JAIN DEEMED-TO-BE UNIVERSITY | ONLINE

Self-Learning Material

Program: MCA
Specialization: Artificial Intelligence
Semester: 2
Course Name: Predictive Analytics using Machine Learning
Course Code: 21VMC6S205
Unit Name: Introduction to Machine Learning

SELF LEARNING MATERIAL

UNIT 03: INTRODUCTION TO MACHINE LEARNING

OVERVIEW

Machine learning– introduction,significance,use cases**; AI : DEEP LEARNING: MACHINE LEARNING-** definition,differences **;Statistics –** definition, example**,Types of statistics –** inferential and descriptive statistics-range,mean,mode,SD,variance,median.

OBJECTIVES

In this unit you will learn

1. WHAT IS MACHINE LEARNING
2. AI vs ML vs DL
3. SIGNIFICANCE OF ML
4. STATISTICS
5. TYPES OF STATISTICS

LEARNING OUTCOMES

At the end of the unit you would

- A basic understanding of machine learning
- How is machine learning different from other concepts
- Statistics and its types

TABLE OF TOPICS

INTRODUCTION TO MACHINE LEARNING

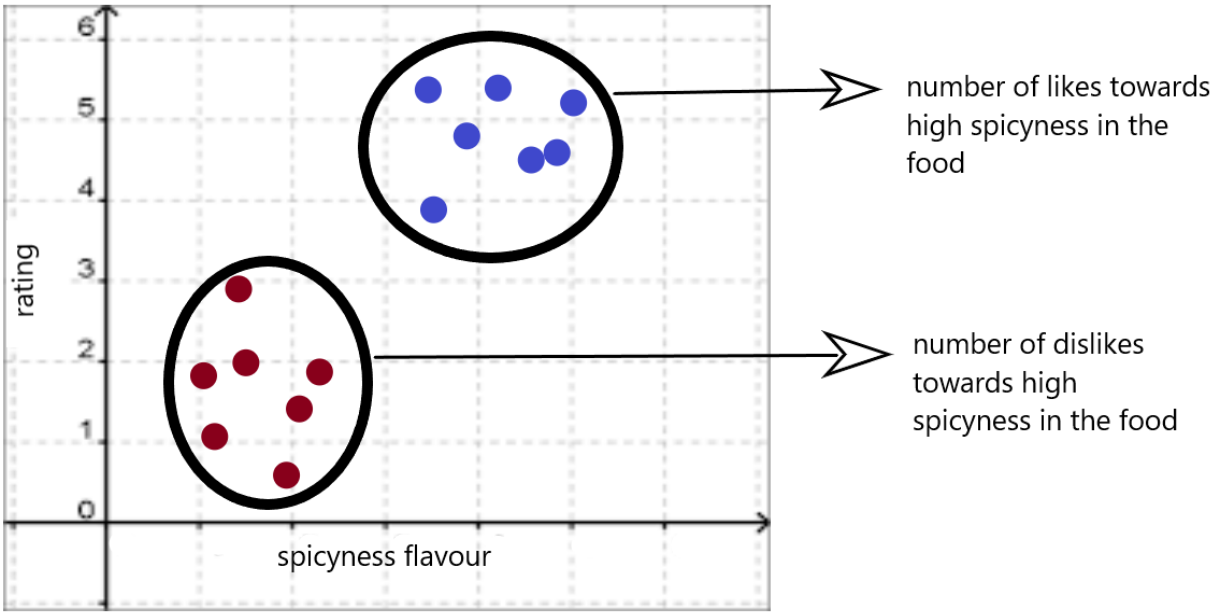
MACHINE LEARNING

A machine is a thing that is created by people to make work easier. A tool or invention multiplies the effect of human effort. A machine when invented reduced the human intelligence.

A learning is defined as the acquisition of knowledge or skills through study, experience, or being taught.

Now, machine learning can be defined as the machine which was invented by human to reduce their work is being made to learn and acquire the required knowledge through experience or being taught by which it does some function with help of an algorithm is know as machine learning. The algorithm by which it does some function is called a machine learning algorithm.

