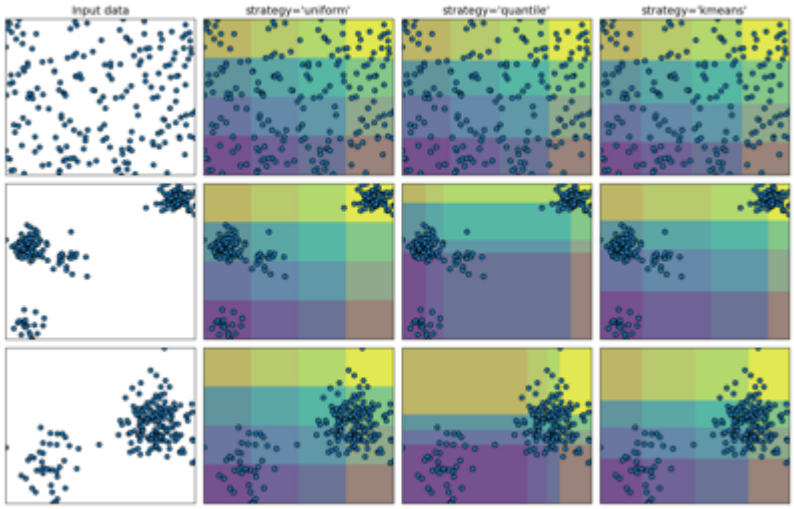
Scikit-Learn

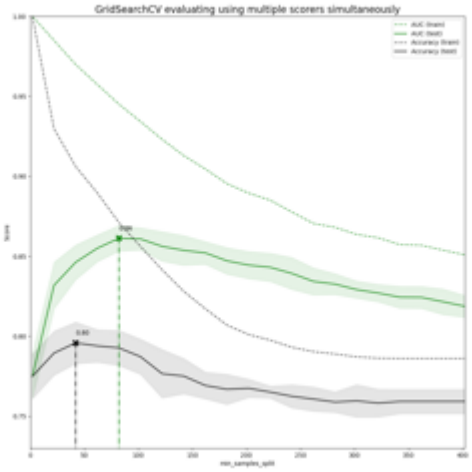
`scikit-learn` (often referred to as `sklearn`) is a powerful Python library widely used in machine learning, data analysis, and statistical modeling. It provides simple and efficient tools for data preprocessing, model selection, and algorithmic implementation, making it a go-to library for developing and testing machine learning models.

Key Features and Modules of `scikit-learn`

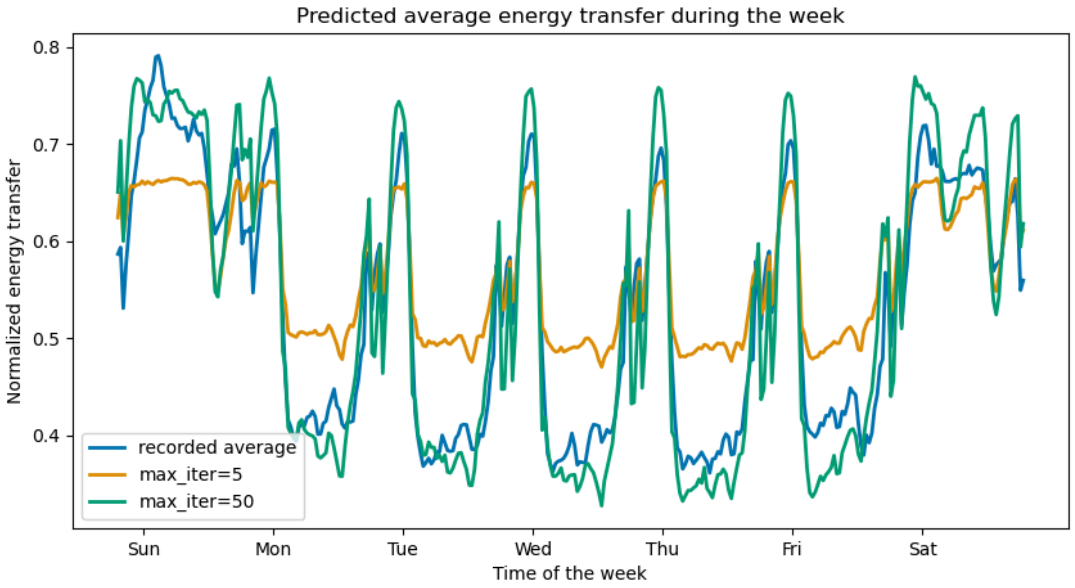
1. **Data Preprocessing**:



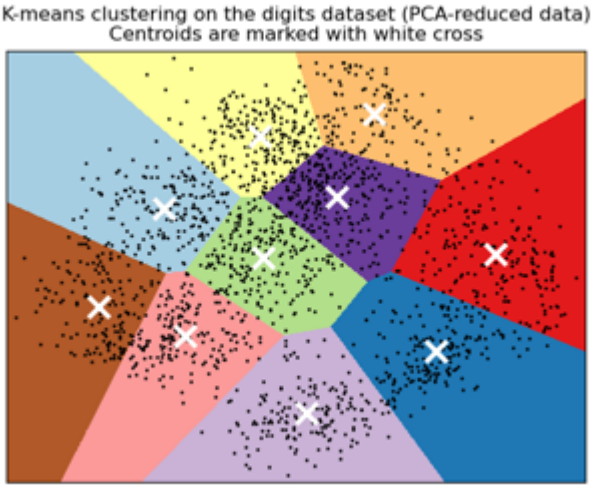
2. **Model Selection**:



