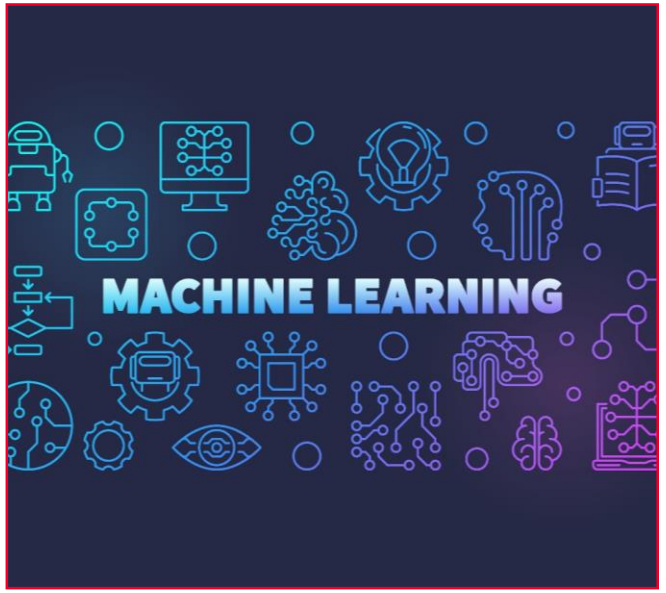
PROAGRO LECTORIUM

Analyzing Big Data with Machine Learning Methods

Vadim Demichev, Candidate of Economic Sciences, Associate Professor of the Department of Statistics and Cybernetics of the Russian State Agrarian University — Moscow Timiryazev Agricultural Academy

Lecture topics

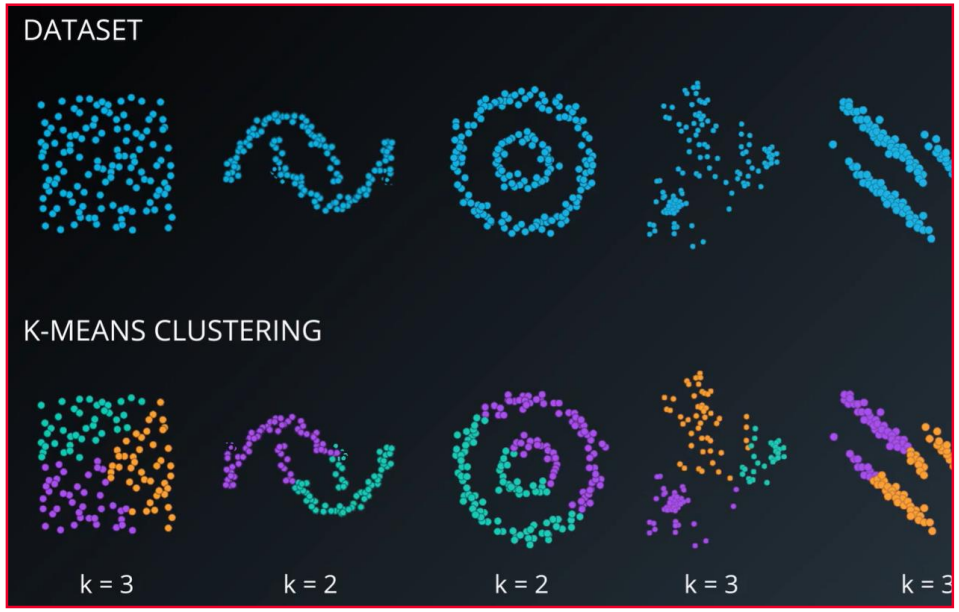
- Introduction to machine learning
- Benefits of using machine learning
- Essential Python libraries for machine learning
- Model building process



Introduction to machine learning

- **Machine learning** is the field of study that gives computers the ability to learn without being explicitly programmed (Arthur Samuel).
- **Machine learning** is the process by which a computer can work more accurately as it collects and learns from the data it is given (Mike Roberts).
- **Machine learning** is a broad subsection of artificial intelligence that studies methods for constructing algorithms capable of learning.

