AiLABS

ARTIFICIAL INTELLIGENCE WITHOUT BORDERS

What is Statistics?

Statistics is a group of methods that are used to **collect**, **organize**, **present**, **analyze**, and **interpret** data to make decisions.

Collection refers to the gathering of information or data.

Organization or **presentation** involves summarizing data or information in textual, graphical, or tabular forms.

Analysis involves describing the data by using statistical methods and procedures.

Interpretation refers to the process of making conclusions based on the analyzed data.

Types of Statistics

1. **Descriptive -** is a statistical procedure concerned with **describing** the **characteristics** and **properties** of a group of persons, places, or things. Involves gathering, organizing, presenting, and describing data.
   Ex - How many students are interested to take Statistics online?
   What are the highest and lowest scores obtained by STENEX applicants this year?
1. **Inferential -** is a statistical procedure that is used to draw inferences or information about the properties or characteristics by a large group of people, places, or things on the basis of the information obtained from a small portion of a large group also called inductive reasoning or inductive statistics.
   Ex - Exit polls in Election, regression

Random Variables

Variables are any entity which holds some value.

Types of Random variables:

1. **Discrete Variables** - is one that can assume a finite number of values. In other words, it can assume specific values only. The values of a discrete variable are obtained through the process of counting.

   Example: the number of chairs in a room

1. **Continuous Variables -** A variable that can assume **any numerical value** over a certain **range** or interval.

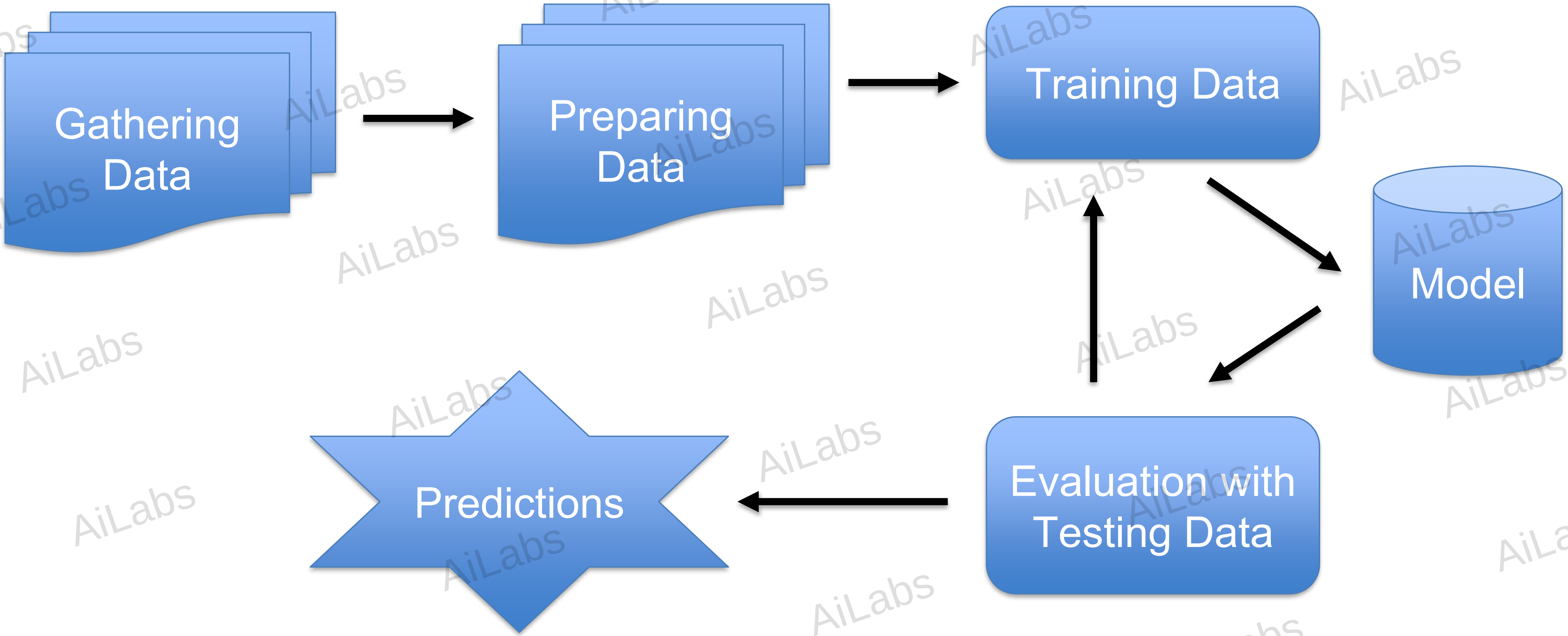
   Example: The height of a person.

1. **Dependent Variable -** which is affected or influenced by another variable.
2. **Independent Variables -** is one which affects or influences the dependent variable.

