Pipeline and Encoding

By Ritthuja Kandasamy

Encoding

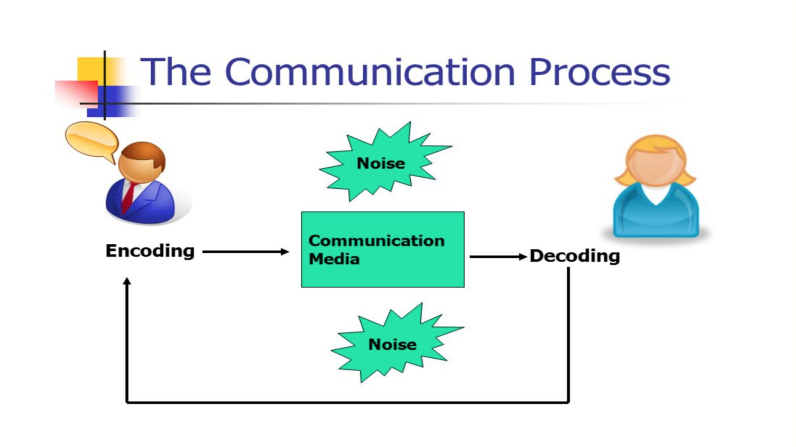
What do you mean encoding?

In computers, encoding is the process of putting a sequence of characters **(letters, numbers, punctuation, and certain symbols)** into a specialized format for efficient transmission or storage. Decoding is the opposite process -- the conversion of an encoded format back into the original sequence of characters.

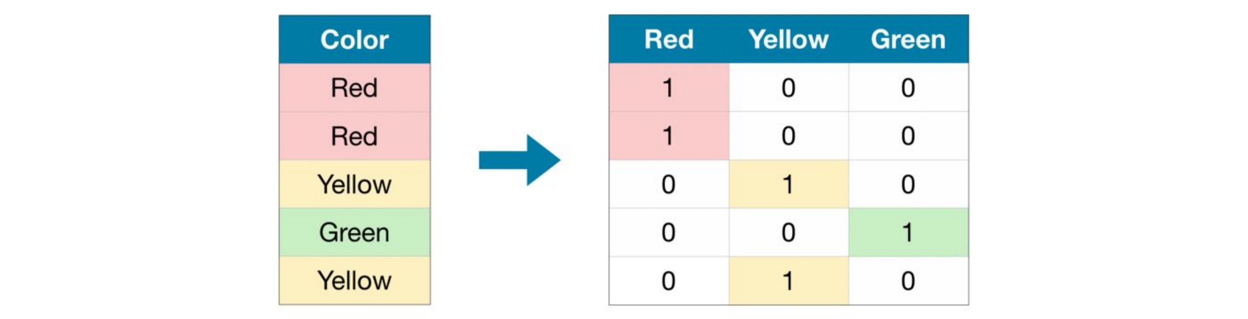
Encoding works similarly to converting temperature from centigrade to Fahrenheit, as it just gets converted in another form, but the original value always remains the same.

What is an example of encoding?

For example, you may realize you're hungry and encode the following message to send to your roommate: **“I'm hungry. Do you want to get pizza tonight?”** As your roommate receives the message, they decode your communication and turn it back into thoughts to make meaning.



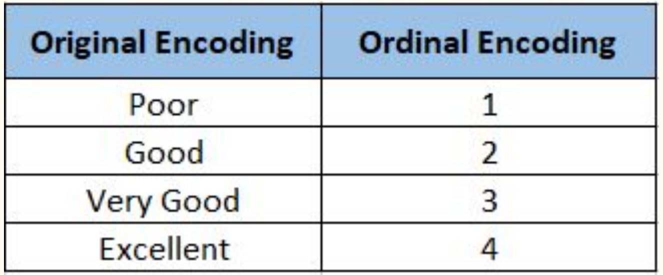
Nominal Variable (Categorical). Variable comprises a finite set of discrete values with no relationship between values.



Color
Red
Red
Yellow
Green
Yellow

Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1
0	1	0

Ordinal Variable. Variable comprises a finite set of discrete values with a ranked ordering between values.

