Информациски системи и големи податоци

Evaluation, Validation techniques

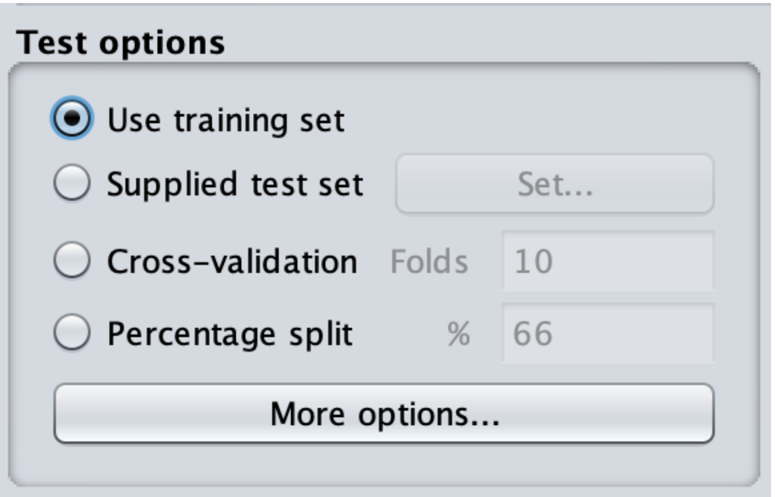
Validation Techniques

- Used to get the error rate of the model
- As close to the true error rate of the population.
- If the data volume is “large enough” to be representative of the population, you may not need the validation techniques. **(never the case)**
- However, in real-world scenarios, we work with samples of data that may not represent the population.
- Thus we use validation techniques
- Once evaluation is complete, all the data can be used to build the final classifier.
- The larger the test data the more accurate the error estimate.

Validation Techniques / Data Split

- Training data
- Hold-out
- K-fold cross-validation (CV)
- Leave one CV
- Random subsampling
- Bootstrapping
- Train-Validation-Test

Training data



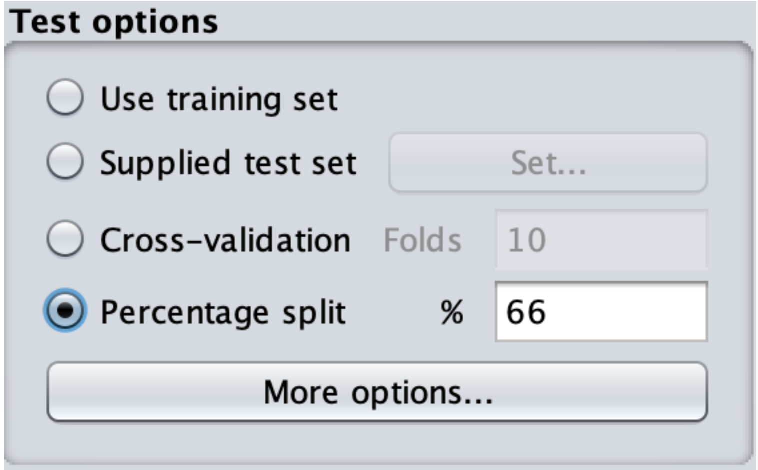
Test options

- ● Use training set
- ○ Supplied test set
- ○ Cross-validation Folds 10
- ○ Percentage split % 66

- Use the training dataset (the whole dataset) to estimate the performance.
- The accuracy/error estimates on the training data are not good indicators of performance on future data.
  - Because new data will probably not be exactly the same as the training data!
- The accuracy/error estimates on the training data measure the degree of classifier’s overfitting.
- *Resubstitution error*. The technique is called the resubstitution validation technique.

Holdout dataset

- The data is split into two different datasets labeled as a **train** and a **test** dataset.
- This can be a 60/40 or 70/30 or 80/20 percentage splits.
- Potential problem: uneven distribution of different classes of data is found in training and test dataset.
- To fix this, the training and test dataset is created with equal distribution of different classes of data. This process is called **stratification**.