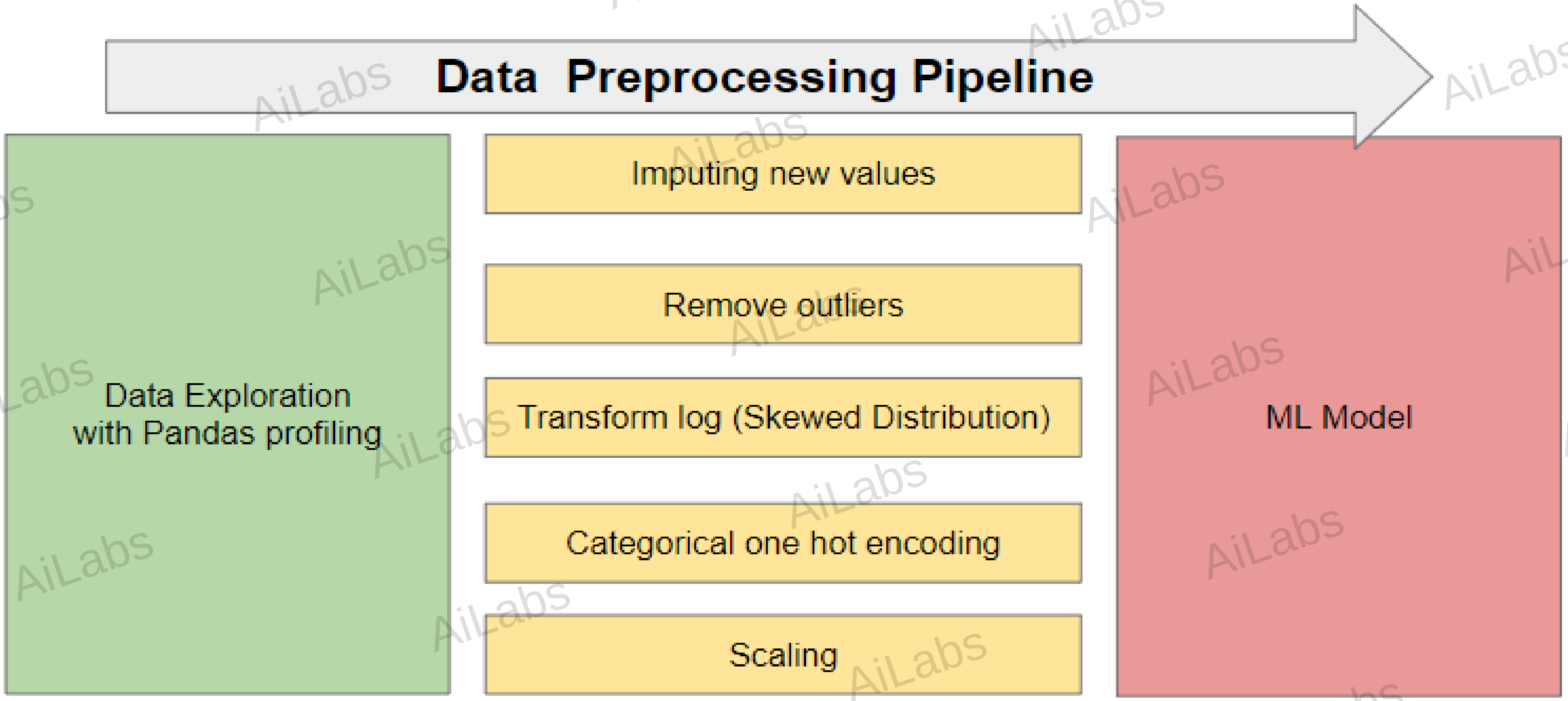
Examples of Data

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Loan ID	Customer	Loan Statu	Current Lc	Term	Credit Sco	Annual In	Years in cu	Home Ow	Purpose	Monthly D	Years of C	Months si	Number o	Number o	Current Cr	Maximum	Bankruptc	Tax Liens
2	14dd8831-	981165ec-	Fully Paid	445412	Short Terr	709	1167493	8 years	Home Mo	Home Imp	5214.74	17.2	NA	6	1	228190	416746	1	0
3	4771cc26-:	2de017a3-	Fully Paid	262328	Short Term			10+ years	Home Mo	Debt Cons	33295.98	21.1	8	35	0	229976	850784	0	0
4	4eed4e6a	5efb2b2b-	Fully Paid	99999999	Short Terr	741	2231892	8 years	Own Hom	Debt Cons	29200.53	14.9	29	18	1	297996	750090	0	0
5	77598f7b-	e777faab-	Fully Paid	347666	Long Terr	721	806949	3 years	Own Hom	Debt Cons	8741.9	12	NA	9	0	256329	386958	0	0
6	d4062e70-	81536ad9-	Fully Paid	176220	Short Term			5 years	Rent	Debt Cons	20639.7	6.1	NA	15	0	253460	427174	0	0
7	89d8cb0c-	4ffe99d3-	Charged C	206602	Short Terr	7290	896857	10+ years	Home Mo	Debt Cons	16367.74	17.3	NA	6	0	215308	272448	0	0
8	273581de-	90a75dde-	Fully Paid	217646	Short Terr	730	1184194	< 1 year	Home Mo	Debt Cons	10855.08	19.6	10	13	1	122170	272052	1	0
9	db0dc6e1-	018973c9-	Charged C	648714	Long Term			< 1 year	Home Mo	Buy House	14806.13	8.2	8	15	0	193306	864204	0	0
10	8af915d9-	af534dea-	Fully Paid	548746	Short Terr	678	2559110	2 years	Rent	Debt Cons	18660.28	22.6	33	4	0	437171	555038	0	0
11	0b1c4e3d-	235c4a43-	Fully Paid	215952	Short Terr	739	1454735	< 1 year	Rent	Debt Cons	39277.75	13.9	NA	20	0	669560	1021460	0	0
12	32c2e48f-:	0de7bcdb	Fully Paid	99999999	Short Terr	728	714628	3 years	Rent	Debt Cons	11851.06	16	76	16	0	203965	289784	0	0
13	fa096848-(	aa0a6a22-	Fully Paid	541970	Short Term			10+ years	Home Mo	Home Imp	23568.55	23.2	NA	23	0	60705	1634468	0	0
14	403d7235-	11581f68-(	Fully Paid	99999999	Short Terr	740	776188	< 1 year	Own Hom	Debt Cons	11578.22	8.5	25	6	0	134083	220220	0	0
15	01d878ae-	900c9191-	Fully Paid	99999999	Short Terr	743	1560907	4 years	Rent	Debt Cons	17560.37	13.3	NA	10	1	225549	496474	1	0
16	2e841c8f-:	2ac05980-	Fully Paid	234124	Short Terr	727	693234	10+ years	Rent	Debt Cons	14211.24	24.7	46	10	1	28291	107052	1	0
17	7cbaa3fa-:	3ec886e7-	Fully Paid	449020	Long Term			9 years	Own Hom	Debt Cons	18904.81	19.4	NA	8	0	334533	428956	0	0
18	c9a16a9d-	abb4c446-	Charged C	653004	Long Term			7 years	Home Mo	Debt Cons	14537.09	20.5	NA	9	0	302309	413754	0	0
19	24e8c8bd-	967e8733-	Fully Paid	666204	Long Terr	723	1821967	10+ years	Home Mo	Debt Cons	17612.24	22	34	15	0	813694	2004618	0	0
20	c6be21f0-	c67b2cb5-	Fully Paid	66396	Short Term			10+ years	Rent	Debt Cons	9898.81	27.1	NA	23	1	9728	402380	1	0
21	41f7dd8d-	422f9b72-	Fully Paid	390390	Short Terr	747	1791738	8 years	Home Mo	Home Imp	2478.55	22.7	NA	6	0	121182	801812	0	0
22	150ebbad	40f729c9-!	Charged C	317108	Long Terr	687	1133274	8 years	Rent	Debt Cons	9632.81	17.4	53	4	0	60287	126940	0	0
23	31ae42f6-	016c5139-	Fully Paid	128238	Short Terr	750	1354073	< 1 year	Rent	Debt Cons	13202.15	11.9	NA	7	0	131936	458788	0	0
24	c7e2b784-	5b53e176-	Charged C	153252	Short Terr	714	1890690	2 years	Rent	Debt Cons	21900.35	15.7	NA	12	0	891594	1081014	0	0