3. **Regression Models**:



4. **Classification and Cluster Models**:



5. **Evaluation Metrics**:

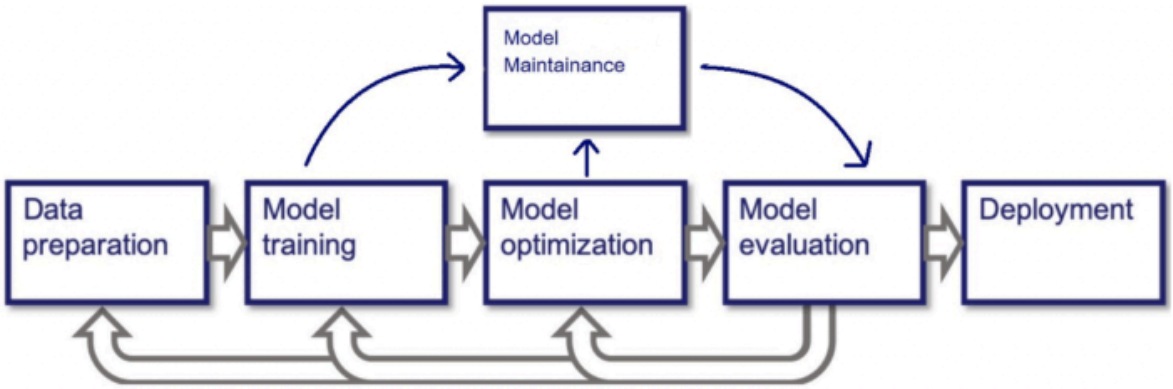
$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

6. **Pipeline Creation**:




```
from sklearn.model_selection import train_test_split
#Scale data for training and testing
from sklearn.linear_model import LinearRegression,LogisiticRegression
#Model we will use for training dataset
from sklearn.preprocessing import StandardScaler
#Used to scale data needed for testing and training
from sklearn.preprocessing import LabelEncoder
#This module is used to convert categorical values into numerical columns for
preprocessing
```

```
from sklearn.metrics import mean_squared_error, r2_score,accuracy_score
#Evaluate model working and accuracy check
```

- `train_test_split` ->Split arrays or matrices into random train and test subsets.
- `LinearRegression` ->Linear regression is a statistical method that is used to predict a continuous dependent variable(target variable) based on one or more independent variables(predictor variables). This technique assumes a linear relationship between the dependent and independent variables, which implies that the dependent variable changes proportionally with changes in the independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value.
- `LogisiticRegression` ->Logistic regression is a supervised machine learning algorithm used for classification tasks where the goal is to predict the probability that an instance belongs to a given class or not. Logistic regression is a statistical algorithm which analyze the relationship between two data factors.

Categorical Variable

Categorical data is the statistical data type "Statistical data type") consisting of categorical variables or of data that has been converted into that form. In other words, categorical variable is a variable that can take on one of a limited, and usually fixed, number of possible values, assigning each individual or other unit of observation to a particular group or on the basis of some.

Correlation Matrix

A correlation matrix is a table that shows the correlation between variables in a machine learning model, and is a useful tool for summarizing data, identifying patterns, and making decisions
A correlation matrix is typically square, with the same variables shown in the rows and columns. The diagonal of the matrix is always a set of ones because the correlation between a variable and itself is always 1.

- `StandardScaler` ->Standardizes features by removing the mean and scaling to unit variance. The standard score of a sample `x` is calculated as: **z = (x - u) / s**
where `u` is the mean of the training samples or zero if `with_mean=False`, and `s` is the standard deviation of the training samples or one if `with_std=False`. Centering and scaling happen independently on each feature by computing the relevant statistics on the samples in the training set.
- `LabelEncoder` ->technique that is used to convert categorical columns into numerical ones so that they can be fitted by machine learning models which only take numerical data. It is an important pre-processing step in a machine-learning project. Label encoding converts the

categorical data into numerical ones, but it assigns a unique number(starting from 0) to each class of data. This may lead to the generation of priority issues during model training of data sets. A label with a high value may be considered to have high priority than a label having a lower value.

- `mean_squared_error` ->metric used to measure the average squared difference between the predicted values and the actual values in the dataset. A lower MSE indicates that the model's predictions are closer to the actual values signifying better accuracy. Conversely, a higher MSE suggests that the model's predictions deviate further from true values indicating the poorer performance.
- `r2_score` ->statistical measure that represents the goodness of fit of a regression model. The value of R-square lies between 0 to 1. Where we get R-square equals 1 when the model perfectly fits the data and there is no difference between the predicted value and actual value. However, we get R-square equals 0 when the model does not predict any variability in the model and it does not learn any relationship between the dependent and independent variables.
- `accuracy_score` ->In multilabel classification, this function computes subset accuracy: the set of labels predicted for a sample must *exactly* match the corresponding set of labels in y_true.

We head to coding implementation now Coding