

## A Comprehensive Study of Data Mining and ITS Techniques

Nithya C<sup>1</sup>, Saravanan V<sup>2</sup>

Department Of Computer Science, Hindusthan College Of Arts And Science, Coimbatore, India.

mail2nithu92@gmail.com

Department Of Information Technology, Hindusthan College Of Arts And Science, Coimbatore, India.

**Abstract:** Data mining is the process of discovering interesting knowledge patterns from large amount of data stored in database. It is an essential process where the intelligent techniques (i.e., machine learning, artificial intelligence, etc ) are used to extract the data patterns (i.e., features). The aim of data mining process is to extract the useful information from dataset and transform it into understandable structure for future use. Machine learning is a subset of artificial intelligence in the field of computer science that often uses statistical techniques to give computers the ability to "learn" (i.e., progressively improve performance on a specific task) with data, without being explicitly programmed. In machine learning, **feature learning** or **representation learning** is a set of techniques that allows a system to automatically discover the representations needed for feature detection or classification from raw data.

**Keywords:** Data Mining, Machine learning, Feature Learning.

### I. INTRODUCTION:

Data mining techniques has the ability to provide the set of useful rules for performing the tasks. Data mining is a process consisting in collecting knowledge from databases or data warehouses and the information collected that had never been known before, it is valid and operational. Nowadays data mining is a modern and powerful IT&C tool, automatizing the process of discovering relationships and combinations in raw data and using the results in an automatic decision support. Machine learning is a data analytics technique that teaches computers to do what comes naturally to humans and animals: learn from experience. Machine learning algorithms use computational methods to "learn" information directly from data without relying on a predetermined equation as a model. The algorithms adaptively improve their performance as the number of samples available for learning increases.

### II. DATA MINING INVOLVES SIX COMMON CLASSES OF TASKS:

- Anomaly detection (outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation.
- Association rule learning (dependency modelling) – Searches for relationships between variables. For example, a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.
- Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
- Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".
- Regression – attempts to find a function which models the data with the least error that is, for estimating the relationships among data or datasets.
- Summarization – providing a more compact representation of the data set, including visualization and report generation.

### MACHINE LEARNING MATTERS:

With the rise in big data, machine learning has become a key technique for solving problems in areas, such as:

- **Computational finance**, for credit scoring and algorithmic trading
- **Image processing and computer vision**, for face recognition, motion detection, and object detection
- **Computational biology**, for tumor detection, drug discovery, and DNA sequencing
- **Energy production**, for price and load forecasting
- **Automotive, aerospace, and manufacturing**, for predictive maintenance
- **Natural language processing**, for voice recognition applications

## WHEN SHOULD YOU USE MACHINE LEARNING?

Consider using machine learning when you have a complex task or problem involving a large amount of data and lots of variables, but no existing formula or equation. For example, machine learning is a good option if you need to handle situations like these:



## FEATURE LEARNING:

In machine learning, **feature learning** or **representation learning** is a set of techniques that allows a system to automatically discover the representations needed for feature detection or classification from raw data. This replaces manual feature engineering and allows a machine to both learn the features and use them to perform a specific task.

Feature learning is motivated by the fact that machine learning tasks such as classification often require input that is mathematically and computationally convenient to process. However, real-world data such as images, video, and sensor data has not yielded to attempts to algorithmically define specific features. An alternative is to discover such features or representations through examination, without relying on explicit algorithms.

Feature learning can be either supervised or unsupervised.

- In supervised feature learning, features are learned using labeled input data. Examples include supervised neural networks, multilayer perceptron and (supervised) dictionary learning.
- In unsupervised feature learning, features are learned with unlabeled input data. Examples include dictionary learning, independent component analysis, autoencoders, matrix factorization<sup>[2]</sup> and various forms of clustering.

## HOW MACHINE LEARNING WORKS:

Machine learning uses two types of techniques: **supervised learning**, which trains a model on known input and output data so that it can predict future outputs, and **unsupervised learning**, which finds hidden patterns or intrinsic structures in input data.

