**Pipeline to follow for a data science project:**

**Step1**: Understand the problem statement from the client

**Step2**: Get the relevant data to solve the given problem(supervised and supervised).

**Step3**: Analyze and visualize the data- using matplot lib and seaborn.

**Step4**: Prepare the data for the algorithm.

**Step5**: Select a Model(supervised or umsupervised or both) and train the data.

**Step6**: Fine tune the model for better accuracy.

**Step7**: Test with real time data.

**Step8**: Deploy in the cloud/Sysytem.

Data: data plays a mojor role in a datascience projet.

Generally client has various kinds of data from various sources.

In my case I access real time data from live sensors for my supply chain management project.

Data files: generally 30 days volume data I used to receive. These files are in various formats. Some are in excel sheets and some are in csv,some are from sql and no sql db or AWS data sets.

Data to the application: fetch the data from various files and load into our application as a dataset or dataframe in python.

Prepare the data: pandas helps in analyzing the data. Data summary helps in understanding mean and median and various features from the data.

Generally data filelds in the form of numerical or categorical or both.

1. Look for standard correlation coefficient(pearson’s)

Or use scatter matrix.

1. Automatic hyperparameter tuning.
2. Feature scaling- normalize the data set. Use Standard scalar.
3. Training set
4. Prediction
5. Cross validation
6. Fine tune the model

a)grid search

b)randomized search

8) Ensemble techniques- to analyze the best model.

9)evaluate on the test data.