Test options

○ Use training set

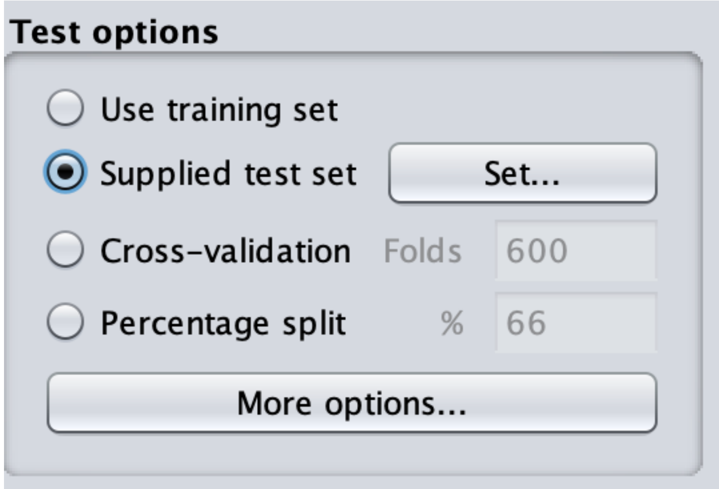
○ Supplied test set

○ Cross-validation Folds 10

● Percentage split % 66

Separate Test set

- Completely Separate test set (separate .arff)



Test options

○ Use training set

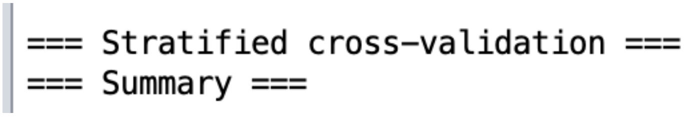
● Supplied test set

○ Cross-validation Folds 600

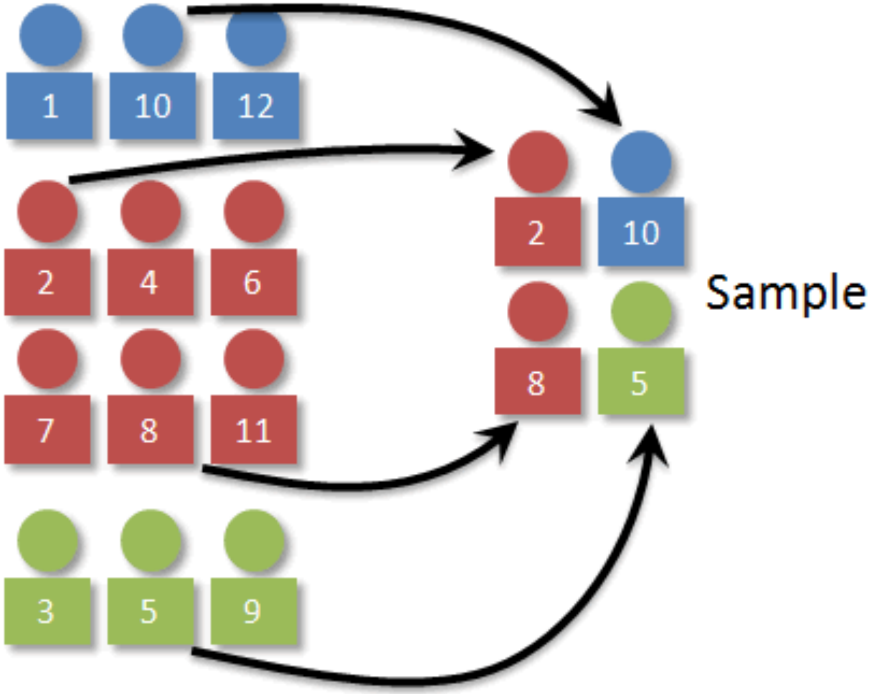
○ Percentage split % 66

Stratified sampling

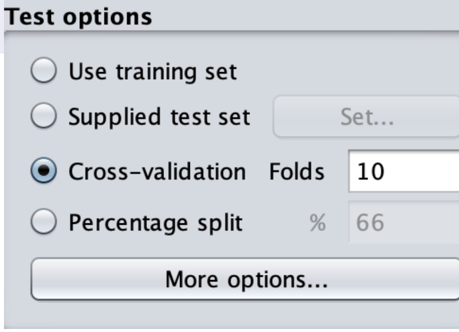
- Keep the class distribution after the sampling
- Make sure that each class is represented with approximately equal proportions in the subsets.
- In Weka by default



```
=== Stratified cross-validation ===
=== Summary ===
```



