Introduction to machine learning: Machine learning is a tool for turning information into knowledge. Machine learning techniques are used to automatically find the valuable underlying patterns within complex data that we would otherwise struggle to discover. The hidden patterns and knowledge about a problem can be used to predict future events and perform all kinds of complex decision making.

A subset of artificial intelligence known as machine learning focuses primarily on the creation of algorithms that enable a computer to independently learn from data and previous experiences.

machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things.

Need for Machine Learning

The demand for machine learning is steadily rising. Because it is able to perform tasks that are too complex for a person to directly implement, machine learning is required. Humans are constrained by our inability to manually access vast amounts of data; as a result, we require computer systems, which is where machine learning comes in to simplify our lives.

By providing them with a large amount of data and allowing them to automatically explore the data, build models, and predict the required output, we can train machine learning algorithms. The cost function can be used to determine the amount of data and the machine learning algorithm's performance. We can save both time and money by using machine learning.

- Solving complex problems, which are difficult for a human
- Decision making in various sector including finance
- Finding hidden patterns and extracting useful information from data

Important Terms of Machine Learning.

• Algorithm: A Machine Learning algorithm is a set of rules and statistical techniques used to learn patterns from data and draw significant information from it. It is the logic behind a Machine Learning model. An example of a Machine Learning algorithm is the Linear Regression algorithm.

• Model: A model is the main component of Machine Learning. A model is trained by using a Machine Learning Algorithm. An algorithm maps all the decisions that a model is supposed to take in order to get the correct output based on the given input

- **Model:** A model is the main component of Machine Learning. A model is trained by using a Machine Learning Algorithm. An algorithm maps all the decisions that a model is supposed to take based on the given input, in order to get the correct output.
- **Predictor Variable:** It is a feature(s) of the data that can be used to predict the output.
- **Response Variable:** It is the feature or the output variable that needs to be predicted by using the predictor variable(s).
- **Training Data:** The Machine Learning model is built using the training data. The training data helps the model to identify key trends and patterns essential to predict the output.
- **Testing Data:** After the model is trained, it must be tested to evaluate how accurately it can predict an outcome. This is done by the testing data set.

Classification of Machine Learning

At a broad level, machine learning can be classified into three types:

1. **Supervised learning**
2. **Unsupervised learning**
3. **Reinforcement learning**

1) Supervised Learning

In supervised learning, sample labeled data are provided to the machine learning system for training, and the system then predicts the output based on the training data.

Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables. More specifically, Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed. It predicts continuous/real values such as **temperature, age, salary, price,** etc.

Regression Regression models are used to predict a continuous value. Predicting prices of a house given the features of house like size, price etc. is one of the common examples of Regression. It is a supervised technique

3) Reinforcement Learning

Reinforcement learning is a feedback-based learning method, in which a learning agent gets a reward for each right action and gets a penalty for each wrong action. The agent learns automatically with these feedbacks and improves its performance. In reinforcement learning, the agent interacts with the environment and explores it. The goal of an agent is to get the most reward points, and hence, it improves its performance.

The robotic dog, which automatically learns the movement of his arms, is an example of Reinforcement learning.

Application of machine learning:

1. Image Recognition:

Image recognition is one of the most common applications of machine learning. It is used to identify objects, persons, places, digital images, etc. The popular use case of image recognition and face detection is, **Automatic friend tagging suggestion**:

Facebook provides us a feature of auto friend tagging suggestion. Whenever we upload a photo with our Facebook friends, then we automatically get a tagging suggestion with name, and the technology behind this is machine learning's **face detection** and **recognition algorithm**.

2. Speech Recognition

While using Google, we get an option of "**Search by voice**," it comes under speech recognition, and it's a popular application of machine learning.

Speech recognition is a process of converting voice instructions into text, and it is also known as "**Speech to text**", or "**Computer speech recognition**." At present, machine learning algorithms are widely used by various applications of speech recognition. **Google assistant**, **Siri**, **Cortana**, and **Alexa** are using speech recognition technology to follow the voice instructions.

3. Traffic prediction:

If we want to visit a new place, we take help of Google Maps, which shows us the correct path with the shortest route and predicts the traffic conditions.

