INSTITUTO TECNOLÓGICO DE TIJUANA

DEPARTAMENTO DE SISTEMAS Y COMPUTACIÓN

SEMESTRE ENERO- JUNIO 2020

CARRERA:

Ingeniería Tecnologías de la Información
y Comunicaciones

MATERIA:

Datos Masivos

UNIDAD POR EVALUAR:

Unidad II

TRABAJO A EVALUAR:

Tarea # 3 Pipeline, Matriz of confusion

NOMBRE DEL ESTUDIANTE

Barraza Sierra Alexis Fernando

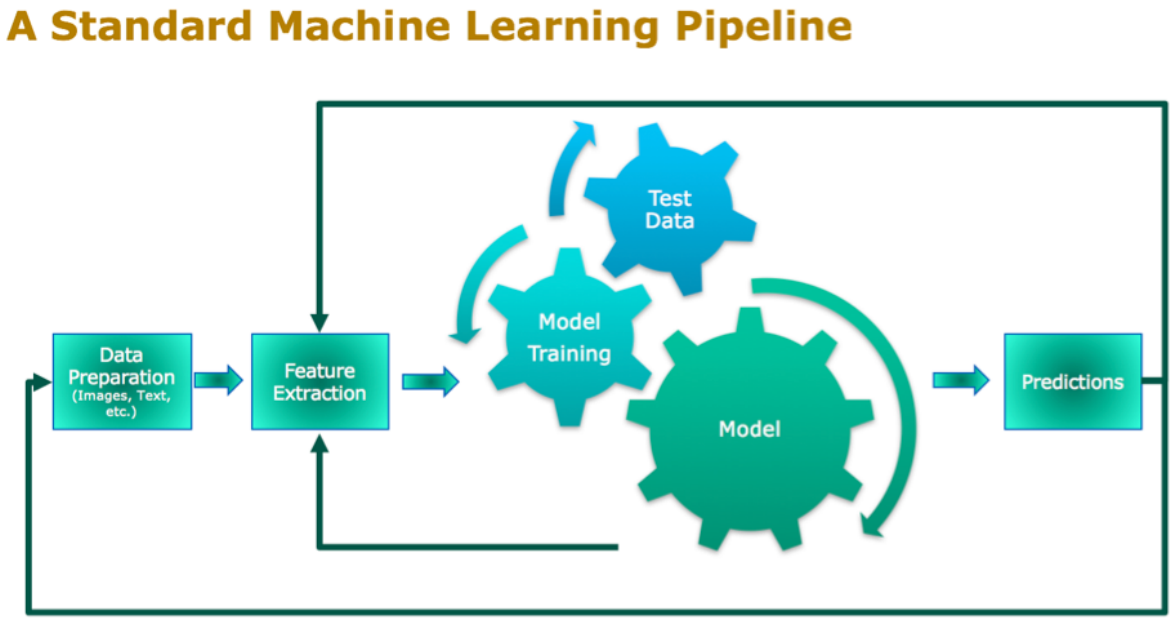
Zamorano Garcia Osvaldo Arturo

PROFESOR:

Dr. Jose Christian Romero Hernandez

Pipeline in Machine Learning

A machine learning pipeline is used to help automate machine learning workflows. They operate by enabling a sequence of data to be transformed and correlated together in a model that can be tested and evaluated to achieve an outcome, whether positive or negative. Machine learning (ML) pipelines consist of several steps to train a model. Machine learning pipelines are iterative as every step is repeated to continuously improve the accuracy of the model and achieve a successful algorithm.



It's not just about storing data any longer, but capturing, preserving, accessing and transforming it to take advantage of its possibilities and the value it can deliver.

- The main objective of having a proper pipeline for any ML model is to exercise control over it. A well-organised pipeline makes the implementation more flexible.

- The main objective of having a proper pipeline for any ML model is to exercise control over it. A well-organised pipeline makes the implementation more flexible.

