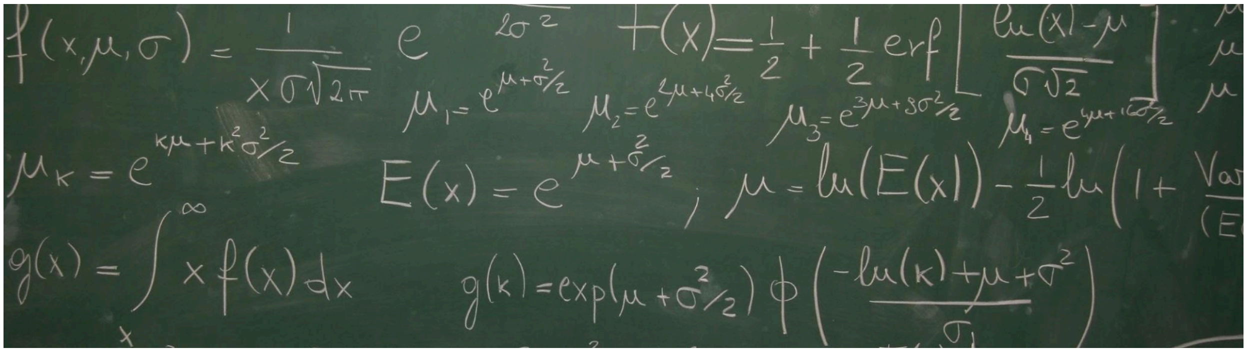

$f(x,\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e$ $2\sigma^2$
$F(x) = \frac{1}{2} + \frac{1}{2} \text{erf}\left[\frac{\ln(x)-\mu}{\sigma\sqrt{2}}\right]$
$\mu_1 = e^{\mu+\sigma^2/2}$ $\mu_2 = e^{2\mu+4\sigma^2/2}$ $\mu_3 = e^{3\mu+9\sigma^2/2}$ $\mu_4 = e^{4\mu+16\sigma^2/2}$
$\mu_K = e^{k\mu+k^2\sigma^2/2}$
$E(x) = e^{\mu+\sigma^2/2}$; $\mu = \ln(E(x)) - \frac{1}{2}\ln(1+$ Var
$g(x) = \int^{\infty} x f(x) dx$
$g(k) = \exp(\mu+\sigma^2/2)\,\phi\left(\frac{-\ln(k)+\mu+\sigma^2}{\sigma}\right)$

Jan 3, 2025 12:00 PM

Introduction to Machine Learning

Table of Contents

Machine Learning

Simplified Explanation of Data Science

Data Science is the art and science of using data to solve problems and make better decisions.

It combines:

- **Math & Statistics**: To analyze data.
- **Programming**: To work with and process large amounts of data.
- **Domain Knowledge**: To understand the context of the data.
- **Storytelling**: To communicate findings in a simple way.

Analogy

Think of Data Science as being like a **detective**:

1. **Collect clues** (data).
2. **Analyze clues** (math, programming).

3. **Solve the mystery** (find insights or solutions).
4. **Explain the case** (visualize and share results).

Why Do We Use Data Science?

