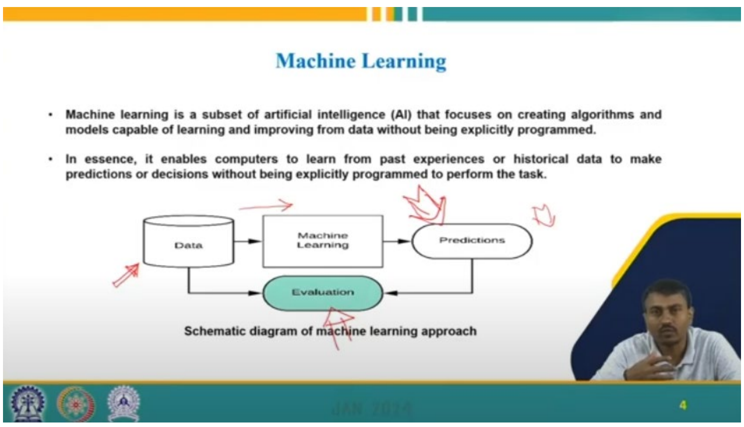
Mine Automation and Data Analytics
Prof. Radhakanta Koner
Department of Mining Engineering
IIT (ISM) Dhanbad
Week - 9
Lecture - 45
Introduction to Machine Learning

Welcome back to my course, Mine Automation and Data Analytics. Today, we will discuss machine learning. And in today's lesson, we will start with the introductory part of machine learning. You have seen in this course that we initially went through the different sensing parts adopted in the Mine Automation strategy. We acquire and collect vast data daily through these sensing tools. And these data we have tried to model the use of these data initially through different statistical methods and statistical tools that we have discussed.

Now, we want to proceed further to take this data for further optimization and get better patterns on the data for optimizing the mining process chain. So today, let us start with the introduction of machine learning. So, in this lesson, we will discuss machine learning and its type and what machine learning is. Then, we will discuss supervised learning, and primarily in supervised learning, we will try to cover the regression and classification with some examples.

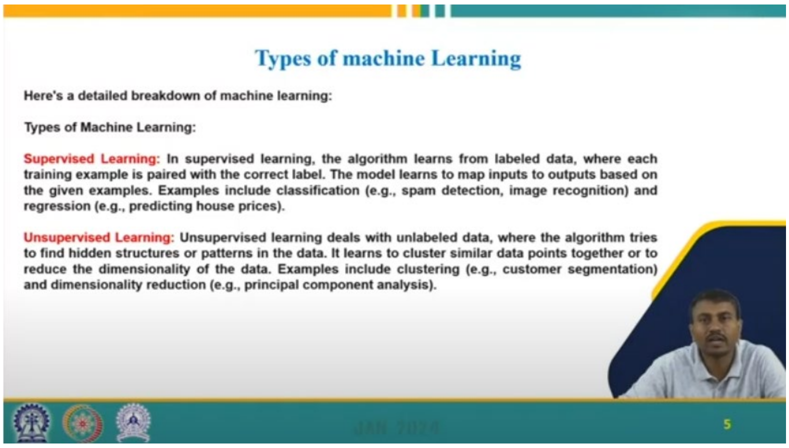
Then, we will apply this machine learning in the mining industry and its potential uses. In the second part, we will discuss the challenges associated with the application of machine learning in the mining industry. Lastly, in today's lesson, we will discuss the machine-learning algorithms for the following lessons. So, we will come across different terminologies associated with machine learning. So, we will introduce you to some of the terminologies we frequently use in the machine learning community.



So, if you look at this particular figure, you will see that it deals with the data. This is the most essential part of the machine learning model, a most vital part of the machine learning model. And data is all. This data is used now for the next level in the machine learning algorithm. And based on that learning and understanding of the data pattern, we will predict from that data, okay? Some output variables and some kinds of phenomena might be in real-world situations. Ultimately, we will predict something from the input data. Now, in the second part of this model, once you are predicting based on the input data, it is essential to assess the quality of this prediction. To determine the quality of this prediction, we are now checking whether this prediction is accurate, whether this prediction requires some amount of correction, and whether this prediction needs further training on this model. So, this is the evaluation performance of the model we are building.

So, this is the closed cycle of machine learning that is going on throughout this process. So, this is a data-driven process. Data is used to learn the model and train the model to predict some output variable or phenomena in real-world examples. Based on the algorithm, we have the mechanism to test the quality of this prediction. So, this machine learning is a subset of artificial intelligence.

So, we are focusing on the algorithm and models. These models can learn, and these algorithms are improving based on learning. So this is the first and second parts; based on this model and experience, we are predicting something in the future, which is a requirement of the day. So here we are using the computers. A sound processing power computer is required, or you can say better computational power is needed daily. So that prediction would be faster and more reliable, and we could process the data very quickly.



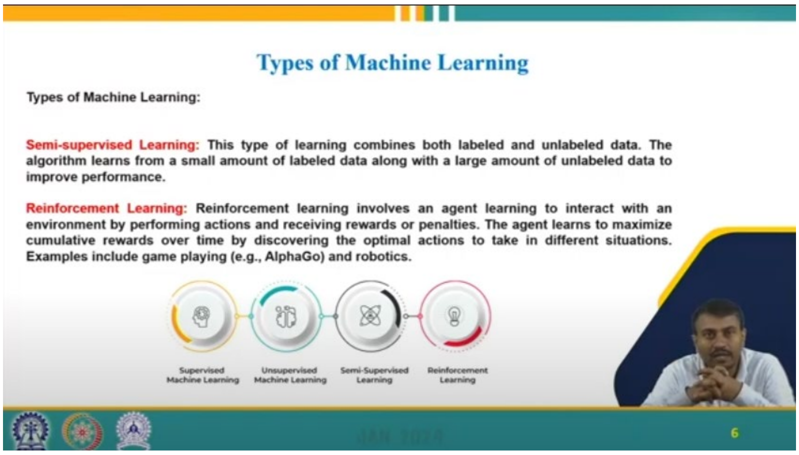
So, this is the central part of machine learning, basically how the data is a cycle, and data is the most essential part of this model. So, let us focus on the different types of machine learning that we use. Today, we will focus on four types of machine learning. The first is the supervised learning. So, from the terminology itself, you can understand that there is some active supervision in this model.