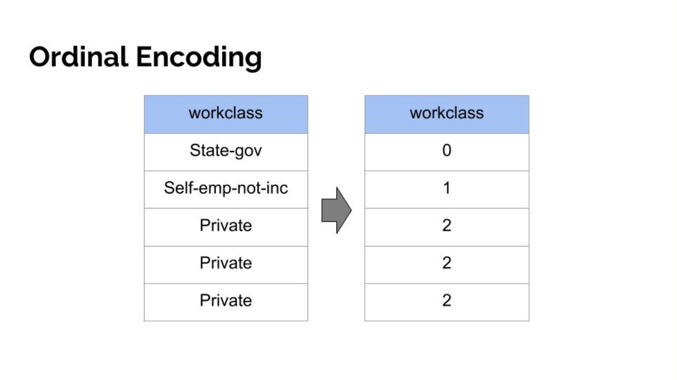


Original Encoding	Ordinal Encoding
Poor	1
Good	2
Very Good	3
Excellent	4

Ordinal Encoding

In ordinal encoding, each unique category value is assigned an integer value.

For example, “*red*” is 1, “*green*” is 2, and “*blue*” is 3.



Ordinal Encoding

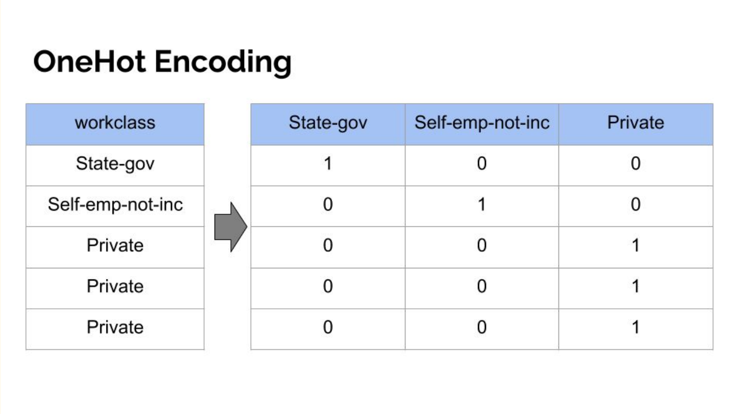
workclass
State-gov
Self-emp-not-inc
Private
Private
Private

workclass
0
1
2
2
2

This is called an **ordinal encoding or an integer encoding** and is easily reversible. Often, integer values starting at zero are used. For some variables, an ordinal encoding may be enough. The integer values have a natural ordered relationship between each other and machine learning algorithms may be able to understand and harness this relationship.It is a natural encoding for ordinal variables. For categorical variables, it imposes an ordinal relationship where no such relationship may exist. This can cause problems and a one-hot encoding may be used instead.

One-Hot Encoding

For categorical variables where no ordinal relationship exists, the integer encoding may not be enough, at best, or misleading to the model at worst.



OneHot Encoding

workclass
State-gov
Self-emp-not-inc
Private
Private
Private

State-gov	Self-emp-not-inc	Private
1	0	0
0	1	0
0	0	1
0	0	1
0	0	1

Forcing an ordinal relationship via an ordinal encoding and allowing the model to assume a natural ordering between categories may result in poor performance or unexpected results (predictions halfway between categories).

In this case, a **one-hot encoding** can be applied to the ordinal representation. This is where the integer encoded variable is removed and one new binary variable is added for each unique integer value in the variable.

PIPELINE

What is Machine Learning Pipeline?

*A **Machine Learning pipeline** is a process of automating the workflow of a complete machine learning task*. It can be done by enabling a sequence of data to be transformed and correlated together in a model that can be analyzed to get the output.

A typical pipeline includes raw data input, **features, outputs, model parameters, ML models, and Predictions**.

Moreover, an ML Pipeline contains multiple sequential steps that perform everything ranging from data extraction and pre-processing to model training and deployment in Machine learning in a modular approach.

