

---

# Machine Learning Process, Associated Areas, Managing Data in ML Models

---

## HIGHLIGHTS

- Need and justification for using Machine Learning
- Machine Learning (ML) Process and Applications
- Machine Learning Problem Solving Steps
- Machine Learning and Associated Areas
- Subsets of Machine Learning
- Machine Learning and Data Mining
- Other Associated Areas of Machine Learning
- A brief overview of Statistical Tools for Machine Learning
- General Statistical Skills
- Machine Learning (ML) and the types and role of data in ML Process
- Data for ML models and Managing Data throughout the ML Process

Training Data Set

Validation Data Set

Hyperparameters

Test Data Set

- Determining the Size of Data Sets
- Machine Learning Problem Solving Steps

## Summary

---

## Need and Justification for the use of Machine Learning Approach

A machine learning approach is used where no fully satisfactory algorithm is available. In cases where no definite solutions exist, or vast numbers of potential answers exist, it is customary to

label some of the correct answers as valid. To improve the algorithm (s) to achieve the best possible result, the algorithm is fed data commonly known as *training data* that the computer uses to improve the algorithm(s) to find correct answers. For example, to train a system for the task of digital character recognition, the [MNIST](#) dataset of handwritten digits has often been used.<sup>[11]</sup>

Machine learning works based on predictive and statistical algorithms that are provided to these machines. The algorithms are designed to learn and improve as more data flows through the system. Fraud detection, e-mail spam, GPS systems are some examples of machine learning applications.

Machine learning applications are based on teaching a computer how to learn. Besides predictive and statistical algorithms, it also uses pattern recognition to draw conclusions from the data.

Machine learning methods are used to develop complex models and algorithms that help to understand and make predictions or find patterns in the data. The analytical models in machine learning allow the analysts to make predictions by learning from the trends, patterns, and relationships in the historical data. Machine learning automates model building. The algorithms in machine learning are designed to learn iteratively from data without being programmed.

According to Arthur Samuel, machine learning gives "computers the ability to learn without being explicitly programmed."<sup>[12][3]</sup> Samuel, an American pioneer in the field of *computer gaming* and artificial intelligence, coined the term "machine learning" in 1957 while at IBM.

Machine learning algorithms are extensively used for data-driven decision-making. Some other applications where machine learning has been used are email filtering, detection of network intruders or detecting a data breach, optical character recognition (OCR), learning to rank, computer vision, signal processing, and a wide range of engineering and business applications. Machine learning is employed in a range of computing tasks. Most of these applications require

huge amounts of data. Machine learning makes it possible to handle and solve these types of problems. Often, designing and programming explicit algorithms that are reproducible and have repeatability with good performance is difficult or infeasible.

## **Machine Learning (ML) Process and Applications**

In machine learning applications, huge amounts of data are needed, and it is challenging to develop programs to accomplish the task due to a large number of potential answers. One approach used in such cases is to label some correct answers as valid answers and use this as *training data*. Using the training data, the computer improves the algorithms to find the correct answer.

Machine Learning uses several approaches to teach computers to learn from huge amounts of high-quality data and focuses on prediction based on *known* properties learned from the *training data*.

Machine learning models are built using training data. In regression models, the training set is a set of labeled examples  $[x, f(x)]$  where  $x$  is the input variables and  $f(x)$  is the label or observed target. The labels can be numerical or categorical. The goal of the machine learning algorithms is given a training set, to find the approximation  $\hat{f}$  of  $f(x)$  that best generalizes or predicts labels for new examples. With high-quality data, machine learning can solve problems with high precision. Note that the algorithms are used to make predictions or draw conclusions without being explicitly programmed to do so.<sup>[2]</sup> Figure 3.1 shows how the training data and ML algorithms are used to achieve the end goal of making predictions.

