

**Report on  
Two Days Workshop on  
“Getting Started with Machine Learning”**

**25-26<sup>th</sup> July, 2021**

**Organized by  
ISA STUDENTS SECTION &  
INSTRUMENTATION AND CONTROL  
ENGINEERING DEPARTMENT,  
L.D COLLEGE OF ENGINEERING**

The pandemic has hastened the entire teaching-learning ecosystem, but it has also marked an increase in resilience in students and faculty community. All the people associated have tried a lot to bridge the gap between the online and offline education. It has showed us a new way of working and coping up with stuffs.

We at ISA believe the same and are towards a common goal of imparting quality insights in the field of control, instrumentation and automation as a whole. From time and again, we have been the poll-bearers of the field and will continue to do so. To start the academic year of ISA, we conducted an event that actually symbolized the new beginnings.

The event was on “Getting started with machine learning”. The machine learning, artificial intelligence has been quite a buzz recently. So, we took this opportunity to demystify this and present the topic for what it really was, especially to the sophomores of the college. It was a very informative session. The event was taken by Ayon Roy. He is an undergrad student which makes him a perfect person to introduce the subject since he too can relate to the experiences and the cries of the student. However, he has achieved a great feat even at this age with mentoring, judging and conducting 150+ data science related events and is really a data geek. Also, he has established the biggest Kaggle community in India.

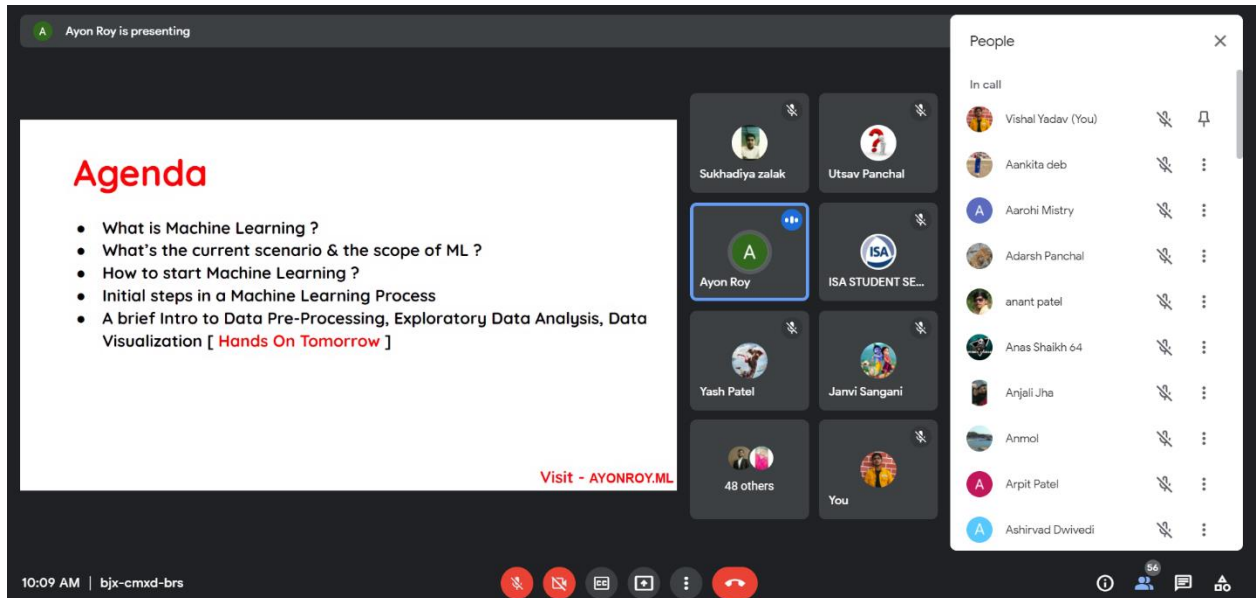
The event was organized on the MS-Teams Platform. We had 100+ registrations from the event from various department of the college. The participants also poured in from other colleges. The session was aimed mainly at the 1<sup>st</sup> and 2<sup>nd</sup> year students who still are not well introduced to the subject. The session would be very beneficial to them.

