Data Modification

Fabian Blasch

02/14/2022

Intro

In this file the PDFs are converted into a Dataset that can be used to train common machine learning models such as but not limited to, Multinomial Elastic-Net Regularized Generalized Linear Models, Random Forests and Boosted Regression Trees.

The Data that will be summarized in the dataset is the actual label of the SRRI, the Bitmap entries with their respective position on the page as well as the corresponding color. Further, a variable will be included that signifies whether the SRRI is depicted vertically or horizontally.

Illustration using a single PDF