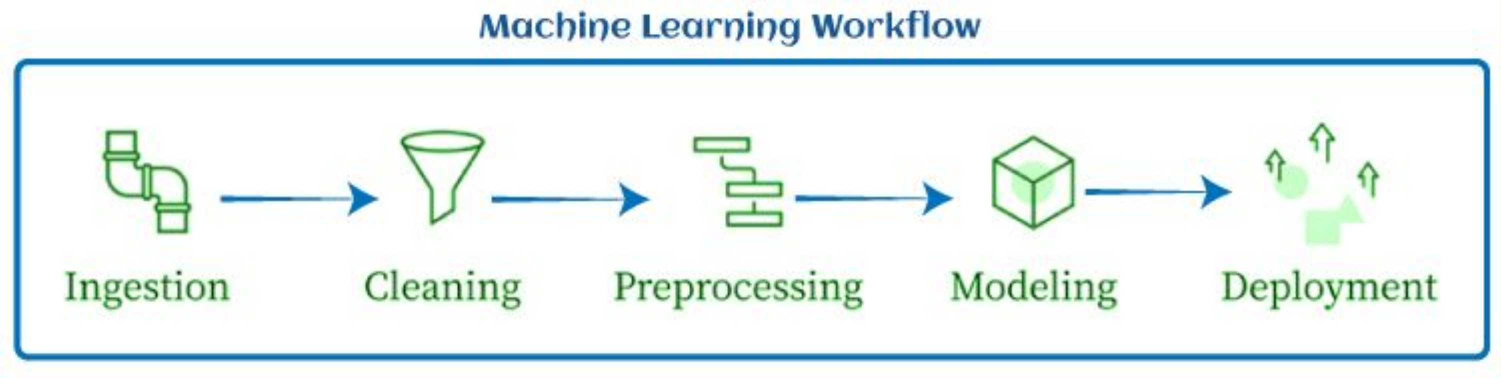
It means that *in the pipeline, each step is designed as an **independent module**, and all these modules are tied together to get the final result.*

Machine Learning Pipeline with simple example

We can understand it with an example.

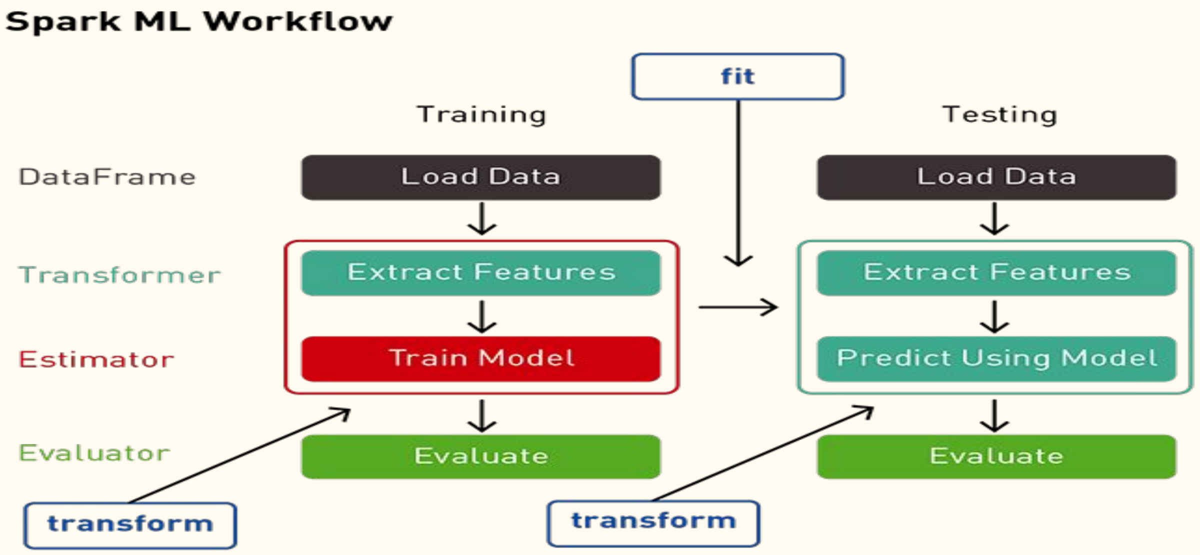
Building any ML model requires a huge amount of data to train the model. As data is collected from different resources, it is necessary to clean and pre-process the data, which is one of the crucial steps of an ML project. However, whenever a new dataset is included, we need to perform the same pre-processing step before using it for training, and it becomes a time-consuming and complex process for ML professionals.

To solve such issues, **ML pipelines** can be used, which can remember and automate the complete pre-processing steps in the same order.



Why do we use 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.



Thank you