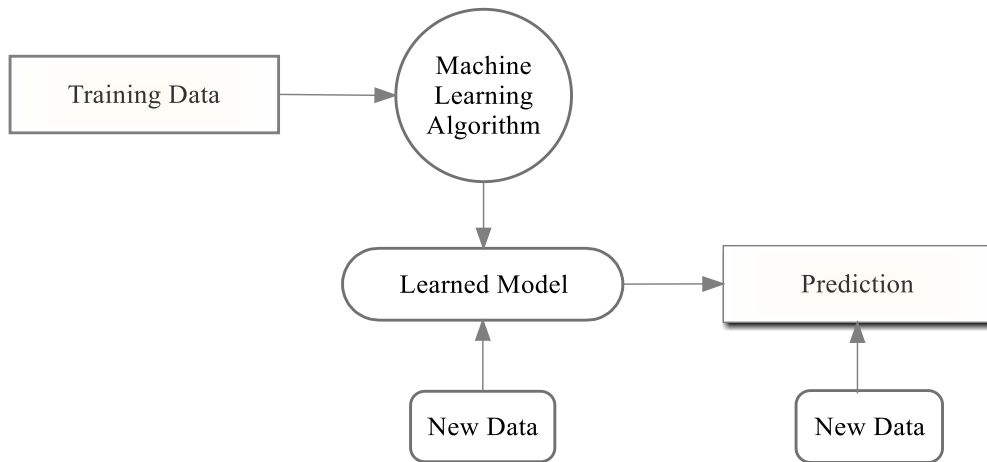


Figure 3.1: Process of achieving the end goal of prediction in Machine Learning

The numerous applications of machine learning include engineering, business, and medical science. Some of the specific applications are cyber security, signal processing, wireless communications, email filtering, speech recognition, agriculture, and computer vision, to name a few. In these applications, it is usually difficult or infeasible to develop conventional programs to perform the needed tasks. [\[3\]](#)[\[4\]](#)

### Machine Learning Problem Solving Steps

Figure 3.2 outlines the overview of the machine learning workflow. The flow in Figure 2.2 explains the machine learning problem-solving steps and approach. It also shows the role of data in achieving machine learning objectives.

Each of the steps in Figure 3.2 - from Data collection to production (deploying the selected model in production) is explained in the subsequent chapters.

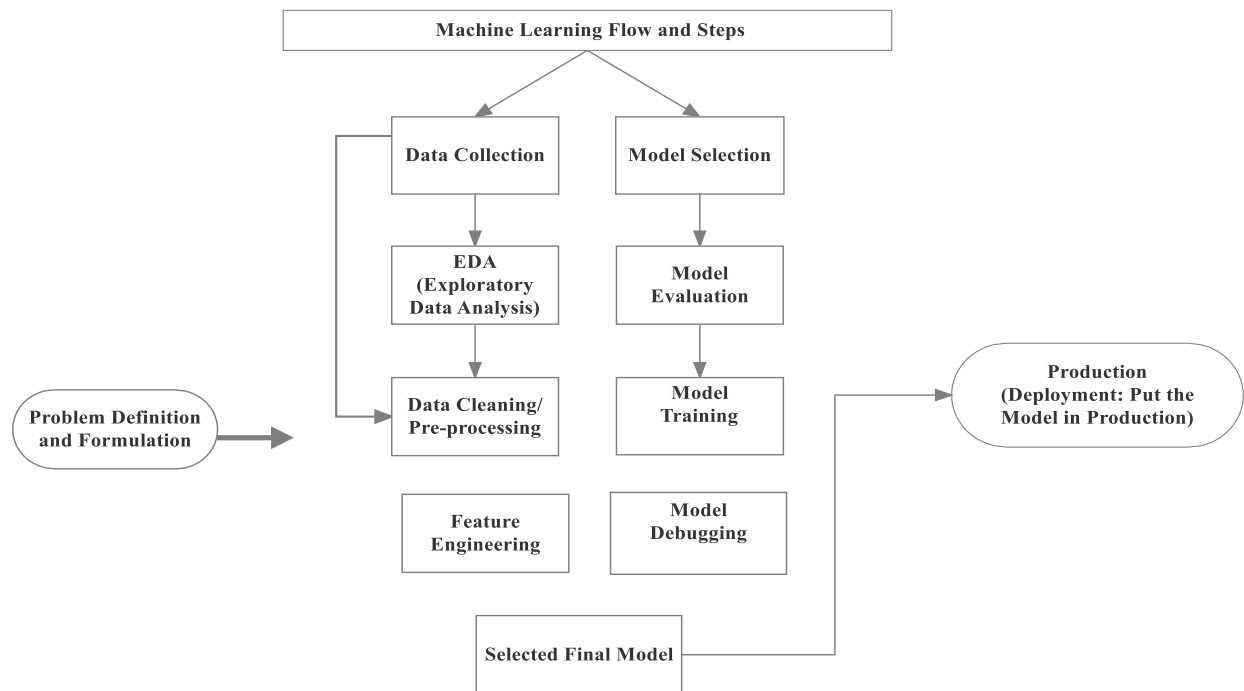


Figure 3.2: Machine Learning Workflow.