For example, let us consider the case where there are people in the area where few likes spicy flavored food and few likes less spicy flavor and some may even dislike the spicy flavored food. By taking a survey and with the collected data a graph is being plotted where the points in blue color are the people who likes the spicy flavor in the food and had given high rating for the food. And the brown points are given by people who dislikes or shown less likings than others towards the spicy flavored food.

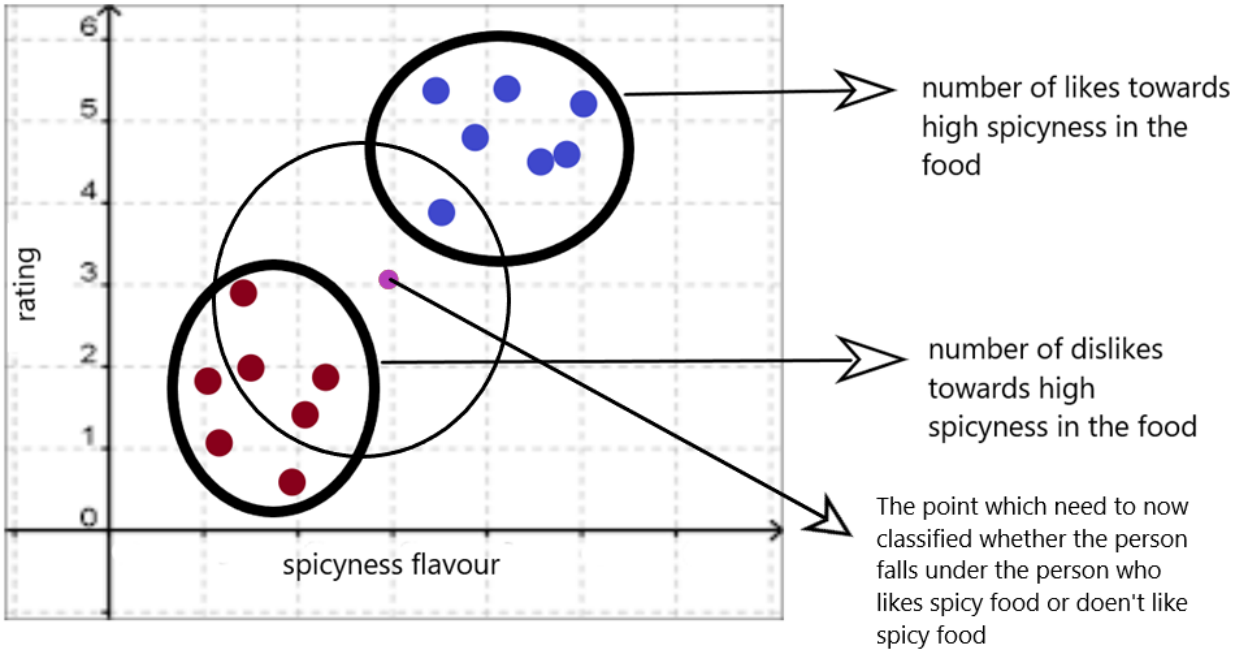


The points that are closer and falls under the category of ‘liking and high rating’ are grouped together which means any points that lies under that category is considered that the person has liking towards the flavor of food. The points that fall under the category of ‘low rating and less liking’ would be considered as the other category.

Now, with this category when a new person is added to the mass and when the survey is again taken and those points are being marked, now the analysis can be easily made and the restaurants to be opened in that area can make dishes accordingly to grow their business according to the people’s choice of taste.

So, if the area people on high range has shown interest towards high spicy foods then the restaurants can cook food which are spicy and if the area people had shown interest to less spicy food then the restaurant can cook food accordingly and alter their food in their menu

This is one such case where machine learning can be applied and is useful to grow a business and improve the growth.



Now, the circle is drawn around the points for a minimum nearby distance to understand the relation which is a famous machine learning algorithm K-means , by using this algorithm we can now predict that the person whose point is depicted as purple falls under ‘disliking spicy food’ as the number of points lying near that point is of those who has given less rating for the spicy flavored food.

Hence, by using this we can predict the needed outcome.

