Machine Learning Cycle

Frame the problem:

The first thing we have to do is to understand the problem and its objective, to be able to know what the benefit from this model you are will build, how we will frame the problem which algorithm we will use, which performance measure to use.

The current solution:

Next, we have to know what the expected result is from building the model. This will help us to know how to solve the problem and the expected performance.

Type of training:

Next, we have to find out which type of training we are going to use. For example, in the house pricing, we will use supervise learning as we have the label in the data. Then, we want to know that is it classification or regression problem. After that, we want to figure out whether are we going to use batch learning or online learning.

Selecting the performance measure:

In the next step, we want to select the performance measure. For example, for supervise and regression problem, the most common measure is RMSE, which will give us the error the model made while predicting, and this gives high weight to the large errors. But, if there are a lot of outliers in the data, using MAE will be preferred.

Starting The Code

Getting/Reading the data:

The first thing we will do is read the data that we are going to work with.

Looking on the data:

After reading the data, we want to see what we have on it and it contains, and we do that by seeing the head of the data or the first five rows.

Info of the data:

Next, we want to see the info of our data, which will give us the total number of rows, types and num of null data.

Summary of data:

We can see a summary of the data by using describe() fun, and this will show the numeric columns. This will show (mean, count, min, max, std, 25%m 50%, 75%).

Hist plot:

We can use hist plot to see the summary as well visually.

Creating a test and train set:

We will create a test set to put it aside for testing, and we do that by taking 20% of our whole data.

Exploring the data:

In this step, we are going to explore our data and know it more by visualizing the exploring it. We can use scatter plot to visualize geographical data.

Looking for correlations in the data:

We can find the correlation between the columns, and we can know that by ranges from -1 to 1. If it is close to 1, it has strong positive correlation, but it is close to -1, it has strong negative correlation, and if it is close to 0, it does not have a correlation.

Preparing the data for Machine learning algorithms:

Cleaning the data:

We want to clean our data from missing values as the ML algorithm does not work with it. We can deal with missing data in three ways, deleting the missing values in the columns, deleting the whole columns that contains missing values, or finding the mean or median to it or fill it with zero.

Handling Text and Categorical data:

We have to convert any text or categorical data to numeric as the model does not work with them. We can use OrdinalEncoder or OneHotEncoder to convert them.

Feature scaling and transformation:

We want to scale our data to be in the same scale, which this will affect the model later if we do not do so. We can do that by using MinMaxScaler or StandardScaler. We should only fit the training data not the testing data and transform both.

Transformation Pipelines:

Creating a pipeline will help us to do many steps at once.

Training a model:

In this step, the data is ready to be trained on the training data, so we are going to build a model and train it on our training data only.

Prediction:

Now, after training the model, we want to make an example to predict it by the model.

Evaluation:

Now, we want to see if the prediction is correct or not, so we use evaluation to figure it out. In the housing example, we can use RMSE to evaluate or model and see the result of our training.

Cross-validation:

We can use cross-validation to evaluate our model.

Fine-Tune:

We can use grid search or randomized search to fine tune the model. This will help to find the best hyperparameters that give good results.

Building other models:

If we do not get a great result from one model, we can try to use another model and go through the same steps.

Finding the best model:

In this step, we are going to compare all the models we have created and find the best one to use in the final model.

Using the final model:

After compromising, we are going to build our best model to use in prediction.

Evaluation on the testing set:

Now, we can use our test set to evaluate our model. So, we are going to use the final model to predict the testing set.

Saving our model:

Lastly, we have to save our model to use it later or use it in deployment stage. If we did not use pipelines, we have to save our scaler as well.