## Machine Learning and Associated Areas

Some of the key areas machine learning is associated with are:

### Data Mining

**Mathematical optimization**, and

**Computational statistics**,

These are considered subsets of machine learning. Machine learning algorithms and models draw heavily from these areas. Computational statistics focus on making predictions using statistical methods and algorithms using computers but not all machine learning is statistical learning. The aim of 'Computational Statistics' is implementing statistical methods on computers [\[sic\]](#).<sup>[3]</sup>

The study of mathematical optimization delivers methods, theory, and application domains to the field of machine learning.

Data mining focuses on exploratory data analysis (EDA) through unsupervised learning.<sup>[6][7]</sup>

## **Subsets of Machine Learning (ML)**

Machine learning programs can perform tasks without being explicitly programmed. It involves computers learning from data fed to them to carry out specific tasks. For simple tasks assigned to computers, it is possible to program algorithms telling the machine how to execute all steps required to solve the problem at hand. In such cases, no learning on the part of the computer is necessary.

For more advanced tasks, it can be challenging for a human to manually create the needed algorithms. In practice, it can turn out to be more effective to help the machine develop its own algorithm, rather than having human programmers specify every needed step.<sup>[11]</sup>

The discipline of machine learning employs various approaches to teach computers to accomplish tasks (find the best possible solutions). The associated areas of machine learning are shown in Figure 3.3. The figure shows the areas that have ties to machine learning. These are instrumental in arriving at ML problem solutions. Each of these areas are discussed in subsequent sections.

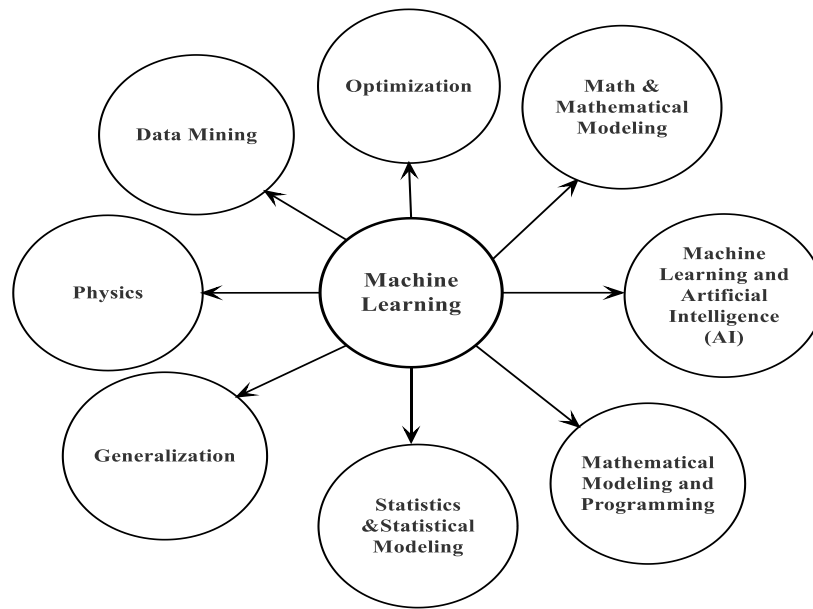


Figure 3.3: Machine Learning and Associated Areas

The associated areas are discussed below.

### Machine Learning and Data Mining

Machine learning and data mining are similar in some ways and often overlap in applications. One of the tasks in machine learning is to make predictions based on *known* properties learned from the training data, whereas data mining algorithms are used for discovery of (previously) *unknown* patterns. Data mining is concerned with knowledge discovery in databases (or KDD).

**Data mining** uses many machine learning methods. On the other hand, machine learning also employs data mining methods as "unsupervised learning" or as a preprocessing step to improve learner accuracy. The goals are somewhat different. The performance of machine learning is usually evaluated with respect to the ability to *reproduce known* knowledge. In data mining, which is knowledge discovery from the data (KDD), the key task is the discovery of previously *unknown* knowledge. Unlike machine learning, which is evaluated with respect to known knowledge, data mining uses uninformed or unsupervised methods that often outperform compared to other

supervised methods. In a typical KDD task, supervised methods cannot be used due to the unavailability of training data. Figure 3.4 outlines the predictive and descriptive data mining learning techniques along with supervised and unsupervised learning techniques.