It predicts the traffic conditions such as whether traffic is cleared, slow-moving, or heavily congested with the help of machine learning

4. Product recommendations:

Machine learning is widely used by various e-commerce and entertainment companies such as **Amazon**, **Netflix**, etc., for product recommendation to the user. Whenever we search for some product on Amazon, then we started getting an advertisement for the same product while internet surfing on the same browser and this is because of machine learning.

Google understands the user interest using various machine learning algorithms and suggests the product as per customer interest.

5. Self-driving cars:

One of the most exciting applications of machine learning is self-driving cars. Machine learning plays a significant role in self-driving cars. Tesla, the most popular car manufacturing company is working on self-driving car. It is using unsupervised learning method to train the car models to detect people and objects while driving

CHALLENGES IN MACHINE LEARNING:

1. Poor Quality of Data

2. Underfitting of Training Data

This process occurs when data is unable to establish an accurate relationship between input and output variables. It simply means trying to fit in undersized jeans. It signifies the data is too simple to establish a precise relationship. To overcome this issue:

- *Maximize the training time*
- *Enhance the complexity of the model*
- *Add more features to the data*
- *Reduce regular parameters*
- *Increasing the training time of model*

3. Overfitting of Training Data

Overfitting refers to a machine learning model trained with a massive amount of data that negatively affect its performance. It is like trying to fit in Oversized jeans. Unfortunately,

this is one of the significant issues faced by machine learning professionals. This means that the algorithm is trained with noisy and biased data, which will affect its overall performance.

4. Machine Learning is a Complex Process

The machine learning industry is young and is continuously changing. Rapid hit and trial experiments are being carried on. The process is transforming, and hence there are high chances of error which makes the learning complex. It includes analyzing the data, removing data bias, training data, applying complex mathematical calculations, and a lot more. Hence it is a really complicated process which is another big challenge for Machine learning professionals.

6. Slow Implementation
This is one of the common issues faced by machine learning professionals. The machine learning models are highly efficient in providing accurate results, but it takes a tremendous amount of time. Slow programs, data overload, and excessive requirements usually take a lot of time to provide accurate results. Further, it requires constant monitoring and maintenance to deliver the best output.

5. Lack of Training Data

The most important task you need to do in the machine learning process is to train the data to achieve an accurate output. Less amount training data will produce inaccurate or too biased predictions. Let us understand this with the help of an example. Consider a machine learning algorithm similar to training a child. One day you decided to explain to a child how to distinguish between an apple and a watermelon. You will take an apple and a watermelon and show him the difference between both based on their color, shape, and taste. In this way, soon, he will attain perfection in differentiating between the two. But on the other hand, a machine-learning algorithm needs a lot of data to distinguish. For complex problems, it may even require millions of data to be trained. Therefore we need to ensure that Machin

- **Underfitting and Overfitting:** If our algorithm works well with the training dataset but not well with test dataset, then such problem is called **Overfitting**. And if our algorithm does not perform well even with training dataset, then such problem is called **underfitting**.

Feature Selection Techniques in Machine Learning

Feature selection is a way of selecting the subset of the most relevant features from the original features set by removing the redundant, irrelevant, or noisy features.

While developing the machine learning model, only a few variables in the dataset are useful for building the model, and the rest features are either redundant or irrelevant. If we input the dataset with all these redundant and irrelevant features, it may negatively impact and reduce the overall performance and accuracy of the model. Hence it is very important to identify and select the most appropriate features from the data and remove the irrelevant or less important features, which is done with the help of feature selection in machine learning.

- **It helps in avoiding the curse of dimensionality.**

- **It helps in the simplification of the model so that it can be easily interpreted by the researchers.**
- **It reduces the training time.**
- **It reduces overfitting hence enhance the generalization.**

Chi-square Test: Chi-square test is a technique to determine the relationship between the categorical variables. The chi-square value is calculated between each feature and the target variable, and the desired number of features with the best chi-square value is selected.

Feature Selection Techniques

There are mainly two types of Feature Selection techniques, which are:

- **Supervised Feature Selection technique**

  Supervised Feature selection techniques consider the target variable and can be used for the labelled dataset.
- **Unsupervised Feature Selection technique**

  Unsupervised Feature selection techniques ignore the target variable and can be used for the unlabelled dataset.

CURSE OF DIMENSIONALITY: The Curse of Dimensionality refers to the various challenges and complications that arise when analyzing and organizing data in high-dimensional spaces (often hundreds or thousands of dimensions). In the realm of machine learning, it's crucial to understand this concept because as the number of features or dimensions in a dataset increases, the amount of data we need to generalize accurately grows exponentially.