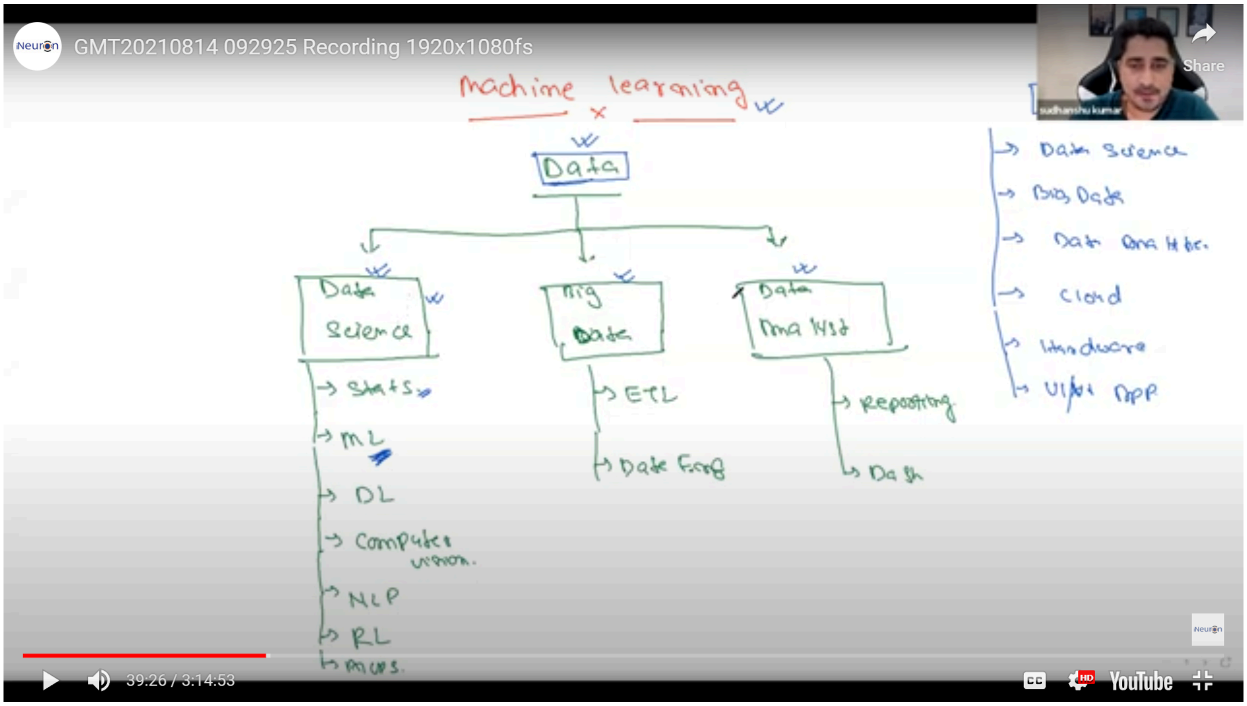
1. To **predict** future outcomes (e.g., weather forecasts, stock prices).
2. To **find patterns** (e.g., customer preferences).
3. To **optimize processes** (e.g., delivery routes, manufacturing).

What Do Data Scientists Do?

- Gather data from various sources.
- Clean and organize messy data.
- Analyze data to find patterns or trends.
- Build models to predict or automate tasks.
- Share results using visuals and reports.

In short, Data Science is about **making sense of data** to solve real-world problems. Simple! 😊

Data Industry:



Artificial Intelligence (AI) is not a technology, It's a system to simulate human intelligence in machines.

E.g: Recommended system, Automatic cars, Google Maps, Targeted ads, etc.

Artificial Intelligence (AI) refers to the simulation of human intelligence in machines, enabling them to perform tasks that typically require human cognition. These tasks include learning, reasoning, problem-solving, understanding natural language, and perception.

Key Concepts in AI

1. **Learning:**
   - The process by which machines improve their performance on tasks using data or experiences.
   - Types:
     - **Supervised Learning:** Learning from labeled data.
     - **Unsupervised Learning:** Finding patterns in unlabeled data.
     - **Reinforcement Learning:** Learning through trial and error with rewards and punishments.

2. **Reasoning:**

   - The ability to solve problems and make decisions based on available information.
   - Example: A chess AI calculates possible moves and their outcomes to decide the best move.

3. **Perception:**

   - Interpreting and making sense of the world through inputs like images, sound, or touch.
   - Example: Facial recognition systems.

4. **Natural Language Processing (NLP):**

   - Enabling machines to understand, interpret, and generate human language.
   - Example: Virtual assistants like Siri or Google Assistant.

5. **Problem Solving:**

   - Finding solutions to complex or unfamiliar tasks.
   - Example: Planning a robot's path through a maze.

Types of AI

1. **Narrow AI (Weak AI):**

   - AI is designed for specific tasks.
   - Examples: Spam filters,and recommendation systems.