Machine Learning Life Cycle



Data Preprocessing Pipeline



How data to be analyzed should look like

Number of times pregnant	Glucose Concentration	Blood Pressure (in mm Hg)	Skin thickness (in mm)	Insulin Concentration (in μU/ml)	BMI	Diabetes Pedigree Function	Age	Out-come
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1

Instead data looks like:

Number of times pregnant	Glucose Concentration	Blood Pressure (in mm Hg)	Skin thickness (in mm)	Insulin Concentration (in µU/ml)	BMI	Diabetes Pedigree Function	Age	Out-come
	148	72/90	35	?	33.6	0.627	-50	1
1	85	66	29		26.6	0.351	31	0
		64	NA	0		0.672	32	1
1	89	66	NA	94	28.1	0.167	-21	0
0	137	40	35	-168		2.288	33	1
5		74	0	0	25.6	0.201	350	0
3	78	135/110	32	88	31	0.248	26	1

Raw Data

	A	B	C	D	E	F	G	H
1	age	job	marital	education	balance	housing_loan	personal_loan	term_deposit_subscription
2	58	management	married	tertiary	2143	yes	no	yes
3	44	technician	single	secondary	29	yes	no	no
4	33	entrepreneur	married	secondary	2	yes	yes	no
5	47	blue-collar	married	unknown	1506	yes	no	no
6	-33	unknown	single	unknown		no	no	no
7	35	management	married		231	yes	no	no
8	33	services	married	secondary	0	yes	no	no
9	28	management	single	tertiary	447	yes	yes	yes

- Some values are **missing**
- Some values are **inconsistent**
- Some columns contain **text entries**

Data Wrangling/Preprocessing

- Refers to set of techniques for resolving several issues such as missing values, correcting values, or cleaning up a dataset
- An essential and integral part of Machine Learning
- Involves transforming raw data into an understandable, process able format
- Prepares raw data to best expose the structure of the problem to the machine learning task
- Process varies from data-to-data and model-to-model

Usual outcomes of a data preprocessing pipeline

- No missing values
- Data from multiple sources have been merged
- All the data is in numerical format
- Outliers have been removed
- Categorical data have been handled
- Data has been scaled
- Dimensionality Reduction

Basics Preprocessing techniques

1. Handling missing data
2. Outlier Detection
3. Categorical Feature handling
4. Feature Scaling
5. Dimensionality Reduction

Missing values

	age	job	marital	education	balance	housing_loan	personal_loan	term_deposit_subscription
0	58	management	married	tertiary	2143.0	yes	no	yes
1	44	technician	single	secondary	29.0	yes	no	no
2	33	entrepreneur	married	secondary	2.0	yes	yes	no
3	47	blue-collar	married	unknown	1506.0	yes	no	no
4	-33	unknown	single	unknown	NaN	no	no	no
5	35	management	married	NaN	231.0	yes	no	no
6	33	services	married	secondary	0.0	yes	no	no
7	28	management	single	tertiary	447.0	yes	yes	yes
8	42	entrepreneur	divorced	tertiary	2.0	yes	no	no
9	58	retired	married	primary	121.0	yes	no	no

- Missing values are listed as ‘**NaN**’
- ‘NaN’ stands for ‘Not a number’
- 0 and ‘NaN’ are **different**

Missing Data Handling

Missing values are representative of the messiness of real world data.

There can be a multitude of reasons why they occur — ranging from human errors during data entry, incorrect sensor readings, to software bugs in the data processing pipeline.

Methods to handle them:

1. **Removal** - it is appropriate only when the proportion of missing data <10%, else you are going to lose a ton of data.
2. **Replacement -** it is one of the most easiest yet dangerous methods, as these values can be misleading. Should be only used when absolutely necessary.
3. **Statistical measure - r**eplace the missing values with mean, median or mode. For numerical values you should go with mean, and if there are some outliers try median

Missing Data

4. **Interpolate -** there are various interpolation methods.

5. Temporal filling – Backward or forward filing, part of the replacement method

6. Categorical values - In case of categorical values being missing, its best to replace them with the most frequent one.

Measure of Central Tendency

Mean - is given by the total of the values of the samples divided by the number of samples.

$$\bar{x} = \frac{1}{n}\Sigma x_i$$

Median - Median is the data point that lies exactly in the centre when the data is sorted in increasing or decreasing order.

$$1,\ 3,\ 3,\ \mathbf{6},\ 7,\ 8,\ 9$$

$$\text{Median} = \underline{\underline{6}}$$

$$1,\ 2,\ 3,\ \mathbf{4},\ \mathbf{5},\ 6,\ 8,\ 9$$

$$\begin{aligned}\text{Median} &= (4+5) \div 2 \\ &= \underline{\underline{4.5}}\end{aligned}$$

Mode - Mode represents the most common value in a data set.