A label dataset does this supervision. A label dataset is one in which you have labeled the data and know the correct pair of outputs based on this data. So, this particular model is instrumental in regression and classification. One of the significant examples that you can see is that nowadays, you are not receiving spam emails in your email box. Because there is a spam filter, the default algorithm is there on the email provider's provider.

So, they have designed the spam filter, and based on that, they are segregating those emails to the spam folder and keeping the inbox. So, this kind of example we use daily is supervised learning. The second kind of machine learning model that we use is unsupervised learning. This unsupervised learning is used mainly for clustering or segmentation. And also for dimensionality reduction.

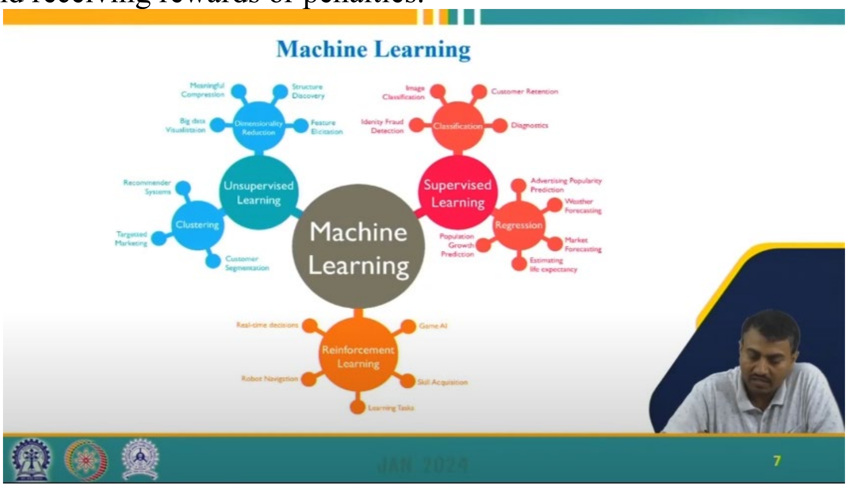
Dimensionality reduction is an essential aspect of mine automation. When we deal with automated mines, we capture different kinds of data and plan to capture a higher level of data. So you have to send that data in real-time through the network. So when the data size is large, and you frequently send or continuously send through the network, that may create a tremendous burden or pressure on the network. And by that, the network might face a failure.

So, to reduce that chance and to handle the data smartly, we sometimes need to reduce the dimension of the data. One of these popular aspects is the principal component analysis. By that, we are principally capturing the essential features and removing the non-essential features, and basically, then we send them. So this kind of example you see every day in your life, mainly if you see a different social media platform; when you send some image, the size is automatically reduced, and it is being sent to the destination to whom you are sending it. Clustering is a typical example also.



So, particularly during the lockdown time, you might have observed identifying the clusters where the corona concentration is high based on the location. We are identifying those spots and hotspots. It is a typical example of unsupervised learning. Another example is semi-supervised learning. So, in supervised learning, you have seen that data is labeled data, and there are paired outputs. In unsupervised learning, data is not labeled, and we try to find the hidden pattern in the data, map the clusters from the data, and find some meaning from the data.

The semi-supervised learning algorithm has some labeled data, a small amount, and the rest unlabeled data, and based on that, we are trying to improve the performance. So this is a kind of example of the semi-supervised learning. The fourth kind of machine learning that we are very much using nowadays is reinforcement learning. So, this reinforcement learning is mainly used in navigation and robotics. So, this model has some agents who learn by interacting with the environment and receiving rewards or penalties.

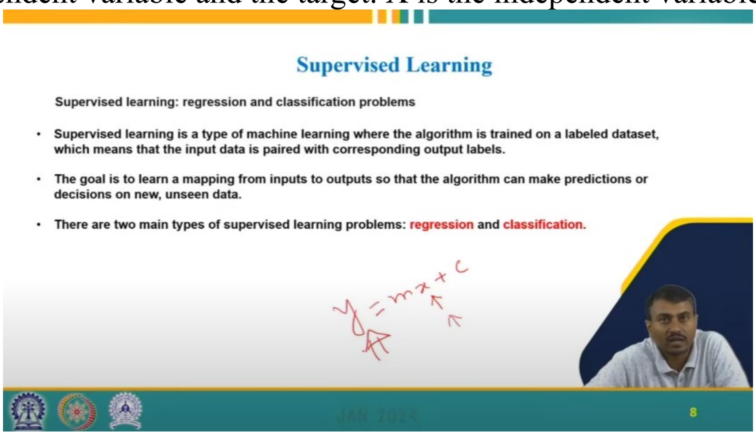


So, this agent maximizes cumulative rewards over time by discovering the optimal action to take in different situations. So these are the total picture on one page, the machine learning. Here, you can see that we have subdivided supervised learning into two parts. One is the classification, and the other is the regression. Regression is a typical example of supervised learning, particularly daily, probably on the mobile. You see the daily alert of the weather, such as that perhaps the rain is coming thunder so that it would be there.

So, this is based on the regression models we are using, which are working well. Nowadays, based on that, different kinds of models are also being developed. Based on these, we are creating the demography and migration of the population and the growth and population growth. Classification, image classification, diagnostics, and customer retention are examples of the classifications. In unsupervised learning, we have two parts again. One is dimension reduction, and the other is clustering.