Introduction to machine learning



There are three key techniques used in machine learning:

- Regression
- Classification
- Clustering

Introduction to machine learning

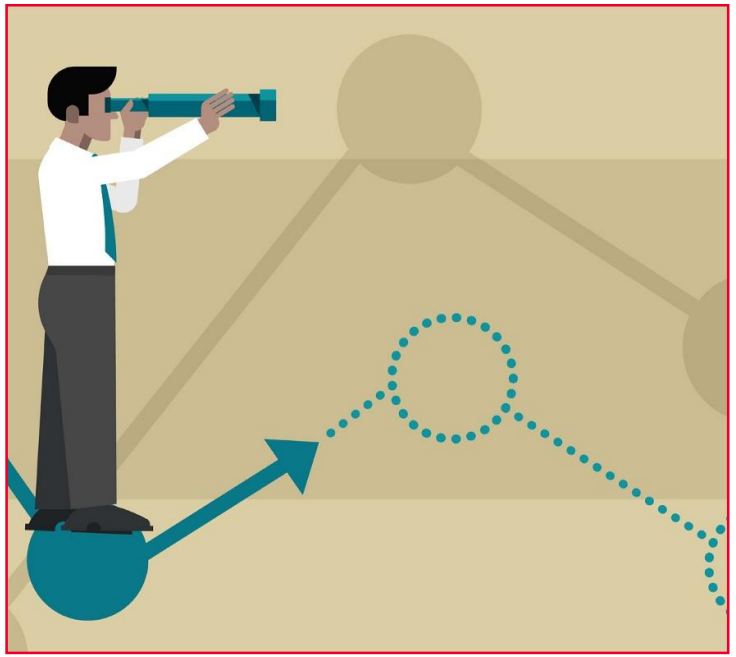
Key concepts of machine learning:

- The label is what we predict, the variable Y in a simple linear regression (also called the target variable).
- The feature is the input variable—variable X in a simple linear regression (they are also called independent variables).
- In machine learning, the notion of a labeled dataset and an unlabeled dataset is also distinguished. The labeled dataset is used to train machine learning models.

Introduction to machine learning

Two stages of the model's life are distinguished:

- Training: it involves creating or learning a model. This means you show the model labeled data and allow it to gradually learn the relationships between the features and the label.
- Prediction: it involves applying a trained model to unlabeled data. This means you use the trained model to make useful predictions (Y). For example, during prediction, you can forecast the value of an apartment for new unlabeled examples.



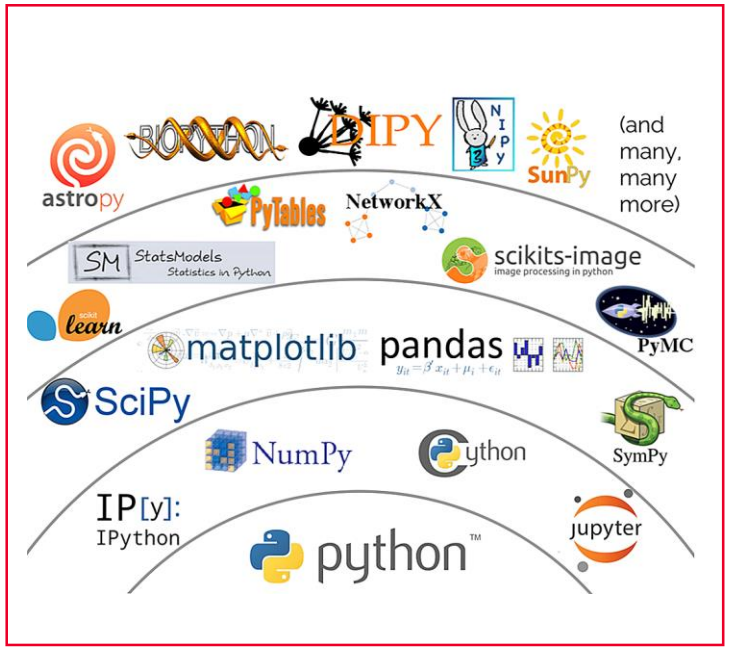
Benefits of using machine learning

- Practical examples of the use of machine learning models include:
- Vegetation analysis (weed identification, vegetation classification, plant disease identification, yield prediction).
- Soil analysis (soil assessment, soil classification, soil fertility prediction).
- Searching for oil fields, gold mines, and archaeological sites based on information about existing sites (classification and regression).
- Finding names of people or places in a text (classification).
- Identifying people from photographs or voice recordings (classification).
- Identification of profitable customers (regression and classification).
- Actively identifying parts of a vehicle where failure is more likely to occur (regression).
- Predicting the amount a person will spend on product X (regression).
- Forecasting a company's annual income (regression).

Essential Python libraries for machine learning

A list of Python libraries for machine learning:

- Libraries for putting data into memory: NumPy, Matplotlib, Pandas, StatModels, Scikit-learn, NLTK, and others
- Operation optimization libraries: Numba, PyCUBA, Blaze, Cython, and others
- Libraries for connecting machine learning tools to data warehouses: PyDoop, PySpark, Hadoopy, and others.