2. **General AI (Strong AI):**

   - AI is capable of performing any intellectual task that a human can do (still hypothetical).

3. **Superintelligent AI:**

   - An AI that surpasses human intelligence in all aspects (theoretical concept).

Key Subfields of AI

1. **Machine Learning (ML):**

   - A subset of AI that enables machines to learn from data and improve over time without explicit programming.
   - Example: Predicting stock prices.

2. **Deep Learning:**
    - A subset of ML that uses neural networks with multiple layers to model complex patterns in data.
    - Example: Image and speech recognition.
3. **Robotics:**
    - Designing intelligent machines that can perform physical tasks.
    - Example: Autonomous drones.
4. **Computer Vision:**
    - Teaching machines to interpret and analyze visual data.
    - Example: Object detection in images.
5. **Expert Systems:**
    - AI systems that emulate the decision-making abilities of a human expert.
    - Example: Medical diagnosis tools.

Applications of AI

1. **Healthcare:** Diagnosing diseases, drug discovery, surgical robots.
2. **Finance:** Fraud detection, algorithmic trading, credit scoring.
3. **Retail:** Recommendation engines, chatbots, inventory management.
4. **Transportation:** Self-driving cars, traffic prediction.
5. **Entertainment:** Content recommendations on platforms like Netflix or Spotify.
6. **Agriculture:** Predicting crop yields, automated harvesting.

Benefits of AI

- Automates repetitive tasks, saving time and resources.
- Enhances decision-making with accurate predictions.
- Provides solutions to complex problems in various industries.

Challenges and Risks

- **Bias in Data:** AI can inherit biases from the data it is trained on.
- **Ethical Concerns:** Privacy, surveillance, and misuse of AI technologies.
- **Job Displacement:** Automation may lead to reduced demand for certain jobs.
- **Dependence:** Over-reliance on AI systems could lead to vulnerabilities.

Simplified Analogy

AI is like a **super assistant**:

- **Narrow AI** is a specialist (e.g., a calculator or language translator).
- **General AI** is a versatile human-like assistant (hypothetical).
- **Superintelligent AI** is an all-knowing guide (future concept).

Summary:
Artificial Intelligence enables machines to think and act intelligently by mimicking human cognitive processes. It has transformative potential across industries but comes with challenges that require careful handling.

Data Science: Data Science is the process of using data to find patterns, gain insights, and solve problems by combining mathematics, statistics, programming, and domain knowledge. It involves collecting, cleaning, analyzing, and interpreting data to make informed decisions.

Data Science is a multidisciplinary field that involves using scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data. It combines elements of statistics, computer science, domain expertise, and machine learning to analyze data and derive actionable insights.

Here's a **single unified approach** to streamline the content while ensuring all key aspects are included in a cohesive and accessible way. The goal is to balance clarity, depth, and structure.

The Machine Learning Pipeline: A Unified Approach

The **Machine Learning (ML) pipeline** or **Data Science Workflow** is a structured framework consisting of sequential phases to develop data-driven solutions. Each step ensures the creation of effective and robust models for solving real-world problems.

---

1. Problem Definition

- Identify the business or research problem.
- Define objectives, success metrics, and constraints.

---

2. Data Collection

- **Purpose**: Gather relevant data for analysis and modeling.
- **Sources**: Databases, APIs, web scraping, sensors, surveys, or manual entry.
- **Types**:
  - Structured (e.g., tabular data).
  - Unstructured (e.g., images, videos, text).

---

3. Data Preprocessing

- **Cleaning**: Handle missing values, remove duplicates, correct inconsistencies.
- **Transformation**: Normalize or standardize data, encode categorical variables.
- **Storage**: Organize data using databases or file systems (e.g., MySQL, AWS S3).

---

4. Feature Engineering

- **Purpose**: Enhance the dataset with meaningful features.
- **Techniques**:
  - Create new features (e.g., extracting age from date of birth).
  - Perform dimensionality reduction (e.g., PCA).
  - Select important features using statistical or algorithmic methods.