RIDGE AND LASSO REGRESSION IN MACHINE LEARNING

ChatGPT

Ridge and Lasso regression are two commonly used techniques in machine learning for regularization, which is a method used to prevent overfitting and improve the performance of regression models.

Ridge Regression:

- **Objective**: Ridge regression aims to minimize the sum of squared errors (like ordinary least squares regression) while adding a penalty term to the coefficients of the features to prevent them from becoming too large.

Lasso Regression:

- **Objective**: Lasso (Least Absolute Shrinkage and Selection Operator) regression, like Ridge, aims to minimize the sum of squared errors but adds a penalty term that encourages some coefficients to become exactly zero.

Lasso Regression Vs Ridge Regression	
Lasso Regression uses L1 regularization (absolute value of coefficients).	Ridge Regression uses L2 regularization (square of coefficients).
Lasso Regression can force them to be exactly zero.	Ridge Regression shrinks coefficients of less significant features towards zero.
Lasso Regression performs both regularization and feature selection, making it more suitable for high-dimensional datasets.	Ridge Regression does not perform feature selection and can only shrink the coefficient values. This makes it more suitable for datasets with highly correlated predictors since it avoids including all of them in the model.
Lasso Regression may be more effective in situations where only a subset of features contribute significantly to the output	Ridge Regression generally works better in scenarios where there are fewer significant features.
Lasso Regression can lead to a sparse model, which means it can create a model with fewer features.	Ridge Regression does not produce sparse models.

Lasso regression can produce many solutions to the same problem.	Ridge regression can only produce one solution to one problem.
Lasso regression usually produces unstable solutions which means the regression line may be uneven and jump by a large amount.	Ridge Regression produces stable solutions which means that even for a small adjustment to a data point, the regression line will not move too much.
Lasso regression has built in feature selection.	Ridge regression does not have built in feature selection
Lasso regression produces sparse outputs	Ridge regression produces non sparse outputs
For Lasso regression , the computation is inefficient for non-sparse cases.	For ridge regression , the computation is efficient because due to analytical solutions.

What is Bias?

Bias is simply defined as the inability of the model because of that there is some difference or error occurring between the model's predicted value and the actual value. These differences between actual or expected values and the predicted values are known as error or bias error or error due to bias. Bias is a systematic error that occurs due to wrong assumptions in the machine learning process.

What is Variance?

Variance is the measure of spread in data from its mean position. In machine learning variance is the amount by which the performance of a predictive model changes when it is trained on different subsets of the training data. Variance errors are either low or high-variance errors.

- **Low variance:** Low variance means that the model is less sensitive to changes in the training data and can produce consistent estimates of the target function with different subsets of data from

the same distribution. This is the case of underfitting when the model fails to generalize on both training and test data.
- **High variance:** High variance means that the model is very sensitive to changes in the training data and can result in significant changes in the estimate of the target function when trained on different subsets of data from the same distribution. This is the case of overfitting when the model performs well on the training data but poorly on new, unseen test data. It fits the training data too closely that it fails on the new training dataset.

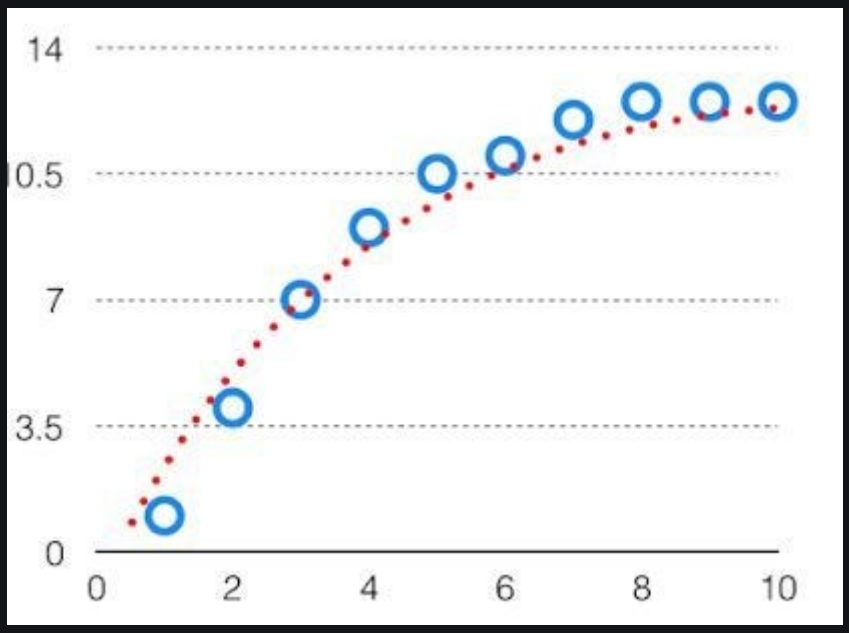
Ways to Reduce the reduce Variance in Machine Learning:

- **Cross-validation:** By splitting the data into training and testing sets multiple times, cross-validation can help identify if a model is overfitting or underfitting and can be used to tune hyperparameters to reduce variance.
- **Feature selection:** By choosing the only relevant feature will decrease the model's complexity. and it can reduce the variance error.
- **Regularization:** We can use L1 or L2 regularization to reduce variance in machine learning models
- **Ensemble methods:** It will combine multiple models to improve generalization performance. Bagging, boosting, and stacking are common ensemble methods that can help reduce variance and improve generalization performance.
- **Simplifying the model:** Reducing the complexity of the model, such as decreasing the number of parameters or layers in a neural network, can also help reduce variance and improve generalization performance.
- **Early stopping:** Early stopping is a technique used to prevent overfitting by stopping the training of the deep learning model when the performance on the validation set stops improving.

Bias Variance Tradeoff

- If the algorithm is too simple (hypothesis with linear equation) then it may be on high bias and low variance condition and thus is error-prone. If algorithms fit too complex (hypothesis with high degree equation) then it may be on high variance and low bias. In the latter condition, the new entries will not perform well. Well, there is

something between both of these conditions, known as a Trade-off or Bias Variance Trade-off. This tradeoff in complexity is why there is a tradeoff between bias and variance. An algorithm can't be more complex and less complex at the same time. For the graph, the perfect tradeoff will be like this.



- We try to optimize the value of the total error for the model by using the Bias-Variance Tradeoff.

- **true Positive(TP):** In this case, the prediction outcome is true, and it is true in reality, also.
- True Negative(TN): in this case, the prediction outcome is false, and it is false in reality, also.
- False Positive(FP): In this case, prediction outcomes are true, but they are false in actuality.
- False Negative(FN): In this case, predictions are false, and they are true in actuality.

II. Confusion Matrix

A confusion matrix is a tabular representation of prediction outcomes of any binary classifier, which is used to describe the performance of the classification model on a set of test data when true values are known.

The confusion matrix is simple to implement, but the terminologies used in this matrix might be confusing for beginners.

n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

III. Precision

The precision metric is used to overcome the limitation of Accuracy. The precision determines the proportion of positive prediction that was actually correct. It can be calculated as the True Positive or predictions that are actually true to the total positive predictions (True Positive and False Positive).

$$Precision = \frac{TP}{(TP + FP)}$$

F-Scores

F-score or F1 Score is a metric to evaluate a binary classification model on the basis of predictions that are made for the positive class. It is calculated with the help of Precision and Recall. It is a type of single score that represents both Precision and Recall. So, ***the F1 Score can be calculated as the harmonic mean of both precision and Recall, assigning equal weight to each of them.***

The formula for calculating the F1 score is given below:

$$F1 - score = 2 * \frac{precision * recall}{precision + recall}$$

Confusion Matrix

It is the easiest way to measure the performance of a classification problem where the output can be of two or more type of classes. A confusion matrix is nothing but a table with two

dimensions viz. “Actual” and “Predicted” and furthermore, both the dimensions have “True Positives (TP)”, “True Negatives (TN)”, “False Positives (FP)”, “False Negatives (FN)” as shown below −

		Actual: 1	Actual: 0
Predicted	1	True Positives (TP)	False Positives (FP)
	0		True Negatives (TN)

Explanation of the terms associated with confusion matrix are as follows −

- **True Positives (TP)** − It is the case when both actual class & predicted class of data point is 1.
- **True Negatives (TN)** − It is the case when both actual class & predicted class of data point is 0.
- **False Positives (FP)** − It is the case when actual class of data point is 0 & predicted class of data point is 1.
- **False Negatives (FN)** − It is the case when actual class of data point is 1 & predicted class of data point is 0.