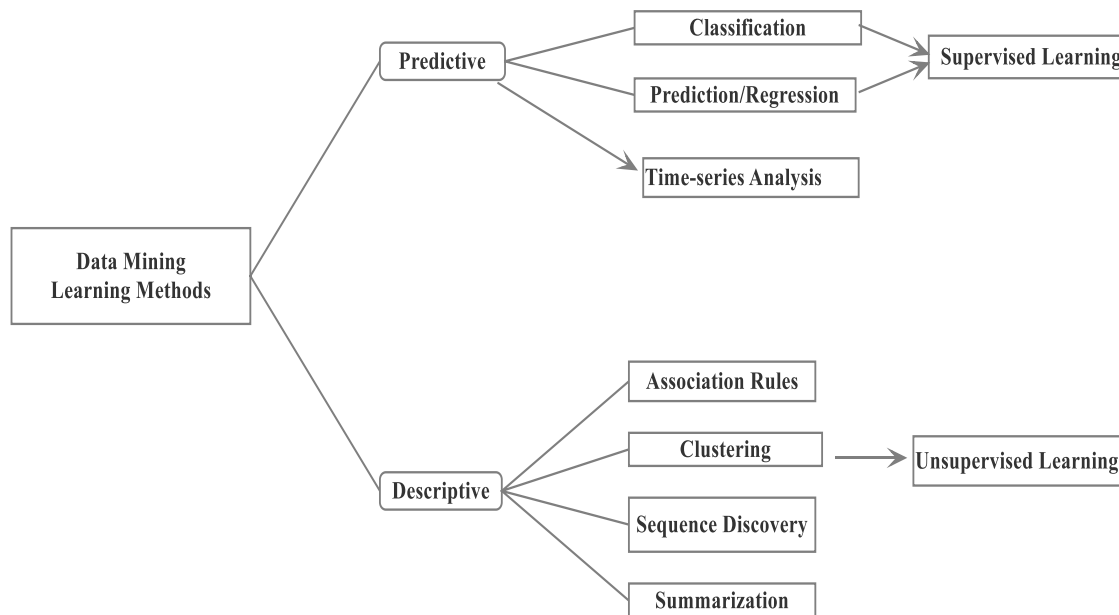


Figure 3.4: Predictive and Descriptive Data Mining

Several data mining and machine learning methods overlap in applications. As indicated earlier, data mining is concerned with knowledge discovery from the data (KDD) where the key task is the discovery of previously *unknown* knowledge. Machine Learning, on the other hand, is evaluated with respect to known knowledge [13] 14][15]16].

Machine learning methods use complex models and algorithms that allow the analysts to make predictions by learning from the trends, patterns and relationships in the historical data. The algorithms are designed to learn iteratively from data without being programmed. In a way, machine learning automates model building.

Machine-learning algorithms have extensive applications in data-driven predictions and are a major decision-making tool. Some applications where machine learning has been used are email

filtering, cyber security, signal processing, fraud detection and others. Machine learning is employed in a range of computing tasks. Although machine-learning models are being used in several applications, it has limitations in designing and programming explicit algorithms that are reproducible and have repeatability with good performance. With current research and the use of newer technology, the field of machine learning and artificial intelligence is becoming more promising.

It is important to note that Data mining uses unsupervised methods that usually outperform the supervised methods used in machine learning. Data mining is the application of knowledge discovery from data where supervised methods cannot be used due to the unavailability of training data. Machine learning may also employ data mining methods such as "unsupervised learning" to improve learner accuracy. The performance of machine learning algorithms depends on its ability to *reproduce known knowledge*.

## **Other Associated areas of Machine Learning:**

### **Optimization**

In many applications, machine learning problems are formulated to minimize some loss function, where the goal is the minimization of some [loss function](#) on a training set of data. Loss functions indicate the difference or the discrepancy between the predictions of the model being trained and the actual problem instances. For example, in classification problems, a label is assigned to the instances and models are trained to correctly predict the pre-assigned labels of a set of examples. [\[33\]](#)

[16]

### **Generalization**

In machine learning, generalization is the ability of the model to adapt to new data that is usually drawn from the unknown probability distribution or from the same distribution as the one used to

create and train the model. Thus, generalization is how accurately the learning machine performs on new previously unseen data or examples. It is about creating a general model that can produce accurate predictions in new cases.