---

5. Exploratory Data Analysis (EDA)

- **Goal**: Understand the data distribution, patterns, and anomalies.
- **Steps**:
  - Summarize the data: `describe()`, `info()`.
  - Visualize: Scatter plots, histograms, heatmaps (tools: Matplotlib, Seaborn).
  - Test hypotheses: Analyze correlations or perform statistical tests.

---

6. Model Building

- **Model Selection**: Choose algorithms based on the problem (e.g., regression for prediction, clustering for grouping).
- **Training**: Train the model on labeled or unlabeled data using libraries like scikit-learn or TensorFlow.
- **Validation**: Split the data into training and testing sets; perform cross-validation.

---

7. Model Evaluation

- **Metrics**:
  - Regression: RMSE, R².
  - Classification: Accuracy, Precision, Recall, F1-score, AUC-ROC.
- **Visualization**: Use confusion matrices, precision-recall curves, and other plots to evaluate performance.
- **Tradeoffs**: Manage bias-variance tradeoffs to ensure generalization.

---

8. Hyperparameter Tuning

- **Goal**: Optimize the model for better performance.
- **Techniques**: Grid Search, Random Search, or Bayesian Optimization.

---

9. Deployment

- **Export Models**: Save models for reuse (e.g., using Pickle or ONNX).
- **Integration**: Deploy using Flask, FastAPI, or cloud platforms like AWS and Google Cloud.
- **Real-World Use**: Enable predictions or insights in live environments via APIs or dashboards.

---

10. Monitoring and Maintenance

- **Performance Monitoring**: Track metrics post-deployment using tools like Grafana or Prometheus.
- **Model Updates**: Retrain the model with new data periodically.
- **Error Analysis**: Investigate failures to improve model robustness.

---

11. Insights and Decision-Making

- **Communication**: Share findings using dashboards, visualizations (e.g., Tableau, Power BI), or reports.
- **Action**: Use the insights to drive data-informed decisions.

Why This Approach?

- Combines theoretical and practical aspects.
- Covers all stages from data collection to decision-making.
- Balances technical depth and accessibility.
- Can be customized based on the specific project or domain.

Applications of Data Science

- **Healthcare:** Predicting diseases, personalized medicine, drug discovery.
- **Finance:** Fraud detection, risk management, customer segmentation.
- **Retail:** Recommendation systems, inventory optimization, pricing strategies.
- **Transportation:** Route optimization, autonomous vehicles.
- **Marketing:** Customer sentiment analysis, targeted advertising.

Key Skills for Data Science

1. **Programming Languages:**
   - Python, R, SQL.
2. **Mathematics and Statistics:**
   - Probability, linear algebra, calculus, statistical modeling.
3. **Machine Learning and AI:**
   - Algorithms, libraries (Scikit-learn, TensorFlow, PyTorch).
4. **Big Data Technologies:**
   - Hadoop, Spark, Kafka.
5. **Data Visualization:**
   - Tools like Tableau, Power BI, or libraries like Matplotlib and Seaborn.

Simplified Analogy

Think of data as **raw ingredients**, and Data Science as the **recipe** and **chef** that transforms it into a **delicious dish**—a meaningful insight that can be used to make decisions or automate processes.