Dealing with numerical missing values

	age	job	marital	education	balance	housing_loan	personal_loan	term_deposit_subscription
0	58	management	married	tertiary	2143.0	yes	no	yes
1	44	technician	single	secondary	29.0	yes	no	no
2	33	entrepreneur	married	secondary	2.0	yes	yes	no
3	47	blue-collar	married	unknown	1506.0	yes	no	no
4	-33	unknown	single	unknown	NaN	no	no	no
5	35	management	married	NaN	231.0	yes	no	no
6	33	services	married	secondary	0.0	yes	no	no
7	28	management	single	tertiary	447.0	yes	yes	yes
8	42	entrepreneur	divorced	tertiary	2.0	yes	no	no
9	58	retired	married	primary	121.0	yes	no	no
10	43	technician	single	secondary	NaN	yes	no	no
11	41	admin.	divorced	secondary	270.0	yes	no	no

- Numerical missing values can be filled in many different ways
- We will discuss two ways:

1) filling with mean
2) filling with median

Filling numerical missing values with mean

	age	job	marital	education	balance	housing_loan	personal_loan	term_deposit_subscription
0	58	management	married	tertiary	2143.000000	yes	no	yes
1	44	technician	single	secondary	29.000000	yes	no	no
2	33	entrepreneur	married	secondary	2.000000	yes	yes	no
3	47	blue-collar	married	unknown	1506.000000	yes	no	no
4	-33	unknown	single	unknown	1362.344253	no	no	no
5	35	management	married	NaN	231.000000	yes	no	no
6	33	services	married	secondary	0.000000	yes	no	no
7	28	management	single	tertiary	447.000000	yes	yes	yes
8	42	entrepreneur	divorced	tertiary	2.000000	yes	no	no
9	58	retired	married	primary	121.000000	yes	no	no
10	43	technician	single	secondary	1362.344253	yes	no	no
11	41	admin.	divorced	secondary	270.000000	yes	no	no
12	-29	admin.	single	secondary	390.000000	yes	no	no
13	53	technician	married	secondary	6.000000	yes	no	yes
14	58	technician	married	NaN	71.000000	yes	no	no
15	57	services	married	secondary	162.000000	yes	no	no
16	51	retired	married	primary	1362.344253	yes	no	no

Filling numerical missing values with median

	age	job	marital	education	balance	housing_loan	personal_loan	term_deposit_subscription
0	58	management	married	tertiary	2143.0	yes	no	yes
1	44	technician	single	secondary	29.0	yes	no	no
2	33	entrepreneur	married	secondary	2.0	yes	yes	no
3	47	blue-collar	married	unknown	1506.0	yes	no	no
4	-33	unknown	single	unknown	448.0	no	no	no
5	35	management	married	NaN	231.0	yes	no	no
6	33	services	married	secondary	0.0	yes	no	no
7	28	management	single	tertiary	447.0	yes	yes	yes
8	42	entrepreneur	divorced	tertiary	2.0	yes	no	no
9	58	retired	married	primary	121.0	yes	no	no
10	43	technician	single	secondary	448.0	yes	no	no
11	41	admin.	divorced	secondary	270.0	yes	no	no
12	-29	admin.	single	secondary	390.0	yes	no	no
13	53	technician	married	secondary	6.0	yes	no	yes
14	58	technician	married	NaN	71.0	yes	no	no
15	57	services	married	secondary	162.0	yes	no	no
16	51	retired	married	primary	448.0	yes	no	no

Median is Robust to Outliers

Name	Monthly Income ($)
Rob	5000
Rafiq	6000
Nina	4000
Sofia	7500
Mohan	8000
Tao	7000
Average	6250

Name	Monthly Income ($)
Rob	5000
Rafiq	6000
Nina	4000
Sofia	7500
Mohan	8000
Tao	7000
Elon Musk	10 million
Average	1.43 million

Name	Monthly Income ($)	Credit Score	Approve Loan?
Rob	5000	650	No
Rafiq	6000	400	No
Nina	4000	780	Yes
Sofia	1.6 million	810	Yes
Mohan	8000	410	No
Tao	7000	850	Yes
Elon Musk	10 million	880	Yes

Name	Monthly Income ($)	Credit Score	Approve Loan?
Rob	5000	650	No
Rafiq	6000	400	No
Nina	4000	780	Yes
Sofia	6500	810	Yes
Mohan	8000	410	No
Tao	7000	850	Yes
Elon Musk	10 million	880	Yes

4000 5000 6000 7000 8000 10 million

Presence of outliers can misguide the missing data.

Dealing with categorical missing values

	age	job	marital	education	balance	housing_loan	personal_loan	term_deposit_subscription
0	58	management	married	tertiary	2143.0	yes	no	yes
1	44	technician	single	secondary	29.0	yes	no	no
2	33	entrepreneur	married	secondary	2.0	yes	yes	no
3	47	blue-collar	married	unknown	1506.0	yes	no	no
4	-33	unknown	single	unknown	448.0	no	no	no
5	35	management	married	NaN	231.0	yes	no	no
6	33	services	married	secondary	0.0	yes	no	no
7	28	management	single	tertiary	447.0	yes	yes	yes
8	42	entrepreneur	divorced	tertiary	2.0	yes	no	no
9	58	retired	married	primary	121.0	yes	no	no
10	43	technician	single	secondary	448.0	yes	no	no
11	41	admin.	divorced	secondary	270.0	yes	no	no

- Filling categorical missing values can be tricky
- We generally fill with highest frequency value

Outliers Detection

Outliers are extreme values that deviate from other observations on data , they may indicate a variability in a measurement, experimental errors or a novelty.

Outliers are generally defined as samples that are exceptionally far from the mainstream of data.

Therefore, Outlier Detection may be defined as the process of detecting and subsequently excluding outliers from a given set of data.

Variance and Standard Deviation

Variance: Measures how far a **set of numbers is spread out.** A variance of zero indicates that all the values are identical. Variance is always **non-negative**, a **small variance** indicates that the data points tend to be **very close to the mean** (Average value), while a **high variance** indicates that the data points are **very spread out** around the mean.

Standard Deviation: The standard deviation gives an idea of **how close the entire set of data is to the average value.** Data sets with a small standard deviation have tightly grouped. Data sets with large standard deviations have data spread out over a wide range of values.

