

# Introduction to Machine Learning in R (NPFL054)

## Easy HW – Clustering and Linear Regression

Contact: Barbora Hladká (hladka@ufal.mff.cuni.cz)

---

### Data

- Titanic data set - <https://ufal.mff.cuni.cz/~hladka/2021/docs/train.csv>
- Movie data set - <https://ufal.mff.cuni.cz/~hladka/2016/docs/mov.development.csv>

### Questions

1. Load the Titanic data set and run the  $K$ -Means algorithm to cluster the passengers into 2 clusters ( $K = 2$ ). Do not use the attribute **Survived**. How much do these two clusters and the **Survived** and **NotSurvived** subgroups overlap? Experiment with different subsets of the given features. You can add some new feature(s).
2. Load the Titanic data set and build a dendrogram using a hierarchical clustering algorithm. Do not use the attribute **Pclass**. Cut the dendrogram into 3 clusters. How much do these three clusters and the **Pclass = 1**, **Pclass = 2**, **Pclass = 3** subgroups overlap? Experiment with different subsets of the given features. You can add some new feature(s). Create nice visualizations.
3. Load the Movie data set and fit a linear regression model with **rating** as a target attribute. If you cannot make computing due to lack of computational power, experiment with a subset of the Movie data set. Experiment with different subsets of the given features. Evaluate your models using Adjusted  $R^2$ .

### Presentation

- Create a 5 min presentation.
- Present your answers. If you want to highlight something in your R code, please do it.
- Explain your answers clearly so that your audience understands your method well.