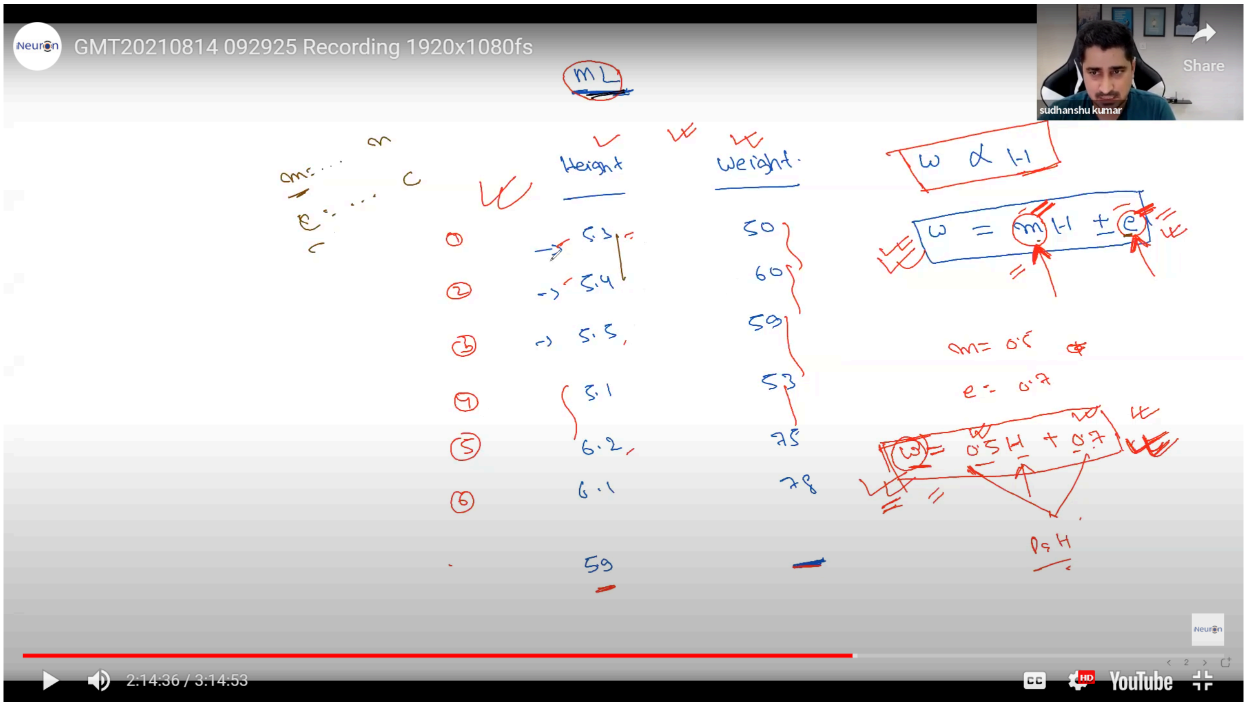
Summary:

Data Science is about turning raw data into knowledge and actionable insights using a blend of mathematics, programming, and domain expertise. It is the backbone of decision-making in many industries today.

Learning is just finding out the pattern, either humans or machines, We just find out the pattern, for instance, A baby can catch a snake, but can we do, it because, We have heard that snakes are dangerous, and it can harm us, since most of the snakes are not dangerous, So, We followed the pattern to not touch the snakes, more examples are tastes of the fruits, wines, veg, non-veg, we can differentiate, once we have tasted it, and if again someone provides us the chicken or banana, we can differentiate the tastes of fruits, and non-veg and so on, even blindly.

Similarly, We do with machines, machines are just pieces of hardware, made up of silicon chips, and similar things, So, How can we can make machines to learn, we can not feed the chicken lollipops or bananas to the machines, So, What can we do is that we can solve the problems in their characteristics, and their characteristic is to do calculations very fast, like addition, subtraction, multiplication, and division, So, We can take advantages of the machines, and do perform these calculations. So, our goal was to make the machine learn, which means the machine has to find the pattern. So, machines learn by algorithms in the form of mathematical equations, and with the algorithms(Mathematical equations) machines learn.

Example:



GMT20210814 092925 Recording 1920x1080fs

Share

sudhanshu kumar

ML

m...

e...