- The learning algorithm finds patterns in the training data that map the input data attributes to the target

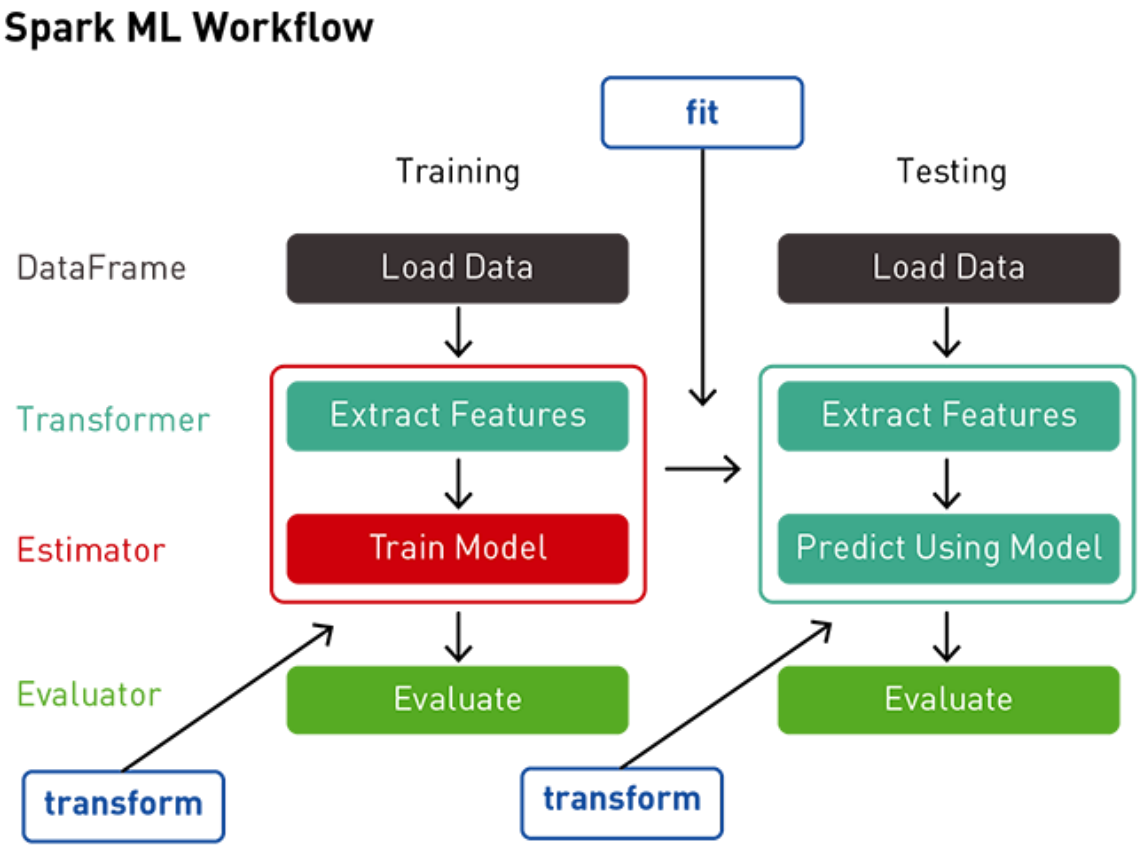
- model can have many dependencies and to store all the components to make sure all features available both offline and online for deployment

- A pipeline consists of a sequence of components which are a compilation of computations.

Challenges Associated with ML Pipelines

A typical machine learning pipeline would consist of the following processes:

- Data collection
- Data cleaning
- Feature extraction (labelling and dimensionality reduction)
- Model validation
- Visualisation



Data collection and cleaning are the primary tasks of any machine learning engineer

who wants to make meaning out of data. But getting data and especially getting the right data is an uphill task in itself. Data quality and its accessibility are two main challenges one will come across in the initial stages of building a pipeline. The captured data should be pulled and put together and the benefits of collection should outweigh the costs of collection and analysis.

Confusion matrix

In the field of machine learning and specifically the problem of statistical classification, a confusion matrix, also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one (in unsupervised learning it is usually called a matching matrix). Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class (or vice versa).

Example

If a classification system has been trained to distinguish between cats and dogs, a confusion matrix will summarize the results of testing the algorithm for further inspection. Assuming a sample of 13 animals — 8 cats and 5 dogs — the resulting confusion matrix could look like the table below:

		Actual class	
		Cat	**Dog**
Predicted class	**Cat**	**5**	2
	Dog	3	**3**

\

In this confusion matrix, of the 8 actual cats, the system predicted that three were dogs, and of the five dogs, it predicted that two were cats. All correct predictions are located in the diagonal of the table (highlighted in bold), so it is easy to visually inspect the table for prediction errors, as they will be represented by values outside the diagonal.

In abstract terms, the confusion matrix is as follows:

		Actual class	
		P	**N**
Predicted class	**P**	**TP**	FP
	N	FN	**TN**

where: P = positive; N = Negative; TP = True Positive; FP = False Positive; TN = True Negative; FN = False Negative.