The “Overall Survival Status” variable is the target variable of the model

Project steps

* Start with a challenge
* Explore the data
* Prepare the data
* Create the ML model
* Evaluate the model
* Deploy, run and monitor

Start with a challenge

* Predict the overall survival status in patients with Cholangiocarcinoma

Data source

* Data was obtained from The Cancer Genome atlas (TCGA)

<https://www.cbioportal.org/study/clinicalData?id=ihch_msk_2021>

Explore the data

* Using the variables that we are using for the to predict the model, describe the dataset using each of those variables
* For the continuous variables do a statistical analysis ( Mean, Max, Min, St. dev)
* For the Categorical variables do a count for each categorical variable

Prepare the data

* Split the data into a training and test model

Create the ML model

* Random Forest
* Logistic Regression
* Compared results with RAPIDMINER
* The “Overall Survival Status” variable is the target variable of the model