جامعة المستقبل
AL MUSTAQBAL UNIVERSITY

قسم الانظمة الطبية الذكية

المرحلة الثالثة

Third lecture

Subject: Artificial Intelligence AII

Class: Third

Lecturers: Dr. Muneera Abed Hadi, M. Sc. Rouaa Safi , M. Sc. Ansam Ali

University of Information Technology
and Communications
College of Medical Informatics
Intelligent Medical Systems Department

Machine Learning Course

Lecture two

By: Dr. Muneera Abed Hmdi

Data Processing and Visualization

- ❑ ***Data*** is the foundation of machine learning. The quality and quantity of data you have directly impact the performance of your machine learning models. In this section, we will explore various aspects of data and its processing, which are crucial for building robust ML systems.

- ❑ ***Data processing and transformation***

- ➢ ***Data processing*** is a crucial step in the machine learning (ML) pipeline, as it prepares the data for use in building and training models. The goal of data processing is to clean, transform, and prepare the data in a format that is suitable for modeling.

- ➢ ***Data transformation*** is the process of converting, cleansing, and structuring data into a usable format that can be analyzed to support decision making processes. Data transformation is used when data needs to be converted to match that of the destination system.

The main steps of data processing

- ❑ ***Data collection:*** This is the process of gathering data from various sources, such as sensors, databases, or other systems. The data may be structured or unstructured, and may come in various formats such as text, images (Facial Expression Recognizer, needs numerous images having a variety of human expressions), or audio.

- ❑ ***Data preprocessing:*** This step involves cleaning, filtering, and transforming the data to make it suitable for further analysis. This may include removing missing values, scaling or normalizing the data, or converting it to a different format image (n*n matrix).

- ❑ ***Data analysis:*** The data is analyzed using various techniques such as statistical analysis, machine learning algorithms, or data visualization. The goal of this step is to derive insights or knowledge from the data.

- ❑ ***Data interpretation:*** This step involves interpreting the results of the data analysis and drawing conclusions based on the insights gained. It may also involve presenting the findings in a clear and concise manner, such as through reports, dashboards, or other visualizations.

The main steps involved in data processing

- ❑ ***Data storage and management:*** Once the data has been processed and analyzed, it must be stored and managed in a way that is secure and easily accessible. This may involve storing the data in a database, cloud storage, or other systems, and implementing backup and recovery strategies to protect against data loss.

- ❑ ***Data visualization and reporting:*** The results of the data analysis are presented to stakeholders in a format that is easily understandable and actionable. This may involve creating visualizations, reports, or dashboards that highlight key findings and trends in the data.

Note: There are many tools and libraries available for data processing in ML, including ***pandas*** for Python, and the Data Transformation and Cleansing tool in ***RapidMiner***. The choice of tools will depend on the specific requirements of the project, including ***the size and complexity of the data and the desired outcome***.

Advantages of Data processing in ML

- ❑ ***Improved model performance:*** Helps improve the performance of the ML model by cleaning and transforming the data into a format that is suitable for modeling.

- ❑ ***Better representation of the data:*** Allows the data to be transformed into a format that better represents the underlying relationships and patterns in the data, making it easier for the ML model to learn from the data.

- ❑ ***Increased accuracy:*** Helps ensure that the data is accurate, consistent, and free of errors, which can help improve the accuracy of the ML model.

Disadvantages of Data Processing in ML

- ❑ ***Time-consuming:*** Data processing can be a time-consuming task, especially for large and complex datasets.

- ❑ ***Error-prone:*** Data processing can be error-prone, as it involves transforming and cleaning the data, which can result in the loss of important information or the introduction of new errors.

- ❑ ***Limited understanding of the data:*** Data processing can lead to a limited understanding of the data, as the transformed data may not be representative of the underlying relationships and patterns in the data.

Types of Data in Machine Learning

❑ ***Machine learning algorithms*** use data to learn patterns and relationships between input variables and target outputs, which can then be used for prediction or classification tasks.

