LCC-SJCE MACHINE LEARNING
DAY1

DATA PRE-PROCESSING

DATA SET

A data set (or dataset) is a collection of data. Most commonly a data set corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the data set.

Columns : Features/Attribute that describe the data

Rows : Data samples/observations

BASIC TYPES OF DATA

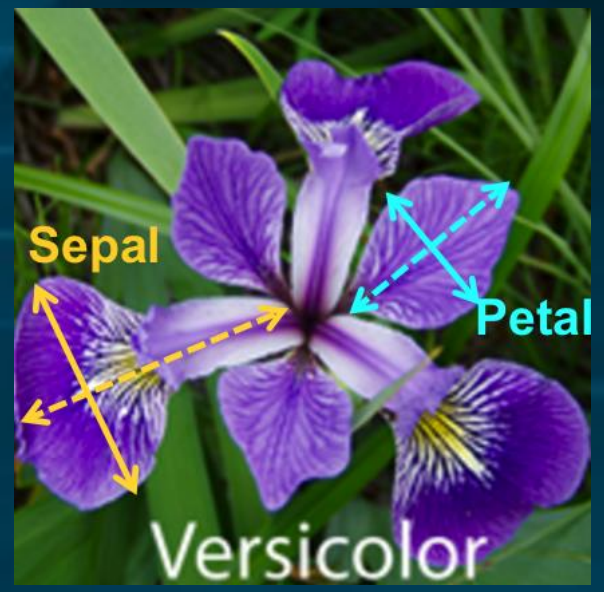
1.Numeric

2.Categorical (Non-Numeric)

Iris dataset

Features

ID/NO.	Sepal length	Sepal width	Petal length	Petal width	Species
1	5.1	3.5	1.4	0.2	*I. setosa*
2	4.9	3.0	1.4	0.2	*I. setosa*
3	4.7	3.2	1.3	0.2	*I. setosa*
4	4.6	3.1	1.5	0.2	*I. setosa*
5	7.0	3.2	4.7	1.4	*I. versicolor*
6	6.4	3.2	4.5	1.5	*I. versicolor*
7	6.9	3.1	4.9	1.5	*I. versicolor*
8	5.5	2.3	4.0	1.3	*I. versicolor*
9	6.3	3.3	6.0	2.5	*I. virginica*
10	5.8	2.7	5.1	1.9	*I. virginica*
11	7.1	3.0	5.9	2.1	*I. virginica*
12	6.3	2.9	5.6	1.8	*I. virginica*



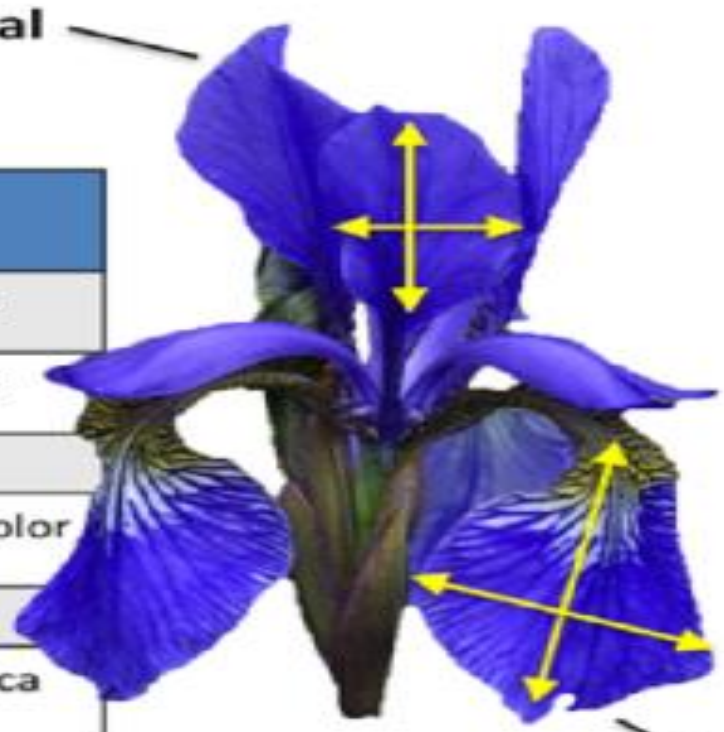
Data samples

TARGET VARIABLE

Target variable, in the machine learning context is the variable that is or should be the output.

Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica



Features
(attributes, measurements, dimensions)

Class labels
(targets)

WHERE TO GET THE DATA FROM ?

http://archive.ics.uci.edu/ml/index.php

kaggle™

https://www.kaggle.com/datasets

CSV FILE

Short for **Comma-separated values**, **CSV** is tabular data that has been saved as plaint

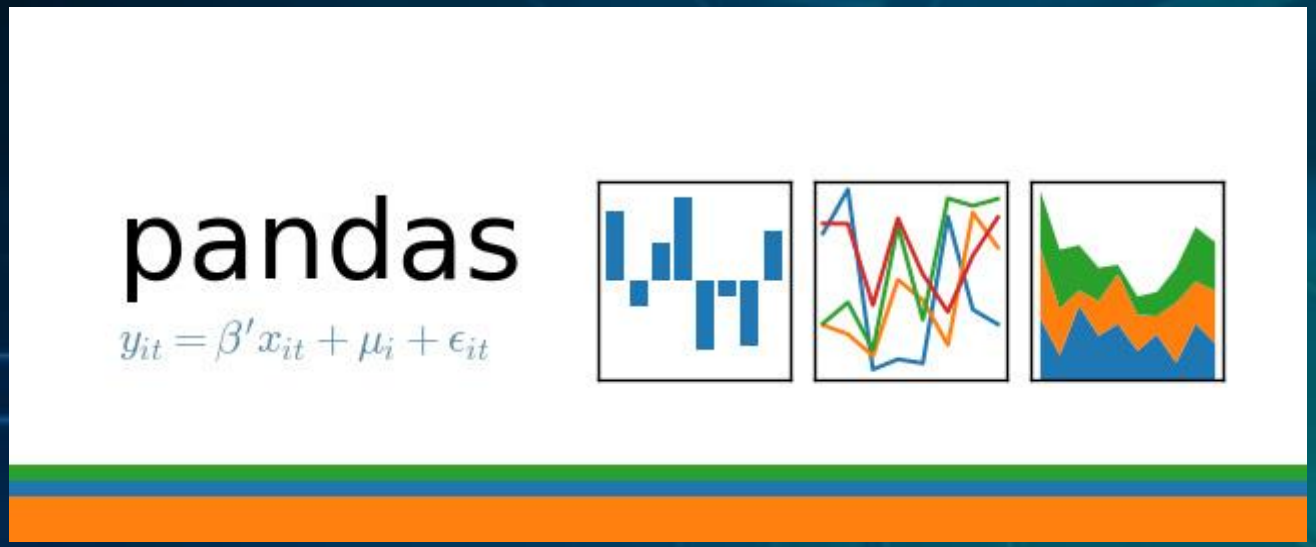
For example, if you had a table similar to the example below, that data would be converted to the CSV data shown below the table data separated by commas.

```
Data1,Data2,Data3
Example1,Example2,Example3
Example1,Example2,Example3
```

Data1	Data2	Data3
Example1	Example2	Example3
Example1	Example2	Example3

IMPORTING DATASET USING PANADS

pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.



HANDLING MISSING DATA

1. IGNORE THE DATA ROW

you'll obviously get **poor performance** if the percentage of such rows is high.

2. USE A GLOBAL CONSTANT TO FILL IN FOR MISSING VALUES

Decide on a new global constant value, like "*unknown*", "*NAN*" or *minus infinity*, that will be used to fill all the missing values. **Cant give it as input to algorithm**

3. USE ATTRIBUTE MEDIAN

Replace missing values of an attribute with the median value for that attribute in the dataset.

4.USE MODE

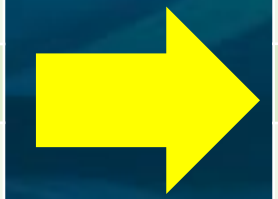
Replace missing values of an attribute with the mode.

5. USE ATTRIBUTE MEAN

Replace missing values of an attribute with the mean value for that attribute in the dataset.

HANDLING MISSING DATA USING MEAN

Country	Age	Salary	Purchased
France	44	72000	No
Spain	27	48000	Yes
Germany	30	54000	No
Spain	38	61000	No
Germany	40		Yes
France	35	58000	Yes
Spain		52000	No
France	48	79000	Yes
Germany	50	83000	No
France	37	67000	Yes



Country	Age	Salary	Purchased
France	44	72000	No
Spain	27	48000	Yes
Germany	30	54000	No
Spain	38	61000	No
Germany	40	**63777.8**	Yes
France	35	58000	Yes
Spain	**39**	52000	No
France	48	79000	Yes
Germany	50	83000	No
France	37	67000	Yes

$$MEAN(AGE) = \frac{\Sigma AGE}{N} = \frac{349}{9} = 38.77$$

$$MEAN(\text{SALARY}) = \frac{\Sigma SALARY}{N} = \frac{574000}{9} = 63777.8$$

HANDLING CATEGORICAL DATA

Categorical data (non numeric) represent types of data which may be **divided into groups**.

Country	Age	Salary	Purchased
France	44	72000	No
Spain	27	48000	Yes
Germany	30	54000	No
Spain	38	61000	No
Germany	40		Yes
France	35	58000	Yes
Spain		52000	No
France	48	79000	Yes
Germany	50	83000	No
France	37	67000	Yes

HANDLING CATEGORICAL DATA

NOMINAL ATTRIBUTE : value of a nominal attribute is just different names. eg:- gender, employee ids

ORDINAL ATTRIBUTE: An attribute for which the possible values are ordered. Eg :-rating : poor ,good ,average ,excellent.

