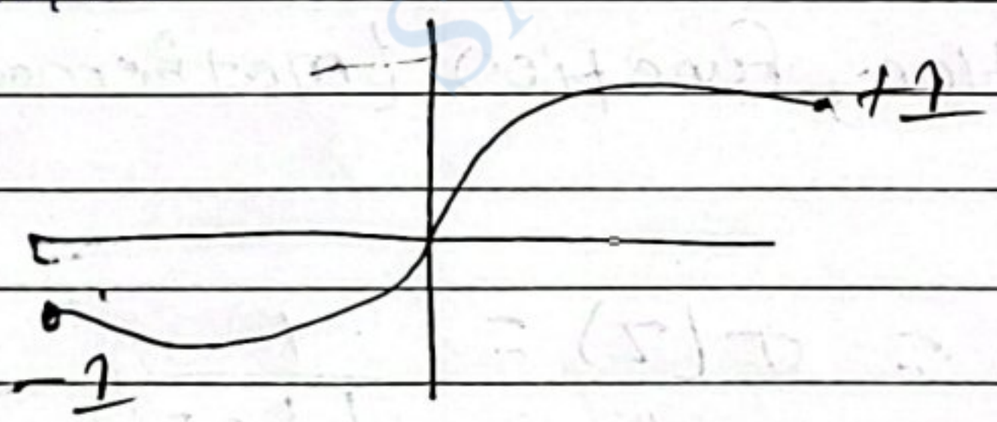


O/P → output ≥ 0.5 → class 1
output < 0.5 → class 0.

* Tanh Activation

Tanh Activation function is an activation function used in neural network especially in hidden layers.

It maps input values to a range between -1 and +1 making it zero-centered.



output range (-1, +1)

If z is large → output ≈ +1
If z is small → output ≈ -1

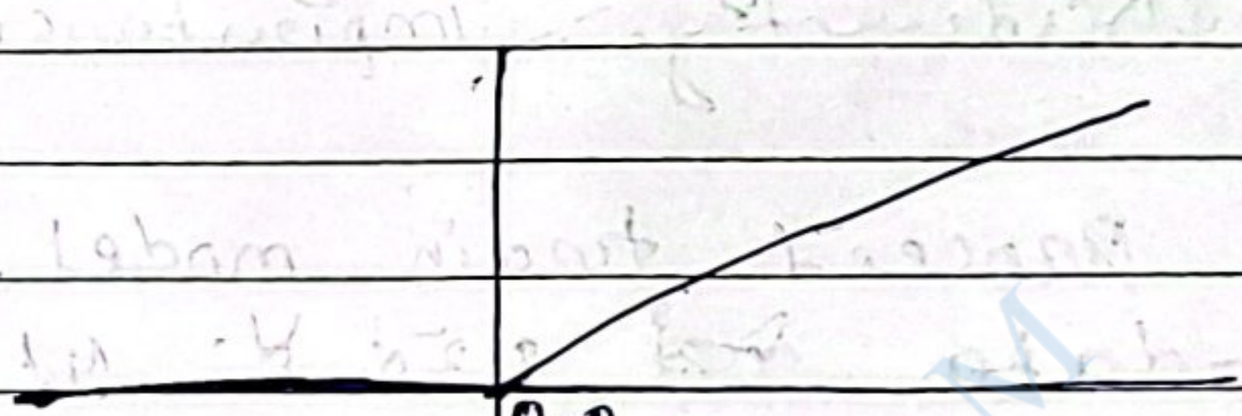
Diplomawallah.in

Date
Page

* ReLU

ReLU (Rectified linear unit) is the most widely used activation function in deep learning especially in hidden layers of neural networks.

It outputs the input directly if it is positive; otherwise, it o/p. zero.

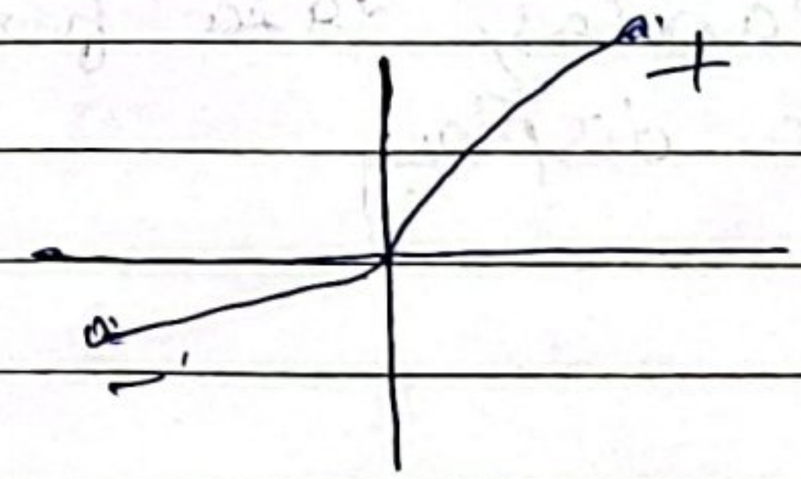


It is the hidden layer of deep neural networks.

* leaky ReLU

Leaky ReLU is a modified version of the ReLU activation function designed to solve the dying ReLU problem.

Instead of outputting zero for all negative inputs, leaky ReLU allows a small, non-zero slope for negative values.



