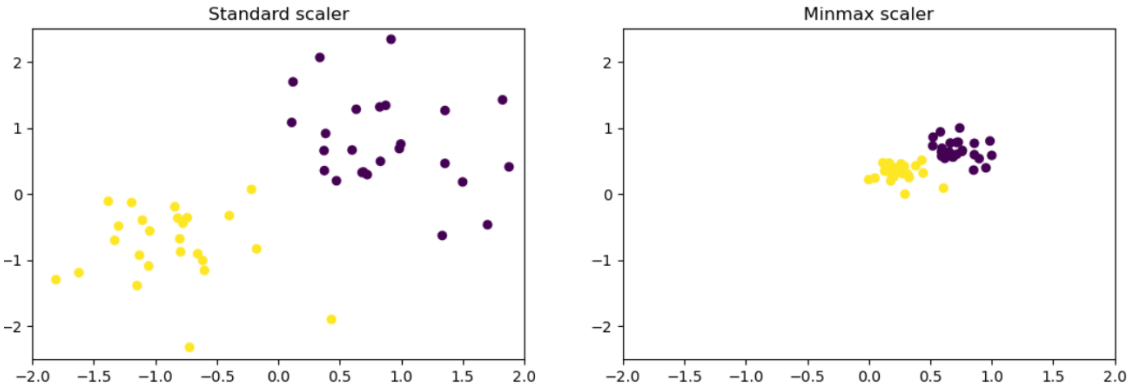
Week 4 Kahoot

Q1



Which of the following is NOT a reason to do data scaling?

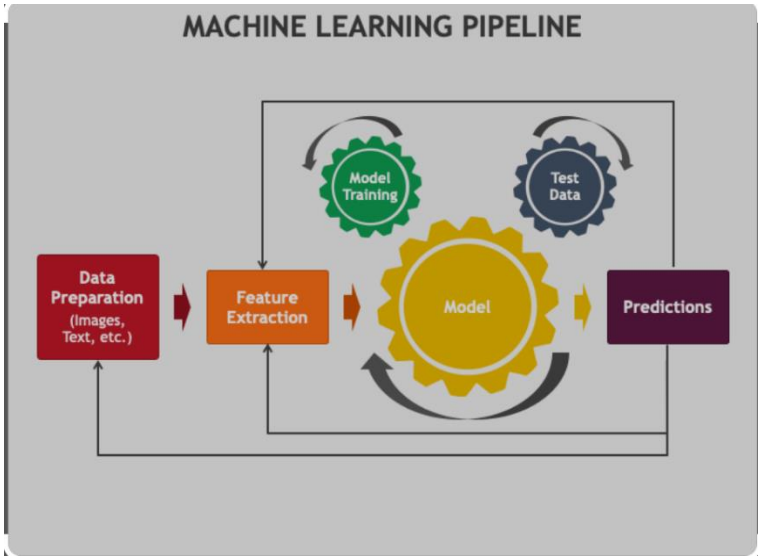
- to makes numerical computations more stable
- **to improve correlation with the target**
- to help gradient descent converge more quickly
- to be able to compare coefficients of linear model in terms of importance

Q2

What does scaler.fit_transform(X) exactly do?

- Scales X based on previously learnt parameters
- **Learns the parameters for scaling of X and scales it**
- Learns the parameters from training data, transforms test data
- Performs cross validation to find the best scaling parameters and scales X

Q3



Generally, we should first do missing value imputation and then data scaling.