MACHINE LEARNING SIGNIFICANCE

Machine learning as said simplifies a large problem, thereby reducing the human intervention and handles to solve the problem at ease. Some of the major problems were solved by applying machine-learning concepts and advancements has been done in many field.

For example, fraud detection which required human intelligence but now using the machine-learning algorithm the fraud detection is done efficiently than human. For this it requires a dataset and a machine-learning algorithm by which a machine learning model can be built to predict the output.

When it comes to machine learning the system is doing the job of human, and how does the system does this? Whether it is being trained or is it being dictated to do?. So, for this question the machine learning can be categorized into below methods of learning .

1. SUPERVISED LEARNING
2. UNSUPERVISED LEARNING
3. SEMI- SUPERVISED LEARNING
4. REINFORCEMENT LEARNING

When is machine learning is being useful?

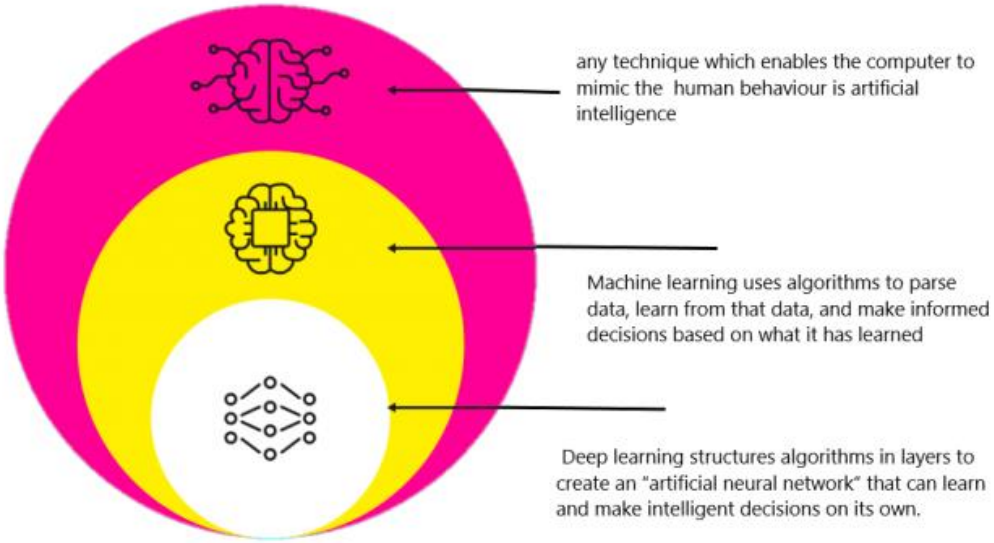
The machine learning concept is used in many fields some are being listed

HEALTH CARE: Now, it is possible to predict the heart attack death rate, birth rate in an area and many such useful insights which can help to take necessary actions based on the outcome.

EDUCATION: To analyze the performance of the students, and to predict the results in the forth coming year with the past history datasets.

FINANCIAL SERVICES: Fraud detection, loan prediction and all such analysis can be made to enhance the respective fields using machine learning algorithm.

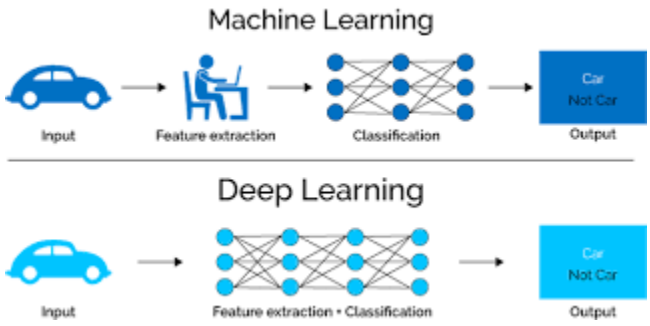
How is machine learning different from AI and Deep learning?



MACHINE LEARNING VS DEEP LEARNING

Deep learning algorithms can be regarded both as a sophisticated and mathematically complex evolution of machine learning algorithms. Deep learning describes algorithms that analyze data with a logic structure similar to how a human would draw conclusions.

- Deep learning is a specialized subset of machine learning.
- Deep learning relies on a layered structure of algorithms called an artificial neural network.
- Deep learning has huge data needs but requires little human intervention to function properly.
- Transfer learning is a cure for the needs of large training datasets.