❑ ***Data is typically divided into two types:***

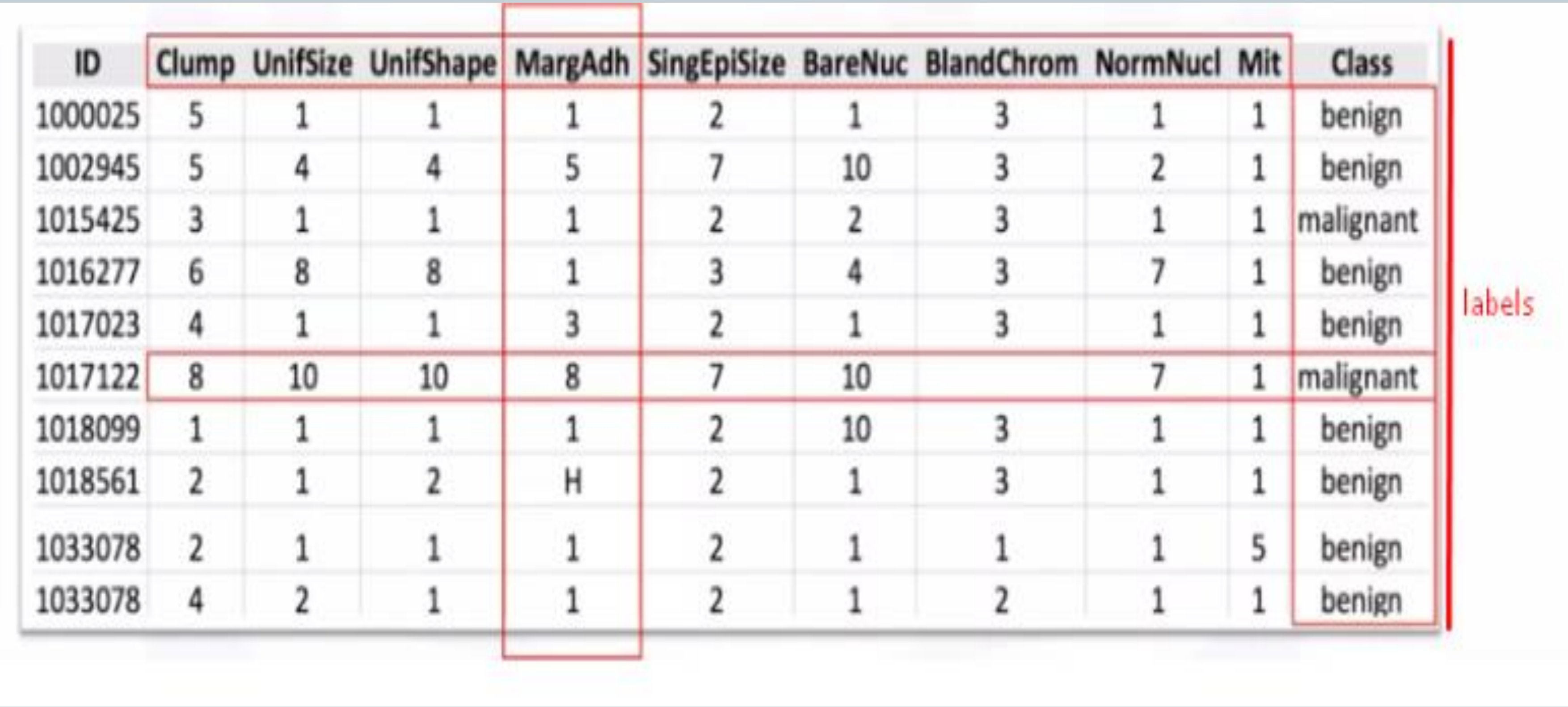
1. Labeled data

2. Unlabeled data

❑ ***Labeled data*** includes a label or target variable that the model is trying to predict, whereas ***unlabeled data*** does not include a label or target variable. The data used in machine learning is typically numerical or categorical. ***Numerical data*** includes values that can be ordered and measured, such as age or income. ***Categorical data*** includes values that represent categories, such as gender or type of fruit.

Types of Data in Machine Learning

- ***Labeled data,*** used by ***Supervised learning*** add meaningful tags or labels or class to the observations (or rows).

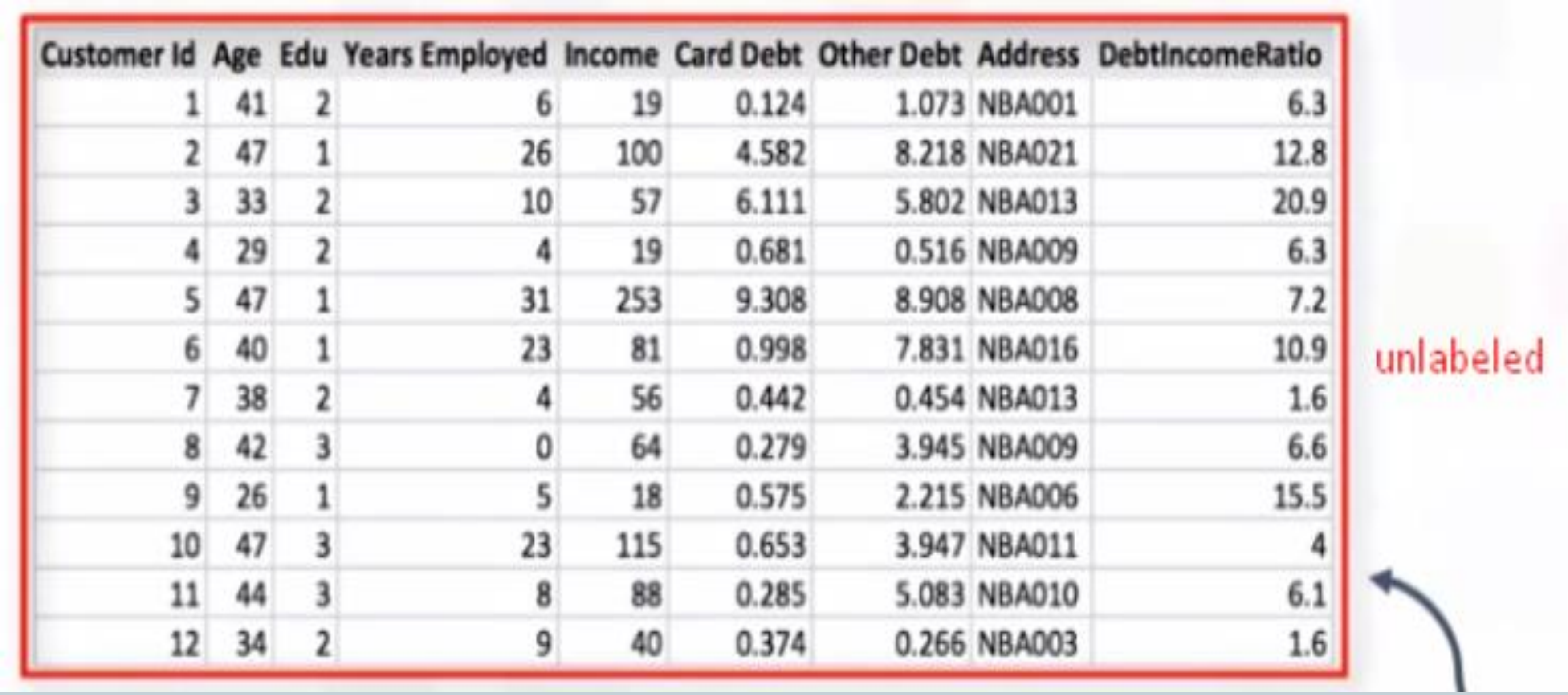


ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

labels

Types of Data in Machine Learning

- ❑ ***Unlabeled data,*** used by ***Unsupervised learning*** however do not have any meaningful tags or labels associated with it.



Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio
1	41	2	6	19	0.124	1.073	NBA001	6.3
2	47	1	26	100	4.582	8.218	NBA021	12.8
3	33	2	10	57	6.111	5.802	NBA013	20.9
4	29	2	4	19	0.681	0.516	NBA009	6.3
5	47	1	31	253	9.308	8.908	NBA008	7.2
6	40	1	23	81	0.998	7.831	NBA016	10.9
7	38	2	4	56	0.442	0.454	NBA013	1.6
8	42	3	0	64	0.279	3.945	NBA009	6.6
9	26	1	5	18	0.575	2.215	NBA006	15.5
10	47	3	23	115	0.653	3.947	NBA011	4
11	44	3	8	88	0.285	5.083	NBA010	6.1
12	34	2	9	40	0.374	0.266	NBA003	1.6

unlabeled

Introduction to Machine Learning: What Is and Its Applications

- ❑ ***Data splitting:*** Data can be divided into ***training and testing sets***. The training set is used to ***train*** the model, and the testing set is used to ***evaluate*** the performance of the model. It is important to ensure that the data is split in a random and representative way.

- ❑ ***Data*** is the most important part of all data analytics, machine learning, and artificial intelligence***. Without data***, we can't train any model, and all modern research and automation will go in vain. ***Big Enterprises*** are spending lots of money just to gather as much certain data as possible.

Introduction to Machine Learning: What Is and Its Applications

- ❑ ***Data preprocessing*** is an important step in the machine learning pipeline. This step can include ***cleaning and normalizing the data, handling missing values, and feature selection or engineering.***

- ❑ ***DATA:*** It can be any unprocessed ***fact, value, text, sound, or picture that is not being interpreted and analyzed***. Data is the most important part of all data analytics, machine learning, and artificial intelligence.

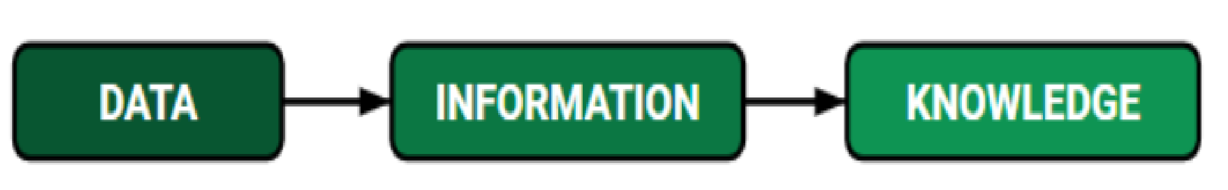
Introduction to Machine Learning: What Is and Its Applications

- ❑ ***Example:*** **Why did Facebook acquire WhatsApp by paying a huge price of $19 billion?**

- ➢ The answer is very simple and logical – it is to have access to the users' information that Facebook may not have but WhatsApp will have. This information about their users is of paramount importance to Facebook as it will facilitate the task of improvement in their services.

- ➢ ***INFORMATION*****:** Data that has been interpreted and manipulated and has now some meaningful inference for the users.

- ➢ ***KNOWLEDGE:*** Combination of inferred information, experiences, learning, and insights. Results in awareness or concept building for an individual or organization.



Introduction to Machine Learning: What Is and Its Applications

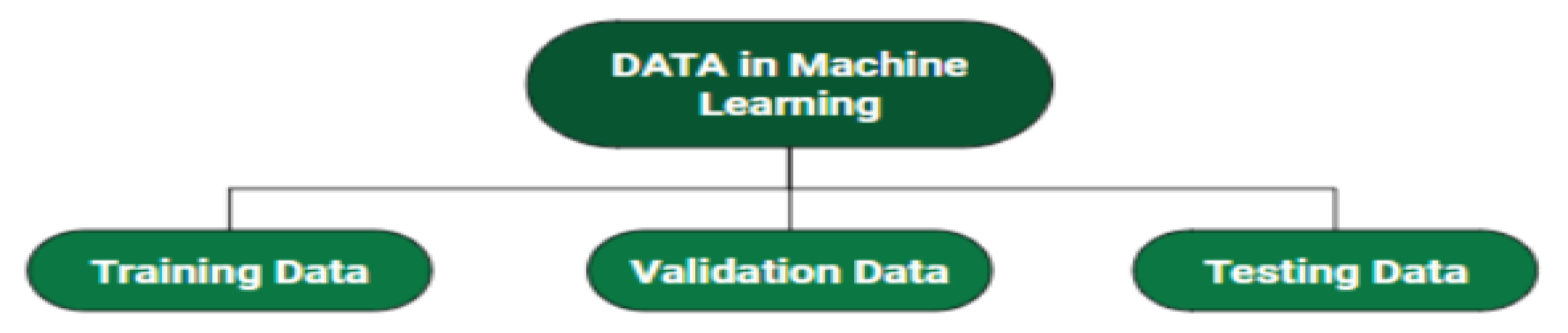
❑ *How do we split data in Machine Learning?*

- ***Training Data:*** The part of data we use to train our model. This is the data that your model sees (both input and output) and learns from.

- ***Validation Data:*** The part of data that is used to do a frequent evaluation of the model, fit on the training dataset along with improving involved hyperparameters (initially set parameters before the model begins learning). This data plays its part when the model is training.

- ***Testing Data:*** Once our model is completely trained, testing data provides an unbiased evaluation. When we feed in the inputs of Testing data, our model will predict some values (without seeing actual output). After prediction, we evaluate our model by comparing it with the actual output present in the testing data. This is how we evaluate and see how much our model has learned from the experiences feed in as training data, set at the time of training.

Data in Machine Learning



Example: There’s a Shopping Mart Owner who conducted a survey for which he has a long list of questions and answers that he had asked from the customers, this list of questions and answers is DATA. Now every time when he wants to infer anything and can’t just go through not question of thousands of customers to find something relevant as it would be time-consuming and not helpful. To reduce this overhead and time wastage and to make work easier, data is manipulated through software, calculations, graphs, etc. This inference from manipulated data is Information. So, Data is a must for Information. Now Knowledge has its role in differentiating between two individuals having the same information. Knowledge is not technical content but is linked to the human thought process.

