Data Science

معالجة اللغات الطبيعية
Natural Language Processing(NLP)

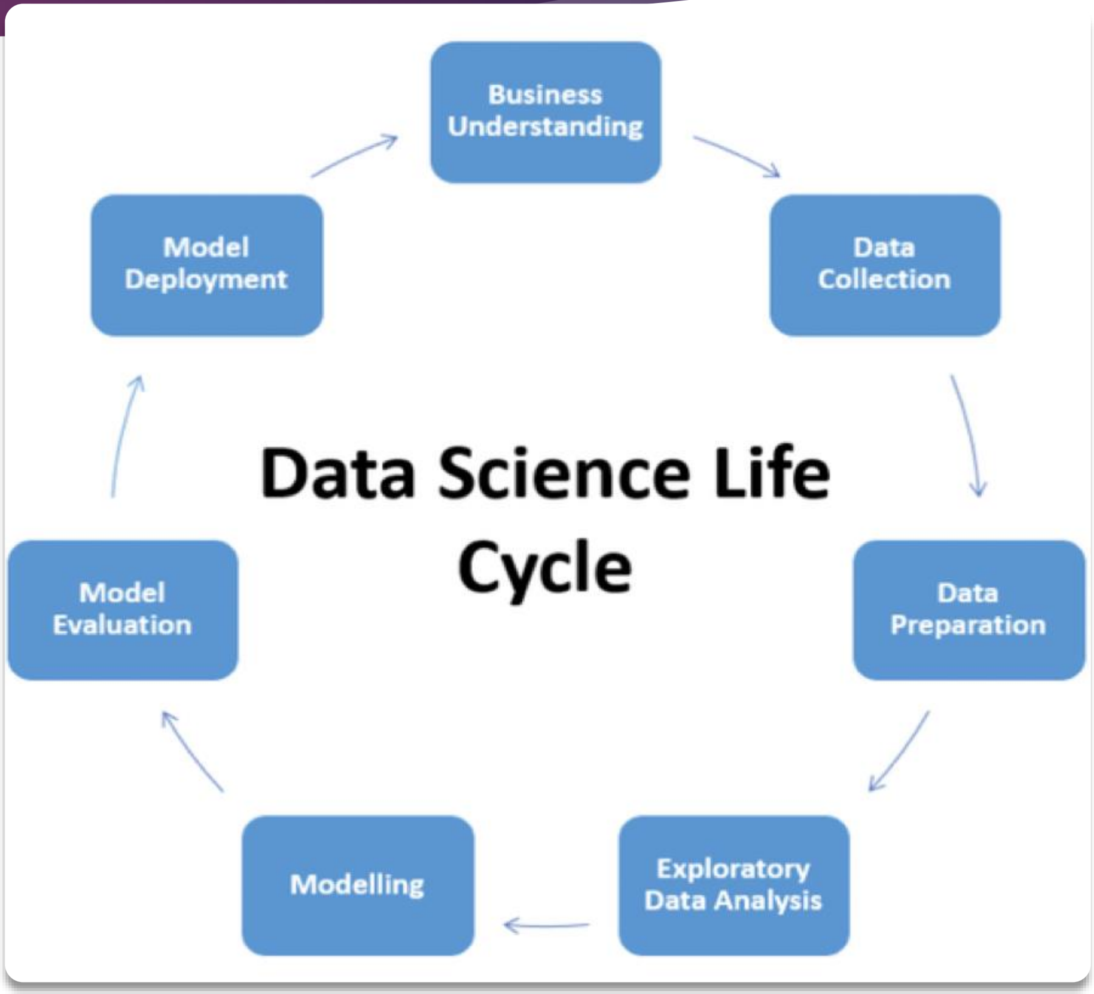
تطبيق عملي
إختبار و تقييم النموذج –(خوارزمية تعلم الاله)
Model Evaluation(1)

SMS SPAM Filtering

AHMAD SHHADEH

Data Science Project life cycle

- Business Understanding
- Data Collection
- Data Preparation
- Exploratory data analytics(EDA)
- Model Building
- **Model Evaluation**
- Model Deployment