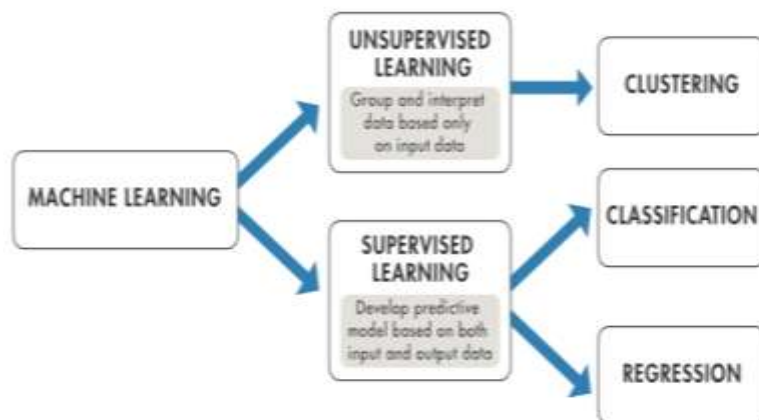


Figure 1. Machine Learning Techniques Include Both Unsupervised And Supervised Learning

## Supervised Learning

Supervised machine learning builds a model that makes predictions based on evidence in the presence of uncertainty. A supervised learning algorithm takes a known set of input data and known responses to the data

(output) and trains a model to generate reasonable predictions for the response to new data. Use supervised learning if you have known data for the output you are trying to predict.

Supervised learning uses classification and regression techniques to develop predictive models.

**Classification techniques** predict discrete responses—for example, whether an email is genuine or spam, or whether a tumor is cancerous or benign. Classification models classify input data into categories. Typical applications include medical imaging, speech recognition, and credit scoring.

Use classification if your data can be tagged, categorized, or separated into specific groups or classes. For example, applications for hand-writing recognition use classification to recognize letters and numbers. In image processing and computer vision, unsupervised pattern recognition techniques are used for object detection and image segmentation.

Common algorithms for performing classification include support vector machine (SVM), boosted and bagged decision trees, k-nearest neighbor, Naïve Bayes, discriminant analysis, logistic regression, and neural networks.

**Regression techniques** predict continuous responses—for example, changes in temperature or fluctuations in power demand. Typical applications include electricity load forecasting and algorithmic trading.

Use regression techniques if you are working with a data range or if the nature of your response is a real number, such as temperature or the time until failure for a piece of equipment.

Common regression algorithms include linear model, nonlinear model, regularization, stepwise regression, boosted and bagged decision trees, neural networks, and adaptive neuro-fuzzy learning.

### Using Supervised Learning to Predict Heart Attacks

Suppose clinicians want to predict whether someone will have a heart attack within a year. They have data on previous patients, including age, weight, height, and blood pressure. They know whether the previous patients had heart attacks within a year. So the problem is combining the existing data into a model that can predict whether a new person will have a heart attack within a year.

### Unsupervised Learning

Unsupervised learning finds hidden patterns or intrinsic structures in data. It is used to draw inferences from datasets consisting of input data without labeled responses.

**Clustering** is the most common unsupervised learning technique. It is used for exploratory data analysis to find hidden patterns or groupings in data. Applications for cluster analysis include gene sequence analysis, market research, and object recognition.

For example, if a cell phone company wants to optimize the locations where they build cell phone towers, they can use machine learning to estimate the number of clusters of people relying on their towers. A phone can only talk to one tower at a time, so the team uses clustering algorithms to design the best placement of cell towers to optimize signal reception for groups, or clusters, of their customers.

Common algorithms for performing clustering include k-means and k-medoids, hierarchical clustering, Gaussian mixture models, hidden Markov models, self-organizing maps, fuzzy c-means clustering, and subtractive clustering.



**Figure 2.** Clustering Finds Hidden Patterns In Your Data

### HOW DO YOU DECIDE WHICH MACHINE LEARNING ALGORITHM TO USE?

Choosing the right algorithm can seem overwhelming—there are dozens of supervised and unsupervised machine learning algorithms, and each takes a different approach to learning.

