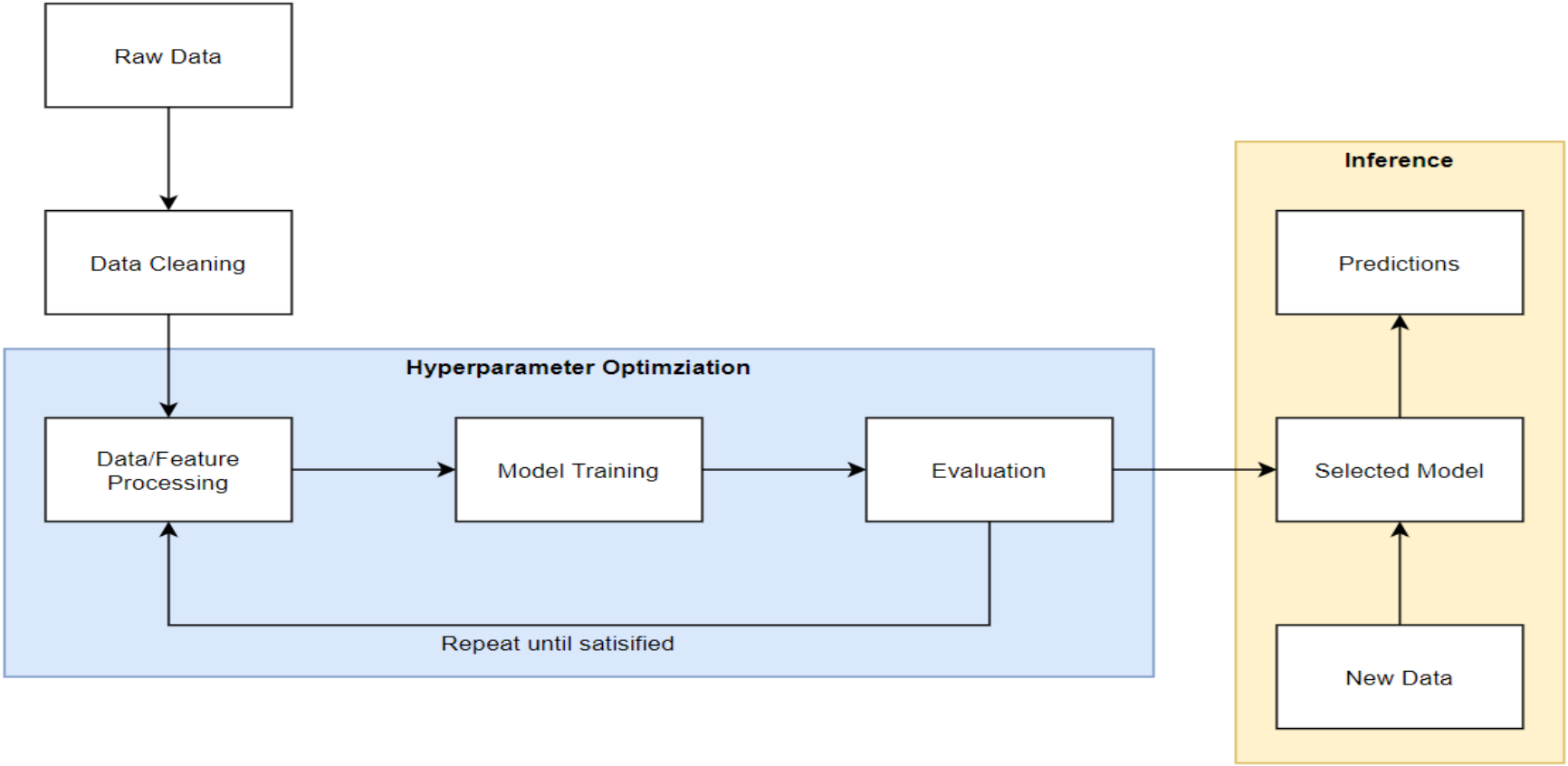
Auto Sklearn

Prerequisite

1. Python 3
2. Machine Learning Algorithms
3. Sklearn, Pandas library

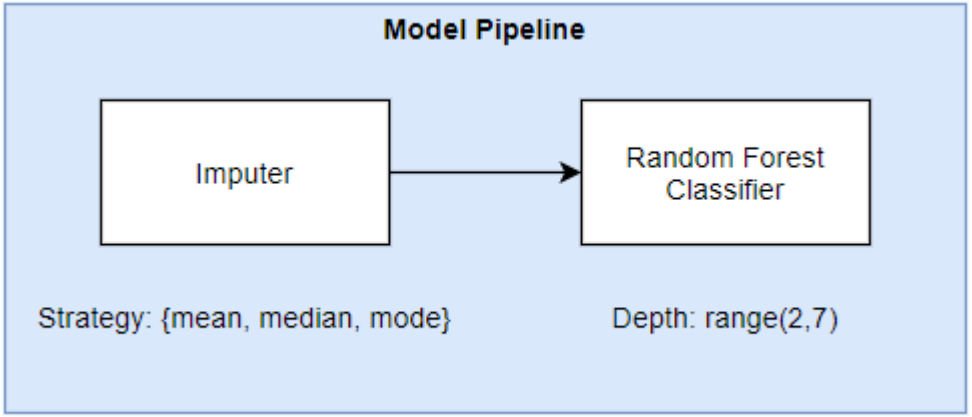
What is Auto-Sklearn?

