SciKit Learn Intro

A Balaji

DataScience Flow

- Exploratory Data Analysis
- Build Model
- Evaluate
- Save the Model

EDA (Exploratory Data Analysis)

- Read the dataset
- Check the data types
- Missing values, Outliers
- Data Distribution (imbalance data)
- Correlation among the variables
- Plot them to gain better understanding

Build Model

- Split the dataset to train and validation set
- Instantiate the model
- Train the model
- Validate the model / Cross validation
- Pipeline etc.

Evaluate Model

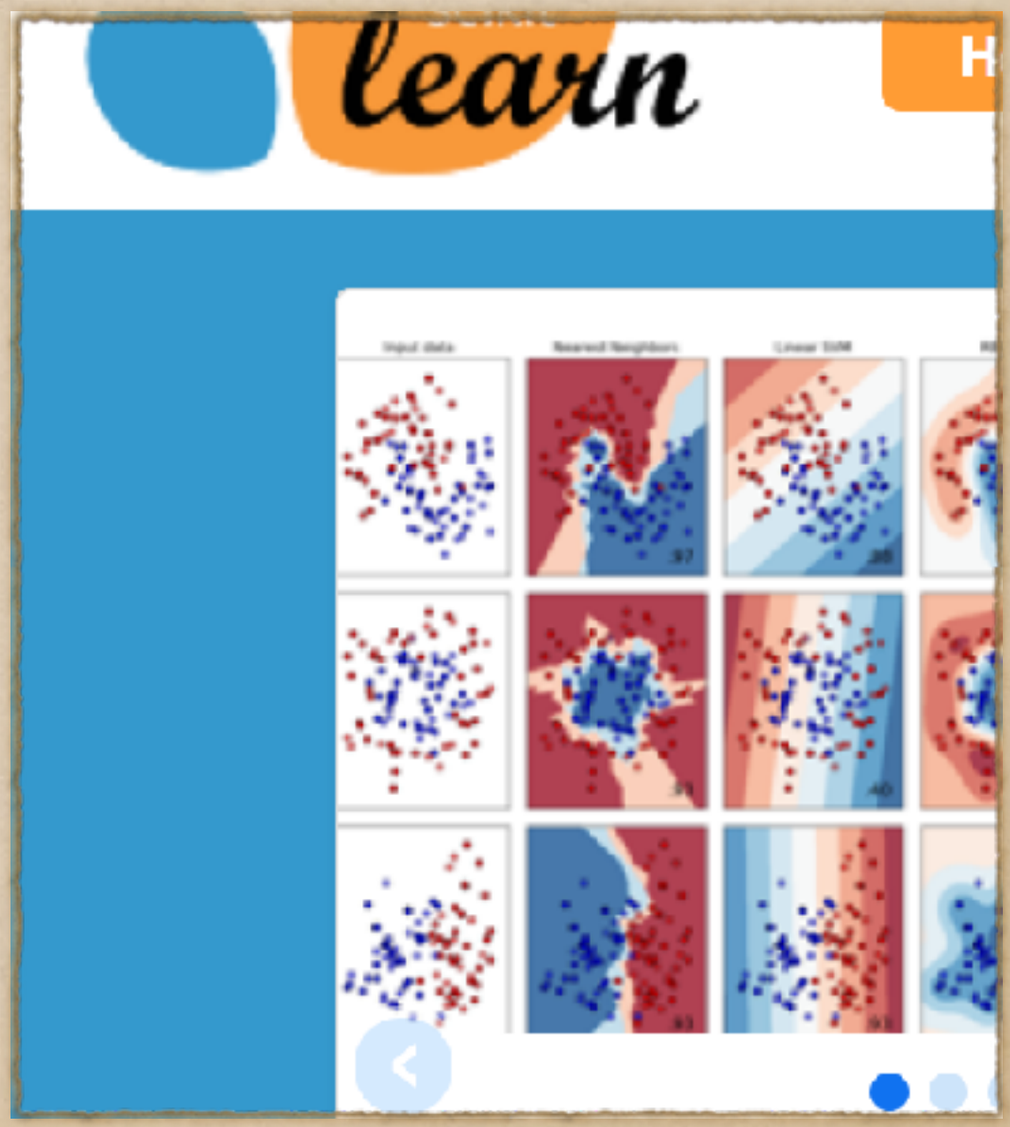
- Get the accuracy
- Other metrics
  - Classification report, Precision, Recall
- Plot the train and test - accuracy and loss

Save Model

- Save the Model
    - For on-demand prediction

SciKit - Getting Started

- www.scikit-learn.org
- pip install scikit-learn
- https://github.com/scikit-learn/scikit-learn



Scikit - EDA

- Sklearn.preprocessing
  - LabelEncoder ~ encode labels from 0 to num_classes-1
  - OneHotEncoder ~ encode as one-hot numeric array

SciKit-Build Model

- Sklearn.model_selection
  - train_test_split()
  - GridSearchCV()
- sklearn.feature_extraction.text
  - CountVectorizer()
  - TfIdfVectorizer()

Scikit-Build Model

- sklearn.linear_model
  - LinearRegression()
  - LogisticRegression()
- sklearn.naive_bayes
  - MultinomialNB()
- Sklearn.neighbors
  - NearestNeighbors()
- sklearn.pipeline
  - make_pipeline()

SciKit-Evaluate Model

- sklearn.metrics
  - accuracy_score(y_true, y_pred)
  - confusion_matrix(y_true, y_pred)
  - classification_report(y_true, y_pred)

Save Model

- SciKit doesn't provide methods to save model
- Alternatively different packages provide serialising and de-serializing the python data structures.
- pickle
- H5py

Save Model -pickle

- import pickle
- Pickle.dump(model, file-object)
- model = Pickle.load(file-object)

Save Model - H5py

- HDF5 is Heterogeneous Data File support version 5

- HDF5 is a container to store two kinds of objects:
  - datasets - array like collections of data
  - Groups - work like dictionaries

Save Model - h5py

- import h5py
- Handle = h5py.File("test.hdf5", 'w')
- dset = handle.create_dataset()
- grp = handle.create_group()

thank you
gracias
merci
obrigado
danke
謝謝
ngiyabonga
teşekkür ederim
спасибо
Баярлалаа
рахмат
dank je
tapadh leat
mochchakkeram
go raibh maith agat
dziękuję
bedankt
hvala
sukriya
kop khun krap
arigatō
takk
dakujem
grazie
terima kasih
sagolun
taiku
мерси
ありがとう
감사합니다
ευχαριστώ
хвала