Formulae

Variance = Var(n) ≈ $\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu\right)^2$

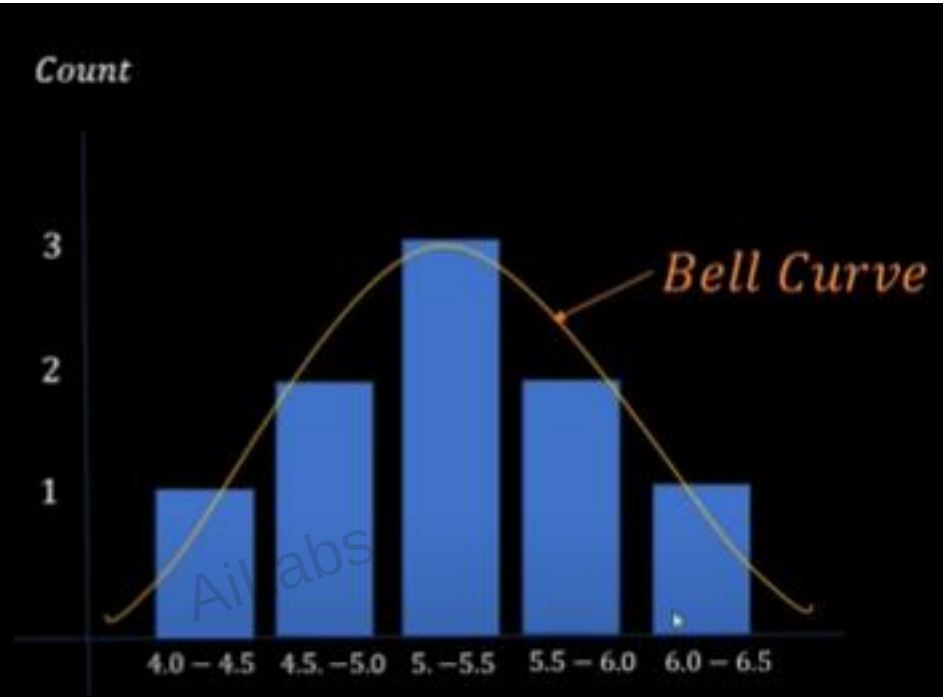
μ is the mean, each observation is subtracted from the mean, there difference is squared and then summed together. Then we take the average of that sum.

Standard Deviation = $\sigma = \sqrt{Var(n)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu\right)^2}$

Normal Distribution

The most well-known continuous distribution is **Normal Distribution,** which is also known as the **Gaussian distribution** or the “Bell Curve.”

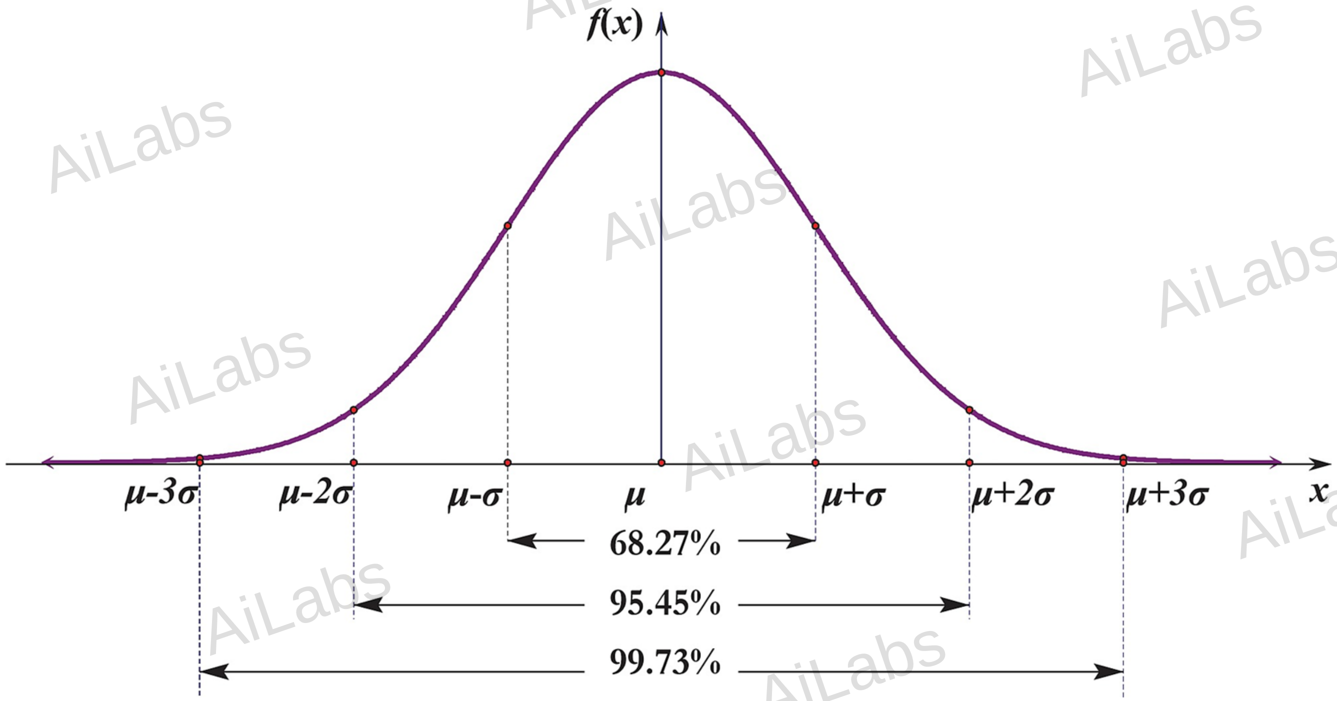
Name	Height (ft)
Rob	6.2
Thomas	5.7
Nina	4.6
Mittal	5.4
Sofia	5.9
Mohan	4.3
Tao	5.1
Deepika	5.2
Rafiq	4.9



Ex - heights of people, price of apartments in an area, Test score, Employee performance.. All these follows normal distribution.

How Normal Distribution is used in machine learning?

Outlier removal: It is estimated that, generally data beyond 3σ are outliers(for smaller dataset, its 2σ).



Method for Outliers Detection

1. **Using standard deviation**
2. **Using Visualization**
3. **Numeric Outlier** - The outliers are calculated by means of the IQR (InterQuartile Range). For example, the first and the third quartile (Q1, Q3) are calculated. An outlier is then a data point xi that lies outside the interquartile range.