There is no best method or one size fits all. Finding the right algorithm is partly just trial and error—even highly experienced data scientists can't tell whether an algorithm will work without trying it out. But algorithm selection also depends on the size and type of data you're working with, the insights you want to get from the data, and how those insights will be used.

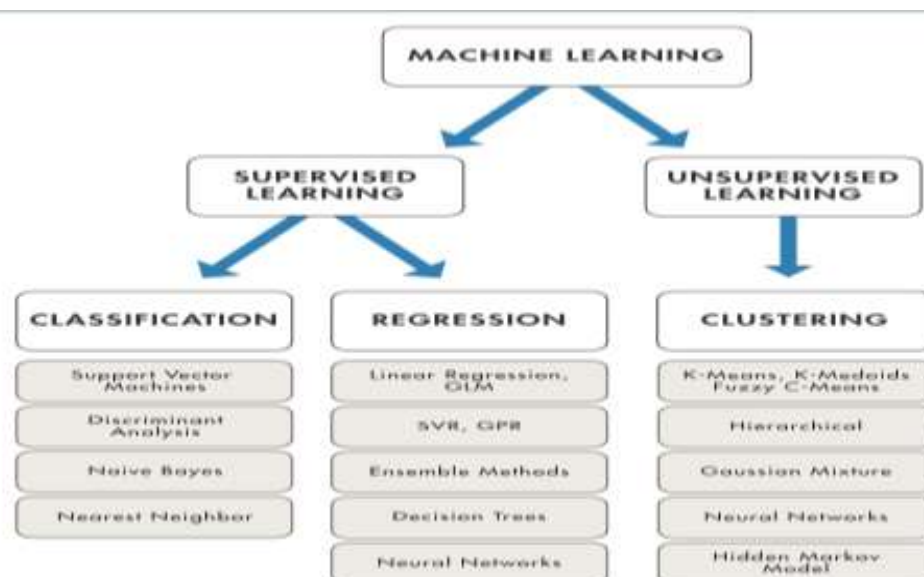


Figure 3. Machine Learning Techniques

Here are some guidelines on choosing between supervised and unsupervised machine learning:

- Choose **supervised learning** if you need to train a model to make a prediction—for example, the future value of a continuous variable, such as temperature or a stock price, or a classification—for example, identify makes of cars from webcam video footage.
- Choose **unsupervised learning** if you need to explore your data and want to train a model to find a good internal representation, such as splitting data up into clusters.

### DATA MINING VS MACHINE LEARNING: WHAT'S THE DIFFERENCE?

Data mining isn't a new invention that came with the digital age. The concept has been around for over a century, but came into greater public focus in the 1930s.

According to Hacker Bits, one of the first modern moments of data mining occurred in 1936, when Alan Turing introduced the idea of a universal machine that could perform computations similar to those of modern-day computers.

*Forbes* also reported on Turing's development of the "Turing Test" in 1950 to determine if a computer has real intelligence or not. To pass his test, a computer needed to fool a human into believing it was also human. Just two years later, Arthur Samuel created The Samuel Checkers-playing Program that appears to be the world's first self-learning program. It miraculously learned as it played and got better at winning by studying the best moves.

We've come a long way since then. Businesses are now harnessing data mining and machine learning to improve everything from their sales processes to interpreting financials for investment purposes. As a result, data scientists have become vital employees at organizations all over the world as companies seek to achieve bigger goals with data science than ever before.

### DATA MINING VS MACHINE LEARNING VS DATA SCIENCE:

With big data becoming so prevalent in the business world, a lot of data terms tend to be thrown around, with many not quite understanding what they mean. What is data mining? Is there a difference between machine learning vs. data science? How do they connect to each other? Isn't machine learning just artificial intelligence? All of these are good questions, and discovering their answers can provide a deeper, more rewarding understanding of data science and analytics and how they can benefit a company.