Cross validation (K-fold CV)



Test options

○ Use training set

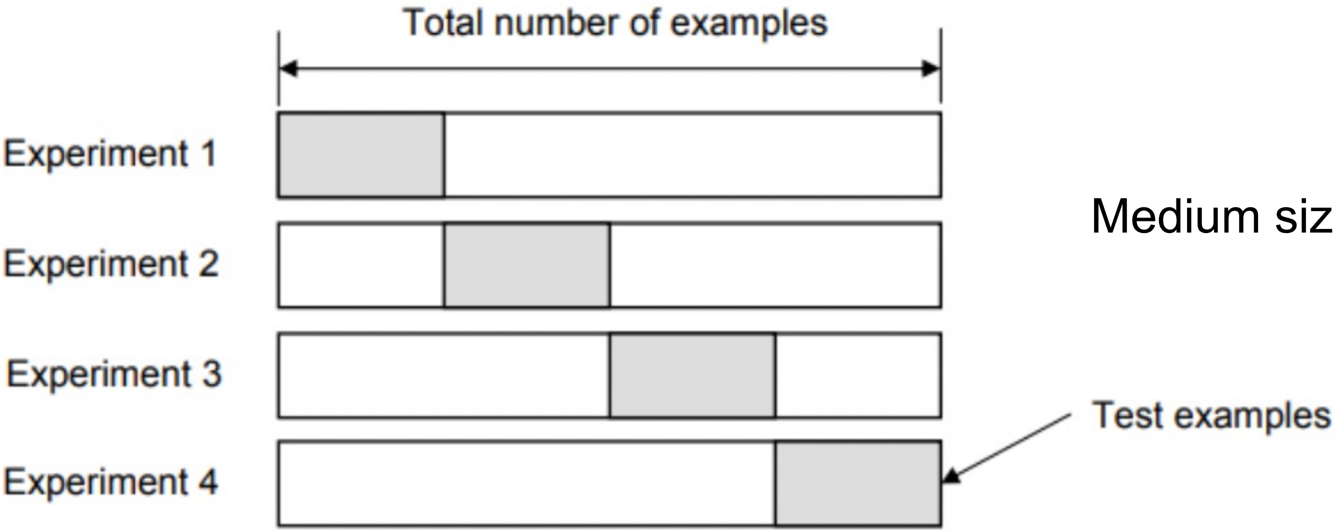
○ Supplied test set Set...

● Cross-validation Folds 10

○ Percentage split % 66

More options...

- k-1 folds are used for training and the remaining one is used for testing. This is repeated for each fold
- The entire data is used for training and testing.
- The error rate of the model is average of the error rate of each iteration.
- The error rate could be improved by using stratification technique.

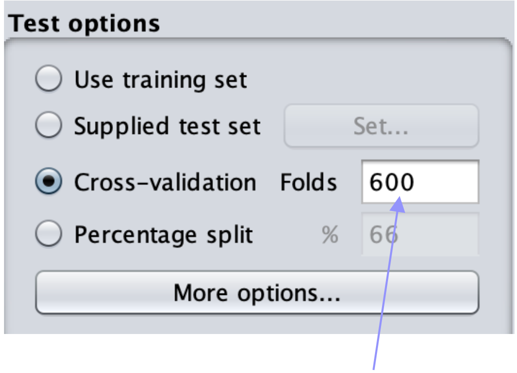


Medium sized datasets

https://vimeo.com/29569892

Leave-One-Out Cross-Validation (LOOCV)

- All the data except one instance is used for training and the left one instance is used for testing. This process is repeated for N times if there are N records.
- The advantage is that entire data is used for training and testing. The error rate of the model is average of the error rate of each iteration.



