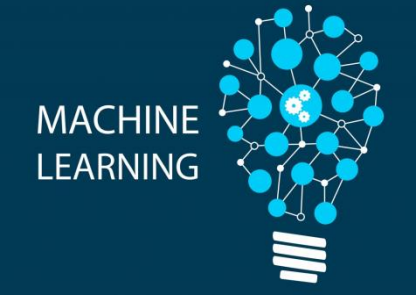
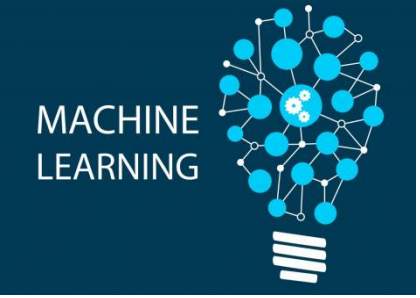


Data Manipulation

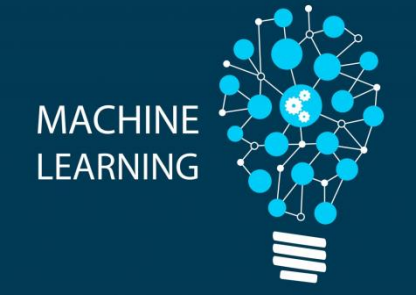
Diptangshu Banik

Twitter - @dipbanik



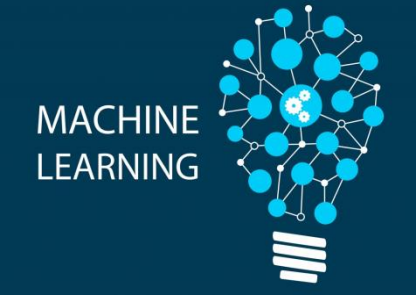
Overview

- Data will almost never be as we need it.
- Used to transform and adjust values in the data to suit the model.



Missing Values

- Delete the row
- Impute(Replace) the missing value with Mean, Median or Mode
- Use algorithms which support missing values - k-Nearest Neighbor.



Drop missing values pandas

Drop the rows where at least one element is missing -

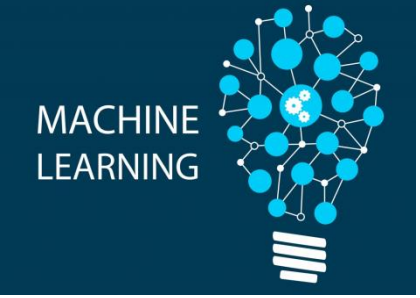
```
>>> df.dropna()
```

- Drop the columns where at least one element is missing –

```
>>> df.dropna(axis='columns')
```

- Drop the rows where all elements are missing-

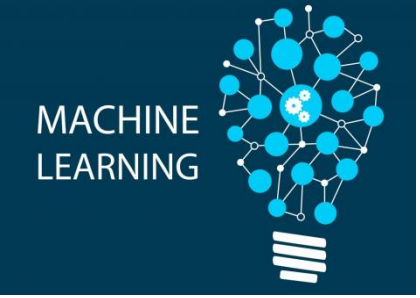
```
>>> df.dropna(how='all')
```



Impute

- Replace values with Mean/Median/Mode

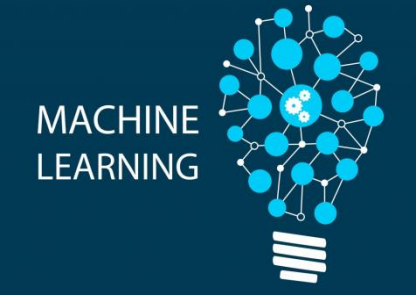
```
>>> from sklearn.preprocessing import Imputer

>>> imputer = Imputer(missing_values = 'NaN', strategy = 'mean', axis=0)
>>> imputer = imputer.fit(X[:, 1:3])
```



Categorical Values Binary

- Where there are binary categorical values – ex. Male, Female

```
>>> from sklearn.preprocessing import LabelEncoder
>>> labelEncoder_X = LabelEncoder())
>>> X[:, 0] =  labelEncoder_X.fit_transform(X[:, 0])
```



Categorical Values
One Hot Encoding

- Where there are multi-categorical values – ex. Multiple countries/Small, meduim, large

```
>>> from sklearn.preprocessing import OneHotEncoder

>>> onehotencoder = OneHotEncoder(categorical_features
= [0])

>>> X = onehotencoder.fit_transform(X).toarray()
```