Percentile -

For example - if a value is 25th percentile of data set that means 25% of data in the dataset is below that value.

minimum value has 0th percentile because all values are greater.

Method for Outliers Detection

IQR - Q3-Q1 (Q3 = 75th percentile, Q1 = 25th percentile)

Lower limit = Q1 - 1.5*IQR
Upper limit = Q3 + 1.5*IQR

So, any value less than lower limit , and more than upper limit is outlier.

Method for Outliers Detection

mohan	1.2
maria	2.3
sakib	4.9
tao	5.1
virat	5.2

khusbu	5.4
dmitry	5.5
selena	5.5
john	5.6
imran	5.6

jose	5.8
deepika	5.9
yoseph	6
binod	6.1
gulshan	6.2

johnson	6.5
donald	7.1
aamir	14.5
ken	23.2
Liu	40.2

1.2	5.35	5.75	6.27	40.2
min	Q1	Q2	Q3	max
	25th Percentile	50th Percentile	75th Percentile	

IQR = Q3 – Q1
= 6.27 – 5.35
= 0.925

lower_limit = Q1 – 1.5*IQR
= 3.96

upper_limit = Q3 + 1.5*IQR
= 7.66

Handling Categorical data

In machine learning, we often encounter some instance which generally include different categories or levels associated with the observation, which are non-numerical and thus need to be converted so the computer can process them.

These features are typically stored as text values which represent various traits of the observations. For example, gender is described as Male (M) or Female (F), product type could be described as electronics, apparels, food etc.

Dealing with text entries (categorical data)

	age	job	marital	education	balance	housing_loan	personal_loan	term_deposit_subscription
0	58	management	married	tertiary	2143.0	yes	no	yes
1	44	technician	single	secondary	29.0	yes	no	no
2	33	entrepreneur	married	secondary	2.0	yes	yes	no
3	47	blue-collar	married	unknown	1506.0	yes	no	no
4	33	unknown	single	unknown	448.0	no	no	no
5	35	management	married	secondary	231.0	yes	no	no
6	33	services	married	secondary	0.0	yes	no	no
7	28	management	single	tertiary	447.0	yes	yes	yes
8	42	entrepreneur	divorced	tertiary	2.0	yes	no	no
9	58	retired	married	primary	121.0	yes	no	no
10	43	technician	single	secondary	448.0	yes	no	no
11	41	admin.	divorced	secondary	270.0	yes	no	no
12	29	admin.	single	secondary	390.0	yes	no	no
13	53	technician	married	secondary	6.0	yes	no	yes
14	58	technician	married	secondary	71.0	yes	no	no
15	57	services	married	secondary	162.0	yes	no	no
16	51	retired	married	primary	448.0	yes	no	no

Types of categorical data

- Nominal – These are variables which are not related to each other in any order such as colour (black, blue, green).
- Ordinal – These are variables where a certain order can be found between them such as student grades (A, B, C, D, Fail).

Label encoding

One of the **simplest** and **most common solutions** advertised to transform categorical variables is **Label Encoding.** It consists of **substituting** each group with a **corresponding numb er** and keeping such numbering consistent throughout the feature.

Categorical Feature	Label Encoding
United States	1
United States	1
France	2
Germany	3
United Kingdom	4
France	2

Label encoding

Disadvantage - Numbers hold relationships. For instance, four is twice two, and, when converting categories into numbers directly, these relationships are created despite not existing between the original categories.
Looking at the example, United Kingdom becomes twice France, and France plus United States equals Germany.

They are best used for ordinal data/categories.

Dealing with categorical data(Non-Ordinal)

- Do these data values have an ‘order’?

NO!!

- We cannot say ‘married’ > ‘single’ or ‘technician’<‘services’

- How to deal with this?

One-hot Encoding!!!

	age	job	marital	education	balance	housing_loan	personal_loan	term_deposit_subscription
0	58	management	married	tertiary	2143.0	1	0	1
1	44	technician	single	secondary	29.0	1	0	0
2	33	entrepreneur	married	secondary	2.0	1	1	0
3	47	blue-collar	married	unknown	1506.0	1	0	0
4	33	unknown	single	unknown	448.0	0	0	0
5	35	management	married	secondary	231.0	1	0	0
6	33	services	married	secondary	0.0	1	0	0
7	28	management	single	tertiary	447.0	1	1	1
8	42	entrepreneur	divorced	tertiary	2.0	1	0	0
9	58	retired	married	primary	121.0	1	0	0
10	43	technician	single	secondary	448.0	1	0	0
11	41	admin.	divorced	secondary	270.0	1	0	0
12	29	admin.	single	secondary	390.0	1	0	0
13	53	technician	married	secondary	6.0	1	0	1
14	58	technician	married	secondary	71.0	1	0	0
15	57	services	married	secondary	162.0	1	0	0
16	51	retired	married	primary	448.0	1	0	0

One-hot encoding categorical data

	age	job	marital	education
0	58	management	married	tertiary
1	44	technician	single	secondary
2	33	entrepreneur	married	secondary
3	47	blue-collar	married	unknown
4	33	unknown	single	unknown
5	35	management	married	secondary
6	33	services	married	secondary
7	28	management	single	tertiary
8	42	entrepreneur	divorced	tertiary
9	58	retired	married	primary
10	43	technician	single	secondary
11	41	admin.	divorced	secondary
12	29	admin.	single	secondary
13	53	technician	married	secondary
14	58	technician	married	secondary
15	57	services	married	secondary
16	51	retired	married	primary

	age	job	divorced	married	single	education
0	58	management	0	1	0	tertiary
1	44	technician	0	0	1	secondary
2	33	entrepreneur	0	1	0	secondary
3	47	blue-collar	0	1	0	unknown
4	33	unknown	0	0	1	unknown
5	35	management	0	1	0	secondary
6	33	services	0	1	0	secondary
7	28	management	0	0	1	tertiary
8	42	entrepreneur	1	0	0	tertiary
9	58	retired	0	1	0	primary
10	43	technician	0	0	1	secondary
11	41	admin.	1	0	0	secondary
12	29	admin.	0	0	1	secondary
13	53	technician	0	1	0	secondary
14	58	technician	0	1	0	secondary
15	57	services	0	1	0	secondary
16	51	retired	0	1	0	primary

One hot encoding

One-Hot Encoding is the most common, correct way to deal with non-ordinal categorical data. It consists of creating an additional feature for each group of the categorical feature and mark each observation belonging (Value=1) or not (Value=0) to that group.