In the above given example ,

Feature extraction refers to the process of transforming raw data into numerical features that can be processed while preserving the information in the original data set

MACHINE LEARNING VS ARTIFICIAL INTELLIGENCE

AI solves tasks that require human intelligence while ML is a subset of artificial intelligence that solves specific tasks by learning from data and making predictions. Artificial intelligence is a technology which enables a machine to simulate human behavior. Machine learning is a subset of AI which allows a machine to automatically learn from past data without programming explicitly. AI system is concerned about maximizing the chances of success. Machine learning is mainly concerned about accuracy and patterns.

ALEXA - AI

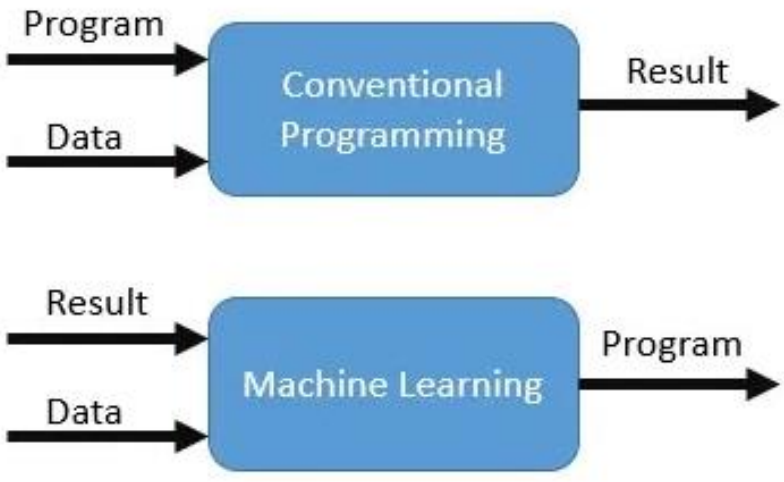
ML ALGORITHM

CONVENTIONAL PROGRAMMING METHODS

Conventional programming is a manual process, which means the programmer creates the logic of the program. Conventional Programming uses conventional procedural language. It could be assembly language or a high-level language such as C, C++, Java, JavaScript, Python, etc. Conventional programming is a manual process, which means the programmer creates the logic of the program.

Here the programmer gives the input and creates the logic for which the output is generated

How is machine learning different from Conventional method ?



In conventional programming, a programmer needs to hard code the logic of the program. In machine learning, it depends a lot on the machine which learns from input data. The computer systems use these statistical models to perform a specific task effectively. Here you don't need to provide explicit instructions; instead, it relies on patterns and inference.

For example let us consider a scenario where a person wins if he scores above 100 and losses of he scores below 100, for this in conventional programming the programmer writes the logic of the code and an input is given to get desired output

In machine learning, a dataset containing of the numbers with output is fitted and trained with a machine learning algorithm after with a model is trained to predict whatever input is given to predict the output.

Unlike traditional programming, machine learning is an automated process. It can increase the value of your embedded analytics in many areas, including data prep, natural language interfaces, automatic outlier detection, recommendations, and causality and significance detection.

STATISTICS

Statistics is the science concerned with developing and studying methods for collecting, analyzing, interpreting and presenting data. Statistical measure are mean, mode, range, and standard deviation. Statistics is a tool that helps us to extract information and knowledge from data.

For example In a classroom there can be many students, and to analyze the performance of the student and make useful insights we can use statistics.

Statistics in machine learning can be classified based on the steps,

STATISTICS IN DATA PREPARATION

A basic understanding of data distributions, descriptive statistics, and data visualization is required to help you identify the methods to choose when performing these tasks.

STATISTICS IN MODEL EVALUATION

Statistical methods are required when evaluating the skill of a machine learning model on data not seen during training.