Diplomawallah.in

Q. Pandas

- why groupby importance, aggregation summarization data.
  explain code.
- why pivot on melt function.
  data processing missing value, role, handle disadvantage, importance.

↑ data Proceed train model, testing
~~Stratified~~ data करते समय 80% और 20% का
20% testing
80% training

- Visualization help EDA.
  Explorarity data Analysis.

Q. Long

- Groupdata by a column Calculate min and display data.

→ Matplotlib or Seaborn का use करके display plot distribution. (numerical data Pandas/ data frame से)
Histogram display

Diplomawallah.in

Date
Page

* Decision Rule

$P \leq \alpha$	α ISO value. ~~If the~~ to check null value * will or not.
$P \geq \alpha$	

* $P \leq \alpha$ · reject null hypothesis.
* $P > \alpha$ fail to reject null · hypothesis.

Q. Average phone usage four hours per day.

write ashy pivot and melt function dataset reshape

How data will be reshape dataset में given है और analysis column के बेसिस पे हमें mean, median Variation, Standard deviation, visualize box plot to detect outliers.

detect outliers using IQR method and print numerical value.

Complete machine program to explic data train test build linear regression Compute R^2 store visualize actua vs Predict value i matplotlib or Scatter plot

Diplomawallah.in

Sagar Sangam | Null hypothesis equal H_0 ≥ 40% then give error

Date
Page

Types of error

① alpha (α) error:- Reject H_0 when it is actually true.
Ex- We claim is our new model is better

(II) β(Beta error):- Failed to reject ~~non~~ null hypothesis when it is false.
Ex- our new model is better but we fail to detect it.

~~Test types~~

Test Statistics:- A test statistics is a number calculate from sample data that tell us how far ~~is~~ data the data from null hypothesis.

Type
① Z → The large samples.
t → Small samples.
x^2 → Categorical data.
f → Testing while comparision.

Diplomawallah.in

Date
Page

* Join probability that two events occur together.

* Based on ~~update~~ helps us update the probability of event when new information is available.

* A Bank ~~one~~ study to customer to analyse loan default or

Hypothesis

Hypothesis testing is formal way to detect whether a p claim about data is likely true or not.

* Null Hypothesis (H_0):— Nothing special is happening.

Ex:- The ML model has the same accuracy as the old one.

* Alternative hypothesis:- Something has changed

Ex:- The new model accuracy is different or better.

Diplomawallah.in

Date
Page

D -1

Customer	default	Income < 50K
1	0	1
2	1	1
3	0	0
4	0	0
5	1	1
6	0	0
7	1	1
8	0	0
9	0	0
10	1	1

Of the event occuring given that another event already occured.

Q: probability of that a Customer default given income is less than 50K?

4, 5, 7, 10

$P(D) = \frac{4}{10} = 0.4$

$P(L) = \frac{5}{10} = 0.5$

$P(L|D) = \frac{P(D \cap L)}{P(D)}$

$= \frac{0.3 \times 0.4}{0.5}$

$= 0.24$

$= \frac{0.4}{0.5} = 0.8$

Diplomawallah.in

Big data

Big data simply means data so large, fast complex that normal tool can't handle it.

Example:- millions of Instagram photo uploaded every minutes.

Google processing millions of search daily.

Why Big data Matter in AI & ML.

→ AI & ML need huge data to learn pattern.

→ Big data include improve model accuracy.

⇒ Big data helps uncover hidden pattern, human miss trends ML algorithm does not.

Example:- Bank finding unsual transaction of fraud detection.

Types of Big data.

i) Structured Data: Data organised in row & Column.

Ex- SQL database, banking transaction

ii) unstructure data: No fix format.

Ex- Image, audio, video etc.

iii) Semi. Structured data:.

Ex- JSON file, XML file, image etc.

Diplomawallah.in

Date
Page

Serverless Services

Serverless doesn't mean without server, It Means we do not manage them.

* The cloud service provider automatically handle Server Setup, Scaling.

* How it works

* we upload our code as a function. where a fogw figure occurs trigger occur the function would automatically.

* you we pay only a time while function run.

* Virtualization

* Virtualization is a core technology behind Cloud computing.

* It allow physical Server/machine to run multiple Virtual machine each acting like a independent computer.

* A Software layer called hypervisor divide digital hardware into multiple virtual environment. Each virtual machine its own OS memory & storage.

Diplomawallah.in

mkdir mini-project

cd mini-project

git init

git status

git add . → all the file in the project folder

git commit -m " changes messages"

Create Github Repository

→ Go to github → New Repo → Name it mini-project

→ Copy the repo URL (.git)

https://github.com/username/min-project.git

→ Linking with local Repo with Github.

git remote add origin "url"

→ push Code

git branch -M Main

git push -u origin main.

From next time we only use.

git add .

git commit -m " Changes messages"

git push.

Diplomawallah.in

Date 19
Page

Application

* E-Commerce :- product recommendation.
* Transportation. Root deviestation.
* Business :- market analysis and sales forecasting.
* Social media :- Train Analysis.

* Difference b/w data science & machine learning.

DataSci - Broader field. data collection, analysis cleaning, visualization and modelling.

Machi - Subset of data science focus only on training model to learn pattern.

Data - Involving statistics visualization

Machine - Involve algorithm and prediction.

Data - Data science find the decision.

Machine - To make accurate prediction.