The event was hands-on workshop which was unfolded in a two-day scheme. Each session was of 1 hours. On the first day, the host Vishal Yadav gave a hearty introduction of Ayon Roy and then the session began. On the first day, the surface of machine learning was scratched with the introduction of the core definitions and explanation of all the jargon used. It gave a big picture view of Machine learning. The students enjoyed this session a lot. The second day of the session was really hands-on and the participants got their hands dirty using the actual tools used in creating machine learning models. The session concluded with the discussion of the steps ahead in the field. Additionally, at the end, best Dr. Manish Thakkar Sir also joined to give

guidance to the participants and expressing his gratitude at the event. The session ended with resolution of participant queries.

It was an altogether a successful event and marked a good start to the year. The feedback we got from the students was also great.

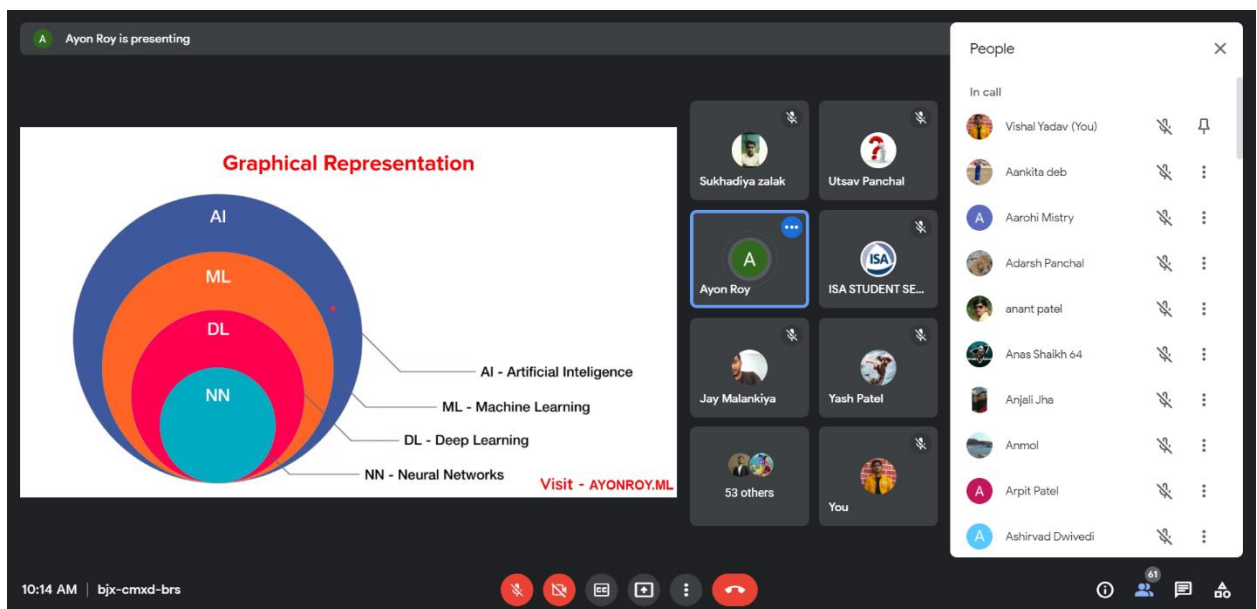
BELOW ARE THE GLIMPSES FROM THE EVENT –



A screenshot of a Zoom meeting interface. The main window displays a slide titled "Agenda" with the following bullet points:

- What is Machine Learning ?
- What's the current scenario & the scope of ML ?
- How to start Machine Learning ?
- Initial steps in a Machine Learning Process
- A brief Intro to Data Pre-Processing, Exploratory Data Analysis, Data Visualization [ **Hands On Tomorrow** ]

At the bottom right of the slide, it says "Visit - [AYONROY.ML](http://AYONROY.ML)". The Zoom interface shows a grid of participants, with "Ayon Roy" highlighted. A "People" list on the right shows 16 participants in the call. The bottom status bar shows the time as 10:09 AM and the meeting ID as bjk-cmxd-brs.



A screenshot of a Zoom meeting interface. The main window displays a slide titled "Graphical Representation" showing a Venn diagram with four overlapping circles representing AI, ML, DL, and NN. The circles are nested, with NN being the innermost and AI being the outermost. Labels with lines pointing to the circles are:

- AI - Artificial Intelligence
- ML - Machine Learning
- DL - Deep Learning
- NN - Neural Networks

At the bottom right of the slide, it says "Visit - [AYONROY.ML](http://AYONROY.ML)". The Zoom interface shows a grid of participants, with "Ayon Roy" highlighted. A "People" list on the right shows 16 participants in the call. The bottom status bar shows the time as 10:14 AM and the meeting ID as bjk-cmxd-brs.

