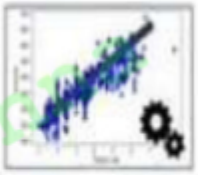
Model evaluation approaches

- Train and Test on the Same Dataset
- Train/Test Split

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	212

Train

Test



	Prediction
6	234
7	256
8	267
9	210

Actual values

Predicted values

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	212

Test

y

Actual values

$$Error = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

	Prediction
6	234
7	256
8	267
9	210

$\hat{y}$

Predicted values

Train and test on the same dataset

High "training accuracy"
Low "out-of-sample accuracy"

What is training & out-of-sample accuracy?

- **Training Accuracy**
  - High training accuracy isn't necessarily a good thing
  - Result of over-fitting
    - Over-fit: the model is overly trained to the dataset, which may capture noise and produce a non-generalized model

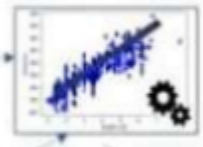
- **Out-of-Sample Accuracy**
  - It's important that our models have a high, out-of-sample accuracy
  - How can we improve out-of-sample accuracy?

Train/Test split evaluation approach

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	212

Train

Test



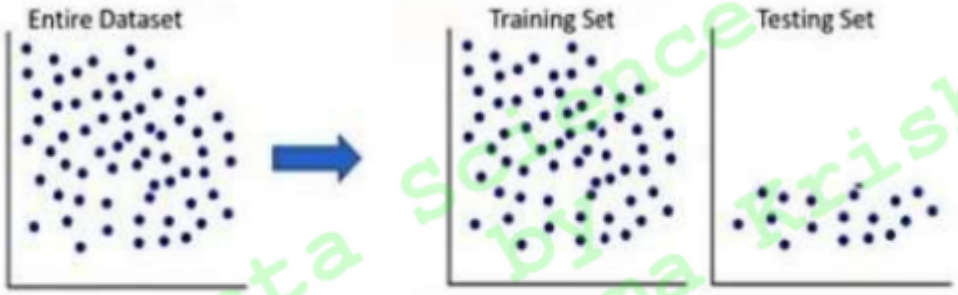
	Prediction
6	234
7	256
8	267
9	210

Actual values

Predicted values

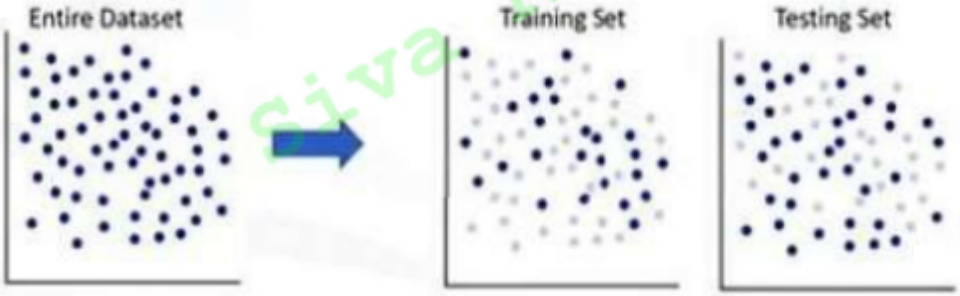
Train/Test split evaluation approach

Test on a portion of train set



- Test-set is a portion of the train-set
- High "training accuracy"
- Low "out-of-sample accuracy"

Train/Test Split



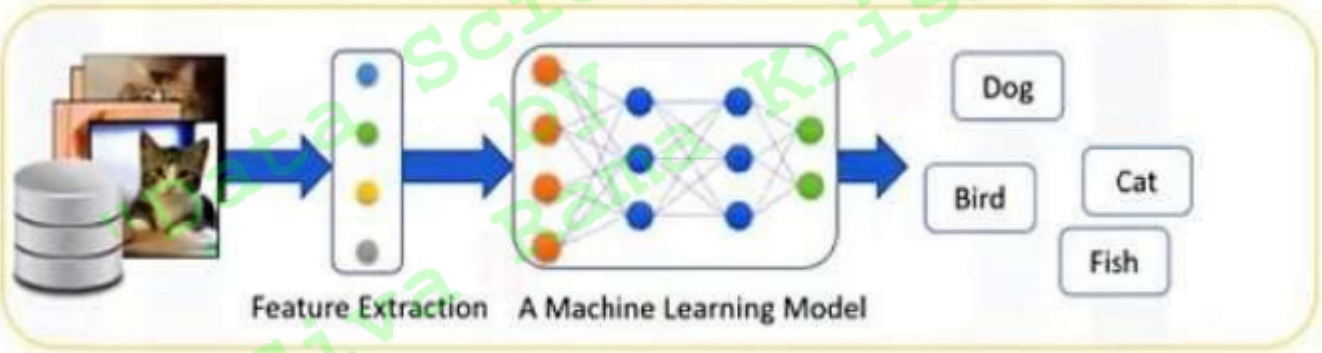
- Mutually exclusive
- More accurate evaluation on out-of-sample accuracy
- Highly dependent on which datasets the data is trained and tested

MACHINE LEARNING

What is machine learning?

Machine learning is the subfield of computer science that gives "**computers the ability to learn without being explicitly programmed.**"

How machine learning works?



