 <b>Marwadi University</b>	<b>Marwadi University</b> <b>Faculty of Technology</b> <b>Department of Information and Communication Technology</b>	
<b>Subject: Artificial Intelligence(01CT0616)</b>	<b>Aim: To learn the Basics of Machine Learning</b>	
<b>Experiment No: 01</b>	<b>Date:</b>	<b>Enrollment No: 92200133003</b>

**Aim:** To learn the Basics of Machine Learning

**IDE:** Google Colab **Theory:**

Machine learning is a subset of artificial intelligence in the field of computer science that often uses statistical techniques to give computers the ability to "learn" (i.e., progressively improve performance on a specific task) with data, without being explicitly programmed. In the past decade, machine learning has given us self-driving cars, practical speech recognition, effective web search, and a vastly improved understanding of the human genome.

### Machine learning tasks


Machine learning tasks are typically classified into two broad categories, depending on whether there is a learning "signal" or "feedback" available to a learning system:

- **Supervised learning:** The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs. As special cases, the input signal can be only partially available, or restricted to special feedback:
- **Semi-supervised learning:** the computer is given only an incomplete training signal: a training set with some (often many) of the target outputs missing.
- **Active learning:** the computer can only obtain training labels for a limited set of instances (based on a budget), and also has to optimize its choice of objects to acquire labels for. When used interactively, these can be presented to the user for labeling.
- **Reinforcement learning:** training data (in form of rewards and punishments) is given only as feedback to the program's actions in a dynamic environment, such as driving a vehicle or playing a game against an opponent.
- **Unsupervised learning:** No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning).

### Machine learning applications

In classification, inputs are divided into two or more classes, and the learner must produce a model that assigns unseen inputs to one or more (multi-label classification) of these classes. This is typically tackled in a supervised manner. Spam filtering is an example of classification, where the inputs are email (or other) messages and the classes are "spam" and "not spam". In regression, also a supervised problem, the outputs are continuous rather than discrete.

In clustering, a set of inputs is to be divided into groups. Unlike in classification, the groups are not known beforehand, making this typically an unsupervised task. Density estimation finds the distribution of inputs in some space. Dimensionality reduction simplifies inputs by mapping them into a lower- dimensional space. Topic

 <b>Marwadi University</b>	<b>Marwadi University</b> <b>Faculty of Technology</b> <b>Department of Information and Communication Technology</b>	
<b>Subject: Artificial Intelligence(01CT0616)</b>	<b>Aim: To learn the Basics of Machine Learning</b>	
<b>Experiment No: 01</b>	<b>Date:</b>	<b>Enrollment No: 92200133003</b>

modeling is a related problem, where a program is given a list of human language documents and is tasked with finding out which documents cover similar topics.

### **Machine learning Approaches**

**Decision tree learning:** Decision tree learning uses a decision tree as a predictive model, which maps observations about an item to conclusions about the item's target value. **Association rule learning** Association rule learning is a method for discovering interesting relations between variables in large databases.

### **Artificial neural networks**

An artificial neural network (ANN) learning algorithm, usually called "neural network" (NN), is a learning algorithm that is vaguely inspired by biological neural networks. Computations are structured in terms of an interconnected group of artificial neurons, processing information using a connectionist approach to computation. Modern neural networks are non-linear statistical data modeling tools. They are usually used to model complex relationships between inputs and outputs, to find patterns in data, or to capture the statistical structure in an unknown joint probability distribution between observed variables.

### **Deep learning**

Falling hardware prices and the development of GPUs for personal use in the last few years have contributed to the development of the concept of deep learning which consists of multiple hidden layers in an artificial neural network. This approach tries to model the way the human brain processes light and sound into vision and hearing. Some successful applications of deep learning are computer vision and speech recognition.

### **Inductive logic programming**


Inductive logic programming (ILP) is an approach to rule learning using logic programming as a uniform representation for input examples, background knowledge, and hypotheses. Given an encoding of the known background knowledge and a set of examples represented as a logical database of facts, an ILP system will derive a hypothesized logic program that entails all positive and no negative examples. Inductive programming is a related field that considers any kind of programming languages for representing hypotheses (and not only logic programming), such as functional programs.

