The 18 categories of knowledge in Scikit-Learn

Machine Learning Tasks:

- Supervised Learning: Training using labeled data, such as classification and regression tasks.
- Unsupervised Learning: Training on unlabeled data, for instance, clustering and dimensionality reduction tasks.
- Semi-supervised Learning: Combining labeled and unlabeled data for training.
- Reinforcement Learning: Learning through interaction with an environment to maximize reward signals by an agent.

Data Preprocessing:

- Feature Scaling: Ensuring that values of different features are on a similar scale, often achieved through standardization or normalization.
- Feature Selection: Choosing the most important features to reduce dimensionality and enhance model performance.
- Handling Missing Data: Managing missing values within the data, typically through mean imputation or interpolation methods.
- Data Encoding: Converting categorical variables into numerical forms, such as one-hot encoding.

Model Selection and Evaluation:

- Cross-Validation: Dividing the dataset into multiple folds to assess the model's performance and prevent overfitting.
- Model Evaluation Metrics: Such as accuracy, precision, recall, F1 score, ROC curves, and AUC.
- Overfitting and Underfitting: Understanding how to address the problems of overfitting and underfitting in a model.
- Hyperparameter Tuning: Methods for selecting the best model hyperparameters, for instance, using grid search or random search.

Supervised Learning Algorithms:

- Linear Regression: A linear model used for regression tasks.
- Logistic Regression: A linear model used for binary classification tasks.
- Decision Trees: Classification and regression models based on tree structures.
- Random Forests: An ensemble method based on multiple decision trees.
- Support Vector Machines: A powerful algorithm used for classification and regression tasks.

Unsupervised Learning Algorithms:

- K-Means Clustering: A clustering algorithm that divides data points into K clusters.
- Principal Component Analysis (PCA): A linear technique used for dimensionality reduction.
- t-Distributed Stochastic Neighbor Embedding (t-SNE): A non-linear dimensionality reduction method used for visualizing high-dimensional data.

Model Saving and Loading:

- Use the joblib or pickle libraries to save and load scikit-learn models.

- Models can be saved as binary files for reuse in different environments.

Custom Estimators and Pipelines:

- You can create custom scikit-learn estimators and transformers to cater to specific task requirements.
- Pipelines can be utilized to chain multiple processing steps together, creating an end-to-end workflow.

Applications in Specific Domains:

- Text Classification and Natural Language Processing (NLP): Methods for handling text data using scikit-learn.
- Image Processing: Employing libraries like scikit-image or OpenCV for image feature extraction and processing.

Feature Engineering:

- Feature Creation: Generating new features based on domain knowledge or characteristics of the data.
- Feature Extraction: Extracting useful information from raw data, such as extracting TF-IDF features from text or extracting edge features from images.
- Feature Combination: Combining multiple features to create new features that enhance model performance.

Model Interpretability:

- Feature Importance: Understanding the contribution of individual features in the model's predictions.
- Local Explanations: Explaining the reasons behind the model's prediction for individual samples, such as using LIME or SHAP methods.
- Global Explanations: Comprehending the behavior of the entire model, for instance, visualizing decision trees.

Ensemble Learning:

- Ensemble Methods: Combining multiple base models to enhance overall performance, such as Bagging and Boosting.
- Randomization and Diversity: Improving ensemble models by introducing randomness or diversity.

Advanced Supervised Learning:

- Multi-label Classification: Handling tasks where a single sample can belong to multiple categories.
- Class Imbalance Issues: Addressing situations where the number of samples in certain classes is significantly lower than others, for example, using oversampling or undersampling methods.

Model Optimization and Acceleration:

- Randomized Search: A method for hyperparameter optimization using random search.

- Model Caching and Distributed Computing: Distributing model training and prediction across multiple machines to expedite processing.

Automated Machine Learning (AutoML):

- Utilizing tools such as AutoML libraries (e.g., TPOT, H2O.ai, Auto-sklearn) to automate the machine learning workflow, including feature selection, hyperparameter tuning, and model selection.

Underlying Implementation in scikit-learn:

- Fundamental data structures in scikit-learn, such as Estimator, Transformer, and Predictor.
- The utilization of NumPy and SciPy in scikit-learn for mathematical computations and optimizations.

Hyperparameter Optimization Algorithms:

- Grid Search: Trying out all possible combinations of hyperparameters.
- Bayesian Optimization: Using Bayesian methods to select the best hyperparameters.

Advanced Unsupervised Learning:

- Gaussian Mixture Models: Used for density estimation and clustering.
- Topic Models: Employed for topic modeling in text data.

Feature Selection Methods:

- Recursive Feature Elimination (RFE): Gradually removing features that contribute less to the model's performance.
- Variance Threshold: Eliminating features with variance below a certain threshold, particularly useful for binary features in classification tasks.