UNIVERSITY OF AVEIRO
Department of Electronics, Telecommunications and Informatics

Regression problem
Prediction of final grades

Foundation of Machine Learning

Group members:

Stefano Petrina - (112056)

Stanisław Franczyk - (112059)

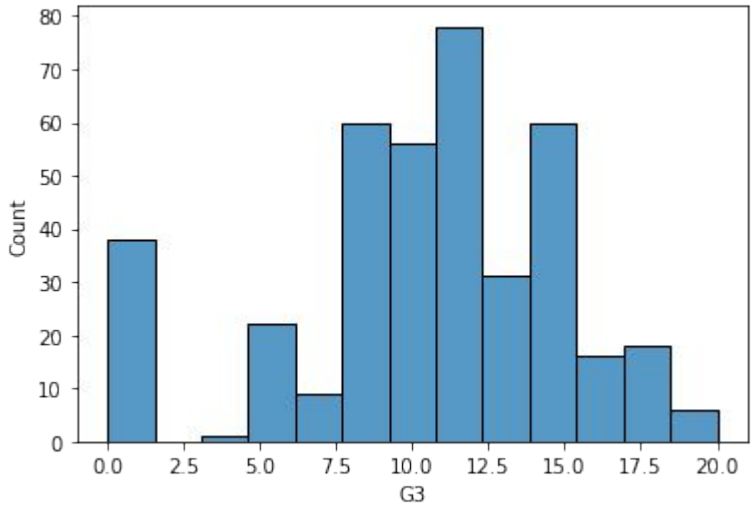
Presentation of the problem

- MATH COURSE
- PORTUGUESE COURSE

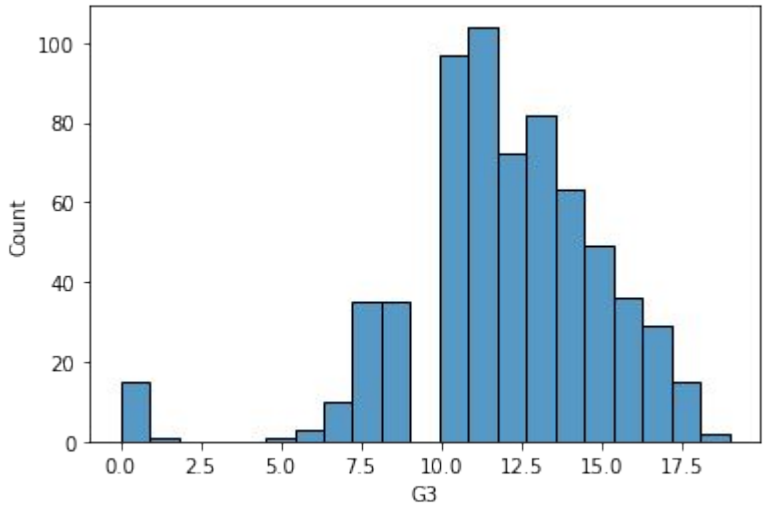
Columns	Description
school	student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
sex	student's sex (binary: 'F' - female or 'M' - male)
age	student's age (numeric: from 15 to 22)
address	student's home address type (binary: 'U' - urban or 'R' - rural)
famsize	family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
Pstatus	parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
Medu	mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
Fedu	father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
Mjob	mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at home' or 'other')
Fjob	father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at home' or 'other')
reason	reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
guardian	student's guardian (nominal: 'mother', 'father' or 'other')
traveltime	home to school travel time (numeric: 1 - ¡15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - ¿1 hour)
studytime	weekly study time (numeric: 1 - ¡2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - ¿10 hours)
failures	number of past class failures (numeric: n if $1 <= n < 3$, else 4)
schoolsup	extra educational support (binary: yes or no)
famsup	family educational support (binary: yes or no)
paid	extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
internet	Internet access at home (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
famrel	quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
freetime	free time after school (numeric: from 1 - very low to 5 - very high)
goout	going out with friends (numeric: from 1 - very low to 5 - very high)
Dalc	workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
Walc	weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
health	current health status (numeric: from 1 - very bad to 5 - very good)
absences	number of school absences (numeric: from 0 to 93)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20, output target)