Minor drawbacks - Increased dimensionality, sparse dataset.

United States	France	Germany	United Kingdom
1	0	0	0
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1
0	1	0	0

Target Encoding

It consists of substituting each group in a categorical feature with the average response in the target variable.
Target Encoding is a powerful solution also because it avoids generating a high number of features, as is the case for One-Hot Encoding, keeping the dimensionality of the dataset as the original one.

For example, if the categories of categorical feature are red, blue and green. Then replace red with **mean of all the target labels** where-ever the feature value is red in training data.

Target Encoding

For example - United states have 5 target variable = 1,0,1,0,0
Mean of all target variable for group “United States’ = (1+0+1+0+0)/5 = 0.40

Country	Target Variable	Target Encoding
United States	1	0.40
Germany	0	0.50
United States	0	0.40
United States	1	0.40
France	1	0.67
Germany	1	0.50
United States	0	0.40
France	1	0.67
United States	0	0.40
France	0	0.67

Feature scaling

- Real-world data comprises of attributes with varying scales.
- Varying scales may cause a model to 'prioritize' one attribute over another.
- Necessary to rescale the attributes so that they all have similar scale

Feature scaling

- In our data, ‘age’ and ‘balance’ have different scales

- We scale them by subtracting the mean and dividing by the standard deviation

age	balance
58	2143.0
44	29.0
33	2.0
47	1506.0
33	448.0
35	231.0
33	0.0
28	447.0
42	2.0
58	121.0
43	448.0
41	270.0
29	390.0
53	6.0
58	71.0
57	162.0
51	448.0

Feature scaling

- Data is comprised of attributes with varying scales.
- Rescale the attributes so that all have the same scale.
- Often referred to as normalization; attributes are rescaled into the range between 0 and 1
- Feature scaling is the statistical operation of using values of features to scale themselves to smaller and similar ranges.
- This is done to equalize the influence of all input features to the machine learning model by scaling them to similar ranges. As machine only understands number and not relationship between them.
- Feature scaling is useful in:
  - optimization algorithms used in the core of machine learning algorithms like gradient descent
  - algorithms that weight inputs like regression and neural networks
  - algorithms that use distance measures like k-Nearest Neighbors

Types of feature scaling

1. **Standardization**
   - a useful technique to transform attributes with a Gaussian distribution and differing means and standard deviations to a standard Gaussian distribution with a mean of 0 and a standard deviation of 1.
   - This technique is more suitable for methods:
     - that assume Gaussian distribution in the input variables
     - designed for discrete data.

Data standardization

Number of times pregnant	Glucose Concentration	Blood Pressure (in mm Hg)	Skin thickness (in mm)	Insulin Concentration (in µU/ml)	BMI	Diabetes Pedigree Function	Age
0.64	0.848	0.15	0.907	-0.693	0.204	0.468	1.426
-0.845	-1.123	-0.161	0.531	-0.693	-0.684	-0.365	-0.191
1.234	1.944	-0.264	-1.288	-0.693	-1.103	0.604	-0.106
-0.845	-0.998	-0.161	0.155	0.123	-0.494	-0.921	-1.042
-1.142	0.504	-1.505	0.907	0.766	1.41	5.485	-0.02

Types of feature scaling

2. **Normalization**
   - Normalization refers to rescaling each observation (row) to have a length of 1.
   - This technique can be useful for sparse datasets (lots of zeros) with attributes of varying scales:
     - when using algorithms that weight input values such as neural networks or
     - with algorithms that use distance measures such as k-Nearest Neighbors

Data normalization

Number of times pregnant	Glucose Concentration	Blood Pressure (in mm Hg)	Skin thickness (in mm)	Insulin Concentration (in µU/ml)	BMI	Diabetes Pedigree Function	Age
0.034	0.828	0.403	0.196	0.	0.188	0.004	0.28
0.008	0.716	0.556	0.244	0.	0.224	0.003	0.261
0.04	0.924	0.323	0.	0.	0.118	0.003	0.162
0.007	0.588	0.436	0.152	0.622	0.186	0.001	0.139
0.	0.596	0.174	0.152	0.731	0.188	0.01	0.144

Extract Transform Load (ETL)

EXTRACT

Retrieve data from homogeneous or heterogeneous sources in its raw state.

TRANSFORM

Conversion of extracted data into proper storage format or structure by using rules or lookup tables or by combining the data with other data.

LOAD

The load phase loads the data into the end target, which can be any data store including a simple delimited flat file or a data warehouse, in this case, an ML algorithm.

Features and Labels

Features					Label
Position	Experience	Skill	Country	City	Salary ($)
Developer	0	1	USA	New York	103100
Developer	1	1	USA	New York	104900
Developer	2	1	USA	New York	106800
Developer	3	1	USA	New York	108700
Developer	4	1	USA	New York	110400
Developer	5	1	USA	New York	112300
Developer	6	1	USA	New York	114200
Developer	7	1	USA	New York	116100
Developer	8	1	USA	New York	117800
Developer	9	1	USA	New York	119700
Developer	10	1	USA	New York	121600