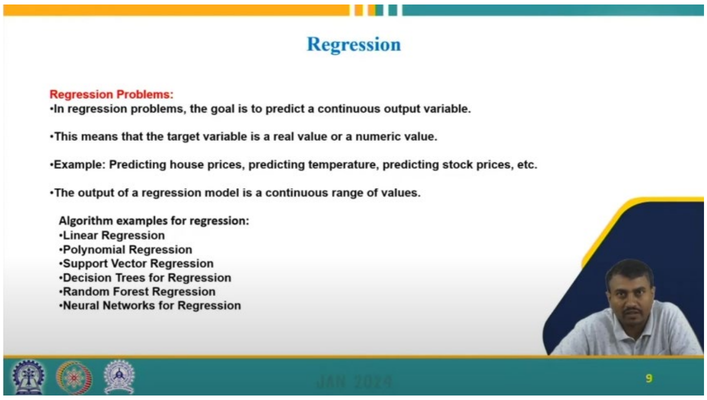
So, because we are dealing with big data, we must reduce the data's dimensions without compromising the data's essential features. Is that okay? And that is what we do in different algorithms under the dimension reduction process. Clustering is a typical example, particularly in today's world; it's all about business and marketing. So based on the customer choice and basically what type of product is coming in the market and based on that. So, this is the kind of thing a clustering algorithm deals with.

And reinforcement learning is using robot navigation and real-time decision making and also in the game applications, okay. So, supervised learning is one of the famous examples of supervised learning, which is regression. You have often encountered this expression: $y = mx + c$, okay? Here, Y is the dependent variable and the target. X is the independent variable, the input.



Based on these inputs, we predict the target variable Y, okay. So here, the goal is based on these inputs, okay? We are mapping this model to predict the Y target variable goal, okay precisely? These models should be such that they can also accommodate unforeseen situations so that some unknown, some other data set of a similar kind X can predict the target variable Y, okay? So this is based on some level data set. Okay, it's based on the level data set. Here, there is a gradient and a constant intercept, okay?

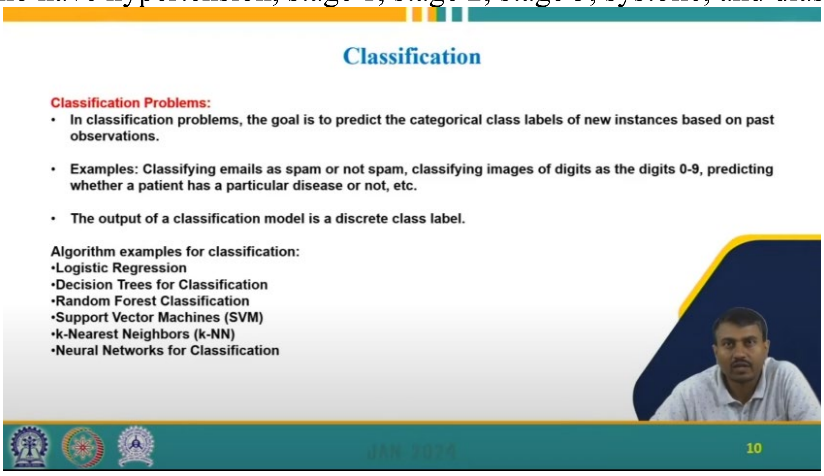
So, we have used this kind of regression model from time to time. I tried to show you this regression model. One is the linear regression. There are several variants of it, including polynomial regression. Also, in supervised learning, we have classifications.



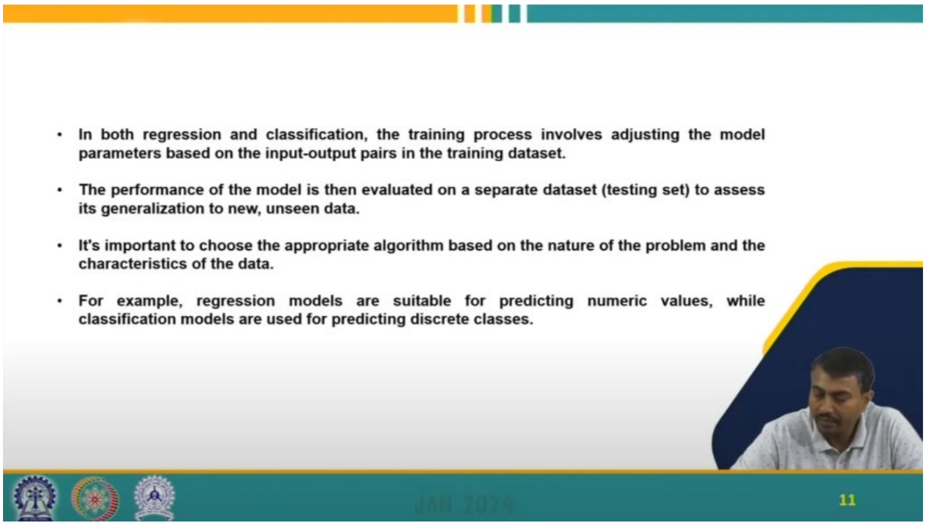
So, in regression, we try to predict the goal based on the data we find in the input variable. Okay. It might be a numerical value, or there may be some real value. One of the famous examples is the house price prediction because nowadays, based on the position, the area where the house is situated, the amenities and the transport facilities, and the quality of the building, everything is clubbed together. Finally, the house price is dependent on these variables. So, to predict those situations, you cannot use only linear regression. You have to go for the higher level of regression model.

The polynomial regression model is one of them. Other variants, such as support vector regression, decision tree for regression, random forest regression, and neural network for regressions, are also used. So, the output of a regression model is a continuous range of values, okay? Classification is also based on the level of data. We are predicting something from past observations of historical data and classifying it. Okay.