Major machine learning techniques

- Regression/Estimation
  - Predicting continuous values
- Classification
  - Predicting the item class/category of a case
- Clustering
  - Finding the structure of data; summarization
- Associations
  - Associating frequent co-occurring items/events

Major machine learning techniques

- Anomaly detection
  - Discovering abnormal and unusual cases
- Sequence mining
  - Predicting next events; click-stream (Markov Model, HMM)
- Dimension Reduction
  - Reducing the size of data (PCA)
- Recommendation systems
  - Recommending items

SUPERVISED VS UNSUPERVISED LEARNING

MACHINE LEARNING

Supervised vs unsupervised learning

Supervised Learning

- **Classification:**
  Classifies labeled data
- **Regression:**
  Predicts trends using previous labeled data
- Has more evaluation methods than unsupervised learning
- Controlled environment

Unsupervised Learning

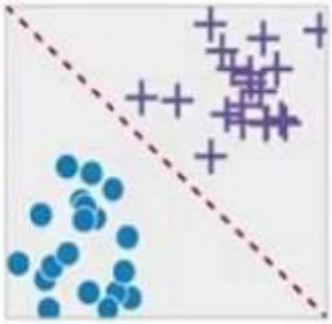
- **Clustering:**
  Finds patterns and groupings from unlabeled data
- Has fewer evaluation methods than supervised learning
- Less controlled environment

What is classification?

Classification is the process of predicting discrete class labels or categories.

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	1	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

Categorical Values

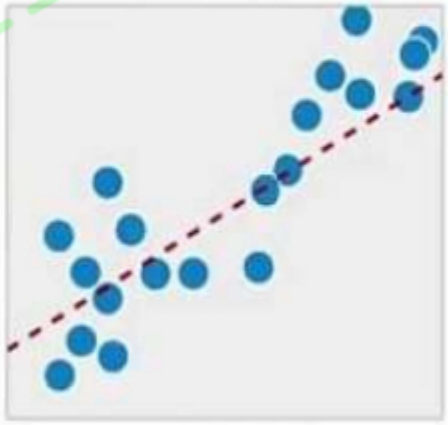


What is regression?

Regression is the process of predicting continuous values.

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

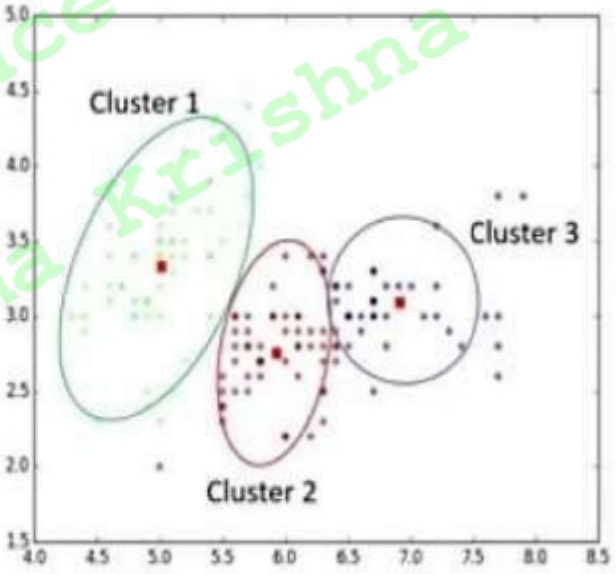
Continuous Values



What is clustering?

Clustering is grouping of data points or objects that are somehow similar by:

- Discovering structure
- Summarization
- Anomaly detection



Python libraries for machine learning

NumPy

SciPy

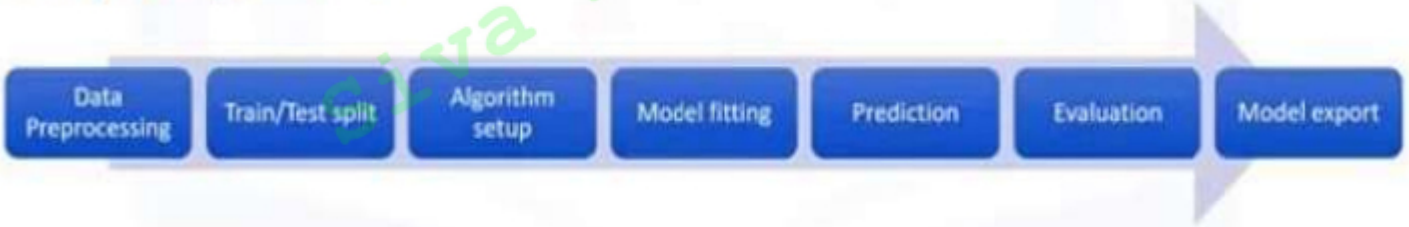
matplotlib

- Scikit-learn is a library containing many machine learning algorithms.
- It utilizes a generalized "estimator API" framework to calling the models.
- This means the way algorithms are imported, fitted, and used is uniform across all algorithms.

- Scikit-learn also comes with many convenience tools, including train test split functions, cross validation tools, and a variety of reporting metric functions.
- This leaves Scikit-Learn as a "one-stop shop" for many of our machine learning needs.
    - Scikit-Learn's approach to model building focuses on **applying models** and **performance metrics**.

More about scikit-learn

- Free software machine learning library
- Classification, Regression and Clustering algorithms
- Works with NumPy and SciPy
- Great documentation
- Easy to implement



scikit-learn functions

```
from sklearn import preprocessing
X = preprocessing.StandardScaler().fit(X).transform(X)
```

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33)
```

```
from sklearn import svm
clf = svm.SVC(gamma=0.001, C=100.)
```

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

```
clf.predict(X_test)
```

```
from sklearn.metrics import confusion_matrix
print(confusion_matrix(y_test, yhat, labels=[1,0]))
```

```
import pickle
s = pickle.dumps(clf)
```