Introduction

This project is mainly about education. I have divided the project to 2 parts.

In the first part of the project, I have analysed the effects of education spendings on Research & Development related factors, GDP per capitas and GINI indexes.

In the second part of the project, I have analysed the factors that affect the outcome of education.

First Part

For the first part of the project, I have used the data from 2 different sources.

I have used education spendings data from OECD and I used World Bank for other data.

Then, I merged all the data and eliminated the data with missing entries.

In the end, I have had a dataset with 26 countries with following attributes:

- country codes
- country names
- years (year is not definite, you can find the details in the blog)
- primary to non-tertiary education spendings (% of GDP)
- public tertiary education spendings (% of GDP)
- private tertiary education spendings (% of GDP)
- PISA scores for reading, math and science
- R&D expenditures (% of GDP)
- researchers in R&D (per million people)
- high technology exports (% of manufactured exports)
- resident and nonresident patent numbers (per 100 people)
- GDP per capitas (/ $1000)
- GINI indexes

I have had the following hypothesises and I have used Linear Regression to test them:

- Primary to non-tertiary education spendings explain the variation in PISA scores.
- PISA scores and education spendings explain the variation in R&D expenditures.
- PISA scores and education spendings explain the variation in researchers in R&D.
- PISA scores, education spendings and R&D data explain the variation in high technology exports.
- PISA scores, education spendings and R&D data explain the variation in patent numbers.
- PISA scores, education spendings, R&D data, high technology exports, patent numbers explain the variation in GDP per capitas.
- PISA scores and education spendings explain the variation in GINI indexes.

(You can find the details of these hypothesises and tests in Part 1 Overview document.)

Second Part

For the second part of the project I have analysed the factors that affect the performance of a children with similar education.

I have used a dataset from UCI ML datasets for this purpose.

Below, you can find the data description:

Description:

“This data approach student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires.”

Attributes:

1) school - student's school
2) sex - student's sex
3) age - student's age
4) address - student's home address type
5) famsize - family size
6) Pstatus - parent's cohabitation status
7) Medu - mother's education
8) Fedu - father's education
9) Mjob - mother's job
10) Fjob - father's job
11) reason - reason to choose this school
12) guardian - student's guardian
13) traveltime - home to school travel time
14) studytime - weekly study time
15) failures - number of past class failures
16) schoolsup - extra educational support
17) famsup - family educational support
18) paid - extra paid classes within the course subject
19) activities - extra-curricular activities
20) nursery - attended nursery school
21) higher - wants to take higher education
22) internet - Internet access at home
23) romantic - with a romantic relationship
24) famrel - quality of family relationships
25) freetime - free time after school
26) goout - going out with friends
27) Dalc - workday alcohol consumption
28) Walc - weekend alcohol consumption
29) health - current health status
30) absences - number of school absences
31) G1 - first period grade
31) G2 - second period grade
32) G3 - final grade

After getting data, I have categorized the numeric variables with equal width binning.

My goal was to predict the performance of a student (G3) given the other attributes.

I have tried to predict 3 different target with several methods. You can find them below:

1) Try to predict G3 given G1 and G2

    — Linear Regression

2) Try to predict G3 in 5 category (0-4, 5-8, 9-12, 13-16, 17-20) w/ and w/o G1, G2

    — Zero Rule
    — Decision Tree
    — Support Vector Machine
    — Logistic Regression

3) Try to predict Top 10%, Top 20%, Bottom 10%, Bottom 20% according to G3

    3.1) No weight (10% Yes, 90% No)

        — Zero Rule
        — Decision Tree
        — Logistic Regression

    3.2) Weight (50% Yes, 50% No)

        — Zero Rule
        — Decision Tree with random “No”s
        — Decision Tree with centroids of “No”s

(You can find the details of these models in Part 2 Overview document.)

Berkan Teber
19080