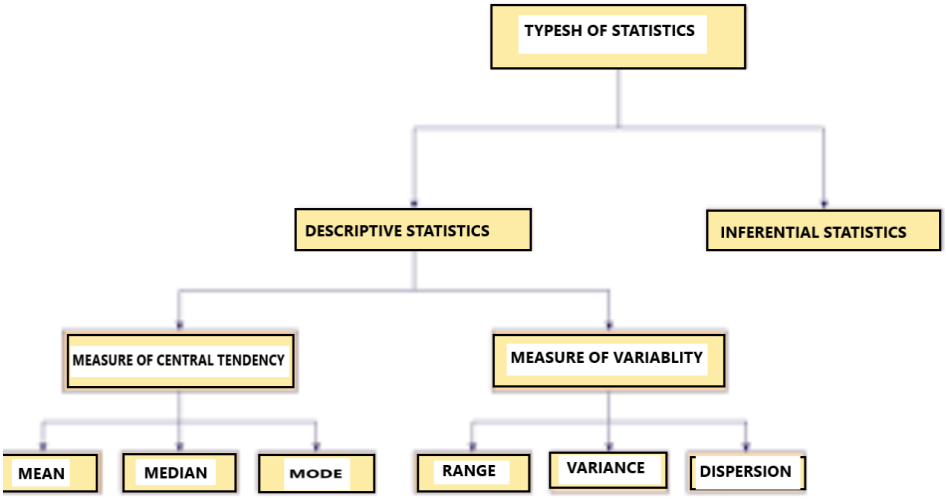
STATISTICS IN MODEL PREDICTION

Statistical methods are required when presenting the skill of a final model to stakeholder. This might include estimation statistics such as prediction intervals.

Statistics can be broadly classified into

Descriptive Statistics: Descriptive statistics refer to methods for summarizing raw observations into information that we can understand and share.

Inferential Statistics: Inferential statistics is a fancy name for methods that aid in quantifying properties of the domain or population from a smaller set of obtained observations called a sample.



DESCRIPTIVE STATISTICS

Descriptive statistics mainly deals with two factors namely,

1. Measure of central tendencies
2. Measure of variability

MEASURE OF CENTRAL TENDANCY

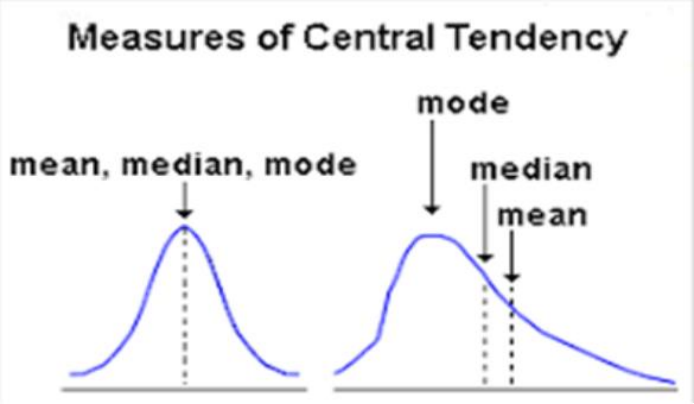
The measure of central tendency calculates a single value that describes the set of data by identifying the central position within the set of data. The valid measures of central tendency are,

1. MEAN
2. MODE
3. MEDIAN

MEAN = SUM OF ALL VALUES
_____
TOTAL NUMBER OF VALUES

MEDIAN = MIDDLE VALUE(WHEN DATA ARRANGED IN ORDER)

MODE= MOST COMMON VALUE



Example

Let us consider a set of data 2,3,4,4,6,7 as the data set of people in a family living in the area A. Now to calculate the mean, mode, and median.

To calculate mean, which is the arithmetic mean of the continuous data

2+3+4+4+6+7/6=4.333 is the mean value

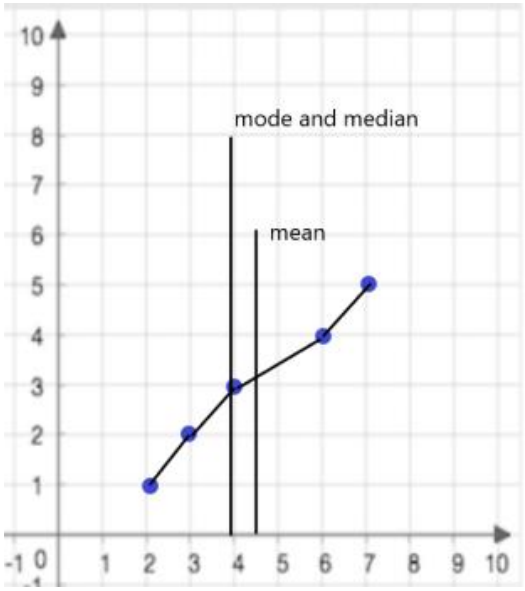
To calculate median, which is the middle value of the data set we need to arrange the data set in order and then calculate, since the data set is in order we can directly infer the median.