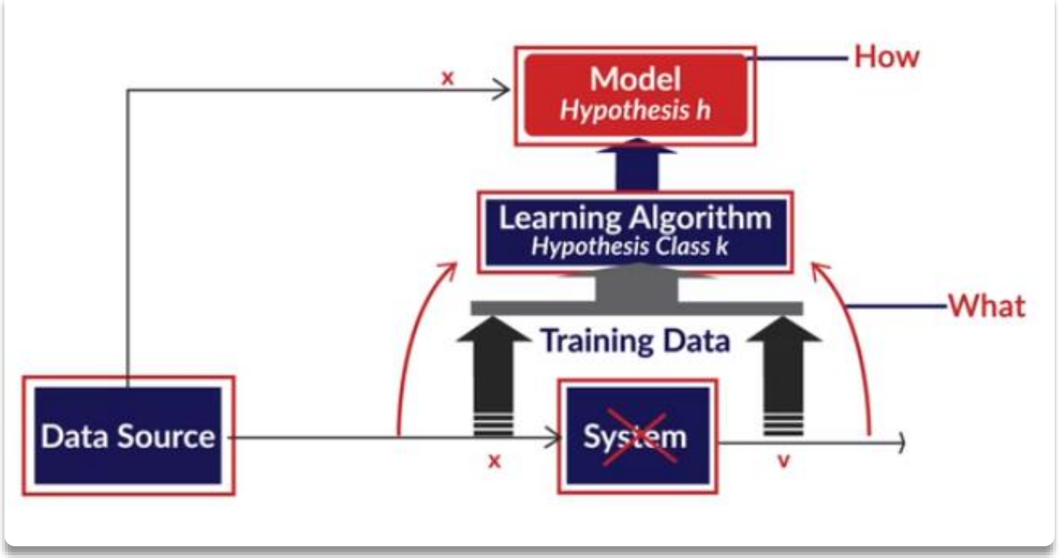


What is Model Evaluation? And why it is matter ?

- **Model Evaluation** is the process through which we **quantify the quality of a system's predictions**.

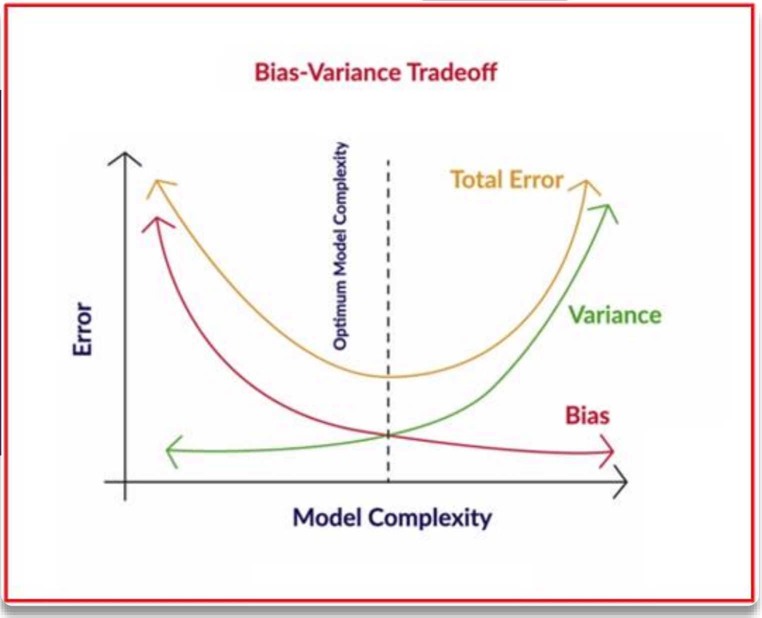
Machine learning Algorithm Vs Model

- **An "*algorithm*"** in machine learning is a procedure that is run on data **to create a machine learning "*model*.**
  - learning algorithm learns from training data and produces a model
- **A "model"** in machine learning is the **output of a machine learning** algorithm run on data.
  - A model represents what was learned by a machine learning algorithm.