Test options

○ Use training set

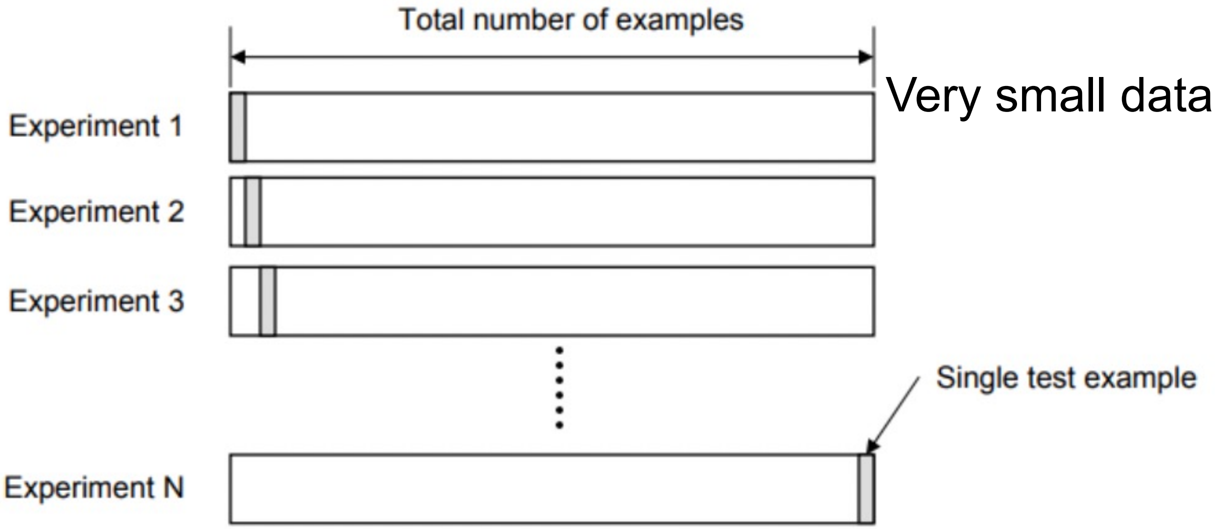
○ Supplied test set Set...

● Cross-validation Folds 600

○ Percentage split % 66

More options...