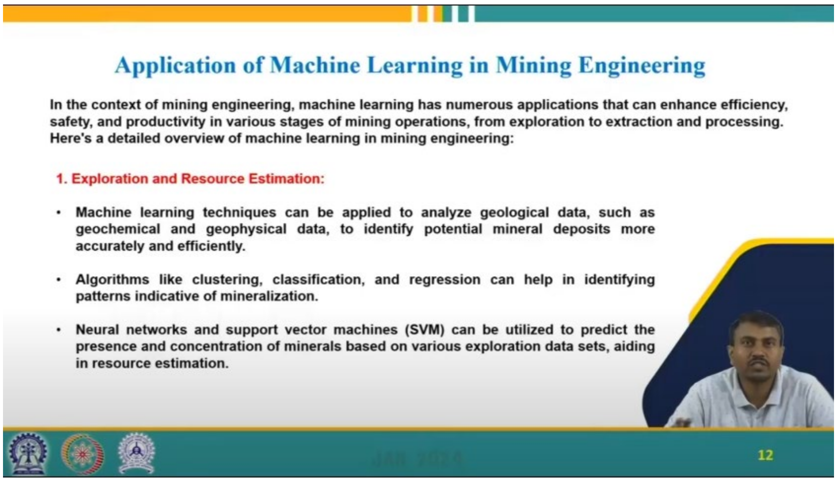
One of the popular ways of using it is designing a spam filter to diagonalize which is spam and which is not and routing that in the inbox and the spam folder. Also, classifying the image based on the digit from 0 to 9 and predicting a disease is okay. You are being grouped into this kind of disease or thing based on these specific symptoms. Or classification you might have seen in the patient, those who have hypertension, stage 1, stage 2, stage 3, systolic, and diastolic.



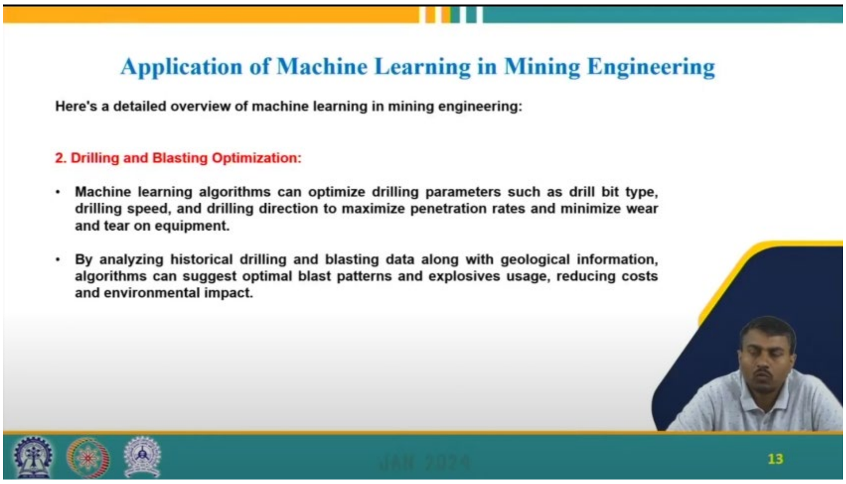
Based on that, we have the range. So, this is a kind of example of classification. So here we deal with the discrete class-level data. So here are the examples: logistic regression, decision trees for classification, random forest classification, support vector machines, k nearest neighbor, and neural network for classifications. This is an example of the algorithm used for classifications. So, in regression and classification, the training process involves adjusting the model parameters based on the input and output pairs in the training dataset. These performances are evaluated using a separate dataset to assess their predictability and generalization capability in unseen data.



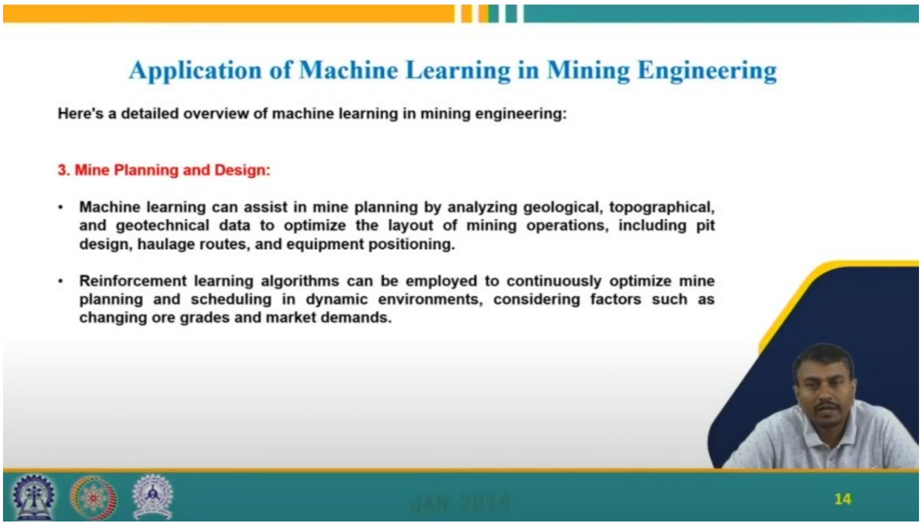
So, choosing the appropriate algorithm based on the data is essential. So, based on the data, we have to select which models we will use, okay? So, let us discuss some examples or the areas where we can use this machine learning in the mining industry, okay? One of the significant areas in which we can use machine learning is exploration and resource estimation. Mining is all about dealing with geology, nature, and exploration, and resource estimation is the mining's primary part or primary stage.



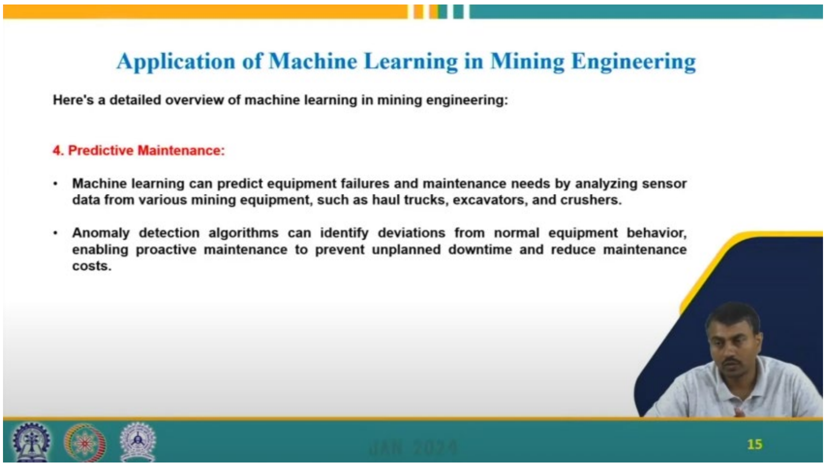
We have to find the mineral's source, which can only be done through the exploration process. We have to estimate the resource and its value and find out its feasibility for mining based on the available technology, and based on that, we will invest a large amount of money in that. So, in this whole exercise, we need to process a large amount of data; for that, we need to model and predict its sustainability and profitability. For that, we might have to use machine learning models. So here, a famous example is the neural network and support vector machines that can be utilized to predict the presence and concentration of minerals based on various explosion data sets and add to resource estimations. Another example is drilling and blasting optimization.
You know that in mining, particularly, and more particularly if you consider open pit mining, a substantial amount of the mining cost is involved in drilling and blasting. It is nearly 18 to 20% of the total cost. So, it is necessary to optimize this cost. This process has several parameters: drill bit, tie, square, penetration rate, drilling speed, drilling directions, and all that. Along with the kind of rock you are dealing with and the type of strata you are dealing with.