Model building process

Modeling in machine learning consists of 4 steps:

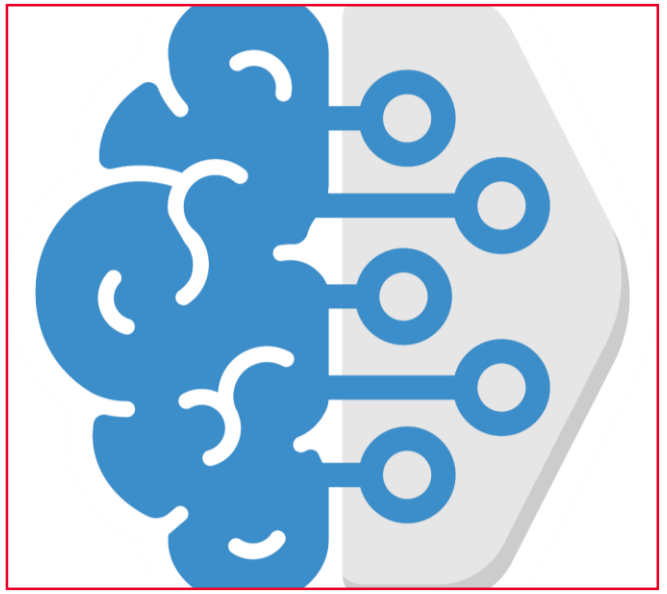
- Indicator planning and model selection
- Model training
- Adequacy testing and model selection
- Applying a trained model to unfamiliar data

Model building process

Indicator planning is defining and identifying possible model variables. This task is one of the important ones because these are the variables used to make forecasts.

Model building process

Model training. With the correct free variables and a planned modeling method in place, model training is started. In this phase, the model is given data from which it could learn.



Model building process

Checking the model adequacy. A good machine learning model has two properties:

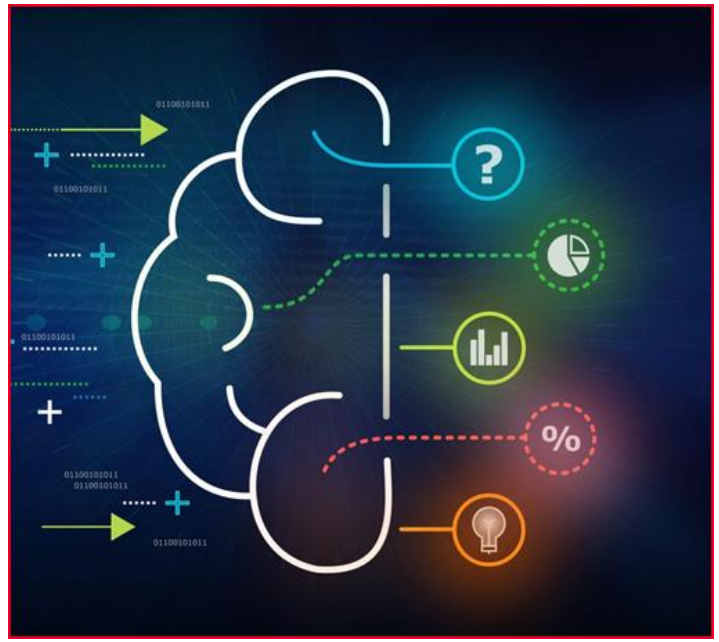
- It should be good at predicting values.
- It should work effectively on data previously unknown to it.

To evaluate these properties, error metrics and adequacy checking strategies are defined.

Model building process

Error metrics in machine learning:

- classification error rate (for classification tasks)
- RMS error (for regression problems)