HANDLING CATEGORICAL DATA

CATEGORICAL ENCODING: Used for Nominal and ordinal types. Give numeric label (value) to each type of the categorical data.

Country	Age	Salary	Purchased
France	44	72000	No
Spain	27	48000	Yes
Germany	30	54000	No
Spain	38	61000	No
Germany	40	63777.8	Yes
France	35	58000	Yes
Spain	39	52000	No
France	48	79000	Yes
Germany	50	83000	No
France	37	67000	Yes

Country	Age	Salary	Purchased
France	44	72000	0
Spain	27	48000	1
Germany	30	54000	0
Spain	38	61000	0
Germany	40	63777.8	1
France	35	58000	1
Spain	39	52000	0
France	48	79000	1
Germany	50	83000	0
France	37	67000	1

YES	**1**
NO	**0**

HANDLING CATEGORICAL DATA

DUMMY OR ONE-HOT ENCODING: preferred Non ordinal types, can be used for both ordinal and nominal. Create separate column for each type of variable.

Country
France
Spain
Germany
Spain
Germany
France
Spain
France
Germany
France

FRANCE	SPAIN	GERMANY
1	0	0
0	1	0
0	0	1
0	1	0
0	0	1
1	0	0
0	1	0
1	0	0
0	0	1
1	0	0

COUNTRY	AGE	SALARY	PURCHASED
FRANCE	44	72000	0
SPAIN	27	48000	1
GERMANY	30	54000	0
SPAIN	38	61000	0
GERMANY	40	63777.8	1
FRANCE	35	58000	1
SPAIN	39	52000	
FRANCE	48	79000	
GERMANY	50	83000	
FRANCE	37	67000	

FRANCE	SPAIN	GERMANY	AGE	SALARY	PURCHASED
1	0	0	44	72000	0
0	1	0	27	48000	1
0	0	1	30	54000	0
0	1	0	38	61000	0
0	0	1	40	63777.8	1
1	0	0	35	58000	1
0	1	0	39	52000	0
1	0	0	48	79000	1
0	0	1	50	83000	0
1	0	0	37	67000	1

FEATURE SCALING

Feature scaling is a method used to standardize the range of independent variables or features of data. It is generally performed during the data preprocessing step.

WHY ? Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without feature scaling. For example, the majority of classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance. Another reason why feature scaling is applied is that gradient descent converges much faster with feature scaling than without it.

FEATURE SCALING

TYPES:

1.STANDARDIZATION

$$X_{stand} = \frac{X - mean(X)}{standard\_deviation(X)}$$

2.NORMALIZATION

$$X_{norm} = \frac{X - min(X)}{max(X) - min(X)}$$

FEATURE SCALING

FRANCE	SPAIN	GERMANY	AGE	SALARY	PURCHASED
1	0	0	44	72000	0
0	1	0	27	48000	1
0	0	1	30	54000	0
0	1	0	38	61000	0
0	0	1			
1	0	0			
0	1	0			
1	0	0			
0	0	1			
1	0	0			

FRANCE	SPAIN	GERMANY	AGE	SALARY	PURCHASED
1.22474	-0.654654	-0.654654	0.758874	0.749473	-1
-0.816497	-0.654654	1.52753	-1.7115	-1.43818	1
-0.816497	1.52753	-0.654654	-1.27555	-0.891265	-1
-0.816497	-0.654654	1.52753	-0.113024	-0.2532	-1
-0.816497	1.52753	-0.654654	0.177609	6.63219e-16	1
1.22474	-0.654654	-0.654654	-0.548973	-0.526657	1
-0.816497	-0.654654	1.52753	0	-1.07357	-1
1.22474	-0.654654	-0.654654	1.34014	1.38754	1
-0.816497	1.52753	-0.654654	1.63077	1.75215	-1
1.22474	-0.654654	-0.654654	-0.25834	0.293712	1

FEW IMPORTANT TERMINOLOGIES

ML model: The ML program that is created for specific application.

Training: Learning process of ML model , it achieved using ML algorithms.

Testing: Testing the trained ML model whether it is giving output with expected accuracy.

SPLITTING THE DATASET

Splitting dataset into TEST and TRAIN set

The dataset will be split into two parts, one is called training set and another one is called test set.

Training set is used for training the ML model.

Test set is used for testing the trained ML model.

Splitting is done using suitable split ratio (eg:70% for training and 30% for testing).

PREFERED ENVIRONMENT TO WORK ON ML /DATA SCIENCE

https://www.anaconda.com/download/

Anaconda is a freemium open source distribution of the Python and R programming languages for large-scale data processing, predictive analytics, and scientific computing, that aims to simplify package management and deployment.

THANK YOU!