
IT496: Introduction to Data Mining



Lecture - 01

Introduction

[Definition, Tasks, and Case Studies]

Arpit Rana

25th July 2023

Course Logistics

Syllabus and Evaluation Scheme

Course Logistics

Instructor	Arpit Rana Room-3105, Faculty Block-3 Email: arpit_rana@daiict.ac.in
TA Contact Info	Himanshu Beniwal (himanshubeniwal@iitgn.ac.in) - Head TA Dhara Shah (202211008@daiict.ac.in) Naiya Patel (202211075@daiict.ac.in)
Prerequisites	Programming in Python, Linear Algebra, Probability and Statistics
Eligibility	<ul style="list-style-type: none">• B.Tech. V / VII Semester (ICT/MnC)• M.Tech. ICT - I Semester (ML/SS specialization)• M.Sc. DS - III Semester

Course Logistics

Credit Weighting	3-0-2-4 (L-T-P-Cr)
Lectures [CEP-108]	Tuesday: 10:00 – 11:00, Thursday, Friday: 11:00 – 12:00
Lab/Tutorial [M.Sc. DS Lab]	Friday, 14:00 – 16:00
Private Study	At least 5 hrs per week
Potential Outcome	<ul style="list-style-type: none">• Learn how to solve Data-driven Decision-Making Problems;• Learn how to work on structured and unstructured (e.g., text, image, sequential) data;• Targeted Jobs: <i>Data Analyst, Data Engineer, Data Scientist, ML Engineer, Research Engineer</i>

Course Logistics

Assessment	<ul style="list-style-type: none">• Surprise Quizzes: 25%• End Term: 25%• Course Projects: 40% (10% + 15% + 15%)• Case Study: 10% <p><u>Extra Credits</u>: ML Challenges, Participate on Course Stream</p>
How to Fail	Skip lectures; avoid private study; cram just before the exam; expect the exam to be a memory test; copy project assignments; be inactive on the course stream
How to Pass	Attend lectures; summarize the notes; expect a problem-solving exam; do your project yourself; <u>be active and accurate in the class and on the course stream</u>

Preliminary Schedule

Week	Lecture	Lab	Due ¹
Week-1 [24 July 2023]	Introduction	- No lab -	-
Week-2 [31 July 2023]	Statistics for Data Mining	- No lab -	-
Week-3 [07 Aug 2023]	Data Preprocessing	End-to-End ML Project in Python	-
Week-4 [14 Aug 2023]	Fundamentals of Predictive Analytics - I Holidays: 15 Aug (Tues), 16 Aug (Wed)	CP - 1	Sunday, 10 Sept. 2023
Week-5 [21 Aug 2023]	Fundamentals of Predictive Analytics – II		
Week-6 [28 Aug 2023]	Regression Techniques Holidays: 30 Aug (Wed) 28 Aug (Mon) to be treated as Tues		
Week-7 [04 Sept 2023]	Dimensionality Reduction Holidays: 07 Sept (Thurs)		
Week-8 [11 Sept 2023]	First In-Semester Exam Week	Evaluation: CP - 1	

Preliminary Schedule

Week	Lecture	Lab	Due ¹
Week-9 [18 Sept 2023]	Eager Classifiers – I: Support Vector Machines and Decision Trees <i>Holidays: 19 Sept (Tues)</i>	CP – 2	Sunday, 15 Oct 2023
Week-10 [25 Sept 2023]	Eager Classifiers – II: Neural Networks <i>Holidays: 28 Sept (Thurs) 29 Sept (Fri) to be treated as Thurs</i>		
Week-11 [02 Oct 2023]	Eager Classifiers – III: Neural Networks Contd. <i>Holidays: 02 Oct (Mon)</i>		
Week-12 [09 Oct 2023]	Lazy Classifiers and Ensemble Techniques		
Week-13 [16 Oct 2023]	Second In-Semester Exam Week	Evaluation: CP – 2	

Preliminary Schedule

Week	Lecture	Lab	Due ¹
Week-14 [23 Oct 2023]	Cluster Analysis – I <i>Holidays: 24 Oct (Tues)</i>	CP - 3	Sunday, 19 Nov 2023
Week-15 [30 Oct 2023]	Cluster Analysis – II <i>Holidays: 31 Oct (Tues) 03 Nov (Fri) to be treated as Tues</i>		
Week-16 [6 Nov 2023]	Outlier Analysis		
Week-17 [13 Nov 2023]	In-Semester Break		
Week-18 [20 Nov 2023]	Association Rule Mining	Evaluation: CP - 3	
Week-19 [27 Nov 2023]	End-semester Examination		

1 – Course Projects (CPs) are due at 11:59 PM on the due date listed.