### **Support vector machines**

Support vector machines (SVMs) are a set of related supervised learning methods used for classification and regression. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that predicts whether a new example falls into one category or the other.

### **Clustering**

Cluster analysis is the assignment of a set of observations into subsets (called clusters) so that observations within the same cluster are similar according to some pre designated criterion or criteria, while observations

 <b>Marwadi</b> University	<b>Marwadi University</b> <b>Faculty of Technology</b> <b>Department of Information and Communication Technology</b>	
<b>Subject: Artificial Intelligence(01CT0616)</b>	<b>Aim: To learn the Basics of Machine Learning</b>	
<b>Experiment No: 01</b>	<b>Date:</b>	<b>Enrollment No: 92200133003</b>

drawn from different clusters are dissimilar. Different clustering techniques make different assumptions on the structure of the data, often defined by some similarity metric and evaluated for example by internal compactness (similarity between members of the same cluster) and separation between different clusters. Other methods are based on estimated density and graph connectivity. Clustering is a method of unsupervised learning, and a common technique for statistical data analysis.

### **Bayesian networks**

A Bayesian network, belief network or directed acyclic graphical model is a probabilistic graphical model that represents a set of random variables and their conditional independencies via a directed acyclic graph (DAG). For example, a Bayesian network could represent the probabilistic relationships between diseases and symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases. Efficient algorithms exist that perform inference and learning.

### **Reinforcement learning**

Reinforcement learning is concerned with how an agent ought to take actions in an environment so as to maximize some notion of long-term reward. Reinforcement learning algorithms attempt to find a policy that maps states of the world to the actions the agent ought to take in those states. Reinforcement learning differs from the supervised learning problem in that correct input/output pairs are never presented, nor sub-optimal actions explicitly corrected.

### **Similarity and metric learning**


In this problem, the learning machine is given pairs of examples that are considered similar and pairs of less similar objects. It then needs to learn a similarity function (or a distance metric function) that can predict if new objects are similar. It is sometimes used in Recommendation systems.

### **Genetic algorithms**

A genetic algorithm (GA) is a search heuristic that mimics the process of natural selection, and uses methods such as mutation and crossover to generate new genotype in the hope of finding good solutions to a given problem. In machine learning, genetic algorithms found some uses in the 1980s and 1990s. Conversely, machine learning techniques have been used to improve the performance of genetic and evolutionary algorithms.

### **Rule-based machine learning**

Rule-based machine learning is a general term for any machine learning method that identifies, learns, or evolves "rules" to store, manipulate or apply, knowledge. The defining characteristic of a rule-based machine learner is the identification and utilization of a set of relational rules that collectively represent the knowledge captured by the system. This is in contrast to other machine learners that commonly identify a singular model that can be universally applied to any instance in order to make a prediction. Rule-based machine learning approaches include learning classifier systems, association rule learning, and artificial immune systems.

 <b>Marwadi</b> University	<b>Marwadi University</b> <b>Faculty of Technology</b> <b>Department of Information and Communication Technology</b>	
<b>Subject: Artificial Intelligence(01CT0616)</b>	<b>Aim: To learn the Basics of Machine Learning</b>	
<b>Experiment No: 01</b>	<b>Date:</b>	<b>Enrollment No: 92200133003</b>

### Feature selection approach

Feature selection is the process of selecting an optimal subset of relevant features for use in model construction. It is assumed the data contains some features that are either redundant or irrelevant, and can thus be removed to reduce calculation cost without incurring much loss of information. Common optimality criteria include accuracy, similarity and information measures.

### Pre Lab Exercise:

a. Write 5 applications, that you encounter daily, where Machine Learning is used?

---



---



---

b. What do you mean by Supervised Learning?

---



---



---

c. What do you mean by Unsupervised Learning?