c

	Height	Weight
1	5.3	50
2	5.4	60
3	5.5	59
4	5.1	53
5	6.2	75
6	6.1	78
	59	

$w \propto H$

$w = mH \pm e$

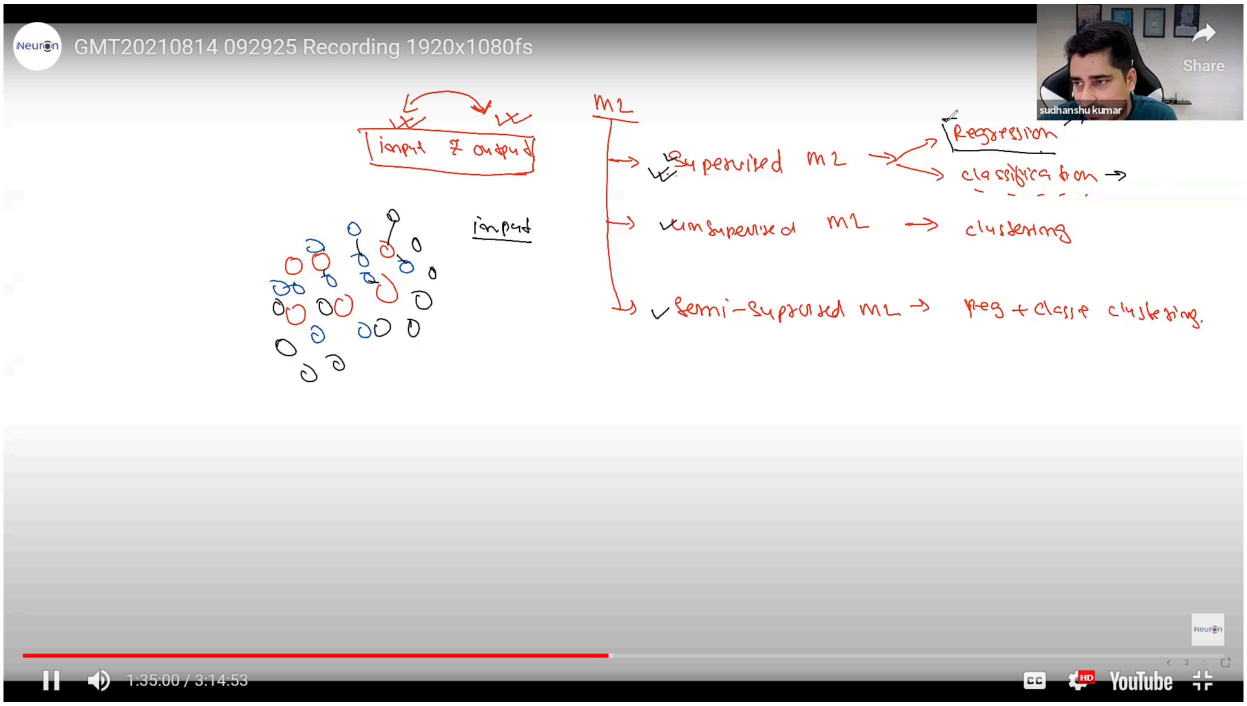
$m = 0.5$

$e = 0.7$

$w = 0.5H + 0.7$

$0.5H$

2:14:36 / 3:14:53



GMT20210814 092925 Recording 1920x1080fs

Share