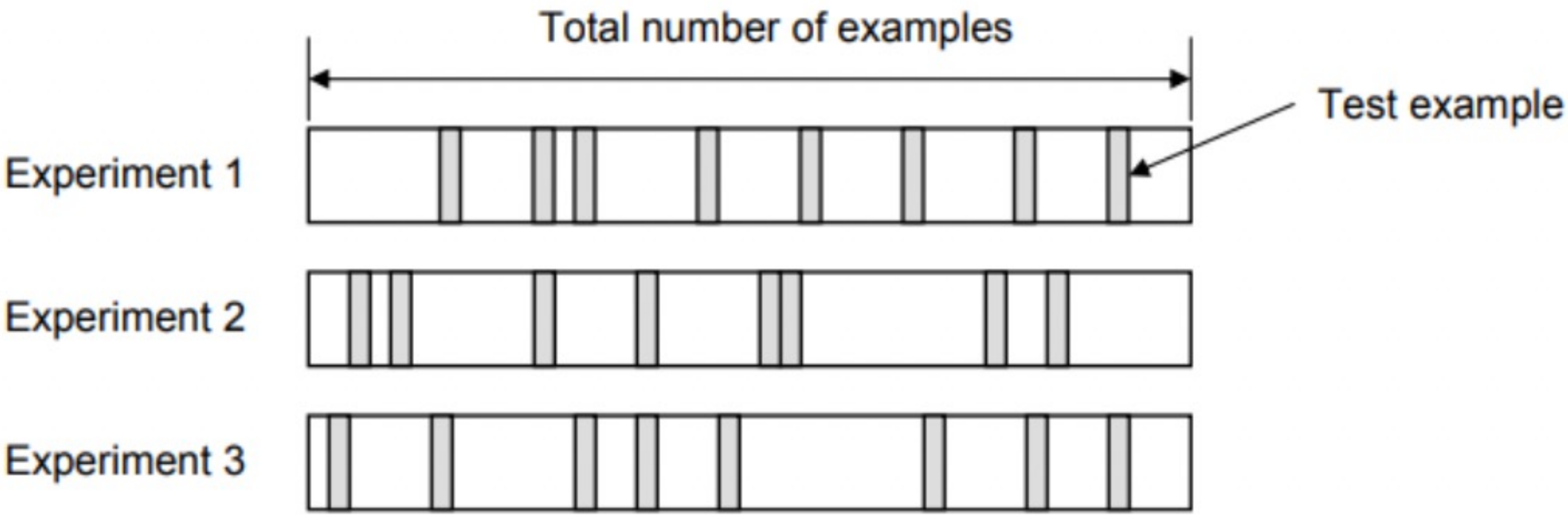
Number of instances



Very small datasets

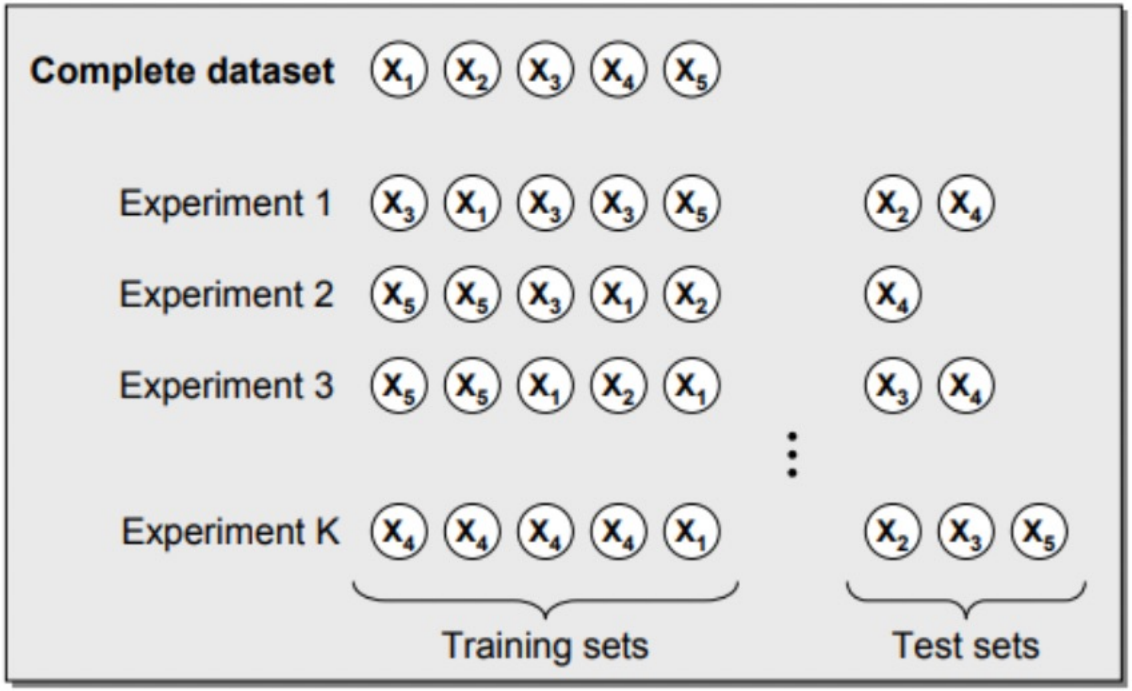
Random Subsampling

- Multiple instances are randomly chosen from the dataset and combined to form a test dataset. The remaining data forms the training dataset.
- The error rate of the model is the average of the error rate of each iteration.