- **True**
- False

Q4

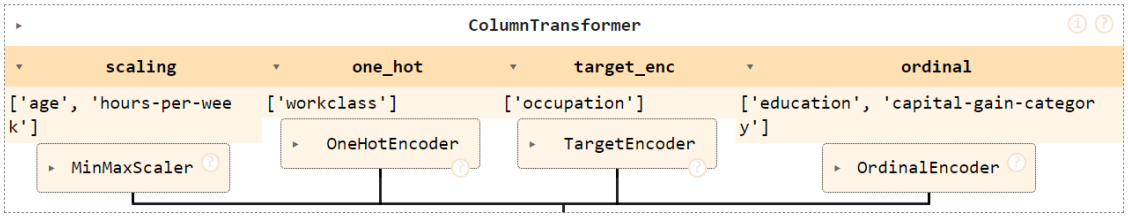
Scikit-learn
https://scikit-learn.org › stable › modules › generated › s...

sklearn.preprocessing.OneHotEncoder

When should we use one hot encoding?

- When we have a variable with few categories that can be ranked
- When we have a variable with many categories that can be ranked
- **When we have a variable with few categories that cannot be ranked**
- When we have a variable with many categories that cannot be ranked

Q5



What happens by default to the features that we did not explicitly name in the sklearn Column Transformer

- nothing, they remain unchanged
- **they will be dropped**
- an error will be raised when the transformer is fitted
- they will be transformed with the default transformation

Q6

Why do we do hyperparameter tuning?

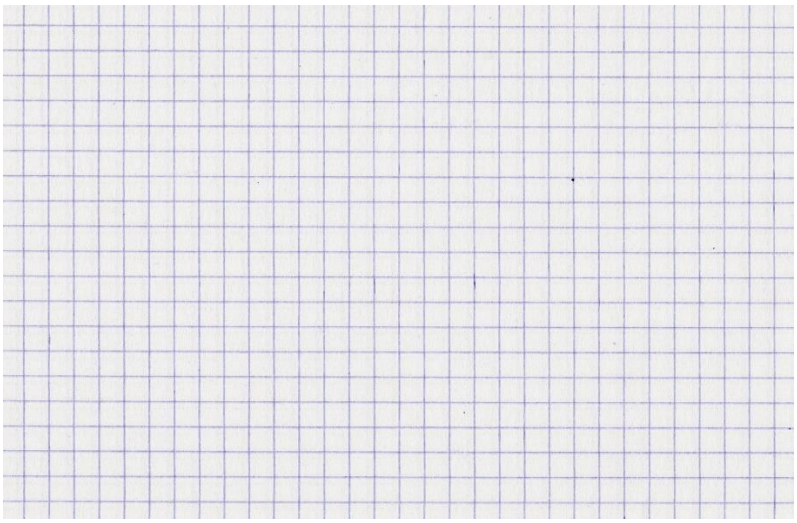
- To minimize the training time
- To improve the data quality
- **To find the parameters that minimize the generalization error**
- To make sure we use the most complex model

Q7

Why do we use Cross Validation in the context of hyperparameter tuning?

- **To more reliably estimate model performance for hyperparameter values**
- To increase the speed of the tuning process
- To eliminate the need for a test set
- To reduce the computational cost by simplifying the model

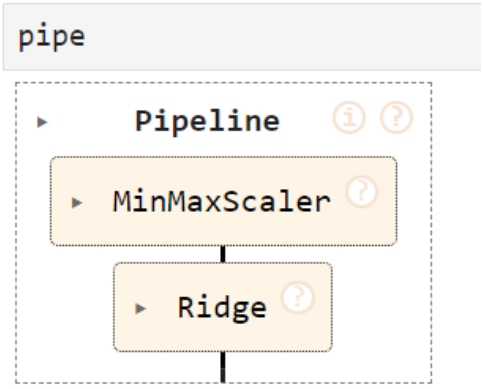
Q8



Using grid search cross validation ensures that we will always find optimal values of the hyperparameters.

- True
- **False**

Q9



Which of the following is NOT an advantage of integrating grid search with sklearn Pipeline?

- We avoid data leakage
- We simplify the code, bundling a sequence of processing and modeling steps
- **We reduce the need for data cleaning**
- We can optimize the parameters of all transformers and the final estimator

Q10

Regarding homework 1

- what homework?
- there is still one week left, I will do it later
- I got started, but it's too hard
- **I got started and I am pretty ok**