Since there are two middle value 4 and 4 we need to add those and divide by 2 to find the median value.

4+4/2=4 is the median value of the data set

To calculate mode, which shows the number that occurs frequently

The number 4 occurs frequently in the given data set hence 4 is the mode.



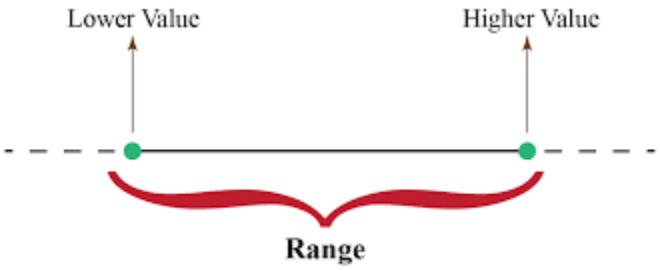
MEASURE OF VARIABLITY

Variability in statistics is the mathematical measure of the spread of a data set. Variability describes how far apart data points lie from each other and from the center of a distribution. Variability is measured on basis of,

1. Range
2. Standard deviation
3. Variance

RANGE

In statistics, the range of a set of data is the difference between the largest and smallest values. Difference here is specific, the range of a set of data is the result of subtracting the sample maximum and minimum.



RANGE = MAXIMUM VALUE-MINIMUM VALUE

STANDARD DEVIATION

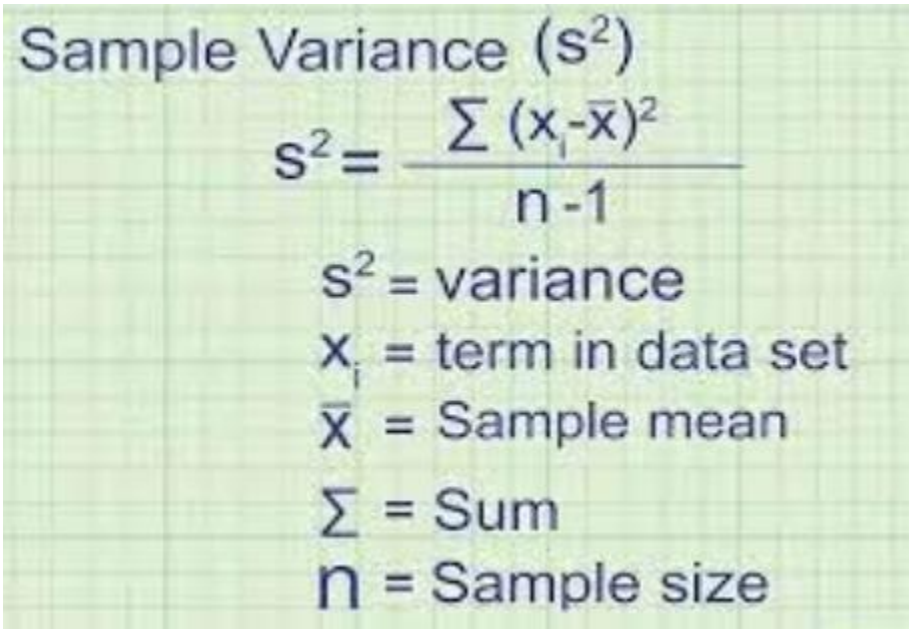
the standard deviation is a measure of the amount of variation or dispersion of a set of values.

