6️⃣ Random forests

Learning steps

- ☑️ ~~yt video~~
- ☑️ ~~notebook/own implementation~~
- ☑️ ~~book chapter~~

Resources

- website 💻
  - lesson 6
- notebooks 🗒️
  - How random forests really work
  - Road to the top, part 1
- book 📘
  - chapter 9
  - solutions to exercises

Quick notes

YouTube video (→ link)

- titanic wreck
  - good model using binary splits
  - we re-created the OneR model

    → what if we did a TwoR?
    - remove 'sex' and do a next split
      - men: age; women: pclass
    - create a decision tree
      - 💡 use DecisionTreeClassifier from sklearn to do it for us

        - use Gini index
    - random forest
        - idea: make different trees using different records (creating independant from one another) ⇒ averaging the errors will lead to an average error of 0
        - out of the bag error (OOB): to test a particular model, use unused records as validation set
        - ❗ to have an idea of the quality of a prediction, take a look at the variance between the different models
    - gradient boosting
        - use model to predict the residuals from the previous one, then finally sum everything
            - can overfit, unlike random forest
- fastkaggle (see notebook)
    - allows to download the data for a competition regardless of the platform used
    - fastcore.parallel
- how to win kaggle competitions fast
    - iterate!
    - use fast models to understand the data more
    - even submitting to kaggle has to be easy
    - AutoML
        - test for multiple hyperparameter values
    - data augmentation
    - tta (test time augmentation)