**Model Selection**

After we built our Machine Learning models, some questions remained unanswered:

1. How to deal with the bias variance tradeoff when building a model and evaluating its performance ?
2. How to choose the optimal values for the hyperparameters (the parameters that are not learned) ?
3. How to find the most appropriate Machine Learning model for my business problem ?

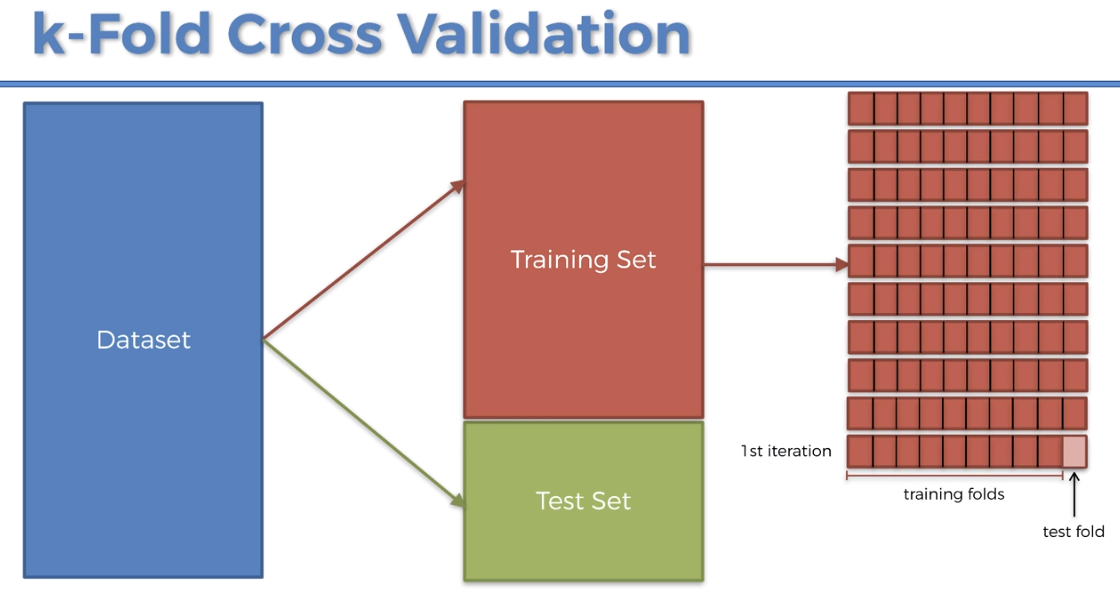
In this part we will answer these questions thanks to Model Selection techniques including:

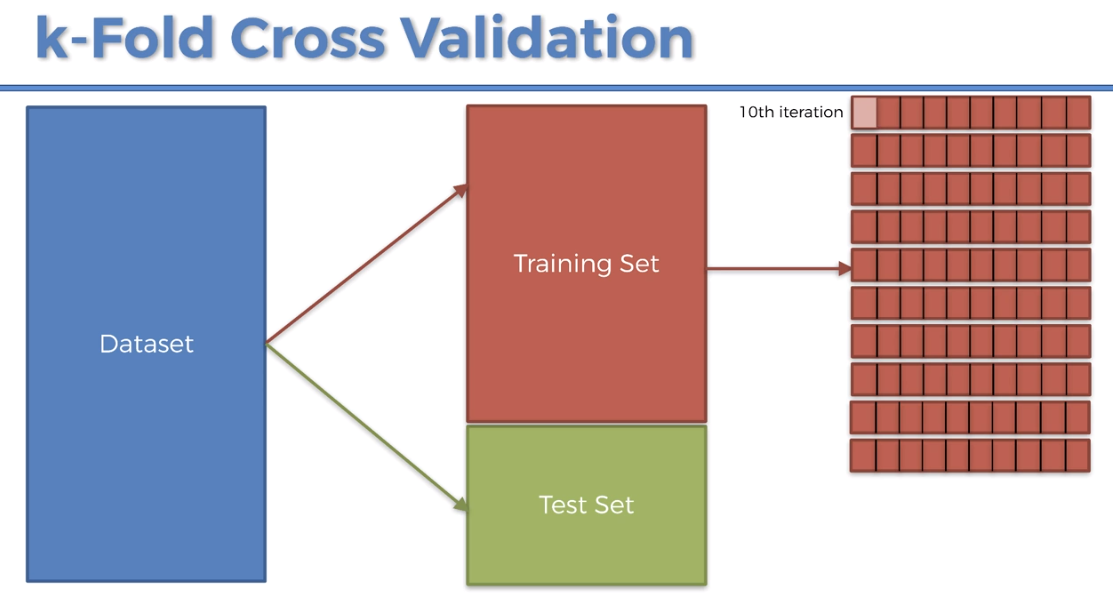
1. k-Fold Cross Validation
2. Grid Search

Eventually we will finish this course by a last bonus section included in this part, dedicated to one of the most powerful Machine Learning model, that has become more and more popular: XGBoost.

**K-Fold Cross Validation:**

It will fix the variance problem. If you change the dataset of training , the accuracy will differ. So we use K-Fold to fix it.



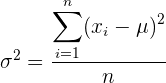


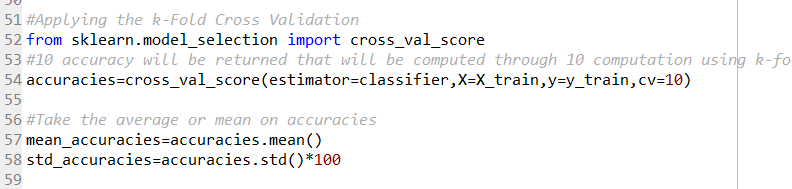
10 iteration will be done.