Model Evaluation

- Machine learning is **training an algorithm on a set of known examples** with a clear goal of **generalizing to unseen examples.**
- In other words; While **training a model is a key step**, on another hand **how the model generalizes on unseen data** is an equally important aspect that should be considered in every machine learning pipeline.
- We need to know whether it works and, consequently, if we can **trust its predictions**. Could the model be **merely memorizing** the data it is fed with, and therefore unable to make good predictions on future samples, or samples that it hasn't seen before?

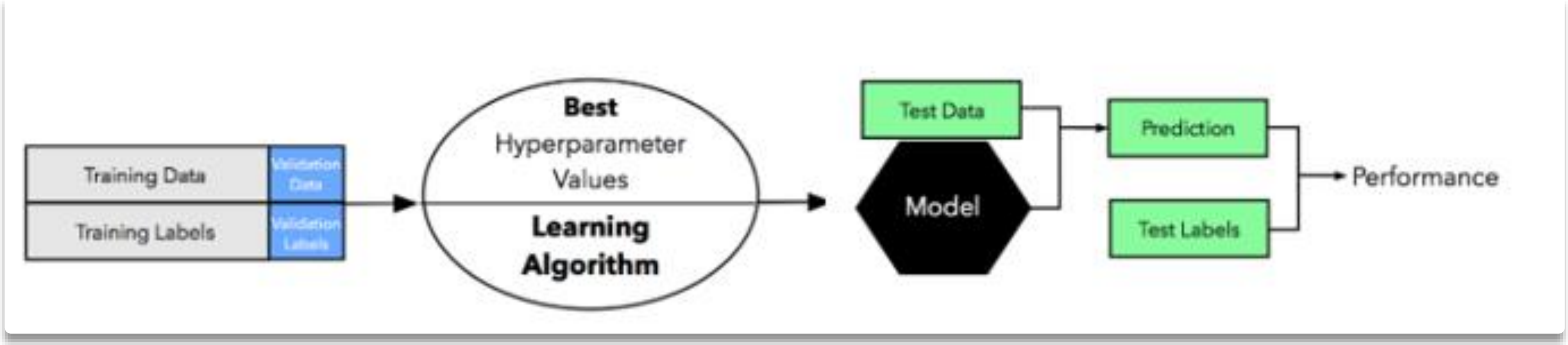


What is bias?
Bias is the difference between the average prediction of our model and the correct value which we are trying to predict. Model with high bias pays **very little attention to the training data** and oversimplifies the model. It always leads to high error on training and test data.
What is variance?
Variance is the variability of model prediction for a given data point or a value which tells us spread of our data. Model with high variance pays a **lot of attention to training data** and does not generalize on the data which it hasn't seen before. As a result, such models perform very well on training data but has high error rates on test data.

Model Evaluation

- The idea of **building machine learning models** works on a **constructive feedback principle**. we build a model, get feedback from metrics, make improvements and continue until we achieve a desirable accuracy.

- For model evaluation we usually works on below three areas:
  1. **The Training data**
  2. The Machine learning algorithms
  3. The hyperparameters for each MLA



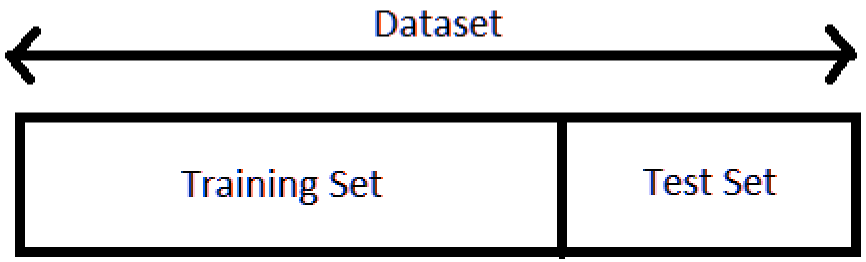
Model Evaluation- Data Splitting

- The key point to remember here is that a model **should never be evaluated on data that it has already seen before**. With that in mind, you will have either one of the following two cases:

