**Chapter 1 Summary: Intro to Machine Learning**

Machine learning, also called predictive analytics or statistical learning, is a field of research that combines statistics, artificial intelligence, and computer science to generate insights from data. This is a valuable alternative to using traditional hardcoded rules to process data when a system is more dynamic with more complex tasks than can be completed directly from patterns in the data.

The two primary types of machine learning algorithms are supervised and unsupervised learning. Supervised learning generalizes known data to automate decision making. To do this, models are trained using inputs and their corresponding outputs and develop a method to identify an output, given an input. This allows models to predict outputs for previously unseen inputs. However, the required input/output pairs can make data collection and labeling time intensive. Conversely, unsupervised learning is trained using inputs without known outputs. This is harder to understand and evaluate than supervised learning.

Both types of algorithms require inputs to be represented in a way that computers can understand. Individual inputs are referred to as samples and output properties describing the inputs as features. Feature extraction/engineering is the process of building a good representation of data and is vital to the success of machine learning algorithms. Another essential step in the preparation process is understanding your data to ensure selection of the best features and algorithms to develop the most relevant solution to the problem at hand. Valuable guiding questions include:

* Can the data collected answer the problem being solved?
* How can the research questions best be phrased as a machine learning problem?
* Is there enough data collected to adequately represent the problem?
* Do the selected features facilitate the right predictions?
* How will success be measured?
* How will the model be used to benefit the bigger picture (i.e. research goals, etc)?

Python is the language most commonly used for data science applications due to its advantages that closely align with the goals and requirements of machine learning. Python includes many libraries such as NumPy, SciPy, pandas, and matplotlib. These built in libraries allow for data loading, visualization, statistics, natural language processing, and image processing. Another key advantage is the ability to easily interact directly with the code and quick iteration through the code. One open-source project that can be accessed and used as a tool to learn machine learning methods is scikit-learn.

Understanding Jupyter Notebook, NumPy, SciPy, matplotlib, and pandas is vital for success in learning machine learning methods using scikit-learn. Jupyter Notebook is an interactive coding environment that includes code, text, and visualizations, making it an ideal tool for exploratory data analysis. NumPy is a package in python used for scientific computing that includes multidimensional arrays, high-level math functions, and pseudorandom number generators. SciPy is a python package for scientific computing including advanced linear algebra, function optimization, signal processing, special math functions, and statistical distributions. Matplotlib is a python library with functions for producing visualizations such as line charts, histograms, and scatter plots. Pandas is a python library for data manipulation and management using a table-like structure called DataFrames, modeled after R DataFrames, that can be accessed with SQL-like queries.

A simple example of a supervised machine learning application is a model that predicts the iris species given measurements of known iris species. The iris species are classes and the individual sample measurements are labels. First the data is cleaned and prepared for training. Since the goal is to identify the species, this is a classification problem that the K Nearest Neighbors (KNN) classification algorithm can be used to build a model for. A predetermined randomized test set of never-before-seen data, as it was held out of the training set, is used to evaluate the model after training. The calculated test set accuracy measures the percent of correct predictions are made for the test set, simulating the accuracy that can be expected in application. Broadly, the 3 core stages of a machine learning workflow presented in this example, are data preparation, model building, and model evaluation.