**Machine Learning: -**Field of study that gives computers the ability to learn without being explicitly programmed.

A computer program is said to *learn* from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E

**Machine learning algorithms:**

* Supervised learning
* Unsupervised learning

Other algorithms are Reinforcement learning, recommender systems etc.

**Supervised learning:** We have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output. (Labeled data).

* **Classification**: When the output variable is a category, such as “red” or “blue” or “disease” and “no disease”.
* **Regression**: When the output variable is a real value, such as “dollars” or “weight”.

**Unsupervised learning:** Unsupervised learning is where you only have input data (X) and no corresponding output variables.

* **Clustering**: Discovering inherent groupings in the data, such as grouping customers by purchasing behavior.
* **Association**:  You want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.

**Machine Learning Use Case:**

1. To show related News based on search.
2. Gene and genome data analysis.
3. Organizing Computer Clusters.
4. Social Network Analysis.
5. Market Segmentation.
6. Astronomical data analysis.

**Feature Selection:** Feature selection is an important step. And good features:

* Lead to Data Compression
* Retain relevant information
* Created based on Exerts knowledge.

Common mistakes:

* Trying to automate feature selection
* Not paying attention to data specific quirks
* Throwing away information unnecessarily.

**Steps for ML Solution Design:**

* Define your error rate.
* Split Data into Training, Test & Validation.
* On the training set pick features
  + Use cross-validation
* On the training set pick prediction function.
  + Use cross-validation
* If you have validation data set
  + Apply to test data set and Refine
  + Now apply to validation set
* If you have validation data set
  + Then apply to test set
* All data sets train, test and validation must reflect structure of the problem.
* All subsets should reflect as much diversity as possible. Random assignment does this, or sometimes we need to balance it –tricky.

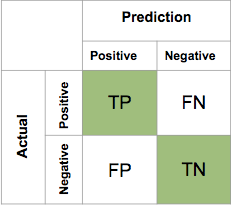
**Type of Errors: Confusion Matrix**

True Positive (TP) = Correctly Identified

False Positive (FP) = Incorrectly Identified

True Negative (TN) = Correctly Rejected

False Negative (FN) = Incorrectly Rejected



Key Quantities:

* **Accuracy**= Pr(Correct Outcome) = (TP+TN)/(TP+TN+FP+FN)
* **Sensitivity/Recall** =Pr(positive Test|disease) = TP/(TP+FN) Also called **True Positive Rate or**
* **Precision** Positive Predicted Value = TP/(TP+FP)
* **Specificity** =Pr(Negative Test|No Disease) = TN/(FP+TN) = 1-False Positive Rate (**FPR**)
* Negative Predicted Value = TN/(FN+TN)

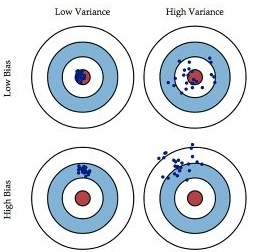
**Bias vs Variance Tradeoff:**

Mathematically, it is defined as

Err(x)=Bias+Variance+Irreducible Error

**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.

**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.



Models with **low bias** (which can learn from the training data well) often have high variance (and therefore an inability to generalize to new data), and this phenomenon is referred to as “overfitting”.

**Types of Error:**

**Type I (alpha)** errors happen when we reject a true null hypothesis (false positive). In terms of the courtroom example, a type I error corresponds to convicting an innocent defendant.

**Type II (beta)** errors happen when we fail to reject a false null hypothesis (false negative). In terms of the courtroom example, a type II error corresponds to acquitting a criminal

**alpha** is also called significance level and

**1-alpha** is called confidence level.

**1-beta** is called power of statistical test and is probability to reject null hypothesis when it is false.