**Introduction**

(This part is taken from Project Overview document and contains the Part 2 section from it)

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 all 3 prediction criteria and model comparisons below:

## 1) Linear Regressions

I have applied linear regression to G1 and G2 to predict G3.

(for Mat)

| | | | |
|---|---|---|---|
| **Dep. Variable:** | G3 | **R-squared:** | 0.822 |
| **Model:** | OLS | **Adj. R-squared:** | 0.821 |
| **Method:** | Least Squares | **F-statistic:** | 906.1 |
| **Date:** | Sat, 29 Apr 2017 | **Prob (F-statistic):** | 1.01e-147 |
| **Time:** | 20:18:45 | **Log-Likelihood:** | -820.11 |
| **No. Observations:** | 395 | **AIC:** | 1646. |
| **Df Residuals:** | 392 | **BIC:** | 1658. |
| **Df Model:** | 2 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [95.0% Conf. Int.] |
|---|---|---|---|---|---|
| **Intercept** | -1.6300 | 0.335 | -4.866 | 0.000 | -2.289 -0.971 |
| **G1** | 0.1533 | 0.056 | 2.728 | 0.007 | 0.043 0.264 |
| **G2** | 0.9869 | 0.050 | 19.908 | 0.000 | 0.889 1.084 |

(for Por)

| | | | |
|---|---|---|---|
| **Dep. Variable:** | G3 | **R-squared:** | 0.848 |
| **Model:** | OLS | **Adj. R-squared:** | 0.847 |
| **Method:** | Least Squares | **F-statistic:** | 1795. |
| **Date:** | Fri, 19 May 2017 | **Prob (F-statistic):** | 9.02e-265 |
| **Time:** | 12:38:25 | **Log-Likelihood:** | -1073.7 |
| **No. Observations:** | 649 | **AIC:** | 2147. |
| **Df Residuals:** | 646 | **BIC:** | 2161. |
| **Df Model:** | 2 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [95.0% Conf. Int.] |
|---|---|---|---|---|---|
| **Intercept** | -0.1713 | 0.215 | -0.796 | 0.426 | -0.594 0.251 |
| **G1** | 0.1489 | 0.036 | 4.186 | 0.000 | 0.079 0.219 |
| **G2** | 0.8971 | 0.034 | 26.448 | 0.000 | 0.831 0.964 |

## 2) Equal Width Binning Classification

Zero Rules

for Mat: 41.3%
for Por: 41.8%

Decision Trees

w/o G1, G2:
for Mat: 41.3%
for Por: 52.2%

w/ G1, G2:
for Mat: 78.0%
for Por: 85.4%

Support Vector Machines

w/o G1, G2:
for Mat: 41.8%
for Por: 52.7%

w/ G1, G2:
for Mat: 60.3%
for Por: 77.3%

Logistic Regressions

w/o G1, G2:
for Mat: 39.4%
for Por: 55.0%

w/ G1, G2:
for Mat: 62.2%
for Por: 69.4%

Best Models

w/o G1, G2:
for Mat: 41.8% — Support Vector Machine
for Por: 55.0% — Logistic Regression

w/ G1, G2:
for Mat: 78.0% — Decision Tree
for Por: 85.4% — Decision Tree

## 3) Predict Top 10%, Top 20%, Bottom 10%, Botom 20%

### 3.1) No weights

Zero Rules

for Mat:

| | |
|---|---|
| Top 10%: | 90% |
| Top 20%: | 80% |
| Bottom 10%: | 90% |
| Bottom 20%: | 80% |

for Por:

| | |
|---|---|
| Top 10%: | 90% |
| Top 20%: | 80% |
| Bottom 10%: | 90% |
| Bottom 20%: | 80% |

Decision Trees

for Mat:

| | |
|---|---|
| Top 10%: | 90.1% |
| Top 20%: | 80.0% |
| Bottom 10%: | 93.4% |
| Bottom 20%: | 80.3% |

for Por:

| | |
|---|---|
| Top 10%: | 90.1% |
| Top 20%: | 80.1% |
| Bottom 10%: | 87.8% |
| Bottom 20%: | 82.4% |

Logistic Regressions

for Mat:

| | |
|---|---|
| Top 10%: | 88.8% |
| Top 20%: | 77.5% |
| Bottom 10%: | 91.1% |
| Bottom 20%: | 79.7% |

for Por:

| | |
|---|---|
| Top 10%: | 89.8% |
| Top 20%: | 77.0% |
| Bottom 10%: | 90.1% |
| Bottom 20%: | 81.2% |

Best Models

for Mat:

| | | | |
|---|---|---|---|
| Top 10%: | 90.1% | — | Decision Tree |
| Top 20%: | 80.0% | — | Zero Rule |
| Bottom 10%: | 93.4% | — | Decision Tree |
| Bottom 20%: | 80.3% | — | Decision Tree |

for Por:

| | | | |
|---|---|---|---|
| Top 10%: | 90.1% | — | Decision Tree |
| Top 20%: | 80.1% | — | Decision Tree |
| Bottom 10%: | 90.1% | — | Logistic Regression |
| Bottom 20%: | 82.4% | — | Decision Tree |

**3.2) Equal weights**

Zero Rules

for Mat:

| | |
|---|---|
| Top 10%: | 50% |
| Top 20%: | 50% |
| Bottom 10%: | 50% |
| Bottom 20%: | 50% |

for Por:

| | |
|---|---|
| Top 10%: | 50% |
| Top 20%: | 50% |
| Bottom 10%: | 50% |
| Bottom 20%: | 50% |

Random DTC

for Mat:

| | |
|---|---|
| Top 10%: | 67.7% |
| Top 20%: | 57.7% |
| Bottom 10%: | 83.8% |
| Bottom 20%: | 59.4% |

for Por:

| | |
|---|---|
| Top 10%: | 57.7% |
| Top 20%: | 68.2% |
| Bottom 10%: | 76.2% |
| Bottom 20%: | 68.1% |

Centroid DTC

for Mat:

| | |
|---|---|
| Top 10%: | 88.4% |
| Top 20%: | 82.9% |
| Bottom 10%: | 90.0% |
| Bottom 20%: | 66.2% |

for Por:

| | |
|---|---|
| Top 10%: | 74.3% |
| Top 20%: | 75.2% |
| Bottom 10%: | 92.3% |
| Bottom 20%: | 78.5% |

Comparison

for Mat:

| | | | |
|---|---|---|---|
| Top 10%: | 67.7% | — | 88.4% |
| Top 20%: | 57.7% | — | 82.9% |
| Bottom 10%: | 83.8% | — | 90.0% |
| Bottom 20%: | 59.4% | — | 66.2% |

for Por:

| | | | |
|---|---|---|---|
| Top 10%: | 57.7% | — | 74.3% |
| Top 20%: | 68.2% | — | 75.2% |
| Bottom 10%: | 76.2% | — | 92.3% |
| Bottom 20%: | 68.1% | — | 78.5% |

**Berkan Teber**
**19080**