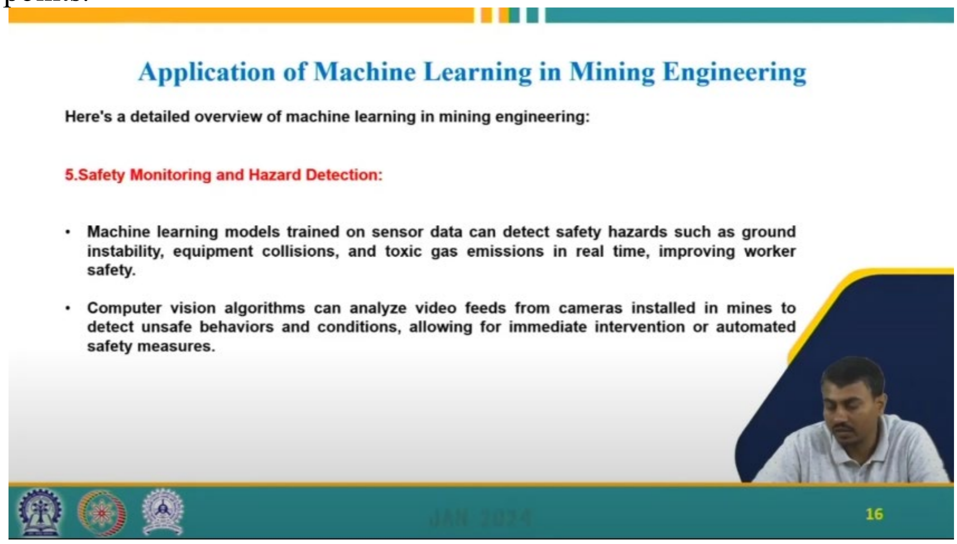
So, geology, geotechnical properties, and geological information are essential. So, based on past data, it is necessary to have a further level of optimization to reduce explosive consumption. By doing so, we can reduce the cost and environmental impact. So, it is one of the key areas we can think of using machine learning models.



Another critical area is the mine planning and design. This is an essential part of the mining industry because, nowadays, planning and design are crucial parts of the mining industry. After all, based on this planning, what kind of machinery is to be deployed, what type of extraction is to be followed, what would be the extraction rate, and everything is done in the mining industry. So, in this particular scenario, we also have to remember that in the mineral industry, particularly in metals, the market value also influences the rate of mining, mining methods, or mining processes. Finally, we are trying to make the maximum profit by extracting minerals. So, keeping all these parameters on track and considering an optimum operation and optimum implementation of the methods and the parameters, we must also apply the machine learning models here. Predictive maintenance is necessary because we utilize many types of machinery for our different operations.

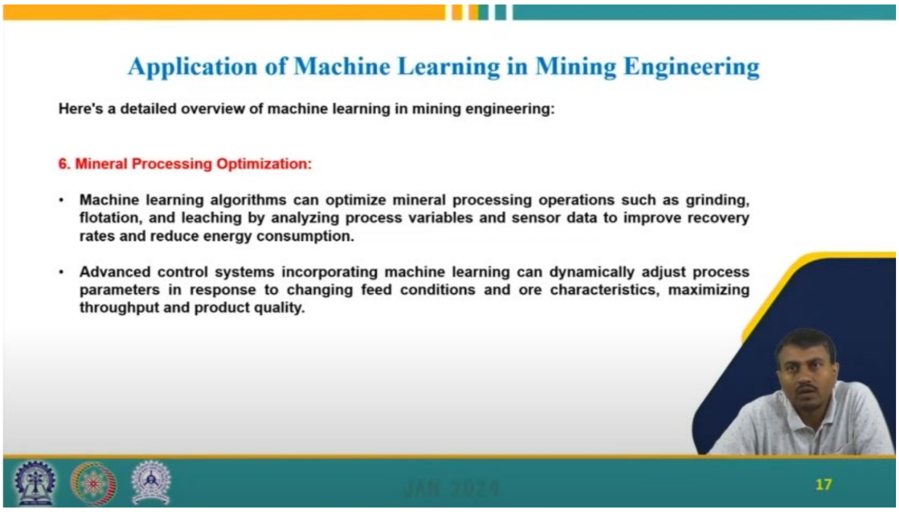


So it is a machine, a mechanical machine. So maintenance is a regular phenomenon, and sometimes, machines break down. So we need to assess what kind of maintenance we need to do or at what frequency we can do it so that the machine can perform optimally. That is a loss because we cannot allow the machine to break down for a reasonable amount of time. So, based on the machine behaviors and different kinds of performance we are getting, we need to build a model for every machine we use in the mining to utilize it in its maximum way. This is the predictive maintenance we are looking at so that all machines can be used thoroughly and to their maximum points.

