**BASICS OF MACHINE LEARNING**

**Training dataset vs Testing dataset:**

Training data is the initial dataset you use to teach a machine learning application to recognize patterns or perform to your criteria, while testing or validation data is used to evaluate your model's accuracy.

**Forecast vs Prediction:**

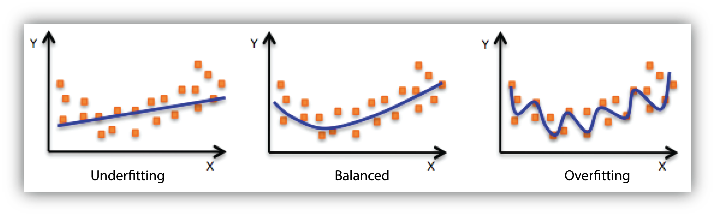
A forecast refers to a calculation or an estimation which uses data from previous events, combined with recent trends to come up a future event outcome. On the other hand, a prediction is an actual act of indicating that something will happen in the future with or without prior information.

**Supervised Learning:**

Supervised learning in machine learning is one method for the model to learn and understand data. There are other types of learning, such as unsupervised and reinforcement learning, but those are topics for another time and another blog post. With supervised learning, a model is given a set of labeled training data. The model learns to make predictions based on this training data, so the more training data the model has access to, the better it gets at making predictions. With training data, the outcome is already known. The predictions from the model and known outcomes are compared, and the model’s parameters are changed until the two align. The point of training is to develop the model’s ability to successfully generalize.

**Generalization vs Overfitting vs Underfitting:**

Generalization is a term used to describe a model’s ability to react to new data. That is, after being trained on a training set, a model can digest new data and make accurate predictions. A model’s ability to generalize is central to the success of a model. If a model has been trained too well on training data, it will be unable to generalize. It will make inaccurate predictions when given new data, making the model useless even though it is able to make accurate predictions for the training data. This is called overfitting. The inverse is also true. Underfitting happens when a model has not been trained enough on the data. In the case of underfitting, it makes the model just as useless and it is not capable of making accurate predictions, even with the training data.

[](https://wp.wwu.edu/machinelearning/files/2017/01/mlconcepts_image5-rnehsa.png)

The figure demonstrates the three concepts discussed above. On the left, the blue line represents a model that is underfitting. The model notes that there is some trend in the data, but it is not specific enough to capture relevant information. It is unable to make accurate predictions for training or new data.  In the middle, the blue line represents a model that is balanced. This model notes there is a trend in the data, and accurately models it. This middle model will be able to generalize successfully. On the right, the blue line represents a model that is overfitting. The model notes a trend in the data, and accurately models the training data, but it is too specific. It will fail to make accurate predictions with new data because it learned the training data too well.

Test Dataset vs Validation Dataset:

A validation dataset is a sample of data held back from training your model that is used to give an estimate of model skill while tuning model’s hyperparameters.

The validation dataset is different from the test dataset that is also held back from the training of the model, but is instead used to give an unbiased estimate of the skill of the final tuned model when comparing or selecting between final models.

**True Positive vs True Negative vs False Positive vs False Negative:**

A true positive is an outcome where the model correctly predicts the positive class.

A true negative is an outcome where the model correctly predicts the negative class.

A false positive is an outcome where the model incorrectly predicts the positive class.

A false negative is an outcome where the model incorrectly predicts the negative class.