Data Categorization

❑ *Different Forms of Data*

- ➢ **Numeric Data:** Numeric features are types of data that consist of numbers, which can be computed mathematically with various standard operators such as add, minus, multiply, divide and more.

- ➢ ***Examples of Numeric Data*** are examination marks, height, weight, the number of students in a class, price of goods, monthly bills, fees and others.

- ➢ **Categorical Data:** A categorical feature is an attribute that can take on one of the limited , and usually fixed number of possible values based on some qualitative property . A categorical feature is also called a nominal feature.

Data Categorization

❑ *Different Forms of Data*

- ***Examples of categorical data*** are hair color (red, blonde, or black), political affiliation (republican, democrat, or other), and gender (male, female, or other). These are considered categorical because one is not more than the other.

- **Ordinal Data:** This denotes a nominal variable with categories falling in an ordered list . Examples include clothing sizes such as small, medium , and large , or a measurement of customer satisfaction on a scale from “not at all happy” to “very happy”.

Properties of Data

- ***Volume (scale of data):*** With the growing world population and technology at exposure, huge data is being generated each millisecond.
- ***Variety of data:*** Different forms of data – healthcare, images, videos, audio clippings.
- ***Speed of data:*** Rate of data streaming and generation.
- ***Value of data:*** Meaningfulness of data in terms of information that researchers can infer from it.
- ***Veracity of data:*** Certainty and correctness in data we are working on.

Properties of Data

- ***Viability of data:*** The ability of data to be used and integrated into different systems and processes.
- ***Security of data:*** The measures taken to protect data from unauthorized access or manipulation.
- ***Accessibility of data:*** The ease of obtaining and utilizing data for decision-making purposes.
- ***Integrity of data:*** The accuracy and completeness of data over its entire lifecycle.
- ***Usability of data:*** The ease of use and interpretability of data for end-users.

Data Visualization

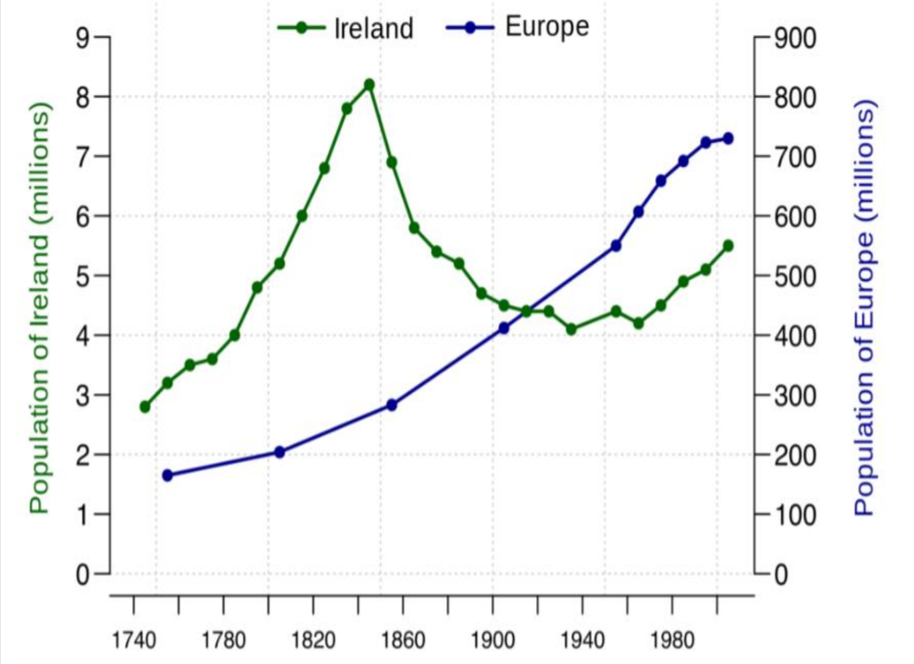
- ❑ ***Data Visualization*** is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

- ❑ ***For example,*** data visualization can be used to check the quality, distribution, and outliers of your data; explore the relationships between features; compare the performance of different models; and understand how a model works.

How do you Visualize data in machine learning?

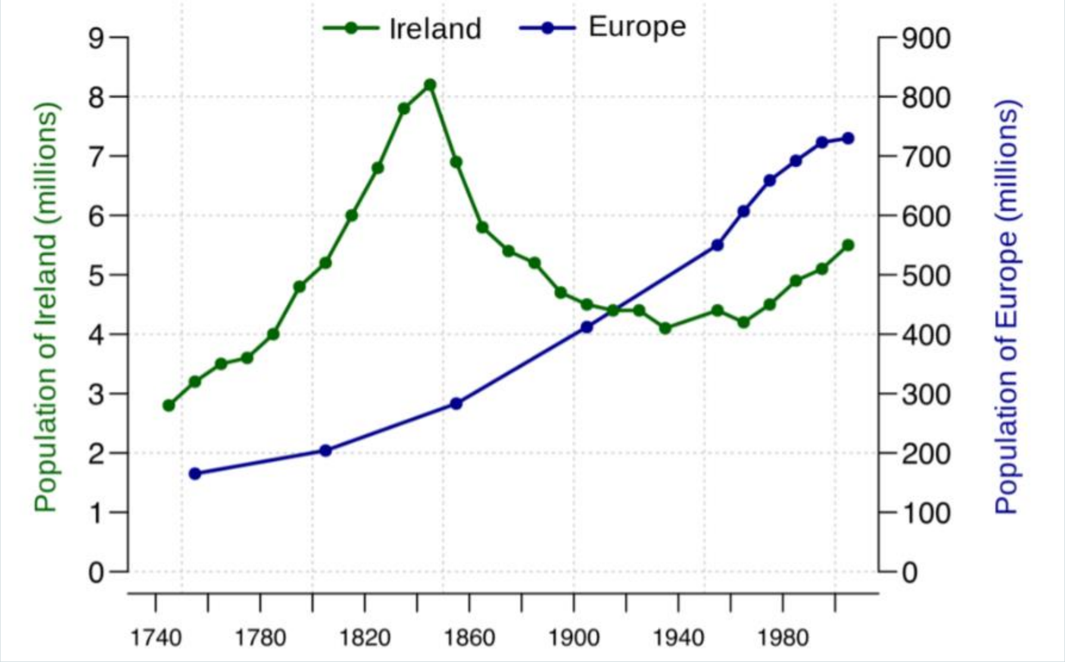
- ❑ ***Line Charts*** *is the simplest way to represent time series data. It is intuitive, easy to create, and helps the viewer get a quick sense of how something has changed over time.* Each data point is represented by a point on the graph, and these points are connected by a line.