Features and Labels

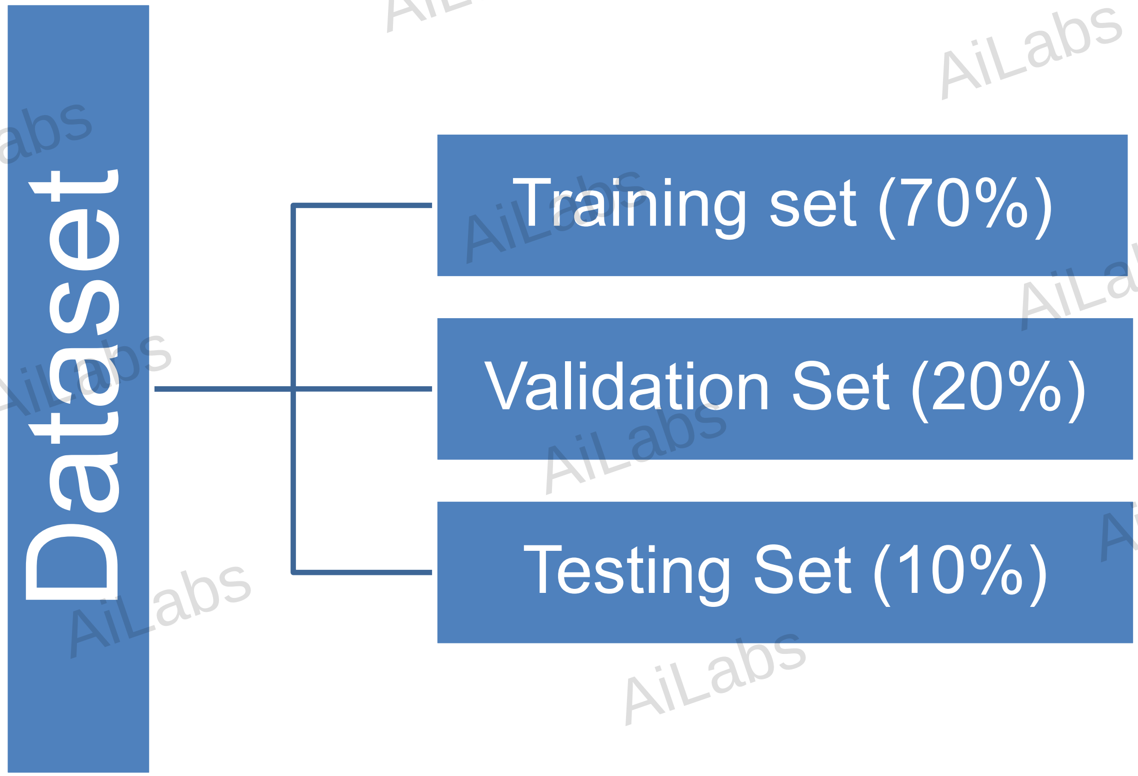
Features

Label

Observations

size	edge	color
small	dotted	green
big	striped	yellow
medium	normal	green

Training testing & validation



Training set and test set

- Data in its entirety is not used for training purpose.
- It is split into two sets:
  - Training Set
  - Testing Set
- Generally, a 70:30 ratio is preferred, between train and test
- Sometimes 80:20 or 90:10 is also used

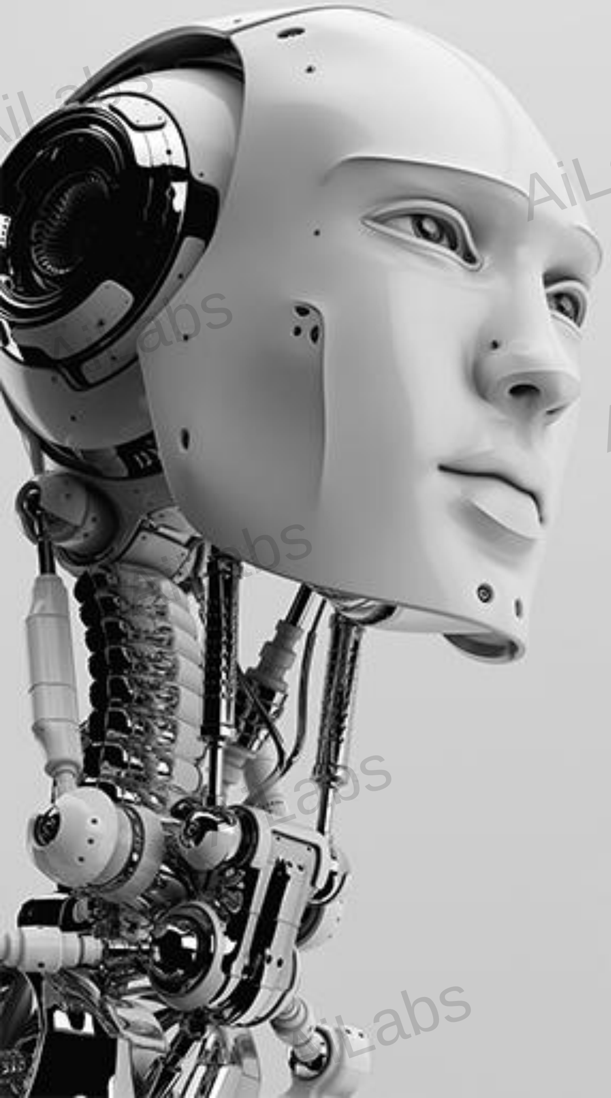
Train Test Split

Subject	t	Feature 1	Feature 2	Target
Paul	1	1000	male	0
Paul	2	1100	male	0
Paul	3	1200	male	1
Paul	4	1300	male	1
Crista	4	20	female	0
Crista	5	100	female	0
Paulina	1	10000	female	0
Paulina	2	100000	female	1
Paulina	3	95000	female	1
Paulina	4	97000	female	1
Paulina	5	99000	female	1
Paulina	6	101000	female	1
George	1	50000	male	1
George	2	50000	male	1
George	3	50000	male	1
George	4	50000	male	1
George	5	50000	male	1
George	6	50000	male	1

Train

Test

Subject	t	Feature 1	Feature 2	Target
Paul	1	1000	male	0
Paulina	1	10000	female	0
George	1	50000	male	1
Paul	2	1100	male	0
Paulina	2	100000	female	1
George	2	50000	male	1
Paul	3	1200	male	1
Paulina	3	95000	female	1
George	3	50000	male	1
Paul	4	1300	male	1
Crista	4	20	female	0
Paulina	4	97000	female	1
George	4	50000	male	1
Crista	5	100	female	0
Paulina	5	99000	female	1
George	5	50000	male	1
Paulina	6	101000	female	1
George	6	50000	male	1



Thank You