**Task Title:** Energy Consumption Prediction using Machine Learning

# Deliverables:

* A Python file with your complete implemented pipeline.
* A report summarizing your findings and discussing the performance of your model. This report should include the same sections as the [detailed requirements section](#bookmark=id.m78mnav35bzu) with the mention of the techniques you applied in each -code snippets are allowed.
  + The report **must contain** your learning curve(involves plotting two key metrics (such as error or accuracy) against the number of training examples or the training iteration. This visualization helps to understand how the model is learning and whether it is overfitting or underfitting.)

# Assignment Deadline:

Friday 15/03/2024 at 11:59 PM

Deliver your work in the “Task Submission “on moodle.

Note that you should submit your results in the Kaggle competition in order to see your scores . You don’t have the true labels in the test data so when you test your model use Kaggle. The highest three scores will get bonus marks.

Kaggle link: <https://www.kaggle.com/t/f7abfc54099846da8aac14d1919140c3>

Each team will name their group on Kaggle by the name given in the teams form.

Video on how to submit your predictions on Kaggle: <https://www.youtube.com/watch?v=3oRFyuj4udI>

Note that : The metric used to evaluate your predictions is R2 Score so all your prediction scores will be from 0 to 1.

# Description:

In this assignment, you will be building a machine-learning model to predict the energy consumption of a steel factory based on various features of the properties.

Your task is to build a regression model that can accurately predict the energy consumption of a steel factory given its features. You will start by preprocessing the data by handling missing values, scaling the features if necessary, and splitting it into training , validation and test sets. Use k-fold technique.

Next, you will implement and train a regression algorithm such as Linear Regression, Polynomial Regression, Ridge Regression or Lasso Regression. You will evaluate the performance of your model using metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared score on the test set.

Finally, you will select the best-performing model and fine-tune its hyperparameters using techniques such as Grid Search and Random Search**(self-study).**You will then evaluate the final model on the test set and report the final performance metrics.

# 

# Dataset: (you will find it on Kaggle competition)

## Attribute Information:

Data Variables Type Measurement

* **Industry Energy Consumption Continuous kWh**
* Lagging Current reactive power Continuous kVarh
* Leading Current reactive power Continuous kVarh
* tCO2(CO2) Continuous ppm
* Lagging Current power factor Continuous %
* Leading Current Power factor Continuous %
* Number of Seconds from midnight Continuous S
* Week status Categorical (Weekend (0) or a Weekday(1))
* Day of week Categorical Sunday, Monday â€¦. Saturday
* Load Type Categorical Light Load, Medium Load, Maximum Load

# Detailed Requirements:

Follow this section thoroughly to satisfy all the needed requirements of this assignment.

## 1 Data Preparation

### 1.1 Missing Data

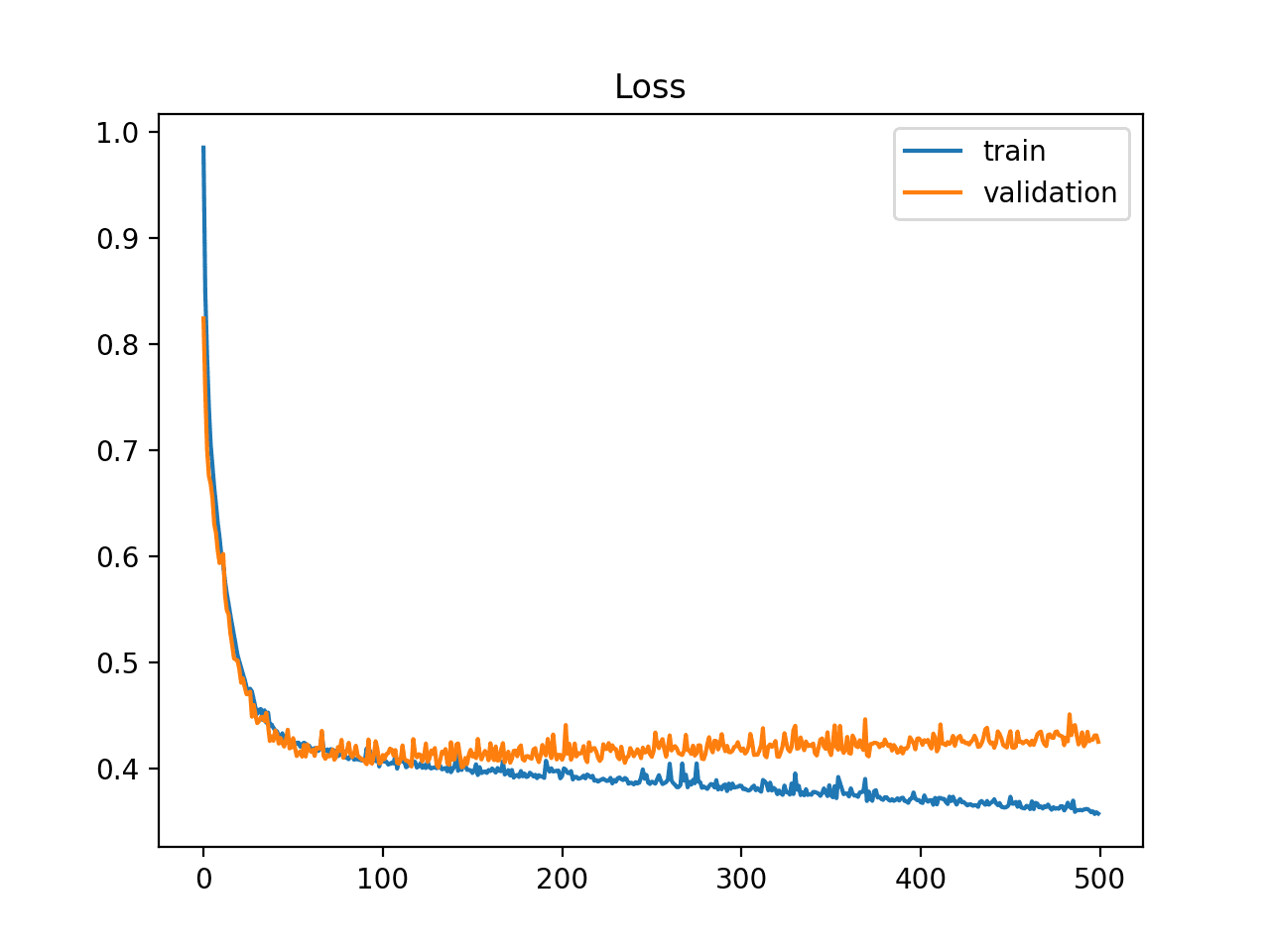
* Report your analysis of the missing data % for each attribute.
* Apply a data imputation technique suitable to each attribute with missing values.

### 1.2 Feature Engineering

* Handling Datetime column.
* Encoding the categorical columns.

## 2 Machine Learning Model

### 2.1 Plotting the Learning Curve



In this requirement, you are tasked to plot the Loss function (MSE) of both your training subset and validation subset over the epochs.

You can either implement a vectorized linear regression model and track your loss function per epoch, or you can utilize the Scikit-learn modules (<https://scikit-learn.org/stable/modules/learning_curve.html>).

## 2.2 Hyperparameter Tuning

Tune at least one of the model’s hyperparameters and report your findings.