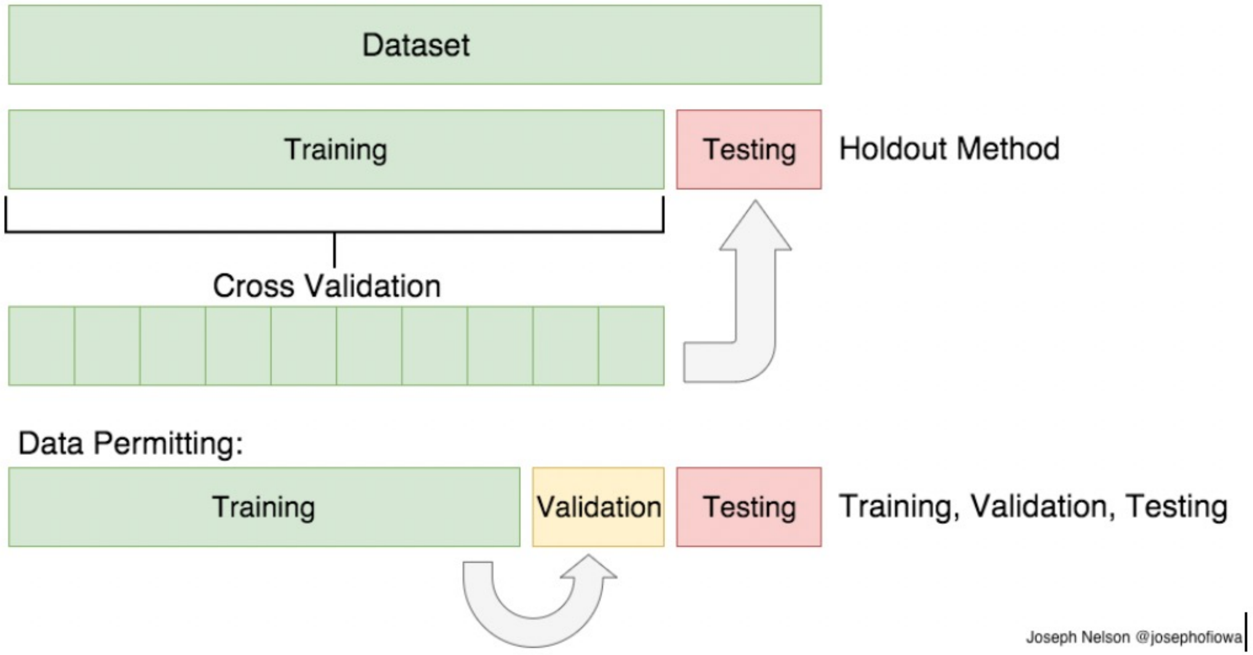
Bootstrapping

- The training dataset is randomly selected with replacement.
- The remaining examples that were not selected for training are used for testing.
- Unlike K-fold cross-validation, the value is likely to change from fold-to-fold.
- The error rate of the model is average of the error rate of each iteration.


Complete dataset x_1 x_2 x_3 x_4 x_5
Experiment 1 x_3 x_1 x_3 x_3 x_5 x_2 x_4
Experiment 2 x_5 x_5 x_3 x_1 x_2 x_4
Experiment 3 x_5 x_5 x_1 x_2 x_1 x_3 x_4
⋮
Experiment K x_4 x_4 x_4 x_4 x_1 x_2 x_3 x_5
Training sets
Test sets

Train, Validation, Test

- The data is split in 3 parts.
  - Train – for training the model
  - Validation – to validate the model (while still training it) and to adjust parameters. This set is commonly used for hyperparameter optimization, Meta-learning, etc.
  - Test – to test the model and estimate the performance

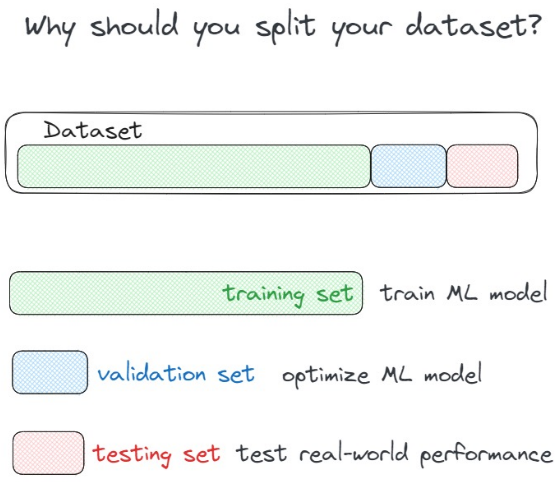


Joseph Nelson @josephofiowa

Data Split

When training a Machine Learning model is essential to **split your available data into training, validation and testing**

