**OPGAVER TIL MACHINE LEARNING PÅ PBA**

**Opgave 1**

Installer Anaconda, (Simon viser, hvordan man bruger jupyter notebooks)

* <https://docs.anaconda.com/anaconda/install/windows/> eller <https://docs.anaconda.com/anaconda/install/mac-os/>
* Vælg Python 3.7

(Simon viser, hvordan man alternativt kan bruge kaggle og colab)

**Opgave 2**

(Workshop, hvor Simon gør det først, og du gør det bagefter)

Brug datasættet titanic.csv, som ligger på itslearning (det er cleanet)

Lav en EDA

* Numerical
  + Head
  + Describe
  + Info
  + keys
* Visual
  + Countplot (på survived)
  + Histogram (på age)
  + Correlation (på alle parametre)
* Hvilke andre slags visualiseringer kan man bruge og til hvilke features?

Hvilken type problem er det (classification/regression, binary/multiclass)?

Brug følgende algoritmer – (med fit, predict og train\_test\_split)

* Knn
* Decision trees
* Overvej hvilke andre supervised algoritmer man kunne have valgt

Mål performance (score(X\_test,y\_test)) på begge algoritmer

Lav 5 fold cross validation på titanic-datasættet, og sammenlign resultatet med det du fik tidligere.

Forklar begreberne overfitting/underfitting

Forklar begrebet k-folds validation

**Opgave 3 (På egen hånd – i jupyter notebooks)**

1. Tag kolonnerne ”area”, ”rooms” og ”new price” fra property-datasættet (ligger på itslearning), og ignorer alle andre kolonner
2. Sørg for at kolonnerne kun indeholder tal
3. Opdel datasættet med train\_test\_split, hvor y er ”new price”
4. Træn en Linear Regression algortime på træningsdelen
5. Og predict på testdelen
6. Udregn RMSE for din prediction
7. Brug get\_dummies på type-kolonnen, så tekstværdierne bliver til talværdier.
8. Gentræn din Linear Regression, og tjek om scoren forbedrede sig.

Forklar begrebet regularization

* Hvilke slags regularizations findes der?

**Opgave 4**

1. Lav lasso regression på bilbasen-datasættet (ligger på itslearning) for at finde den parameter der er vigtigst for prisen.
2. Opsæt en KNN-regressor med en k på 3, på bilbasen-sættet, og lav en confusion matrix

Forklar begreberne Precision og recall

**Opgave 5**

Forklar hvad en ROC-kurve viser

Forklar begrebet hyperparameter

Forklar hvad GridSearchCV gør

Forklar begrebet Normalize

**Opgave 6**

Løs Projektet: “Predicting Credit Card Approvals”, som findes på datacamp under projects