Both data mining and machine learning are rooted in data science and generally fall under that umbrella. They often intersect or are confused with each other, but there are a few key distinctions between the two. Here's a look at some data mining and machine learning differences between data mining and machine learning and how they can be used.

### DATA USE

One key difference between machine learning and data mining is how they are used and applied in our everyday lives. For example, data mining is often used *by* machine learning to see the connections between relationships. Uber uses machine learning to calculate ETAs for rides or meal delivery times for UberEATS.

Data mining can be used for a variety of purposes, including financial research. Investors might use data mining and web scraping to look at a start-up's financials and help determine if they want to offer funding. A company may also use data mining to help collect data on sales trends to better inform everything from marketing to inventory needs, as well as to secure new leads. Data mining can be used to comb through social media profiles, websites, and digital assets to compile information on a company's ideal leads to start an outreach campaign. Using data mining can lead to 10,000 leads in 10 minutes. With this much information, a data scientist can even predict future trends that will help a company prepare well for what customers may want in the months and years to come.

Machine learning embodies the principles of data mining, but can also make automatic correlations and learn from them to apply to new algorithms. It's the technology behind self-driving cars that can quickly adjust to new conditions while driving. Machine learning also provides instant recommendations when a buyer purchases a product from Amazon. These algorithms and analytics are constantly meant to be improving, so the result will only get more accurate over time. Machine learning isn't artificial intelligence, but the ability to learn and improve is still an impressive feat.

Banks are already using and investing in machine learning to help look for fraud when credit cards are swiped by a vendor. CitiBank invested in global data science enterprise Feedzai to identify and eradicate financial fraud in real-time across online and in-person banking transactions. The technology helps to rapidly identify fraud and can help retailers protect their financial activity.

#### **FOUNDATIONS FOR LEARNING:**

Both data mining and machine learning draw from the same foundation, but in different ways. A data scientist uses data mining pulls from existing information to look for emerging patterns that can help shape our decision-making processes. The clothing brand Free People, for example, uses data mining to comb through millions of customer records to shape their look for the season. The data explores best-selling items, what was returned the most, and customer feedback to help sell more clothes and enhance product recommendations. This use of data analytics can lead to an improved customer experience overall.

Machine learning, on the other hand, can actually learn from the existing data and provide the foundation necessary for a machine to teach itself. Zebra Medical Vision developed a machine learning algorithm to predict cardiovascular conditions and events that lead to the death of over 500,000 Americans each year.

Machine learning can look at patterns and learn from them to adapt behavior for future incidents, while data mining is typically used as an information source for machine learning to *pull* from. Although data scientists can set up data mining to automatically look for specific types of data and parameters, it doesn't learn and apply knowledge on its own without human interaction. Data mining also can't automatically see the relationship between existing pieces of data with the same depth that machine learning can.

#### **PATTERN RECOGNITION:**

Collecting data is only part of the challenge; the other part is making sense of it all. The right software and tools are needed to be able to analyze and interpret the huge amounts of information data scientists collect and find recognizable patterns to act upon. Otherwise, the data would largely be unusable unless data scientists could devote their time to looking for these complex, often subtle and seemingly random patterns on their own. And anyone even somewhat familiar with data science and data analytics knows this would be an arduous, time-consuming task.

Businesses could use data to shape their sales forecasting or determine what types of products their customers really want to buy. For example, Walmart collects point of sales from over 3,000 stores for its data warehouse. Vendors can see this information and use it to identify buying patterns and guide their inventory predictions and processes for the future.

It's true that data mining can reveal some patterns through classifications and and sequence analysis. However, machine learning takes this concept a step further by using the same algorithms data mining uses to automatically learn from and adapt to the collected data. As malware becomes an increasingly pervasive problem, machine learning can look for patterns in how data in systems or the cloud is accessed. Machine learning also looks at patterns to help identify which files are actually malware, with a high level of accuracy. All this is done without the need for constant monitoring by a human. If abnormal patterns are detected, an alert can be sent out so action can be taken to prevent the malware from spreading.