Model building process

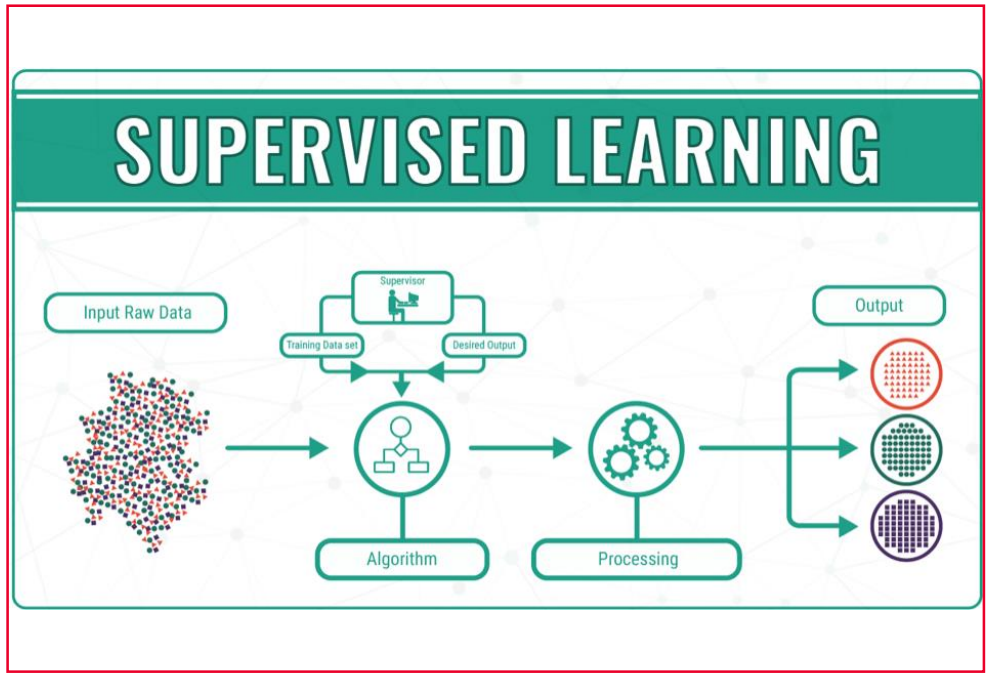
Strategies for adequacy testing:

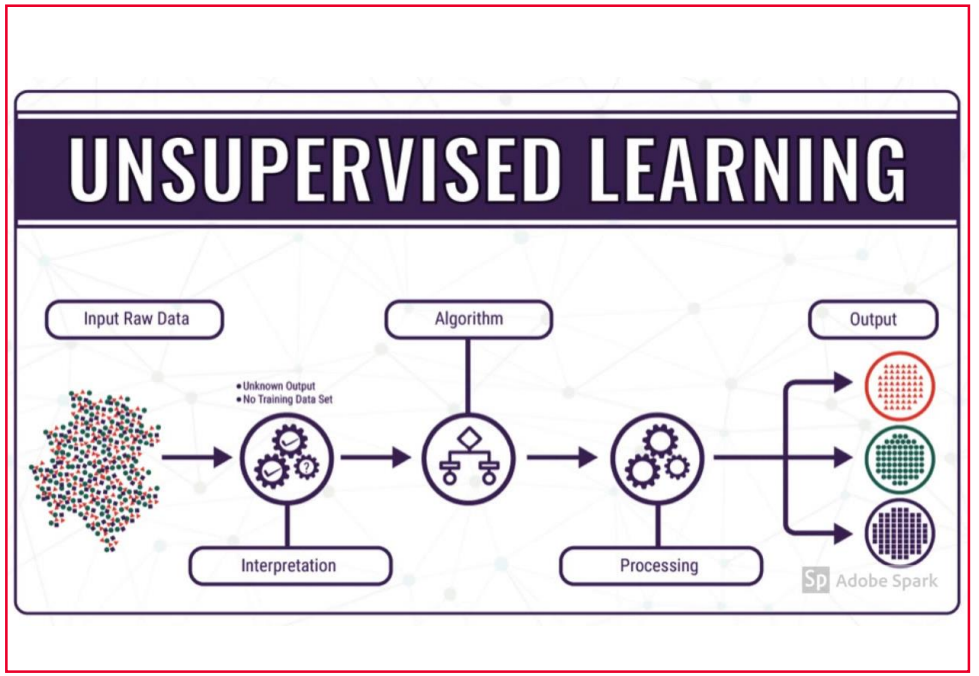
- Splitting the data into a training set with X% of the observations and a control sample with the rest of the data.
- K-fold cross-validation—the data set is divided into k parts. Each part is used once as a test dataset, while the remaining parts form the training dataset.
- Leave-one-out cross-validation—this method is identical to k-fold cross-validation, but with k = 1. One observation is always excluded and the remaining data are used for training.
- Regularization—this involves introducing a penalty for each additional variable used to build the model.
- L1 regularization involves building a model with the minimum possible number of independent variables.
- L2 regularization aims to minimize the discrepancies between the coefficients of the independent variables.

Model building process

Predicting new observations is the final step in the modeling process, which involves applying the model to unfamiliar data.

Model building process





According to the way labeled data is used, we distinguish between supervised learning methods, unsupervised learning methods, and semi-supervised methods.

PROAGRO
LECTORIUM

Thank you!