- ❑ A line graph shows ***how a dependent variable and independent variable changed***. An independent variable remains ***unaffected*** by other parameters, whereas the dependent variable depends on how the independent variable ***changes***. Time is always the independent variable, which is plotted on the horizontal axis and the dependent variable is plotted on the vertical axis.



How do you Visualize data in machine learning?

- ❑ ***Example,*** in this line graph, the populations of Europe and Ireland are the dependent variables, and time is the independent variable. It clearly highlights the sudden drop in Ireland’s population in the 1840s. History books will tell you this was the result of the devastating Irish, a period of mass starvation, disease, and emigration in Ireland between 1845 and 1852.

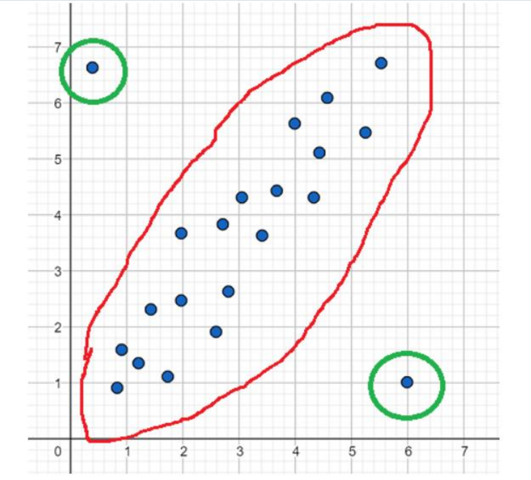


How do you Visualize data in machine learning?

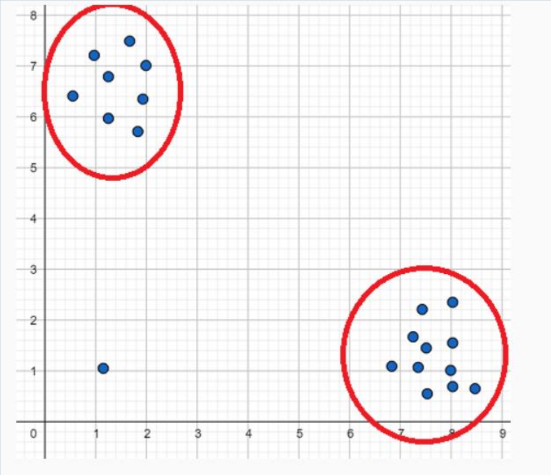
- ***Scatter Plots:*** A quick and efficient method of displaying the relationship between two variables is to use scatter plots. With one variable plotted on the x-axis and the other variable drawn on the y-axis, each data point in a scatter plot is represented by a point on the graph. We may use scatter plots to visualize data to find patterns, clusters, and outliers.

