Intelligent Data Analysis

Exam: Income groups (Project 1)

This project is part of the exam *Intelligent Data Analysis.* Each project assignment is to be resovled by a single student on his/her own. The student is supposed to present the solution as part of the oral exam. The student is required to present a printed version of the Python code together with diagrams, tables, etc. that summarize the results. The specific way of how the project is presented is up to the student's choice.

Problem setting

A polling institute wants to be able to estimate an individual's income from his/her personal data (see einkommen.train). To this aim, 30.000 individuals were interviewed concerning the features summarized below. For some of the individuals, not all features are available. Crucially, the income of only 5.000 of the interviewee's is known.

Your task is to predict the income group of the remaining 25.000 interviewees and to prepare the data such that they can be used for further regression and correlation analyses.

- Age
- Employment type (Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked)
- Weighting factor to compensate for an interview-dependent selection bias
- Level of education (Bachelors, Some-college, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 1st-4th, 5th-6th, 7th-8th, 9th, 10th, 11th, 12th, Masters, Doctorate, Preschool)
- Schooling/training period
- Marital status (Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse)
- Employment area (Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces)
- Partnership (Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried)
- Ethnicity (White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black)
- Gender (Female, Male)
- Gains on financial assets
- Losses on financial assets
- Weekly working time
- Country of birth (United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong Kong, Holand-Netherlands)

- Income ($\leq 50k$, $> 50k$)

Exercise

Load the data into Python and preprocess it. Choose adequate data transformations, normalizations etc. and decide on how to deal with missing values (marked with “?”). Consider which kinds of features the preprocessed data shall contain. Once you have preprocessed the data, train a model to predict a person’s income group and apply it to the 25.000 individuals whose income group is unknown. Identify a suitable learning method and implement it in Python. Train and evaluate the model. Provide a short documentation and motivation of each of your steps.