Process of Data Sciences

Data sciences is generally the process or a technique of prediction. It involves several steps.

They are;

Data collection or Data gathering:

Generally whenever we are going for a project we will have a problem statement. So, regarding the problem statement we should collect the data required.

For Example. In a heart disease prediction dataset we will have data columns such as cp, trestbps, chol, fbs, talach etc.

should collect the data relavant to the problem statement.

Data loading:

We load the dataset using pandas in python. And the code snippet we use is

```
pandas.read_csv("filename.csv")
```

And it encodes the dataset as utf-8 which is set to default. some dataset may not accept it, so by mentioning latin1, windows1252 for encoding variable we could load it

Eg: Terrorist attacks dataset.

Data preprocessing:

As a part of data preprocessing we make the data ready for the further process. As we know that the dataset may contain some null values, garbage or any unwanted values. Generally we use fillna(),

dropna() → central tendencies such as median, mode, mean are used

ffill() → forward fill

bfill() → backward fill

And should remove unwanted which are nothing but outliers.

And we use pandas to get and write some queries.

Queries are structured in such a way which meets and satisfies our conditions.

For example if we need eastern region values from sales dataset we need to apply some conditions. They are structured using pandas.

And upto this Exploratory Data Analysis Part (EDA) is done.

Data Visualisation :

After getting all the required ones we go for visualising the data. So, that we get a clear idea of how data is behaving. It clearly makes us to understand the data variation for different attributes i.e., columns.

We go for Matplotlib and Seaborn in python for visualisation.

For Matplotlib we have scatterplots, lineplots, piecharts e.t.c

In Seaborn we have scatterplots, lineplots, boxplots, swarm plots, Stripplots, heatmaps e.t.c

These libraries help us to have good visualisations and can get a good idea regarding the data.

Detecting X and y :

Here X means independent variables and y is dependent variable which means output.

Eg, In fruits dataset predicting

fruit → label is y

sphericity & weight → labels are X.

Detecting or knowing X and y is the crucial part and is known through the Data-Analytics part.

Machine learning:

After detecting x and y, we divide the dataset into train dataset and test dataset. We do it with a function called train_test_split. We import it with the following code snippet.

```
from sklearn.model_selection
        import train_test_split
```

The records are not splitted in the indices from 0 to n and from n to end of records into train and test datasets. But they are splitted in a random order.

After doing it we feed the model with training dataset. We will have some algorithms such as Decision Tree, KNN, linear regression, logistic regression.

And according to the need we go for algorithm.

We initialise a model and feed it with the records and gets all the information from the data.

We test the accuracy and other metrics with test dataset. We feed a model using the following code snippet.

```
model.fit(X_train, y_train).
```

Cross validation (CV):

Going for CVS we can go for Kfold cross validation. K folds are divided that means K groups and each fold can go for testing.

If 5 groups are divided then if

1st group for testing remaining for training.

2nd group for testing remaining all for training.

and so on.

We had different metrics such as accuracy score, balanced accuracy score, recall score, precision score, AUC e.t.c

These metrics says how well our model is performing and says about our model predictions and all.

Deep learning:

Deep learning is that which is images, audios e.t.c And for deep learning we should have a good understanding about regression models that is linear regression, logistic regression e.t.c.

We use artificial neural networks which will have no. of neurals.

We will apply an activation function to the output of the neuron.

And these are the steps followed and the process of datascience.

Thanks for shortlisting.

I hope for the best.

Thanking you.