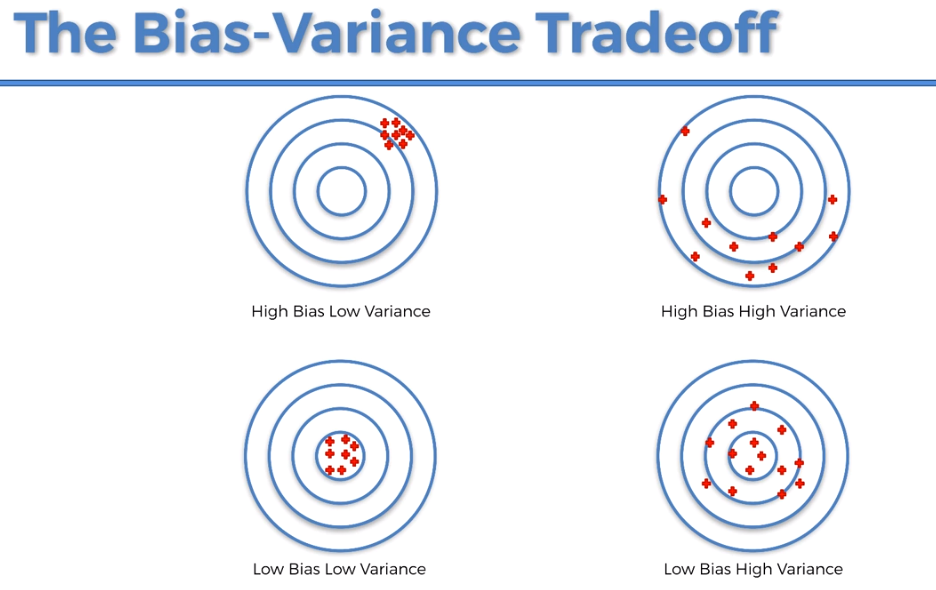
What is high bias in machine learning?

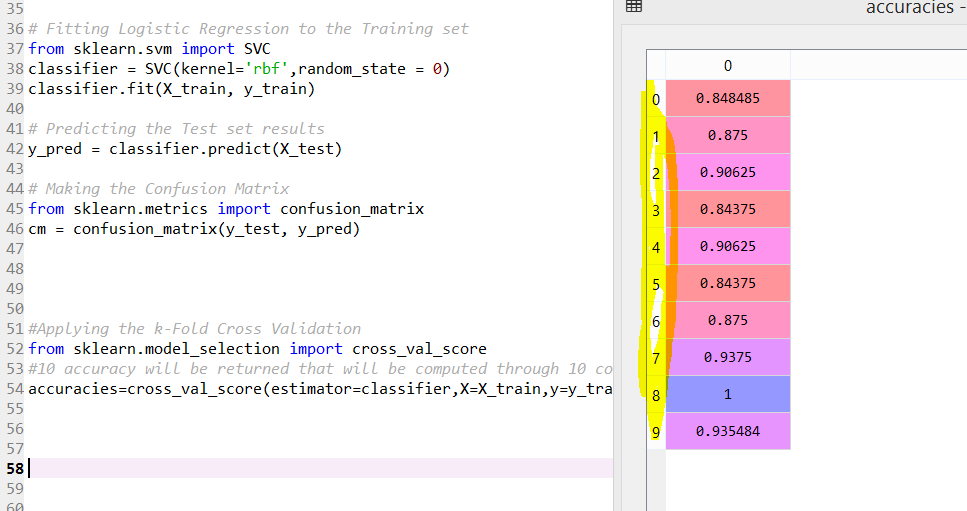
The **bias** is an error from erroneous assumptions in the **learning** algorithm. **High bias** can cause an algorithm to miss the relevant relations between features and target outputs (underfitting). The variance is an error from sensitivity to small fluctuations in the training set.

[Variance](https://www.investopedia.com/terms/v/variance.asp) and covariance are mathematical terms frequently used in statistics, and despite the similar-sounding names they actually have quite different meanings. A [covariance](https://www.investopedia.com/terms/c/covariance.asp) refers to the measure of how two random variables will change together and is used to calculate the correlation between variables. The variance refers to the spread of the data set—how far apart the numbers are in relation to the mean, for instance. Variance is particularly useful when calculating the probability of future events or performance.  
**Variance:** The average of the squared differences from the mean. Here is the formula which we will use in our python code.













**Grid Search🡪improving model performance**

How do I know which parameter to select while building the model?

🡪Use grid search

How do I select which model to solve the business problem?

* + Classify based on classification and regression
  + Ask it is linear(SVM) or non-linear(Kernal SVM) problem
    - * Grid search helps in finding linear or non-linear

**Definition :** SVC**(**self**,** C**=**1.0**,** kernel**=**'rbf'**,** degree**=**3**,**gamma**=**'auto'**,** coef0**=**0.0**,** shrinking**=**True**,**probability**=**False**,** tol**=**1e-3**,** cache\_size**=**200**,**class\_weight**=**None**,** verbose**=**False**,** max\_iter**=**-1**,**decision\_function\_shape**=**'ovr'**,** random\_state**=**None**)**

**Include the parameters of the algorthim to the grid search.**