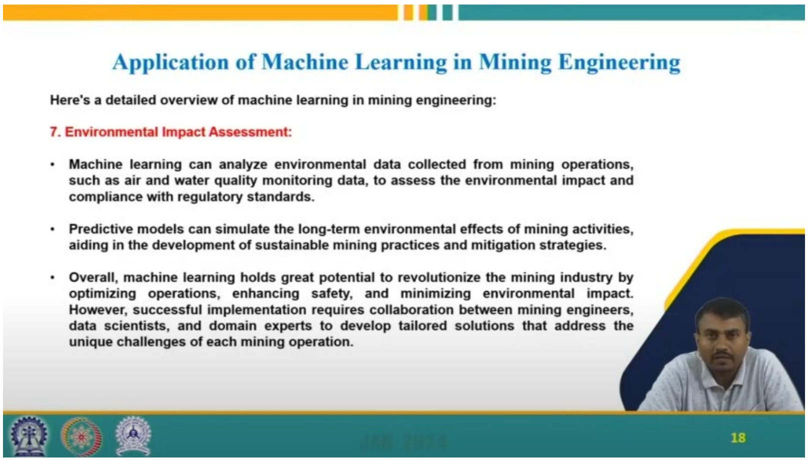


By doing so, we are reducing the maintenance cost and utilizing the machine at the maximum level. Safety monitoring and hazard detection. It is essential because worker safety and the safety of the mines are paramount. Different sensors are fitted in the mines, machines, and other mine areas based on specific parameters. So we can avoid the collisions of the equipment.

We can avoid the roof fall based on the data by pattern. Based on the sagging data or roof displacement data, we can predict what kind of failure might occur and when the roof fall might occur. Similarly, we can expect the emission level in the open-pin mine slope stability issues or the coal mines, particularly the emission of toxic gas based on specific data. So when will it reach the climax or the peak? So based on that, we also need to change the ventilation patterns. So for all these, we need to apply machine learning models so that these models can accurately predict and we can reap the benefit of it for getting a better safety notion and better safety features in the mine.

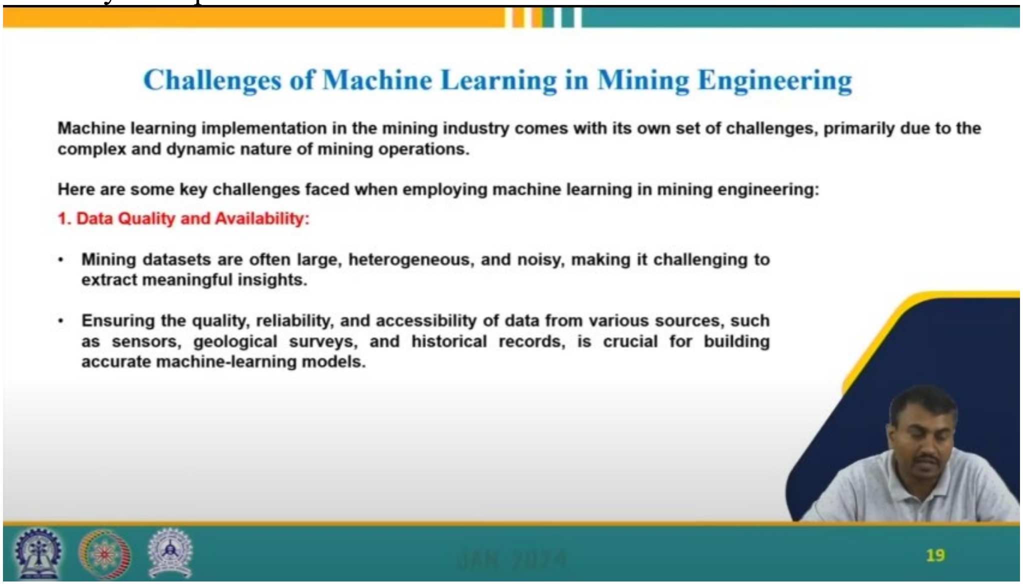


Mineral processing optimization. The processing part is also an essential part of the mining, and it is if we increase the extraction, okay, or row recovery if we can maximize that. So by that we are going to reduce the cost and also we are reducing the energy consumption. So, this is a dynamic process. Based on the different ore characteristics, what kind of processing is to be followed, and what kind of parameters are to be adjusted so that this process can dynamically adjust based on the distinct change in the feeds? Okay. So, this is a critical consideration in the mining industry. Here, we have an excellent potential to use a machine learning model to optimize this process further.

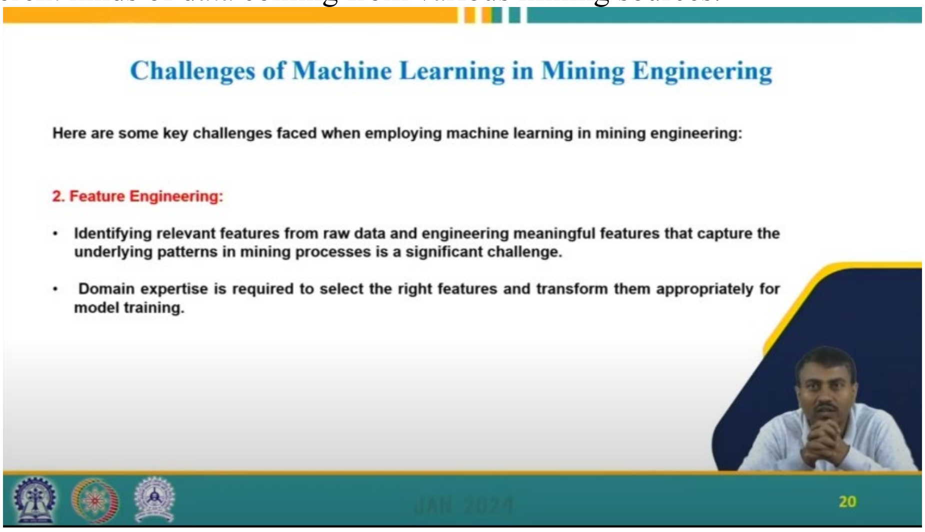


Environmental impact assessment. Nowadays, it is a very critical and sensitive issue because mining affects the environment. So it is the responsibility of the mining industry and the stakeholders of the mining industry to operate the mine sensibly so that we can reduce this impact, okay? And we should do something based on this impact to minimize it, okay? So for that, we required long-term planning of what kind of impact this mining will have on the environment. So, for that, we need to predict what kind of long-term impact this environment for this mining parameter will have on the environment.

