Selenge Tuvshin, Iftikhar Fakhar, Haider Tobias Abraham
Machine Learning
8th of October 2023

Exercise 0
Selection of datasets for classification

Dataset 1: Absenteeism at work

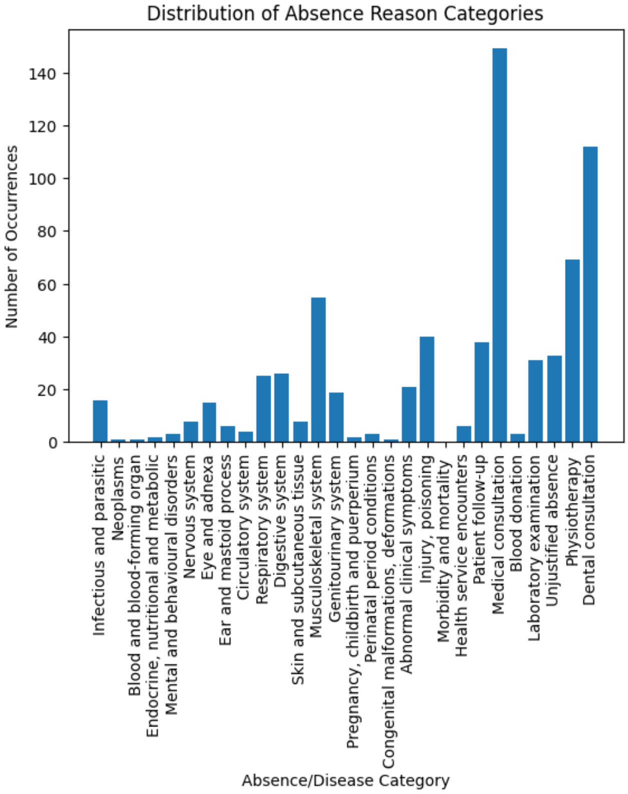
The dataset has 740 entries detailing employee absences at a Brazilian courier company. It can be used for constructing classifiers that determine the reasons for employee absences (target feature) by using information about their physical, social, and work circumstances. Each entry is associated with a unique employee identifier, so that grouping of samples by individuals is possible. The samples have 21 features that were manually inputted into a spreadsheet in the form of mostly integer numbers. There are no missing values, but some cleaning is helpful in order to map the features encoded in integers to their correct data type.

Categorical data

- Day of week
- Month
- Season
- Disciplinary failure
- Education
- Social drinker/smoker

Numerical data

- Transportation expense
- Distance from residence to work
- Service time
- Age
- Work load average/day
- Hit target
- Number of children/pets
- Weight
- Height
- Body mass index
- Absenteeism time in hours



Dataset 2: Predict students' dropout and academic success

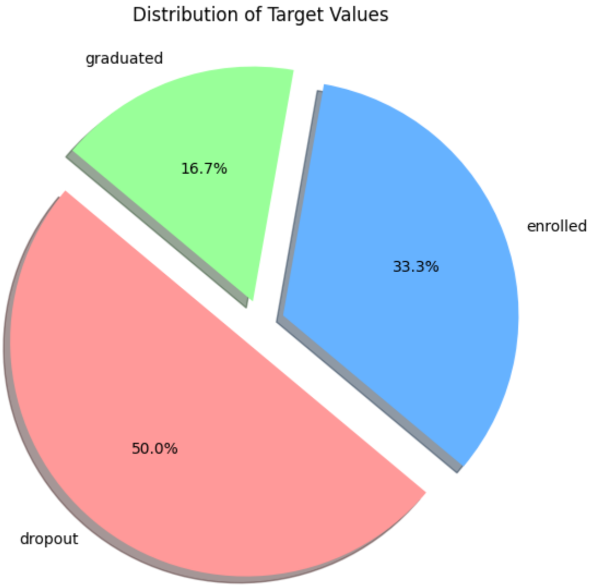
The dataset in question has been obtained from a well-established higher education institution, compiled from multiple distinct databases. It encompasses data concerning students registered in a variety of undergraduate programs, spanning diverse disciplines such as agronomy, design, education, nursing, journalism, management, social service, and several technological fields.

The primary ambition driving the assembly and analysis of this dataset is the creation of a classification model. This model aims to predict a student's probable academic trajectory, categorizing them as either a dropout, graduate, or currently enrolled. The predictors for this classification leverage insights drawn from socio-economic indicators, demographic details, and the student's prior academic journey.

This dataset is comprehensive, containing 4,424 records and 36 distinct attributes. These attributes include variables like admission grades, which vary between 0 and 200, a detailed account of the family background, and the academic scores achieved by students in both their first and second semesters. With this wealth of information, we are well-equipped to forecast the target classifications.

Categorical data

- Marital status
- Father's qualification
- Mother's qualification
- Gender
- Target



Numerical data

- Admission grade
- Curricular units 1st sem (grade)
- Curricular units 2nd sem (grade)
- Age at enrollment

Works Cited

Martiniano, Andrea, and Ricardo Ferreira. “Absenteeism at work.” *UCI Machine Learning Repository*, 4 April 2018, https://archive.ics.uci.edu/dataset/445/absenteeism+at+work. Accessed 8 October 2023.

“Predict students' dropout and academic success.” *UCI Machine Learning Repository*, https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success. Accessed 8 October 2023.