There are two points that do not fit the pattern

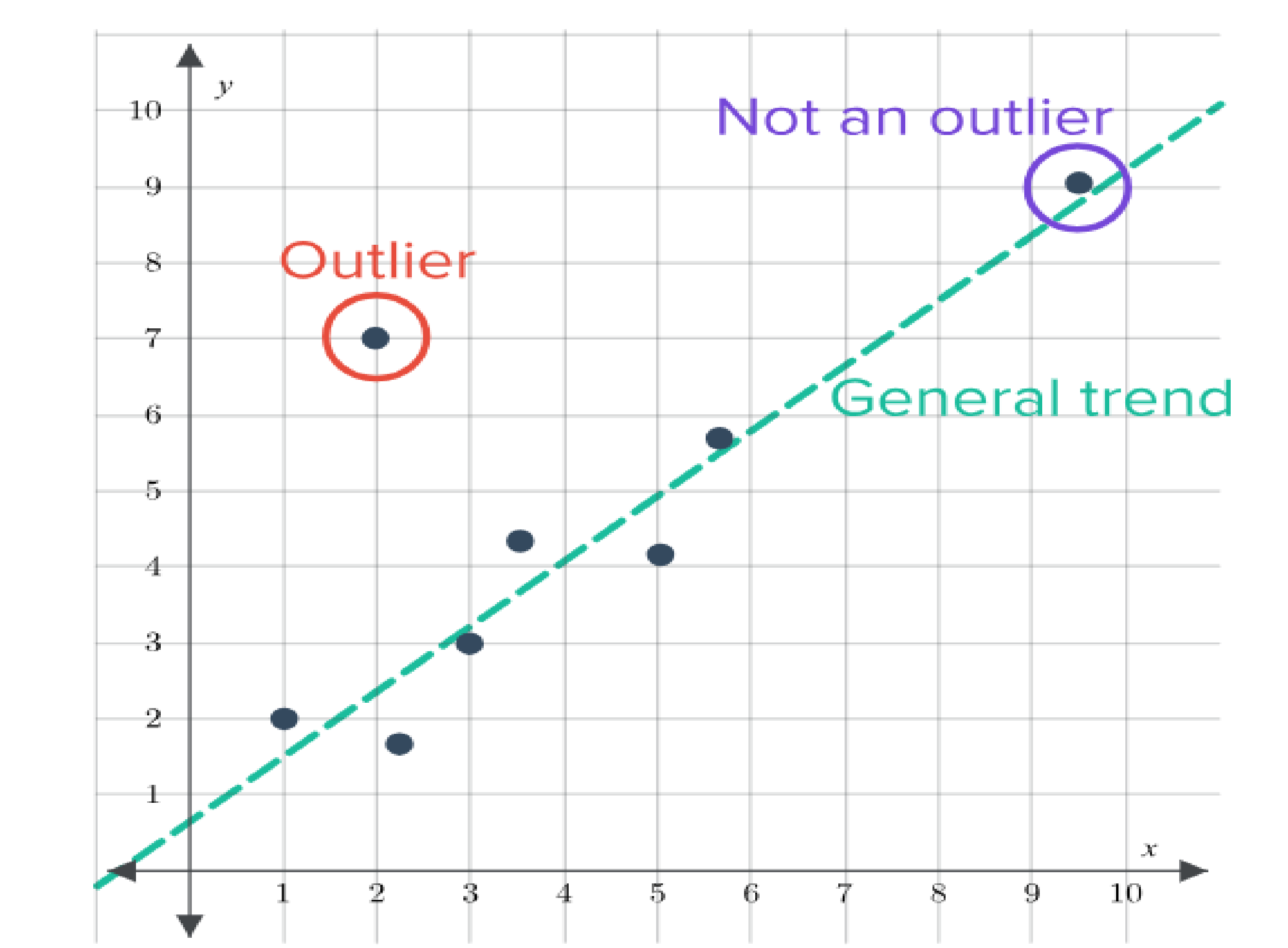
There are two clusters



linear pattern

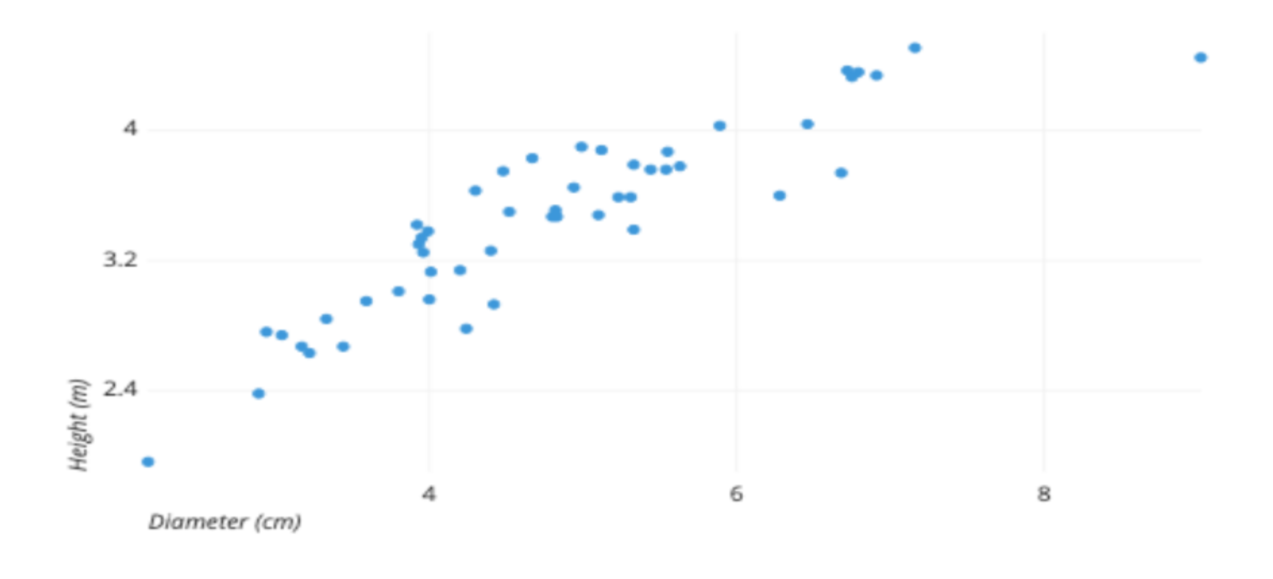


Scatter Plots Example



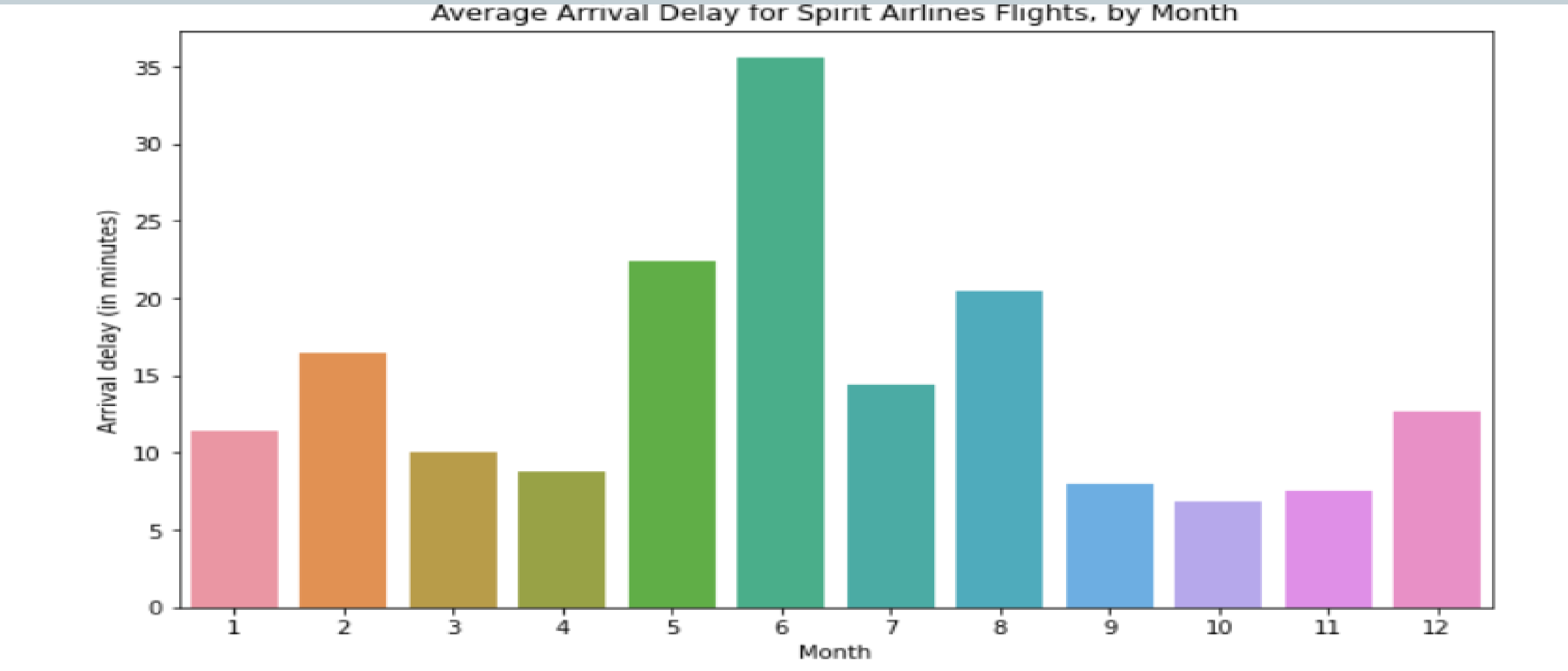
How do you Visualize data in machine learning?