Generalization of learning algorithms is an active topic of research in machine and **deep learning**.\*\*

## Physics

Analytical and computational techniques of statistical physics have been applied in machine learning. Statistical physics has been applied in [medical diagnostics](#).<sup>[37]</sup>

Other machine learning tools including variation inference and maximum entropy are derived from statistical physics. Although statistical physics is not a pre-requisite for Machine Learning and AI, a background in Physics is helpful.

## Statistics

Statistics is a major component of machine learning and data analytics. Statistics and machine learning are closely related fields. Statistics and statistical tools are widely used in machine learning. Both descriptive and inferential statistics tools are applied to visualize data to extract unseen patterns and draw inference. The broader analytics tools including the descriptive, predictive, and prescriptive analytics have applications in machine learning. A career in machine learning requires a deep understanding of Statistics and programming. The popular computer software to apply statistical concepts and programming are R-statistical programming language and Python. The overall field of machine learning and statistics is termed Data Science that we discussed in Chapter 1. The term Data Science was suggested by [Michael I. Jordan](#),

We have briefly presented the tools and techniques of statistics below – the knowledge of which is essential in learning machine learning and its applications.

For statistical applications, Leo Breiman suggested two types of statistical modeling: data model and algorithmic model.<sup>[29]</sup> The "algorithmic model" means more or less the machine learning algorithms like Random Forest.

## **A Brief Overview of Statistical Tools for Machine Learning:**

Businesses generate and consume massive amounts of data and rely on data for decision making. The data must be processed and analyzed using appropriate tools. Statistics is about making sense from data. Data scientists are usually not formally trained in statistics. Since Statistics is at the core of machine learning, machine learning professionals should be trained in statistics and data analysis to properly understand, analyze, and interpret statistical models. Statistics and data analysis constitute the tools and methods that help understand, analyze and interpret results.

Statistics is divided into two broad categories – Descriptive Statistics and Inferential statistics.

Descriptive statistics are about describing data. It uses charts and graphs or visual tools and numerical methods. Graphical methods are very effective in visualizing the patterns in the data and using these patterns to describing distributions. Probability distributions are powerful tools in making decisions from the data. Recently, several application software like Tableau is widely used to visualize patterns in big data.

Inferential statistics tools are used to conclude sample data. It is the process of concluding the population using the sample data. In statistics, we almost always rely on the sample data to draw conclusions about the population as in many cases, the population from which the samples are drawn is huge, and it is not usually possible to study the entire population.

Statistics studies variation. Almost all data shows variation, and statistics is the tool that studies variation in the data. The two main reasons for studying statistics are: (1) statistics allows us to study variation and (2) the tools of statistics are used to draw conclusions using limited data.

Statistics and machine learning are two closely related areas of study. Machine learning uses huge amounts of data and statistical tools and models to solve problems. It is usually not possible to solve real-world problems with machine learning without a good understanding of statistical fundamentals.

The major tools of statistics used to solve machine learning problems include exploratory data analysis (EDA), probability and probability distributions, sampling and estimation theory, hypothesis testing, design of experiment models, and others. These tools play an integral role in solving problems across many machine learning applications.

The other major tools used in solving machine learning problems are regression and classification methods. The regression models – from simple regression, multiple regression, ridge regression to many other types of regression are widely used. These methods are used in training machine learning data depending on the applications.

## **General Statistics Skills**

Here is a summary of the general statistical skills required for machine learning.

- Defining problems that are sound and are statistically answerable.
- Be able to use, calculate, and interpret simple statistical tools, including descriptive statistics, exploratory data analysis tools (EDA), and data visualization techniques to explain and communicate the findings.
- Understand mathematical statistical tools and their applications.

- Understand the concepts behind the central limit theorem and the law of large numbers.
- Apply the major inferential statistics tools, including sampling theory, estimation, hypothesis testing, and analysis of variance (ANOVA)
- Be able to establish relationships between the target variable and the independent variables or the features using appropriate statistical models.
- Design statistical hypothesis testing experiments, A/B testing, and so on, and design experiments using design of experiment techniques.
- Be able to understand and interpret various regression and classification models.
- Be able to calculate and interpret performance metrics, for example p-value, alpha, typeI and typeII errors, power of tests.