Ayon Roy is presenting

## Renaming the Columns

```
# Renaming the column names
df = df.rename(columns={"Engine HP": "HP", "Engine Cylinders":
"CYLinders", "Transmission Type": "Transmission", "Driven_wheels":
"Drive Mode", "Highway MPG": "MPG-H", "City mpg": "MPG-C", "MSRP":
"Price" })
df.head(5)
```

Make	Model	Year	HP	Cylinders	Transmission	Drive Mode	MSRP	Price
BMW	1 Series M	2011	335.0	6.0	MANUAL	rear wheel drive	28	19.48130
BMW	1 Series	2011	300.0	6.0	MANUAL	rear wheel drive	28	19.40890
BMW	1 Series	2011	300.0	6.0	MANUAL	rear wheel drive	28	20.34590
BMW	1 Series	2011	230.0	6.0	MANUAL	rear wheel drive	28	19.29450
BMW	1 Series	2011	230.0	6.0	MANUAL	rear wheel drive	28	19.34550

Renaming the column names.

Visit - [AYONROY.ML](http://AYONROY.ML)

10:15 AM | bjx-cmxd-brs

Ayon Roy is presenting

## How to preprocess a mixed dataset for machine learning

What is preprocessing?  
Preprocessing describes the process of cleaning and converting a 'raw' (i.e. unprocessed) dataset into a clean dataset.

```
graph LR
  A[raw dataset] -- preprocessing --> B[clean dataset]
```

When given a dataset, the preprocessing can have various steps depending on  
a) what type of data you're looking at (text, images, time series...)  
b) what models you want to train.

In this notebook we will look at a dataset with both numerical and categorical attributes.

**Possible preprocessing steps**

As mentioned already, the preprocessing steps you will need for your dataset depend on the nature of the dataset and models you want to train. Possible preprocessing steps are:

TAPAN CHAVDA has left the meeting

10:16 AM | bjx-cmxd-brs

Ayon Roy is presenting

many attributes is 0. For the "pregnancies" attribute this makes sense. However, for attributes like "blood pressure" or "insulin" is doesn't. This indicates that certain values are missing.

The rows named 25%, 50% and 75% represent the corresponding *percentiles*. For example, 25% of the woman had less than 1 pregnancy, 50% had less than 3 pregnancies and 75% had less than 6 pregnancies.

### Visualizing aspects of the dataset

Another way to get familiar with the dataset is to look at all numerical attributes and plot a histogram for each of them.

What is a histogram?  
**A histogram is a graphical representation of data that uses bars of different heights.** It groups data points into value ranges. The exact shape of the value ranges depends on the number of bars. With a lot of bars, value ranges will be small, with only a few bars they will be wider. The height of a bar (y-axis) represents the number of data points that fall into the corresponding value range (x-axis).

Since our dataset has only numerical values we can simply call the `DataFrame.hist()` function (see section "Encoding categorical attributes" for more details on how to convert categorical to numerical attributes).

`DataFrame.hist()` calls `matplotlib.pyplot.hist()` on each series in the `DataFrame`, resulting in one histogram per column (i.e. one histogram per attribute).

The histograms will also make it easier to detect outliers or erroneous values (see section "Handling noisy data" for more details).

```
In [8]: df.hist(bins=50, figsize=(25, 20))
plt.show()
```

10:41 AM | bjk-cmxd-brs

Ayon Roy is presenting

```
In [11]: correlation_matrix = train_df.corr(method='pearson')
correlation_matrix
```

Out[11]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI
Pregnancies	1.000000	0.141610	0.124869	-0.090113	-0.074484	0.003832
Glucose	0.141610	1.000000	0.138061	0.055264	0.323325	0.210250
BloodPressure	0.124869	0.138061	1.000000	0.220584	0.094920	0.291821
SkinThickness	-0.090113	0.055264	0.220584	1.000000	0.441852	0.415479
Insulin	-0.074484	0.323325	0.094920	0.441852	1.000000	0.215458
BMI	0.003832	0.210250	0.291821	0.415479	0.215458	1.000000
DiabetesPedigreeFunction	0.004180	0.165979	0.086961	0.195016	0.155286	0.151589
Age	0.523424	0.280893	0.246683	-0.130566	-0.046646	0.017820
Outcome	0.245579	0.465803	0.060411	0.082470	0.136964	0.273198

Let's take a look at how each attribute correlates with the final diagnosis

10:45 AM | bjk-cmxd-brs