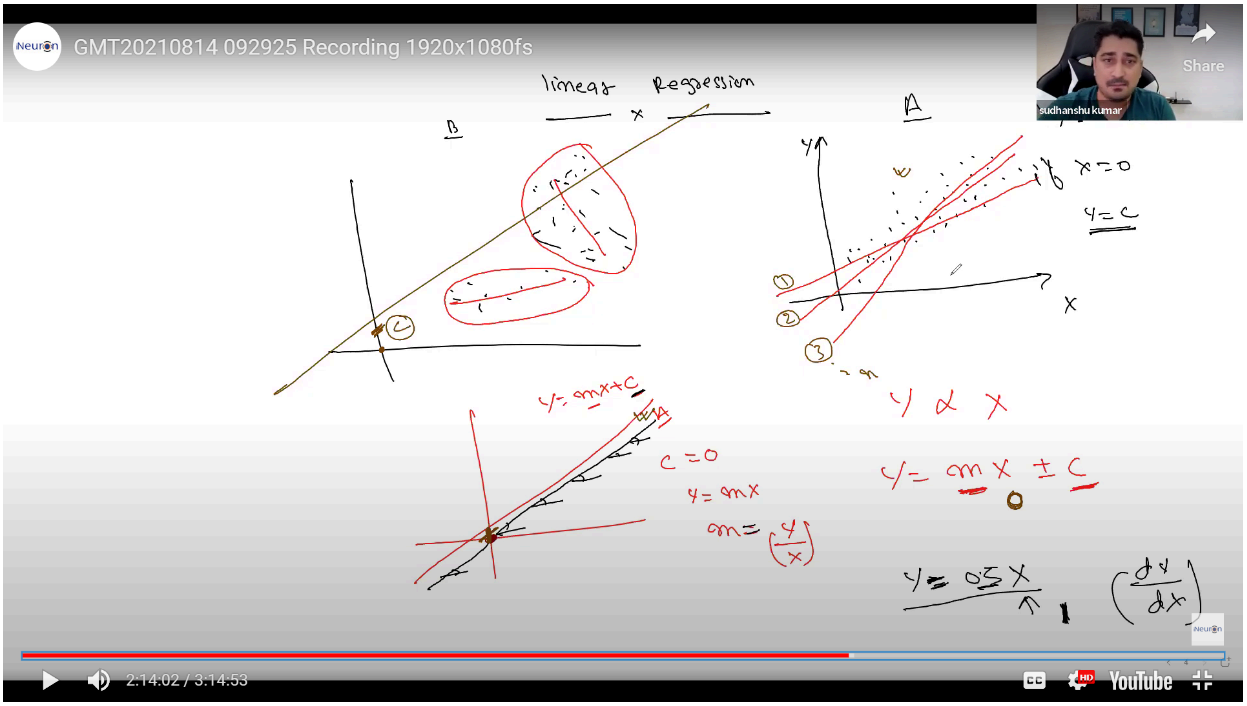
sudhanshu kumar

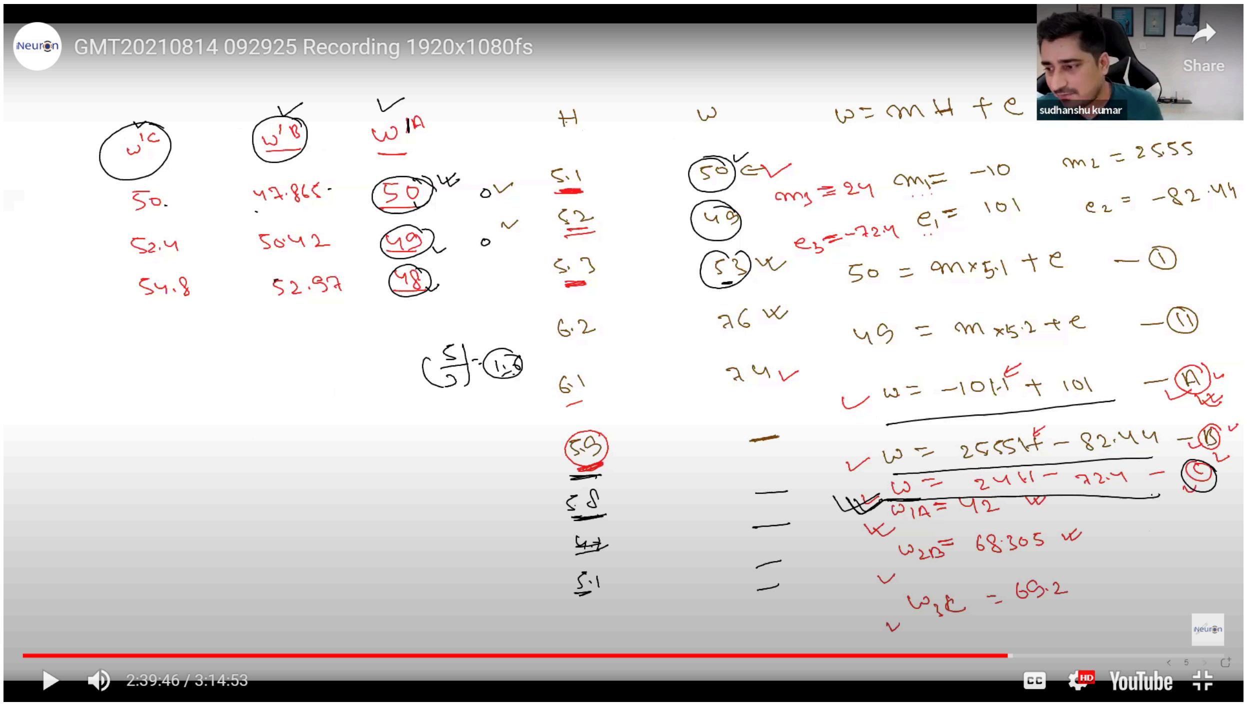
input ≠ output

input

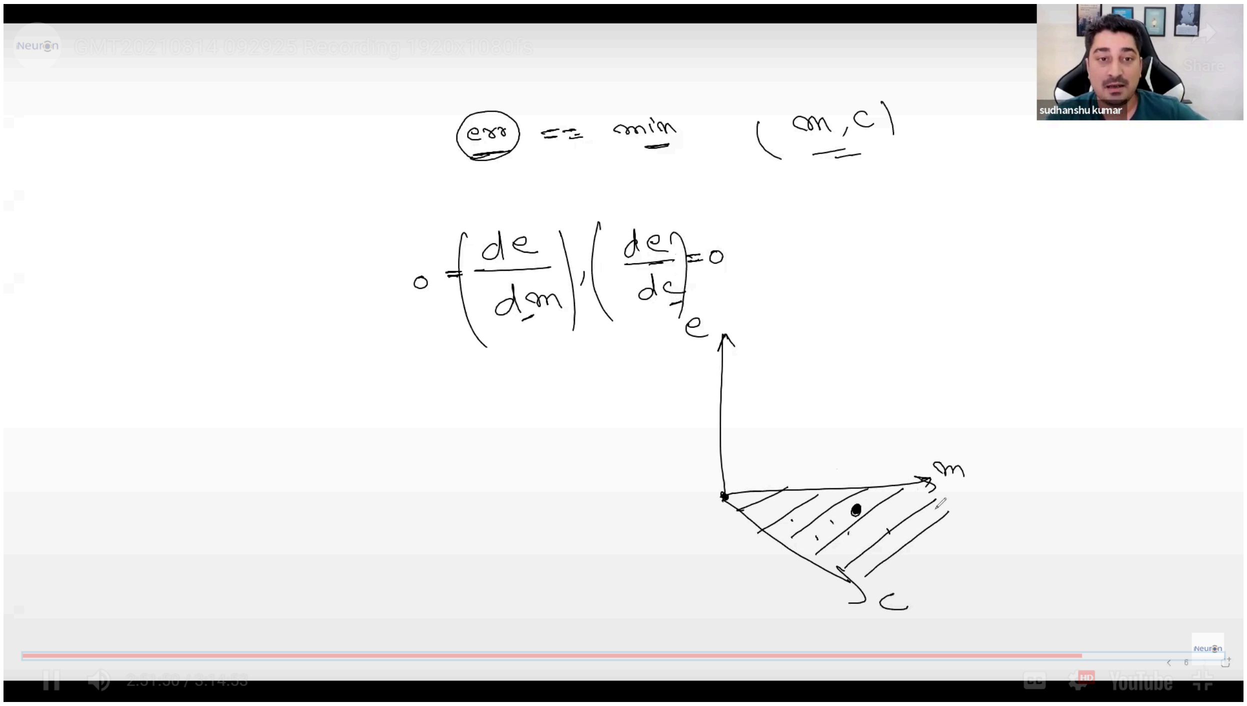
ML
- → supervised ML → Regression, classification →
- → unsupervised ML → clustering
- → semi-supervised ML → reg + classif clustering.

1:35:00 / 3:14:53





Learning: The way in which we try to find out the optimal value of ‘m’, and ‘c’ is called as learning.



$$\boxed{\underline{err}} == \underline{min} \quad (\underline{m}, c)$$

$$0 = \left(\frac{de}{dm}\right), \left(\frac{de}{dc}\right) = 0$$