## **Machine Learning (ML) Process, and the types of Data in ML Process**

ML process heavily depends on data and data is key to machine learning. Managing and using data is critical to machine learning.

Machine learning solutions are as good as the data and the ML process heavily depends on the quality of data. It is important to learn how to use and manage data throughout the ML process.

One of the major goals of machine learning is to construct or design algorithms that are used to make and improve predictions by learning from large quantities of data. ML models are effective when high-quality data is used. Thus, ML is a data-driven decision-making process where the predictions are made using *mathematical models* from the input data. The input data used to build the models are divided into three data sets and are used in different stages of the model creation and testing.

## Data for ML models and Managing Data throughout ML Process

In the machine learning model building process, the data is divided or partitioned into three categories and are used in different phases of model building. The data sets are:

- **Training Data Set**
- **Validation Data**
- **Test Data**

### Training Data Set

The training data set is a set on which the model is fit initially. The training data set consists of pairs of input and output vectors (or scalars). The inputs are often termed as independent variables or *attributes* or *features*, and the output or the answer key is termed as a *label*, or *target*, or dependent (response) variable.

The training data is a set of examples used to fit the *parameters*\* of the models. These parameters may be the weights of connections between the neurons in *artificial neural networks* or a *classifier* or a *naïve Bayes classifier* that are trained using the training data set using a *supervised learning* method or using optimization methods (e.g., *gradient descent*).

Parameters are internal to the model and are learned or estimated purely from the data during training as the algorithm used tries to learn the mapping between the input features and the labels or targets.

In the initial phase of the training, parameters are initialized to some values (random values or set to zeros). The initial values are updated as the training progresses using an optimization algorithm like gradient descent. The learning algorithm continuously updates the parameter values in the learning process. At the end of training or learning, the updated parameters are the model parameters. Some examples of parameters are the coefficients of the regression modes (linear,

multiple, or logistics regression models), weights in the neural networks, or the centroid of the clusters in clustering algorithms.

In machine learning and deep learning, the parameters are the values the learning algorithm can change independently during the learning process. It is important to note that the parameter values are affected by the choice of *hyperparameters* set before training begins and the learning algorithm uses them to learn the parameters.

As indicated, the parameters are internal to the model and are learned or estimated purely from the data during the training. These parameters are continuously updated, and the final parameters are the parameters at the end of the training process. The hyperparameters are set before the training process. Setting the right hyperparameter values is critical as it directly affects the model's performance. The process of setting the hyperparameters is called hyperparameter tuning.

The initial model is run using the training dataset. The result produced by the model is compared with the target for each input vector in the training set. Based on the result of comparison and the type of learning algorithm used, the model parameters are adjusted. The model fitting is an iterative process and involves both *variable selection* and *parameter estimation*.  
[<https://towardsdatascience.com/parameters-and-hyperparameters-aa609601a9ac>].

## **Validation data set.**<sup>[3]</sup>

In the second phase, the model fitted in the first stage using the training data is used to predict the responses using another set of observations known as the **validation data set**. This new data set known as the validation set provides an unbiased evaluation of the model fit used by the training set by tuning the *hyperparameters*.

## Hyperparameters

Hyperparameters are parameters whose values control the learning process and determine the values of model parameters that a learning algorithm ends up learning. The prefix 'hyper\_' suggests that they are 'top-level' parameters that control the learning process and the model parameters that result from it.

In machine learning models, the hyperparameter values are set by the machine learning engineer usually before the model training begins. These parameters are external to the model and the model cannot change these values during the model training or learning.

Hyperparameters are used by the model in the model training process, but these are not the part of resulting model that results at the end of the training process. At the end of the learning, we have the trained model. It is usually not known what hyperparameter values were used to train the model, we only know the trained model parameter values.

It is important to note that hyperparameters are values we decide before the model training begins and they remain the same at the end of the training. Some examples include number of hidden layers in neural networks, number of iterations in training a neural network, number of clusters in a clustering algorithm, learning rate in optimization algorithms (for example, gradient descent), the cost function or loss function the model will use, and others.

It is important to understand the difference between hyperparameters and parameters. The hyperparameters affect the parameters the model learns but the hyperparameters are unchanged.

The difference between the test and validation data set is that the training data set is a set of data used to fit the model. A validation dataset is a subset of data used to provide an objective assessment of a model's fit on the training data while changing hyperparameters.

