

Titanic dataset

Zadání úlohy

- <https://www.kaggle.com/c/titanic>
- cíl: vytvořit model predikující, kteří pasažéři přežili ztroskotání Titanicu
 - ačkoliv svoji roli hrála i náhoda, zjevně některé skupiny pasažérů měly větší šanci na přežití
 - máme k dispozici jméno, věk, pohlaví, socio-ekonomickou třídu apod.

Úkoly

- vyzkoušejte přidat další atributy
 - sečíst příbuzné na palubě
 - s počtem příbuzných pracovat i ve formě kategorií
 - věk převést na kategoriální proměnnou ve formě skupin (intervalů)
 - zkuste pracovat s proměnnou Cabin (např. z písmena udělat kategorii)
 - zkuste pracovat se jménem a vytáhnout z něj titul (např. “Countess”) nebo počet slov ve jméně (delších než např. 3 znaky)
- vyzkoušejte další klasifikátory (logistickou regresi, SVM, XGBoost) a porovnejte výsledky s ostatními klasifikátory
 - XGBoostClassifier nutno nainstalovat skrz pip `install xgboost`
- vyzkoušejte grid search a/nebo randomized search pro nalezení nejlepších hyperparametrů