  1. The training data is **abundant**
     - In this case is straightforward because you can use **as many observations as per your preference** to train and test the model.
  2. The training data is **limited**.
     - In this case, however, you will need to find some ‘hack’ so that the model can be evaluated on unseen data and, **simultaneously, does not eat up the data available for traini**ng.
     - In this case will have below three scenarios :
       I. **Split into train and test sets**: Tuning a hyperparameter makes the model 'see' the test data. Also, the results are dependent on the **specific train-test split.**
       II. **Split into train, validation and test sets**: The validation data would **eat into the training** set.
       III. **Applying cross-validation, split the data into train and test sets and train** multiple models by sampling the train set. Finally, you only use the test set to test the hyperparameter once.

Train-Test split technique

- The train-test split is a **technique for evaluating the performance** of a machine learning algorithm, It can be used for classification or regression problems and can be used for any supervised learning algorithm.
- The procedure involves taking a dataset and **dividing it into two subsets**.
  - **Train Dataset**: Used **to fit** the machine learning model.
  - **Test Dataset**: Used **to evaluate** the fit machine learning model: not used to train the model
- The objective is to estimate the performance of the machine learning model on new data: data not used to train the model.

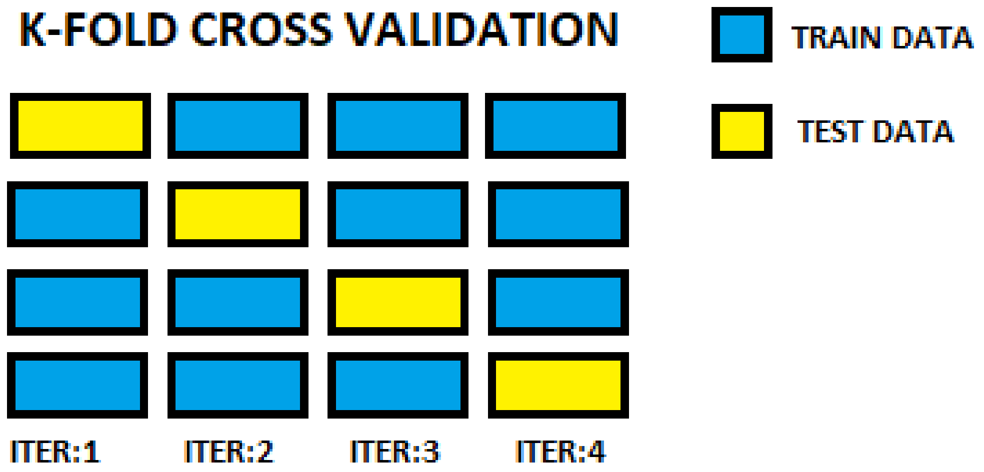


Cross-Validation technique

- **Cross-validation** is a technique that involves **partitioning the original observation** dataset into a training set, used to train the model, and an **independent set used to evaluate** the analysis.
- The various types of cross-validation are as follows:
  - k-fold Cross-Validation
  - Leave One Out (LOO)
  - Leave p-Out (LPO)
  - Stratified k-Fold
- The most common cross-validation technique is k-fold cross-validation,

k-fold Cross-Validation

- **In the k-fold CV** : The original dataset is partitioned into **k equal size subsamples**, called folds.
- The k is a **user-specified number**, usually with 5 or 10 as its preferred value.
- This is repeated k times, such that each time, **one of the k** subsets is used as the test set/validation set and the other **k-1 subsets are put** together to form a training set.
- The error estimation is averaged over all k trials to get the total effectiveness of our model.
- For the example in the right side :we divided the training data into k-groups of samples. If k=4 (say), you use k-1 folds to build the model and test the model on the **k^{th} fold**.



Note : Every data point gets to be in a test set exactly once and gets to be in a training set k-1 times. This significantly **reduces bias**, as we're using most of the data for fitting, and it also significantly **reduces variance**, as most of the data is also being used in the test set. Interchanging the training and test sets also adds to the effectiveness of this method.

Thank You