ASSIGNMENT 2 - POLITECNICO DI TORINO

Python project – Dataset analysis and pre-processing for linear regression

Dataset description

This data approach student achievement in secondary education of two Portuguese schools for both Portuguese language and Mathematics courses (you will be given data regarding the Math course). The data attributes include student grades, demographic, social and school related features and it was collected by using school reports and questionnaires[1]. Precise descriptions of each attribute can be found in **A2_attributes_description.txt** file.

GOAL: analyse data and prepare the dataset for a REGRESSION task on the target variable G3 (i.e. final year grade for Math course) following the instructions described in **A2_Outline.docx** file.

Note: the target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades. It is more difficult to predict G3 without G2 and G1, but such prediction is much more useful. For this reason, the exercise will require to drop G1 and G2 features.

[1] Citation : P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.