Project Week 5

Picking mushrooms

Building a full Machine Learning Classification Project

The goal of this challenge is to do a complete machine learning project. We are using a public mushroom dataset. We predict if the mushrooms are edible or poisonous from their color and shape.

TASK 1 - Data Gathering

- Download the following Kaggle dataset https://www.kaggle.com/uciml/mushroom-classification
- Import it into a pandas dataframe
- Check for missing values

TASK 2 - Basic Data Exploration & Train-Validation-Split

- Explore the data and try to understand the features
- Look for the most important features
- Encode the categorical data columns with an encoder of your choice
- Split the dataset into a training and a validation data set

TASK 3 - Basic Model Training and Tuning

- Train a classification model of your choice
- Pick a performance metric and print the results for the training data set as well as the validation data set
- Repeat this process for at least two other algorithms of your choice
- Take a couple of minutes to try different hyperparameters for each of the algorithms
- Add a 5-fold Cross-Validation to each of the algorithms

TASK 4 - Column Transformer & Feature Engineering

- Add a columntransformer to your model structure
- Now you can directly see the different effects of feature engineering. Try to change the encoding and explore the effects. Try skipping columns that you think to be not that important and check if you are right
- Don’t forget to document what you tried and what you learned from these tries

TASK 5 - Present your findings

- On Friday, you’ll present your problem as well as your solution
- Tell us how you build the models and why you chose certain algorithms, feature engineering steps and how you tuned and tried to optimize it
- It should contain at least 5 visualizations for the data exploration part and 2 on how the different algorithms perform on the problem
- Tell us how your training and validation scores compared to the test scores on Kaggle
- You can use Slides and/or your jupyter notebook

BONUS

- Test one of the wrapper methods of feature selection that we used
- Perform a grid search to really find the optimal set of hyperparameters for each of your algorithms