FOR DATA SCIENCE BEGINNERS

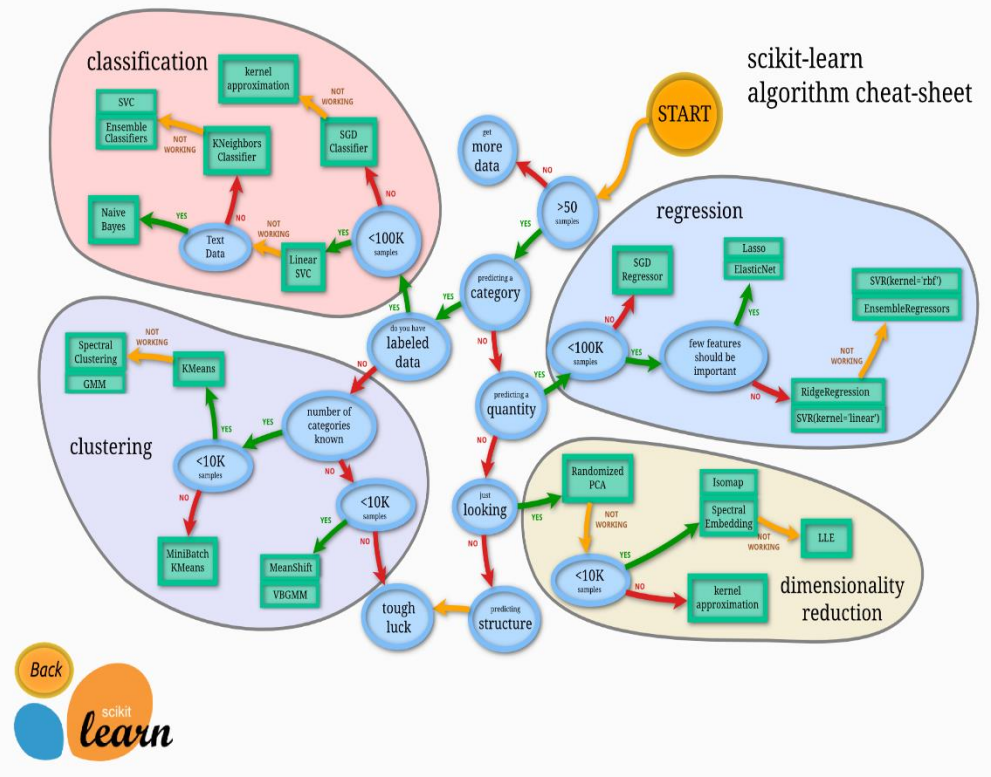
Steps to Follow

1. Collect the data from different source

For practices take a sample data from Kaggle or elsewhere.

2. Choose the model suitable for your data

Scikit-learn provides cheat-sheet which will help you for choosing the right model/algorithm.



3. Import the required library

Scikit-learn cheat sheet will guide you which library to import and get going.

Code:

```
from sklearn.ensemble import RandomForestClassifier
```

4. Setup the random seed

Random seed is used in data science for reproducibility and consistency in data analysis.

Code:

np.random.seed(42)

5. Now make your data ready for prediction

1. Drop the column that you want to predict and take this data as 'X'
2. Take the column that you want to predict separately

6. Split the data into train and test

'train_test_split()' use this function for splitting the data

Code:

x_train,x_test,y_train,y_test = train_test_split(x,y,test_size = 0.2)

here test_size = 0.2 refers to percentage of data that will be taken for train and test

7. Now fit your data into your model

Code:

model = RandomForestClassifier()

model.fit(x_train,y_train)

8. Calculate the score

Code:

model.score(x_test,y_test)

If your data scores good number then its good if not then go for step-9

9. Trying another model for better result

If your score is not good then you need to change the model and try it again and again and filter your data for better result and for choosing your model you can go to scikit-learn choosing the right model docs where it shows you which model will fit your data

Tip From My Side: Ensemble Regressor for regression model is best to go with, Similarly

Ensemble classifier for classification problem is good. Also with Clustering problem spectral clustering is good or GMM.