---



---



---

d. What do you mean by Reinforcement Learning?

---



---




---

### Program (Code):

Perform the following tasks:

1. Load a dataset in your IDE

 <b>Marwadi University</b>	<b>Marwadi University</b> <b>Faculty of Technology</b> <b>Department of Information and Communication Technology</b>	
<b>Subject: Artificial Intelligence(01CT0616)</b>	<b>Aim: To learn the Basics of Machine Learning</b>	
<b>Experiment No: 01</b>	<b>Date:</b>	<b>Enrollment No: 92200133003</b>

+ Code

+ Text

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

df = pd.read\_csv('/content/googleplaystore\_v2.csv')

df

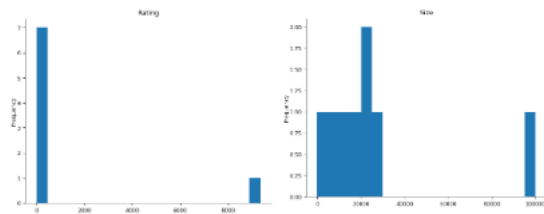
	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Last Updated	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159	19000.000000	10,000+	Free	0	Everyone	Art & Design	January 7, 2018	1.0.0	4.0.3 and up
1	Coloring book moana	ART_AND_DESIGN	3.9	967	14000.000000	500,000+	Free	0	Everyone	Art & Design;Pretend Play	January 15, 2018	2.0.0	4.0.3 and up
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510	8700.000000	5,000,000+	Free	0	Everyone	Art & Design	August 1, 2018	1.2.4	4.0.3 and up
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644	25000.000000	50,000,000+	Free	0	Teen	Art & Design	June 8, 2018	Varies with device	4.2 and up
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967	2800.000000	100,000+	Free	0	Everyone	Art & Design;Creativity	June 20, 2018	1.1	4.4 and up
...	...	...	...	...	...	...	...	...	...	...	...	...	...
10836	Syaâa Maroc - FR	FAMILY	4.5	38	53000.000000	5,000+	Free	0	Everyone	Education	July 25, 2017	1.48	4.1 and up
10837	Fr. Mike Schmitz Audio Teachings	FAMILY	5.0	4	3600.000000	100+	Free	0	Everyone	Education	July 6, 2018	1.0	4.1 and up
10838	Parkinson Exercices FR	MEDICAL	NaN	3	9500.000000	1,000+	Free	0	Everyone	Medical	January 20, 2017	1.0	2.2 and up
10839	The SCP Foundation DB fr nn5n	BOOKS_AND_REFERENCE	4.5	114	21516.529524	1,000+	Free	0	Mature 17+	Books & Reference	January 19, 2015	Varies with device	Varies with device
10840	iHoroscope - 2018 Daily Horoscope & ...	LIFESTYLE	4.5	108787	15000.000000	10,000,000+	Free	0	Everyone	Lifestyle	July 26, 2018	Varies with device	Varies with device

## 2. Observe the statistics of all the features

```
#Identify the statistic of dataset
df.describe()
```

	Rating	Size
count	9367.000000	10841.000000
mean	4.193338	21516.529524
std	0.537431	20746.537567
min	1.000000	8.500000
25%	4.000000	5900.000000
50%	4.300000	18000.000000
75%	4.500000	26000.000000
max	19.000000	100000.000000


### Distributions



### 2-d distributions

100000

## 3. Obtain the shape of the dataset

 <b>Marwadi University</b>	<b>Marwadi University</b> <b>Faculty of Technology</b> <b>Department of Information and Communication Technology</b>	
<b>Subject: Artificial Intelligence(01CT0616)</b>	<b>Aim: To learn the Basics of Machine Learning</b>	
<b>Experiment No: 01</b>	<b>Date:</b>	<b>Enrollment No: 92200133003</b>

```
# check shape of dataset
df.shape
```

```
(10841, 13)
```