So for that kind of situation, a machine learning kind of model would be better, and nowadays, you can see that different science organizations, particularly the association of other scientists, are predicting how much this impact will be on the global climate so that we are going to face that. So, they are issuing alerts and alarms to the different governments globally so that we can reduce this impact. Mainly, I am talking about climate change. So mining is also a potential area where we can use this model and develop some solution based on this input from the model so that we can do something in the mining itself, during the mining itself, so that this kind of effect is minimized. This requires collaboration with academia, data scientists, environmental scientists, people, and different domain experts so that we can face this kind of challenge and do what is necessary to improve our environment.

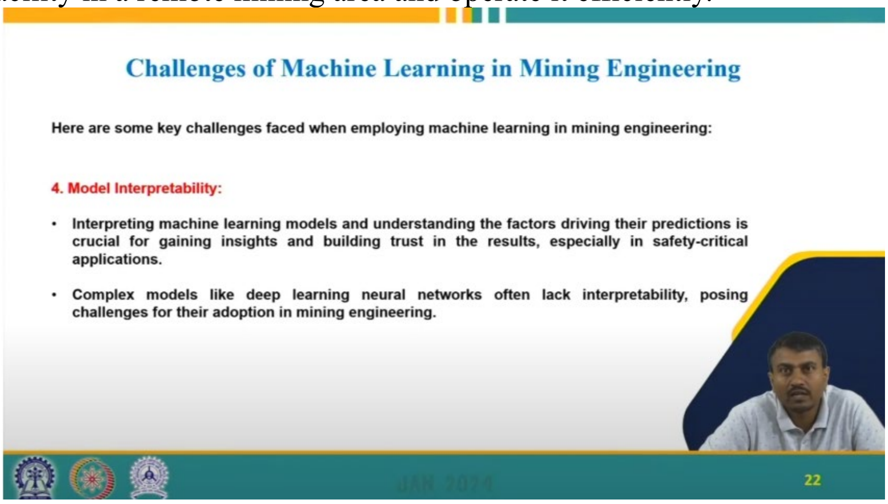


So now, we will focus on some of the challenges of applying these machine learning models in the mining industry. One of the significant challenges we will face in the mining industry is the data quality and availability. When we collect data, this data is noisy and heterogeneous and is a massive amount of data. Sometimes, it is tough to extract some meaningful sense from this data. These data may be geological type data, geological survey data, ventilation data, environmental data, or different kinds of data coming from various mining sources.

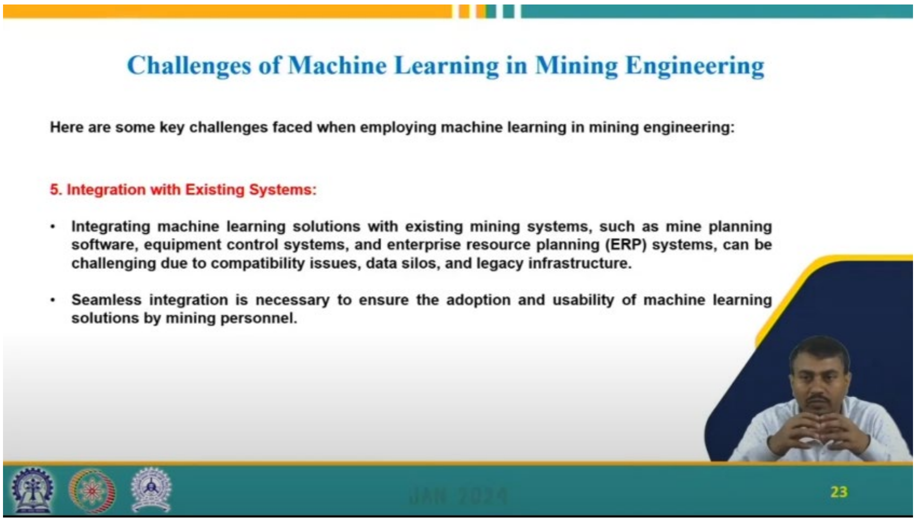


So, ensuring the quality of the data is very difficult in the mining scenarios. So, we have to face the challenge of what kind of models we will use to train this model to use this data meaningfully. The second part is feature engineering. So this requires a domain expert in the mining who can assess what parameter is critical from this particular raw data. Not only will the data scientist be able to say this, but we require good collaboration with the domain expert, mining engineers, and those dealing in the field who have ideas about different features in the data.

Meaningful convergence is required to use this data to extract the specific features from the data, and it is a challenge: scalability and computational resources. Mining operations generate a large amount of data in real time, and a scalable machine-learning algorithm and computational resources are required to process and analyze the data efficiently. This is also a considerable investment in terms of its operation and implementation. So, it is also a challenge to establish that kind of facility in a remote mining area and operate it efficiently.



It's a challenge. Model interpretability is also a challenge because there are different kinds of applications and other types of insights from the data. Then, we have to trust the models, and based on the prediction, we have to see in a real-world scenario that it matches that. So, the complex model may be required to use this data meaningfully and predict a good solution. So, here is the role of mining engineers, data scientists, academia, and researchers together: we must think because mining is all about dealing with complex situations.

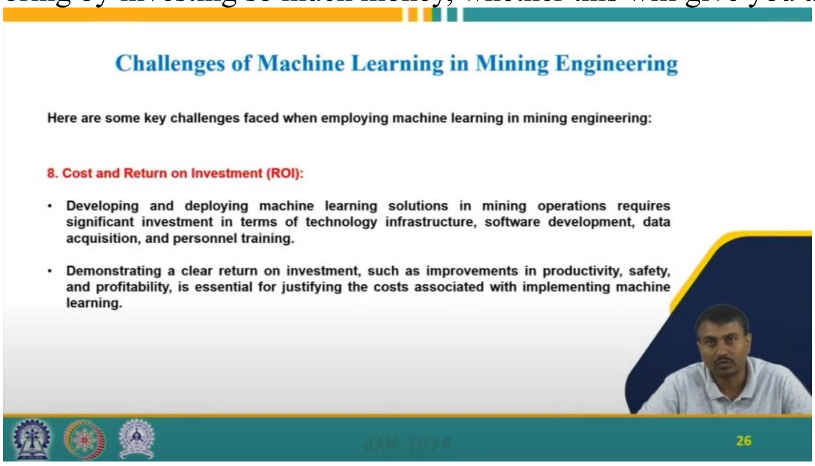


So, a complex kind of situation means a complex type of data, and based on that, you are building a model that will predict well that requires a very complex model, which is very difficult for the engineers to build. Integration with the existing system is essential because, in one go, you will not abolish the existing system. You have to integrate with the existing system. You have to fit well within the scope. And for that, you have to face the challenge of whether this integration can be made successful and how you can make this system seamless.