Formula	Explanation
$\sigma = \sqrt{\frac{\Sigma (X-\mu)^2}{N}}$	• σ = population standard deviation • Σ = sum of... • X = each value • μ = population mean • N = number of values in the population

DATA	STANDARD DEVIATION S.D = √ (DATA-MEAN)/N		
3	3-4	-1	
3	3-4	-1	$\frac{\sqrt{(-1+(-1)+0+2)}}{4}$
4	4-4	0	= √0=0
6	6-4	2	

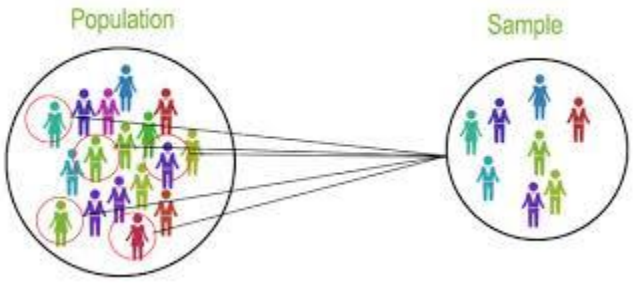
VARIANCE

variance is the expectation of the squared deviation of a random variable from its population mean or sample mean.



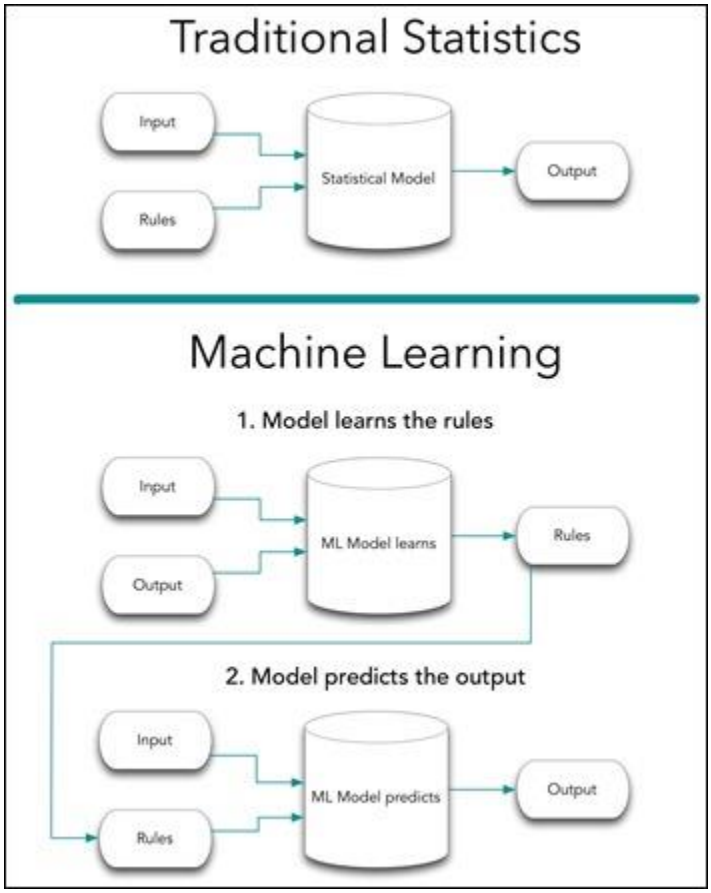
INFERENTIAL STATISTICS

This takes data from a sample and makes inferences and predictions about the larger population from which sample is drawn. It is difficult to analyze the entire data individually so we make inferential statistics to analyze a sample which is drawn from the population and made some predictions.



MACHINE LEARNING VS STATISTICAL MODEL

The major difference between machine learning and statistics is their purpose. Machine learning models are designed to make the most accurate predictions possible. Statistical models are designed for inference about the relationships between variables



FORMULAS USED

MEAN = SUM OF ALL VALUES
_____
TOTAL NUMBER OF VALUES

MEDIAN = MIDDLE VALUE(WHEN DATA ARRANGED IN ORDER)

MODE= MOST COMMON VALUE

Formula	Explanation
$\sigma = \sqrt{\frac{\Sigma (X-\mu)^2}{N}}$	• σ = population standard deviation • Σ = sum of... • X = each value • μ = population mean • N = number of values in the population

Sample Variance (s^2)

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

s^2 = variance
x_i = term in data set
$\bar{x}$ = Sample mean
$\sum$ = Sum
n = Sample size