#### 4. Separate all the features

```
df.info()
```


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10841 entries, 0 to 10840
Data columns (total 13 columns):
 #   Column              Non-Null Count  Dtype  
---  -
 0   App                 10841 non-null  object 
 1   Category            10841 non-null  object 
 2   Rating              9367 non-null   float64
 3   Reviews             10841 non-null  object 
 4   Size                10841 non-null  float64
 5   Installs            10841 non-null  object 
 6   Type                10840 non-null  object 
 7   Price               10841 non-null  object 
 8   Content Rating      10840 non-null  object 
 9   Genres              10841 non-null  object 
10   Last Updated        10841 non-null  object 
11   Current Ver         10833 non-null  object 
12   Android Ver         10838 non-null  object 
dtypes: float64(2), object(11)
memory usage: 1.1+ MB
```

#### 5. Fill the missing values, if any, using the statistically relevant value

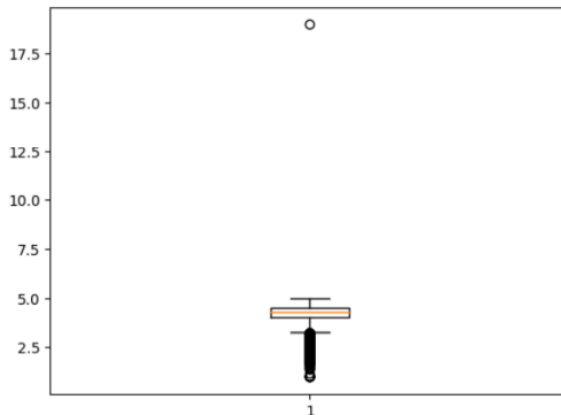
```
df.isnull().sum()
```

```
App                 0
Category            0
Rating              1474
Reviews             0
Size                0
Installs            0
Type                1
Price               0
Content Rating      1
Genres              0
Last Updated        0
Current Ver         8
Android Ver         3
dtype: int64
```

#### 6. Observe the Box-Plot of each feature

 <b>Marwadi University</b>	<b>Marwadi University</b> <b>Faculty of Technology</b> <b>Department of Information and Communication Technology</b>	
<b>Subject: Artificial Intelligence(01CT0616)</b>	<b>Aim: To learn the Basics of Machine Learning</b>	
<b>Experiment No: 01</b>	<b>Date:</b>	<b>Enrollment No: 92200133003</b>

```
plt.boxplot(df["Rating"])
```



7. Comment on the IQR range and outliers

### Results:

To be attached with

### Observation and Result Analysis:

a. Statistics of the dataset

---



---



---

b. IQR and Outliers

---




---



---

### Post Lab Exercise:

a. Write 5 applications, that you encounter daily, where Machine Learning is used?

 <b>Marwadi</b> University	<b>Marwadi University</b> <b>Faculty of Technology</b> <b>Department of Information and Communication Technology</b>	
<b>Subject: Artificial Intelligence(01CT0616)</b>	<b>Aim: To learn the Basics of Machine Learning</b>	
<b>Experiment No: 01</b>	<b>Date:</b>	<b>Enrollment No: 92200133003</b>

---



---



---

b. What do you mean by Supervised Learning?

---



---



---

c. What do you mean by Unsupervised Learning?

---



---



---

d. What do you mean by Reinforcement Learning?

---



---



---

e. Which is a better mean of replacing the missing value-mean, median or mode?

---



---



---

f. How IQR range helps to understand the nature of the distribution of the data?

---



---



---

g. Which are the ways to replace the categorical data to the numerical ones?

---



---



---


h. When can you replace the categorical data with (i) one hot encoded value and (ii) numerical equivalent continuous value?

---



---



 <b>Marwadi</b> University	<b>Marwadi University</b> <b>Faculty of Technology</b> <b>Department of Information and Communication Technology</b>	
<b>Subject: Artificial Intelligence(01CT0616)</b>	<b>Aim: To learn the Basics of Machine Learning</b>	
<b>Experiment No: 01</b>	<b>Date:</b>	<b>Enrollment No: 92200133003</b>

---