So, data can be transferred seamlessly from one operation to the other, and we can adopt the machine learning model. That is relatively challenging because, in mining itself, we use some kinds of software in some of the machinery we use. We are following some systems. So we have to comply with all these things and finally integrate all these data on one platform so that you can use it meaningfully.

That is a big challenge. The next part is regulatory compliance and stakeholder acceptance. Whatever you are doing in mining, you must comply with the prevalent regulations in that particular country's conditions. The stakeholders' approval for applying that change in the existing model is also needed. So, addressing all these concerns requires a good amount of engagement from all these stakeholders and the agencies dealing with this kind of operation.

Skill gap and training. That is, we have to fulfil and address this particular issue. And that is a requirement of the data, and all these stakeholders should come together and address this issue so that all the miners or the mining people, those working in the industry, should be trained; they should be imparted a good amount of training. So that their quality is up and they are valued members of this community. They can utilize this new technology better. Cost and return on investment. It is a critical aspect, finally. The return on investment is the crucial factor in the change you will bring by investing so much money, whether this will give you a good return.



And all these stakeholders will give you the target. So, you have to perform, and you have to show that, yes, these are the ways that we are going to perform, and finally, we will achieve this amount of return at the end of the day. So, it is also essential to prove that we are necessarily bringing more returns with this kind of change. So, addressing these challenges requires collaboration between mining engineers, data scientists, technology providers, and other stakeholders to develop tailored solutions that address each mining operation's specific needs and constants. So let us see some of the terminologies that we will often come across in machine learning, and we need to introduce these terminologies to you so that frequently, whenever we are using these terminologies, you can understand, okay, this terminology means this.

One is the feature. Features are the input variables or attributes used to make predictions in a machine-learning model. So, features can be numerical, age, salary or categorical, gender, city, etc. Second is the level. Level of the correct output corresponding to the input data in supervised learning. In classification, task levels represent the categories or classes to which input data belongs.

So, in regression, task levels are continuous values to be predicted—training data. Training data is the data set used to train the machine learning model, and it consists of input-output pairs in supervised learning or just input data in unsupervised learning. Validation data. Validation data is a separate data set used to fine-tune the hyperparameters and evaluate the model's performance during training. So, it helps prevent overfitting by providing an independent data set for testing.

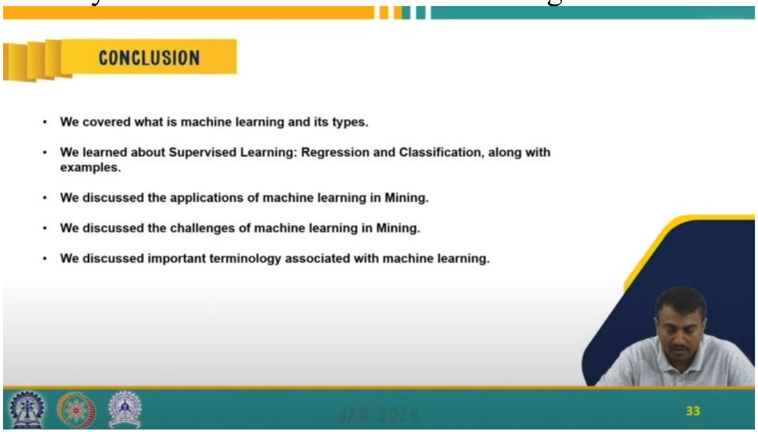
Testing data. Testing data is a separate data set used to evaluate the trained model's final performance and assess how well the model generalizes to unseen data. Model. A model is a mathematical representation of a real-world process or system learned from data, and it captures patterns and relationships between input features and output levels. Algorithm. An algorithm is a set of rules or procedures followed to solve a problem, and in machine learning, algorithms are used to train models based on the provided data.

Examples include decision trees, neural networks, and support vector machines—hyperparameters. Hyperparameters are set before the training begins, and the learning process is controlled. Examples include the learning rate in gradient descent or the number of hidden layers used in the neural networks—loss function. A loss function measures the difference between the predicted value and the actual value in the training data, and it quantifies the model performance during training and guides the optimization process.

Optimization. Optimization refers to the process of adjusting model parameters to minimize loss function. This process involves techniques like gradient descent and iteratively updates the model performance to improve the model. Overfitting. It is a kind of challenge and a problem. So, overfitting occurs when the model learns to capture noise or violent patterns. So, we need to take some algorithm that drops out L1, and L2 regulation is used to prevent overfitting.

So, this overfitting indicates that the model performance is poor, and the model cannot get a proper picture of the complexity in the model or the data—similarly underfitting. So, underfitting happens when the model is too simplistic to capture the underlying pattern in the data, leading to poor performance. Cross-validation.

Cross-validation is a technique used to assess the performance of a machine-learning model. It involves splitting the data into multiple subsets, training the model on different subsets, and evaluating its performance with the remaining data—bias-variance trade-off. The bias-variance trade-off refers to the balance between the model's ability to capture the underlying pattern in the bias data and its sensitivity to variation in the variance training data. So these are the references.



So, let me conclude in a few sentences what we have covered in this lecture. We have introduced machine learning and what machine learning is and learned the two different aspects or models of supervised learning, regression, and classifications. We have discussed the application of machine learning in the mining industry and its challenges, and we have introduced some of the fundamental terminologies we use in machine learning. Thank you.