#### **IMPROVED ACCURACY:**

Both data mining and machine learning can help improve the accuracy of data collected. However, data mining and how it's analyzed generally pertains to how the data is organized and collected. Data mining may include using extracting and scraping software to pull from thousands of resources and sift through data that

researchers, data scientists, investors, and businesses use to look for patterns and relationships that help improve their bottom line.

One of the primary foundations of machine learning is data mining. Data mining can be used to extract more accurate data. This ultimately helps refine your machine learning to achieve better results. A person may miss the multiple connections and relationships between data, while machine learning technology can pinpoint all of these moving pieces to draw a highly accurate conclusion to help shape a machine's behavior.

Machine learning can enhance relationship intelligence in CRM systems to help sales teams better understand their customers and make a connection with them. Combined with machine learning, a company's CRM can analyze past actions that lead to a conversion or customer satisfaction feedback. It can also be used to learn how to predict which products and services will sell the best and how to shape marketing messages to those customers.

### III. CONCLUSION:

In this paper, a survey is carried out about Machine learning in data mining.

The future is bright for data science as the amount of data will only increase. By 2020, our accumulated digital universe of data will grow from 4.4 zettabytes to 44 zettabytes, as reported by Forbes. We'll also create 1.7 megabytes of new information every second for every human being on the planet.


As we amass more data, the demand for advanced data mining and machine learning techniques will force the industry to evolve in order to keep up. We'll likely see more overlap between data mining and machine learning as the two intersect to enhance the collection and usability of large amounts of data for analytics purposes.

According to reporting from *Bio IT World*, the future of data mining points to predictive analysis, as we'll see advanced analytics across industries like medical research. Scientists will be able to use predictive analysis to look at factors associated with a disease and predict which treatment will work the best.

We're just scratching the surface of what machine learning can do and how it will spread to help scale our analytical abilities and improve our technology. According to reporting from Geekwire, as our billions of machines become connected, everything from hospitals to factories to highways can be improved with IoT technology that can learn from other machines.

But some experts have a different idea about data mining and machine learning altogether. Instead of focusing on their differences, you could argue that they both concern themselves with the same question: "How we can learn from data?" At the end of the day, how we acquire and learn from data is really the foundation for emerging technology. It's an exciting time not just for data scientists but for everyone that uses data in some form.

### REFERENCES:

- [1]. <https://www.mathworks.com/discovery/machine-learning.html>
- [2]. <https://www.import.io/post/data-mining-machine-learning-difference/>
- [3]. <https://neilpatel.com/blog/data-mining/>
- [4]. <https://importioioweb.staging.wpengine.com/solutions/machine-learning/>
- [5]. [https://en.wikipedia.org/wiki/Machine\\_learning](https://en.wikipedia.org/wiki/Machine_learning)
- [6]. Coates, Adam; Lee, Honglak; Ng, Andrew Y. (2011). *An analysis of single-layer networks in unsupervised feature learning* (PDF). *Int'l Conf. on AI and Statistics (AISTATS)*.
- [7]. Csurka, Gabriella; Dance, Christopher C.; Fan, Lixin; Willamowski, Jutta; Bray, Cédric (2004). *Visual categorization with bags of keypoints* (PDF). *ECCV Workshop on Statistical Learning in Computer Vision*.
- [8]. Daniel Jurafsky; James H. Martin (2009). *Speech and Language Processing*. Pearson Education International. pp. 145–146.
- [9]. Y. Bengio; A. Courville; P. Vincent (2013). "Representation Learning: A Review and New Perspectives". *IEEE Trans. PAMI*, special issue *Learning Deep Architectures*. **35**: 1798–1828. [arXiv:1206.5538](https://arxiv.org/abs/1206.5538) . doi:10.1109/tpami.2013.50.
- [10]. Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). *"From Data Mining to Knowledge Discovery in Databases"* (PDF). Retrieved 17 December 2008.