Chair of Computational Logistics

RWTH AACHEN UNIVERSITY

Machine Learning and Data Analytics **Sommer semester**

Dr. Rossana Cavagnini mlda@cl.rwth-aachen.de

Project proposals

For the following projects, a task to be performed is suggested. However, some datasets are suitable both for regression and classification tasks. The students may freely decide to work on one of the two tasks based on their interest.

Project 1 (Can you predict how many bikes will be rented in a bikesharing system on a day?):

This dataset contains 731 observations spanning two years for the Capital Bikeshare system, Washington D.C., USA. The label (output variable) represents the number of daily rental events. The features correspond to seasonal and weather information. Can you predict the daily rental number of bikes based on the features? Click here for the dataset (To work on this project, use only the "day" dataset and consider only the "training.csv" file as if it was the only file available)

Project 2 (Will a person accept the coupon recommended to them in different driving scenarios?):

This dataset contains 12684 observations. The label (output variable) is binary and represents whether a person accepts or not the coupon. Can you predict whether the person will accept the coupon or not? Click here for the dataset (This paper provides detailed information about the dataset)

Project 3 (Can you predict the concrete compressive strength?):

This dataset contains 1030 observations. The label (output variable) is continuous and represents the concrete compressive strength. Can you predict the concrete compressive strength? Click here for the dataset

Project 4 (Can you predict whether a data scientist will change job?):

This dataset contains 19158 observations. The label (output variable) is binary and represents whether the data scientist will change job or not. Can you predict whether the data scientist will change job? Click here for the dataset (use only the dataset "aug_train" as if it was the only file available)

Project 5 (Can you predict product backorder?):

Material backorder is a common problem in a supply chain system, impacting an inventory system's service level and effectiveness. Identifying parts with the highest chances of shortage prior to their occurrence can present a high opportunity to improve an overall company's performance.
In this project, you will train classifiers to predict future back-ordered products and generate predictions for a test set.
Click here for the dataset (if you cannot access it, send an email to mlda@dpo.rwth-aachen.de).

Project 6 (Can you predict the absenteeism duration?):

The dataset contains 740 absenteeism events at a courier company and 20 features for each of them. The label (output variable) indicates the absenteeism time in hours. You can decide to perform either a regression or a classification task (or both) on this dataset. If you perform a regression task, then you should predict the time duration of the absenteeism events. If you perform a classification task, then you should predict in which category the time duration of the absenteeism events falls.
Click here for the dataset.

Project 7 (Can you predict employees' productivity?):

The goal of this project is to predict the productivity of employees based on a number of features contained in the dataset.
Click here for the dataset.

Project 8 (Can you predict the daily rental price?):

This dataset contains 1623 Sevillian holiday rentals extracted from Booking.com with 28 features. The label corresponds to the daily rental price.
Can you predict the daily rental price based on the features?
Click here for the dataset. Additional details about the dataset description can be found in the paper that you can find at the provided link.

Project 9 (Predicting health expenses incurred by insurances):

Predict health expenses (in dollars) from basic personal characteristics of insured people.
Click here for the dataset.

Project 10 (Can you predict the operational cost of a vertical farm?):

Given a vertical farming system with a hybrid microgrid configuration, can you predict the operational cost of this vertical farm?
The dataset consists of 200 observations. The total number of features is 343 (the base features are 14. Of these, 7 features are constant over time, so only their "average" value can be considered. The remaining 7 features have been aggregated on a monthly base, and for each of the 12 months, their average, minimum, maximum, and variance values are provided).
(Note: This dataset has been developed by the Chair of Computational Logistics. Send an email to mlda@cl.rwth-aachen.de to get this dataset. The dataset also requires a minimum preparation before being used.)