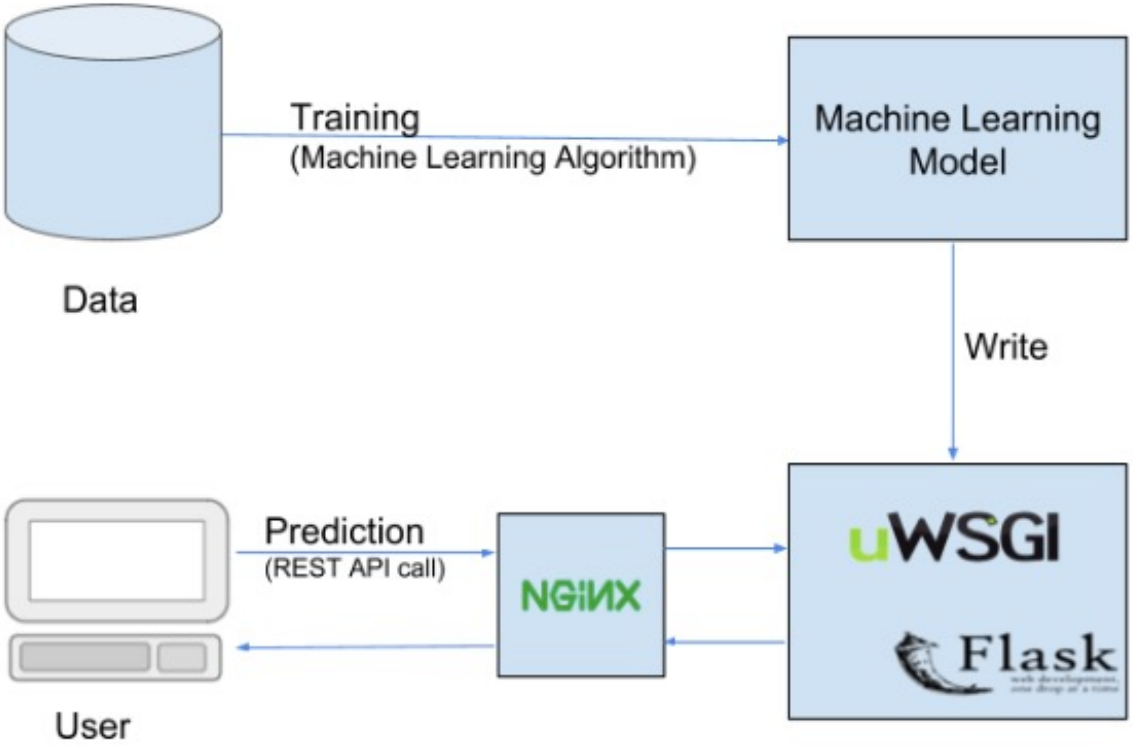
- Preventing **Data Leakage**: occurs when information from the **validation** or **test** set unintentionally influences the training process, leading to **optimistic performance** estimates. By keeping the validation and test sets separate from the training data, you ensure fair evaluations and avoid data leakage.
- Preventing **Overfitting**: The validation and test sets **help detect/prevent** overfitting by providing an independent evaluation since these sets are not seen by the model during training.
- **Iterative Model Improvement**: splitting data into sets allows you to iterate and optimize your model based on the performance metrics observed on the validation set, enhancing the model's capabilities.
- **Unbiased evaluation**: by keeping the test set unused, you can assess the performance of your final model simulating a real-world scenario.



Summary

- Once evaluation is complete, all the data can be used to build the final classifier.
- The larger the test data the more accurate the error estimate.
- Use test sets and the hold-out method for “large” data;
- Use the cross-validation method for “middle-sized” data;
- Use the leave-one-out and bootstrap methods for small data;
- Don’t use test data for parameter tuning - use separate validation data.

ML Model deployment



https://cloudxlab.com/blog/deploying-machine-learning-model-in-production/

WEKA

- 10-fold CV
- Use training set
- Percentage split (Random Seed)
- Supplied Test set