Preliminary Schedule


Lab Projects and Case Study:

- Projects will be allocated to groups, with **each group consisting of four members**.
- The assigned group will be responsible for delivering a **case study** on an AI-focused startup that addresses issues of the nation's priority and global interest.
- Further instructions related to the labs and case study presentations will be provided later on the classroom portal.



Introduction

Definition and Tasks



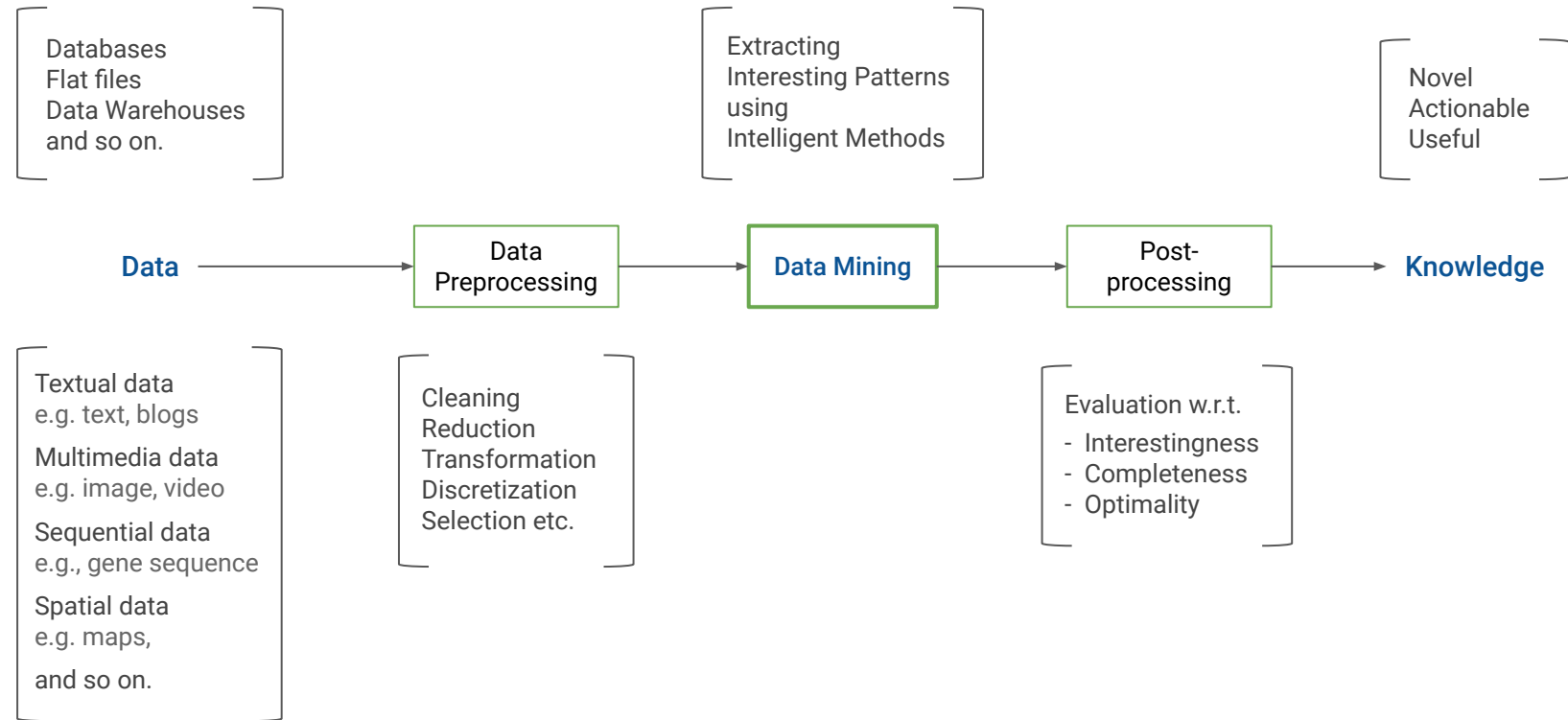
What is Data (Knowledge) Mining?

The process of automatically (or semi-automatically) discovering *interesting patterns* from large amounts of data.

- Implicit (somewhat hidden),
- Non-trivial (not obvious),
- Previously unknown (novel), and
- Potentially useful (for consumers / sellers / stakeholders)



Data (Knowledge) Mining: Knowledge Discovery from Data



Data Mining vs. Machine Learning

The process of automatically (or semi-automatically) discovering *interesting patterns* from large amounts of data.

It uses methods at the intersection of *machine learning*, *statistics*, and *database systems*.

E.g., customer churn



Machine learning (ML) is focused on understanding and building methods that *'learn'*.