We can use confusion_matrix function of sklearn.metrics to compute Confusion Matrix of our classification model.

Classification Accuracy

It is most common performance metric for classification algorithms. It may be defined as the number of correct predictions made as a ratio of all predictions made. We can easily calculate it by confusion matrix with the help of following formula −

$$Accuracy=TP+TNTP+FP+FN+TN$$

We can use accuracy_score function of sklearn.metrics to compute accuracy of our classification model.

Classification Report

This report consists of the scores of Precisions, Recall, F1 and Support. They are explained as follows −

Precision

Precision, used in document retrievals, may be defined as the number of correct documents returned by our ML model. We can easily calculate it by confusion matrix with the help of following formula −

$$Precision=TPTP+FP$$

Recall or Sensitivity

Recall may be defined as the number of positives returned by our ML model. We can easily calculate it by confusion matrix with the help of following formula −

$$Recall=TPTP+FN$$

Specificity

Specificity, in contrast to recall, may be defined as the number of negatives returned by our ML model. We can easily calculate it by confusion matrix with the help of following formula −

$$Specificity=TNTN+FP$$

Support

Support may be defined as the number of samples of the true response that lies in each class of target values.

F1 Score

This score will give us the harmonic mean of precision and recall. Mathematically, F1 score is the weighted average of the precision and recall. The best value of F1 would be 1 and worst would be 0. We can calculate F1 score with the help of following formula −

$$F1 = 2 * (precision * recall) / (precision + recall)$$

F1 score is having equal relative contribution of precision and recall.

We can use classification_report function of sklearn.metrics to get the classification report of our classification model.

Linear Regression

Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features.

Logistic Regression

Logistic regression is a supervised machine learning algorithm mainly used for classification tasks where the goal is to predict the probability that an instance of belonging to a given class. It is used for classification algorithms its name is logistic regression. it's referred to as regression because it takes the output of the linear regression function as input and uses a sigmoid function to estimate the probability for the given class. The difference between linear regression and logistic regression is that linear regression output is the continuous value that can be anything while logistic regression predicts the probability that an instance belongs to a given class or not.

- Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value.
- It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.
- Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.

Linear Regresssion	Logistic Regression	
1	Linear regression is used to predict the continuous dependent variable using a given set of independent variables.	Logistic regression is used to predict the categorical dependent variable using a given set of independent variables.
2	Linear regression is used for solving Regression problem.	It is used for solving classification problems.
3	In this we predict the value of continuous variables	In this we predict values of categorical varibles
4	In this we find best fit line.	In this we find S-Curve .
5	Least square estimation method is used for estimation of accuracy.	Maximum likelihood estimation method is used for Estimation of accuracy.
6	The output must be continuous value,such as price,age,etc.	Output is must be categorical value such as 0 or 1, Yes or no, etc.
7	It required linear relationship between dependent and independent variables.	It not required linear relationship.
8	There may be collinearity between the independent variables.	There should not be collinearity between independent varible.

More specific examples

- **Area Map:** A form of geospatial visualization, area maps are used to show specific values set over a map of a country, state, county, or any other geographic location. Two common types of area maps are choropleths and isopleths. Learn more.
- **Bar Chart:** Bar charts represent numerical values compared to each other. The length of the bar represents the value of each variable. Learn more.
- **Box-and-whisker Plots:** These show a selection of ranges (the box) across a set measure (the bar). Learn more.
- **Bullet Graph:** A bar marked against a background to show progress or performance against a goal, denoted by a line on the graph. Learn more.
- **Gantt Chart:** Typically used in project management, Gantt charts are a bar chart depiction of timelines and tasks. Learn more.
- **Heat Map:** A type of geospatial visualization in map form which displays specific data values as different colors (this doesn't need to be temperatures, but that is a common use). Learn more.
- **Highlight Table:** A form of table that uses color to categorize similar data, allowing the viewer to read it more easily and intuitively. Learn more.
- **Histogram:** A type of bar chart that split a continuous measure into different bins to help analyze the distribution. Learn more.
- **Pie Chart:** A circular chart with triangular segments that shows data as a percentage of a whole. Learn more.
- **Treemap:** A type of chart that shows different, related values in the form of rectangles nested together. Learn more.