## Test Data Set

The test dataset is a subset of the training dataset that is utilized to give an objective evaluation of a final model. It is a set of examples used to assess the performance or generalization of a fully-specified classifier.<sup>[8][9]</sup> A test data set is independent of the training data. It should follow the same probability distribution as the training data set. If the model fits well to the training set and also fits the test data set, it is an indication of minimal *overfitting*.

Overfitting occurs when the model provides accurate predictions for the training data but not for new data. The reasons for overfitting are (i) the training data has too much noise, (ii) the model being fitted is too complex and learns the noise in the training data set, (iii) the training data set is too small. One of the methods of testing for overfitting is **K-fold cross-validation**.

Underfitting is another undesirable condition in models. This occurs when the model cannot establish a meaningful relationship between the input and output data. Underfit models provide inaccurate results for both the training and test data sets, and they have high bias. Overfit models have high variance. Overfit models provide accurate results for the training data set but not to the test set. In model training it is important to find a tradeoff between bias and variance.

## Determining the Size of Data Sets

Dividing the data set accurately is critical to creating a successful and effective machine learning model. Dividing the data set in training, validation, and test sets is about determining the size of

each set. While more training data should lead to a more accurate model, a smaller amount of test and validation data may provide a more biased estimate of the model and its performance. So, there is a need for a balance between these sets. If a large quantity of data is available, a 60-20-20 split is suggested in the literature. There is also an 80-10-10 split used. When the data set is small. A smaller test may not provide a good enough representation, so in practice a K-fold Cross-validation may be used. Suggestions regarding splitting of the data sets into training, validation, and test sets can be found in the reference below.

1. [<https://glassboxmedicine.com/2019/09/15/best-use-of-train-val-test-splits-with-tips-for-medical-data/#:~:text=Common%20ratios%20used%20are%3A,20%25%20val%2C%2020%25%20test>]
2. <https://deepchecks.com/training-validation-and-test-sets-what-are-the-differences/>

In machine learning algorithm we split the dataset in a certain ratio training, validation, and test data set. The model is built on a training dataset. The test dataset is used for the evaluation of the model efficiency.

## Machine Learning Problem Solving Steps

Almost every machine learning project goes through the following steps. These are:

**(1) Defining the Problem and the Problem Statement:** This is usually the first step in a machine learning project. The tools used in this step are Exploratory Data Analysis (EDA) and Data mining.

**Data Exploration:** The major statistical tools used in this step are descriptive statistics, exploratory data analysis, and data visualization tools.

**(2) Data preparation and Data cleaning:** These involve dealing with missing values, dealing with data corruption and data

errors (from a bad sensor), and unformatted data (observations with different scales). Some of the cleaning methods include **outlier detection** and **missing value imputation**.

### **(3) Data Preparation, Transformation, and Standardization**

If data has errors and contains inconsistencies, it cannot be used to train machine learning models. In such cases, the techniques used are data transformation and standardization of the data. The data might need to go through a set of transformations to change its shape or structure, making it more suitable for the algorithms being used.

The techniques used in this case are data **sampling, feature selection methods, data transformation, scaling, encoding** and **standardization**.

### **(4) Model Selection & Evaluation**

A key step in solving a predictive analytics problem in machine learning is selecting and evaluating the appropriate model and learning method. There are several statistical models and modeling techniques used to approach these problems.

Some of the statistical tools that are helpful in the process are **Experimental design and Design of Experiments (DOE)**. Experimental design is a part of statistics that is very useful in the initial stage of the model. It is also used in the feature selection and evaluation process of a model. The other tools used in conjunction with DOE are **Statistical Hypothesis Testing** and **Estimation Theory**. The use of all these tools require a good foundation and understanding of Statistics.