❑ ***Scatter Plots:*** The example scatter plot above shows the diameters and heights for a sample of fictional trees. Each dot represents a single tree; each point’s horizontal position indicates that tree’s diameter (in centimeters) and the vertical position indicates that tree’s height (in meters). From the plot, we can see a generally tight positive correlation between a tree’s diameter and its height. We can also observe an outlier point, a tree that has a much larger diameter than the others. This tree appears short for its girth, which might warrant further investigation.



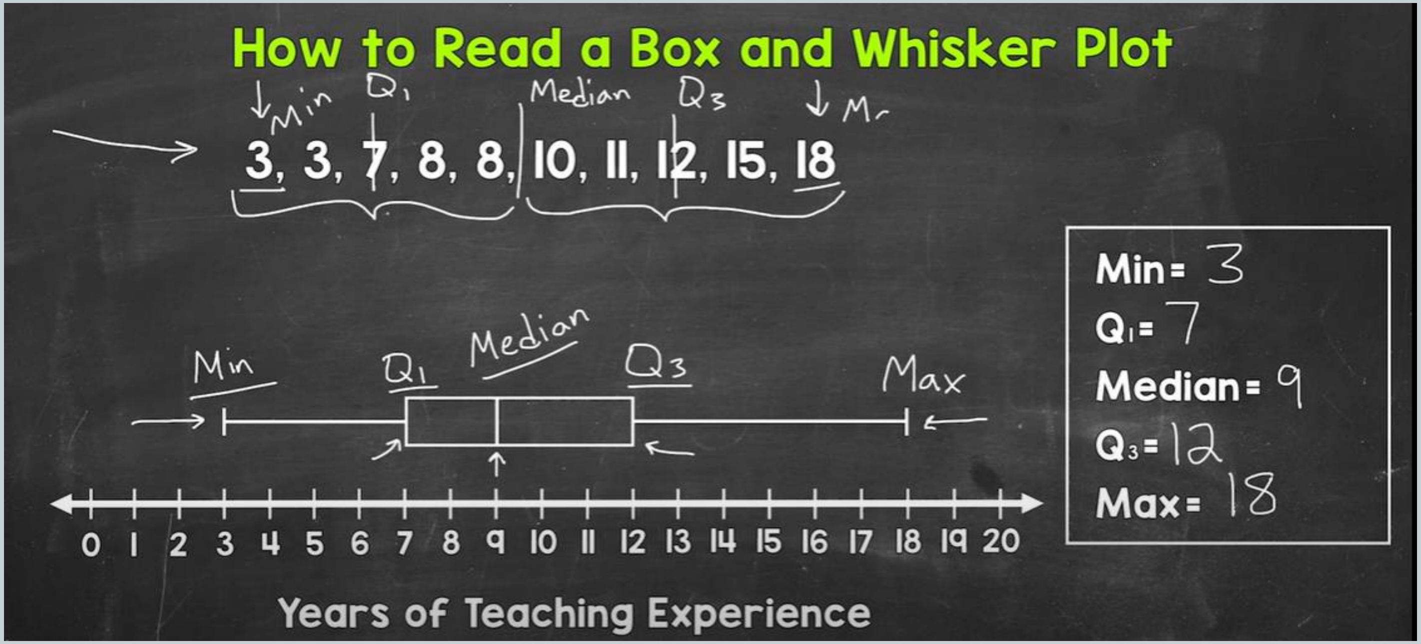
How do you Visualize data in machine learning?

- ***Bar Charts:*** Bar charts are a common way of displaying categorical data. In a bar chart, each category is represented by a bar, with the height of the bar indicating the frequency or proportion of that category in the data. Bar graphs are useful for comparing several categories and seeing patterns over time.



How do you Visualize data in machine learning?

- ***Box Plots:*** Box plots are a graphical representation of the distribution of a set of data. In a box plot, the median is shown by a line inside the box, while the center box depicts the range of the data. The whiskers extend from the box to the highest and lowest values in the data, excluding outliers. Box plots can help us to identify the spread and skewness of the data.



Advantages of using data in Machine Learning:

- ❑ ***Improved accuracy:*** With large amounts of data, machine learning algorithms can learn more complex relationships between inputs and outputs, leading to improved accuracy in predictions and classifications.

- ❑ ***Automation:*** Machine learning models can automate decision-making processes and can perform repetitive tasks more efficiently and accurately than humans.