A typical Machine Learning Workflow



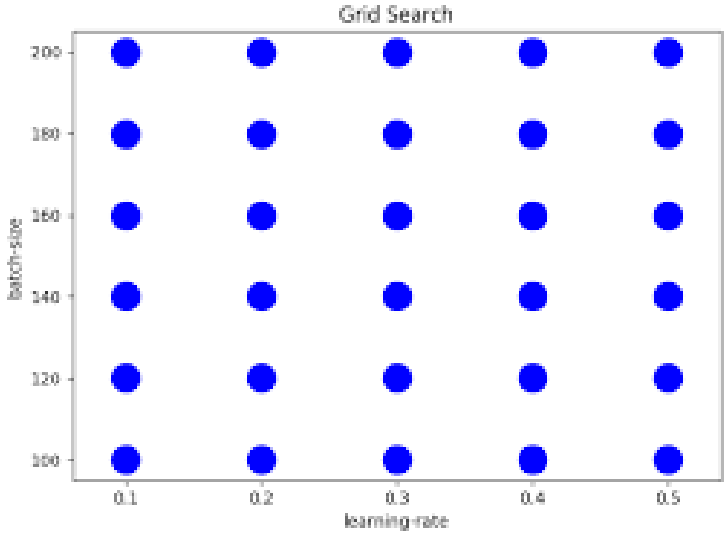
Hyperparameter Optimization

- The objective of hyperparameter optimization is to find the optimal model pipeline components and their associated hyperparameters.



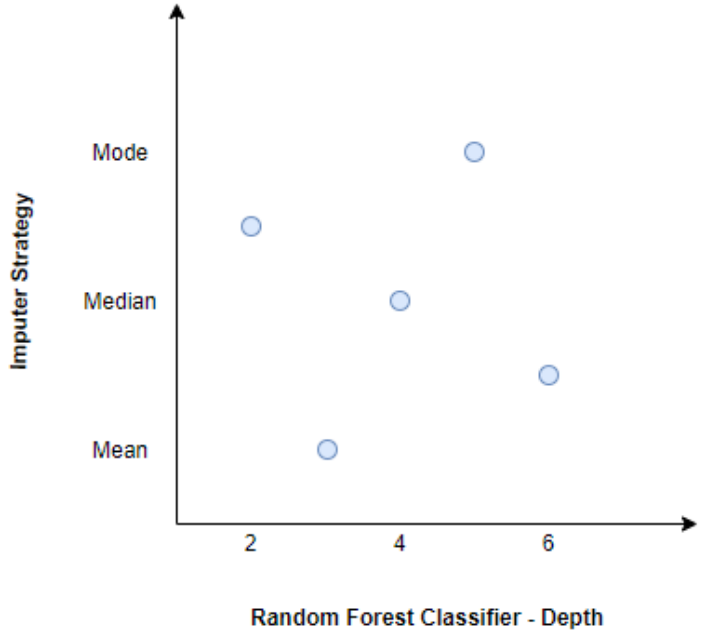
Grid Search

- For each hyperparameter, we make a list of possible values and try all possible combinations of values.



Random Search

- In random search, we define the range and choices for each hyperparameter and the sets of hyperparameters are randomly chosen within these boundaries.

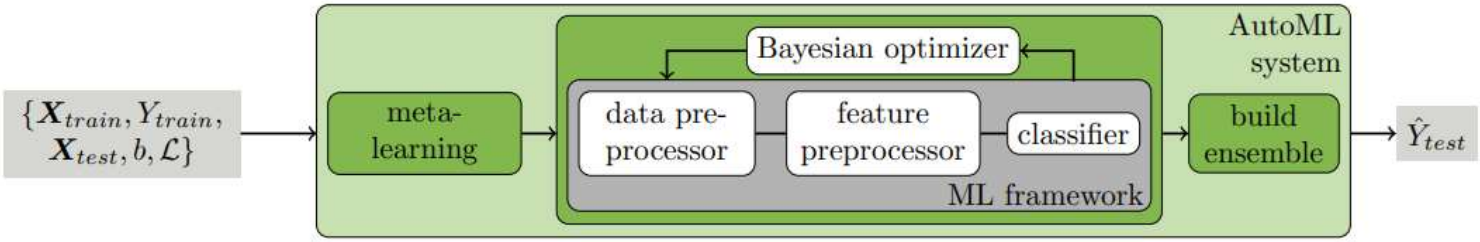


Bayesian Optimization

- Bayesian optimization stores prior searched hyperparameters and results of a predefined objective function (e.g. binary cross entropy loss) and use it to create a surrogate model.
- The purpose of a surrogate model is to quickly estimate the performance of the actual model given a particular set of candidate hyperparameter.
- This allows us to decide if we should use the set of candidate hyperparameter to train the actual model with.
- As the number of trials increases, the surrogate model, updated with additional trial results, improves and starts to recommend better candidate hyperparameters.
- Bayesian optimization suffers from cold start problem as it requires trial data to build the surrogate model before it is able to recommend good candidate hyperparameter for the next trial.
- To overcome the cold start problem, Auto-Sklearn, an open source AutoML library, incorporates warm start, through a process called meta-learning, into Bayesian optimization to get instantiation of hyperparameters that are better than random.

Auto-Sklearn

- Automated Machine Learning (AutoML) is the process of automating tasks in the machine learning pipeline such as data preprocessing, feature preprocessing, hyperparameter optimization, model selection and evaluation.



- Auto-Sklearn uses Bayesian optimization with warm start (meta-learning) to find the optimal model pipeline and build an ensemble from the individual model pipelines at the end.

Meta Learning

- The purpose of meta-learning is to find good instantiation of hyperparameters for Bayesian optimization so that it performs better than random at the start.
- The intuition behind meta learning is simple: datasets with similar meta features performs similarly on the same set of hyperparameter.
- Meta features as defined by Auto-Sklearn authors are *"characteristics of the dataset that can be computed efficiently and that help to determine which algorithm to use on a new dataset"*.

THANK YOU