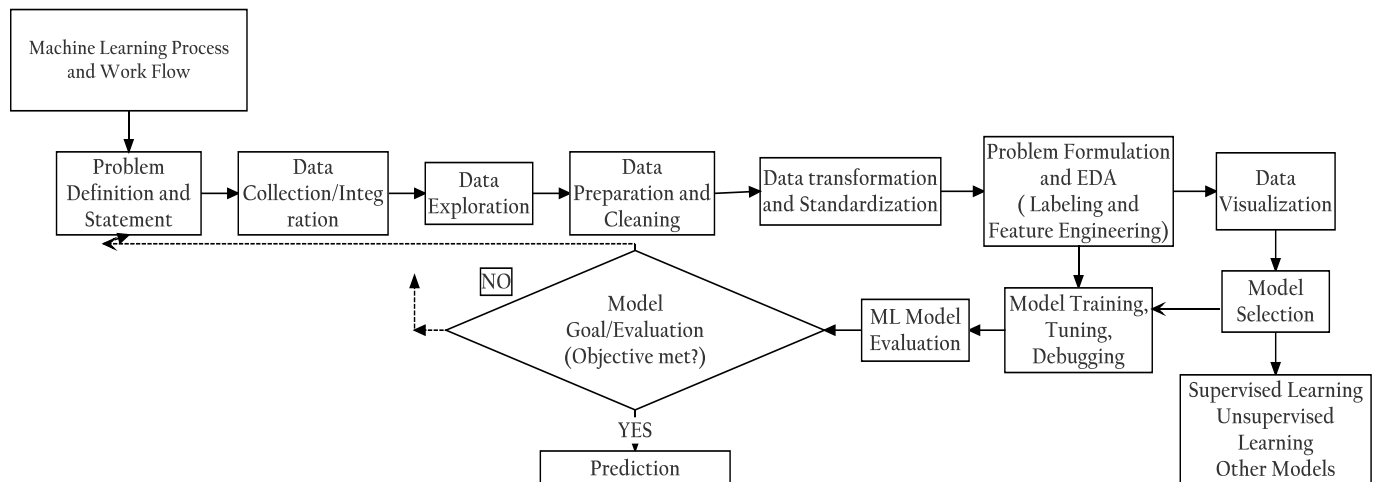
### **Fine-tuning the model**

Almost all machine learning algorithms use hyperparameters that are used to customize the training and refine the models.

The values of hyperparameters **control the learning process and help determine the values of model parameters that a learning or training algorithm uses as improved model**. The 'hyper\_' means that these are 'top-level' parameters that control the learning process and establish the final model parameters that are used by the model..

Some of the examples of hyperparameters are **the number of nodes and the layers in a neural network or the number of branches in a decision tree**. Hyperparameters also determine the model architecture, learning rate, and model complexity.

A model starts the training process with random parameter values and adjusts them throughout whereas hyperparameters are **outside of the model and are set by the analyst before the model training**. The values of hyperparameters are expected to improve the model accuracy but it is not always the case. They may have adverse effects on the model. The hyperparameter tuning is empirical in nature and requires many experiments to evaluate the effect of different hyperparameter settings on the performance of the model. Figure 3.5 outlines the above steps.



**Chapter 3** **Figure 3.5: Steps in Solving Machine Learning Problem.**

## Summary

This chapter discusses the machine learning process, associated areas of machine learning, data used, and the way the data is divided and managed in building ML Models. The major topics of discussion include the need and justification for using machine learning, machine learning (ML) applications, machine learning problem-solving steps, associated areas, and subsets of Machine Learning.

Some of the subsets or the areas machine learning interacts with to solve the problems include computational statistics, mathematical optimization, data mining, statistics, and programming. These areas are explained briefly.

Machine learning uses huge amounts of data and statistical tools and models to solve problems. It is usually not possible to solve real-world machine learning problems without a good understanding of statistics.

The major tools of statistics used to solve machine learning problems include exploratory data analysis (EDA), probability and probability distributions, sampling and estimation theory, hypothesis testing, design of experiments, and others. These tools play an integral role in solving problems across all machine learning applications.

ML heavily depends on data and data is key to building Machine Learning models. Managing and using Data is critical to machine learning. Machine learning solutions are as good as the data. The ML process heavily depends on the quality of data. It is important to learn how to use and manage data throughout the ML process. The chapter discusses the way the data is divided and managed throughout the machine learning process.

ML is a data-driven decision-making process where the decisions are made using *models*. The input data used to build the models are divided into three data sets and are used in different stages of the model creation and testing. The three categories are: Training Data Set, Validation Data Set, and Test Data. Each of these sets, the recommended size and proportion of these sets are discussed.

Finally, this chapter discusses the steps in solving machine learning problems. Every machine learning problem goes through these steps – problem definition and problem statement, data exploration, data preparation and data cleaning, data transformation and standardization, model selection and evaluation, and fine tuning the model. These steps and the role of statistics in these steps are discussed. The chapter also provided an overview of statistical tools in machine learning.