- ❑ ***Personalization:*** With the use of data, machine learning algorithms can personalize experiences for individual users, leading to increased user satisfaction.

- ❑ ***Cost savings:*** Automation through machine learning can result in cost savings for businesses by reducing the need for manual work and increasing efficiency.

Disadvantages of using data in Machine Learning:

- ***Bias:*** Data used for training machine learning models can be biased, leading to biased predictions and classifications.

- ***Privacy:*** Collection and storage of data for machine learning can raise privacy concerns and can lead to security risks if the data is not properly secured.

- ***Quality of data:*** The quality of data used for training machine learning models is critical to the performance of the model. Poor quality data can lead to inaccurate predictions and classifications.

- ***Lack of interpretability:*** Some machine learning models can be complex and difficult to interpret, making it challenging to understand how they are making decisions.

Use of Machine Learning

- ***Machine learning*** is a powerful tool that can be used in a wide range of applications. Here are some of the most common uses of machine learning:

- ***Predictive modeling:*** Machine learning can be used to build predictive models that can predict future outcomes based on historical data. This can be used in many applications, such as stock market prediction, fraud detection, weather forecasting, and customer behavior prediction.

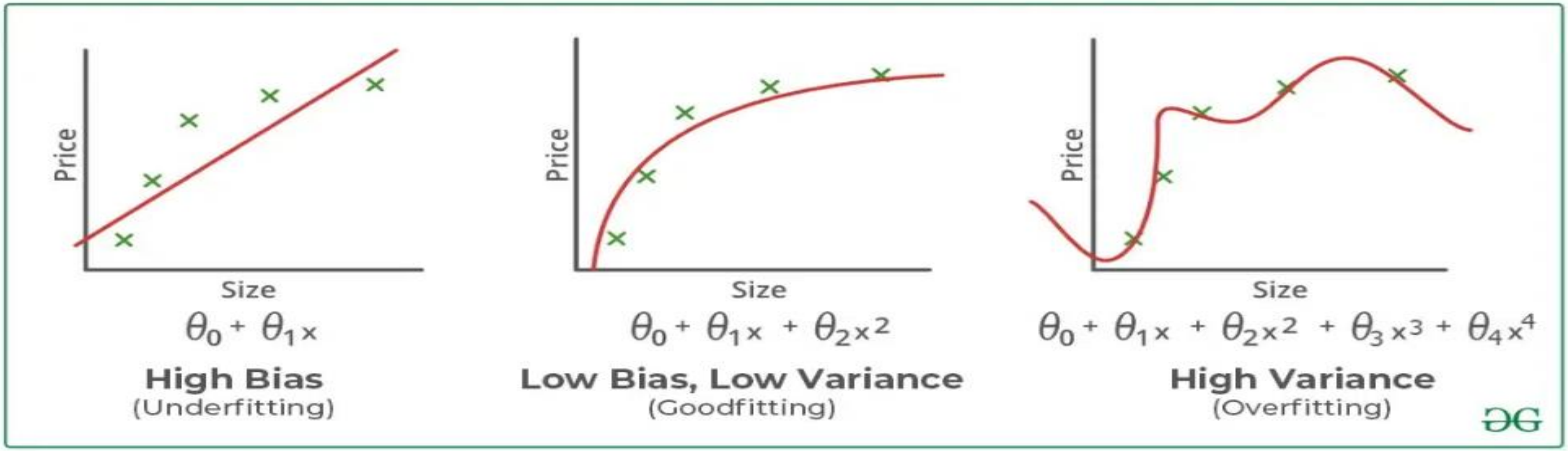
- ***Image recognition:*** Machine learning can be used to train models that can recognize objects, faces, and other patterns in images. This is used in many applications, such as self-driving cars, facial recognition systems, and medical image analysis.

Bias and Variance in Machine Learning

- ***Bias:*** Bias refers to the error due to overly simplistic assumptions in the learning algorithm. These assumptions make the model easier to comprehend and learn but might not capture the underlying complexities of the data. It is the error due to the model's inability to represent the true relationship between input and output accurately. ***When a model has poor performance both on the training and testing data means high bias because of the simple model, indicating underfitting.***

- ***Variance:*** Variance is the error due to the model's sensitivity to variations in the training data. It's the variability of the model's predictions for different instances of training data. High variance occurs when a model learns the training data's noise and random changes rather than the underlying pattern. As a result, ***the model performs well on the training data but poorly on the testing data, indicating overfitting.***

Bias and Variance in Machine Learning



- ❑ ***Simple model – high bias / indicating underfitting***
- ❑ ***Complex model – high variants / indicating overfitting***

Issues of using data in Machine Learning:

- ***Data quality:*** One of the biggest issues with using data in machine learning is ensuring that the data is accurate, complete, and representative of the problem domain. Low-quality data can result in inaccurate or biased models.

- ***Data quantity:*** In some cases, there may not be enough data available to train an accurate machine learning model. This is especially true for complex problems that require a large amount of data to accurately capture all the relevant patterns and relationships.

- ***Bias and fairness:*** Machine learning models can sometimes produce bias and discrimination results if the training data is biased or unrepresentative. Unrepresentative data occurs when the training or fine-tuning data is not sufficiently representative of the underlying population or does not measure the phenomenon of interest. This can lead to unfair outcomes for certain groups of people, such as minorities or women.

Issues of using data in Machine Learning:

- ***Overfitting and underfitting:*** Overfitting occurs when a model is too complex and fits the training data too closely, resulting in poor generalization to new data. Underfitting occurs when a model is too simple and does not capture all the relevant patterns in the data.

- ***Privacy and security:*** Machine learning models can sometimes be used to infer sensitive information about individuals or organizations, raising concerns about privacy and security.

- ***Interpretability:*** Some machine learning models, such as deep neural networks, can be difficult to interpret and understand, making it challenging to explain the reasoning behind their predictions and decisions.