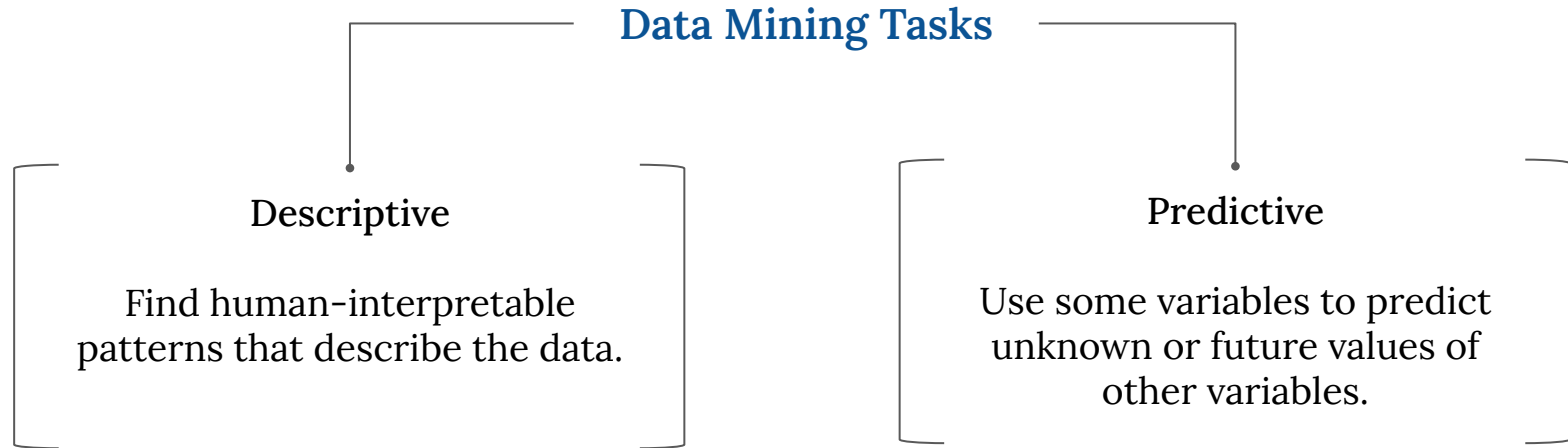
It leverages data to improve performance on some set of tasks.

E.g.: A spam filter (an ML program)



Data Mining Tasks

The actual data mining task is the semi-automatic or automatic analysis of large quantities of data to extract interesting patterns.



Descriptive Tasks

- **Cluster Analysis** (groups of data records),
 - Market Segmentation
 - Document clustering
- **Anomaly Detection | Outlier Analysis** (unusual records), and
 - Credit card fraud detection
 - Stock market manipulation detection
- **Association Rule Mining**, Sequential pattern mining (dependencies)
 - Market-basket analysis for sales promotion, shelf and inventory management
 - Medical informatics to find combination of patient symptoms and test results associated with certain diseases

Predictive Tasks

- **Classification** (predicting the class of a record)
 - Categorizing news stories as finance, weather, entertainment, sports, etc
 - Classifying land covers (water bodies, urban areas, forests, etc.) using satellite image data
- **Regression** (predicting the value of a variable of a record)
 - Predicting sales amounts of new product based on advertising expenditure.
 - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
 - Time series prediction of stock market indices.



Business Case Studies



Fakespot, GoKwik, and Intello Labs

Case Study - I: Fakespot

Problem Identified

- Nearly 93% consumers read reviews before any kind of purchasing decision.
- Out of these, around 91% of 18–34 year olds trust reviews as much as a recommendation from a friend!
- Over 30% of reviews are found to be *fake*.

Target Audience

All e-commerce businesses that allow users to write reviews.

Data-driven Solution

Fakespot reports provide an Adjusted Rating that weighs reviews based on authenticity and then recalculates it.



Courtesy: Fakespot

Case Study - II: GoKwik

Problem Identified

- In e-commerce, more than 30% of orders are returned to origin (RTO, i.e. shipped back to the warehouse) in India.

Target Audience

All *e-commerce* businesses

Data-driven Solution

- Mostly, CoD orders are converted to RTO.
- So, analyzing customer behavioural patterns and disable CoD option for those showing high-risk RTO behaviour.



Courtesy: Gokwik

Case Study - III: Intello Labs

Problem Identified

- One-third of the food produced in the world for human consumption every year gets lost or wasted.
- Mainly (in some countries) at the early stages of the food value chain.

Target Audience

From growers to packers, from exporters to food services

Data-driven Solution

Smart, scalable solutions to *digitize food quality*, achieve fair pricing and reduce food wastage.



Best delivered today



*Best delivered in 2
days*



*Best delivered in 5
days*

Using AI, ML, and Computer Vision technology

Next lecture

Statistics for Data Mining

27th July 2023