Histogram of final grades

MATH COURSE



PORTUGUESE COURSE

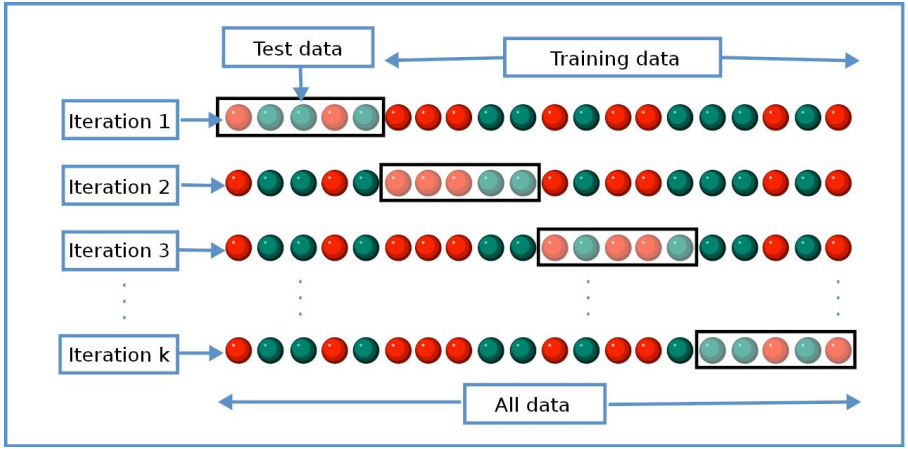


Model training

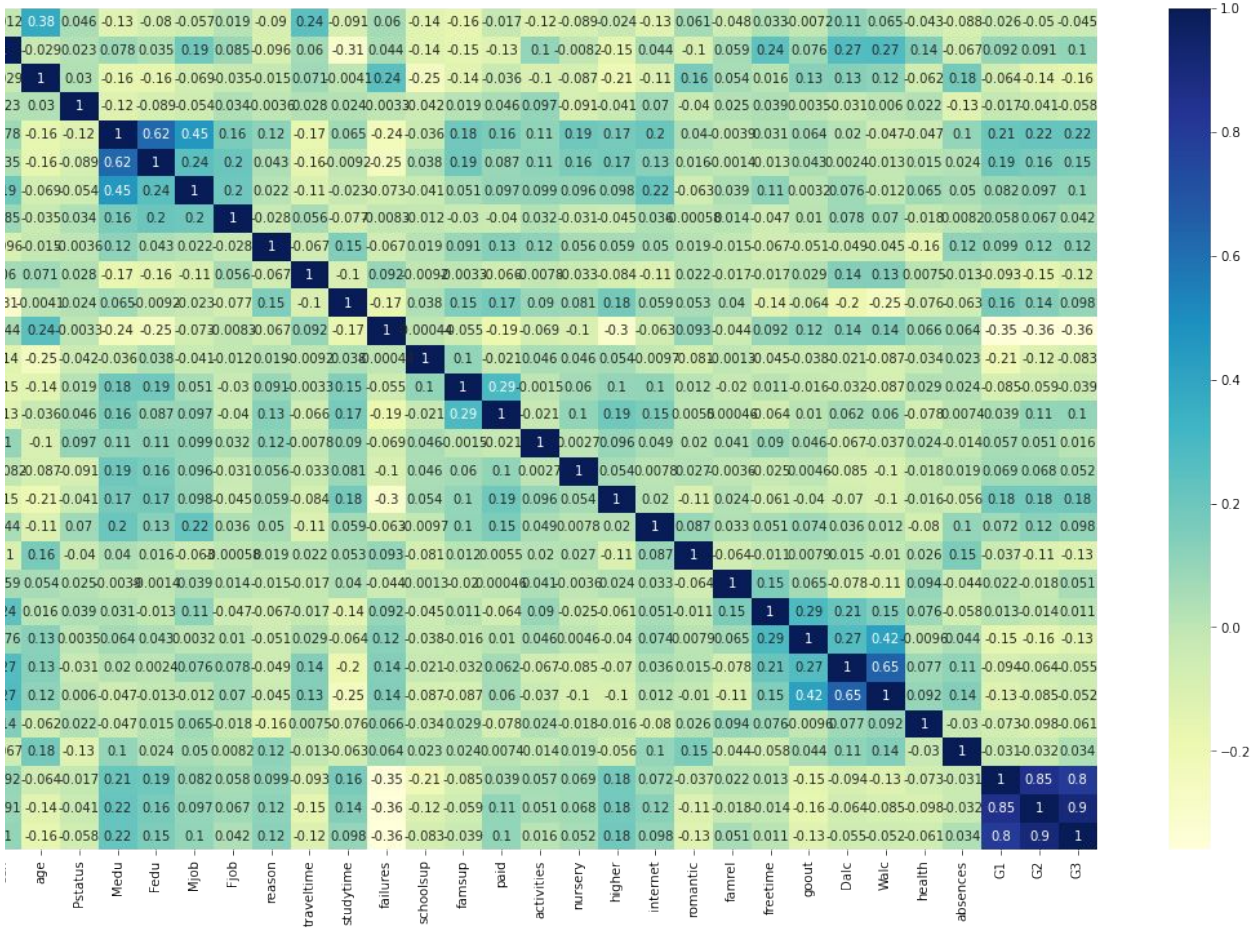
- K-fold Cross Validation
- Data Splitting
- Learning Rate

K-fold Cross Validation

- Divide the students in k=10 groups
- One group at a time is iteratively excluded (test data) and an attempt is made to predict it with the non-excluded group (training data)



HEATMAP of math dataset



Data Splitting

10 best parameters of math dataset

1. failures
2. Medu
3. higher
4. age
5. Fedu
6. goout
7. romantic
8. traveltime
9. Mjob
10. address

10 best parameters of Portuguese dataset

1. failures
2. higher
3. school
4. studytime
5. Medu
6. Fedu
7. Dalc
8. Walc
9. reason
10. address

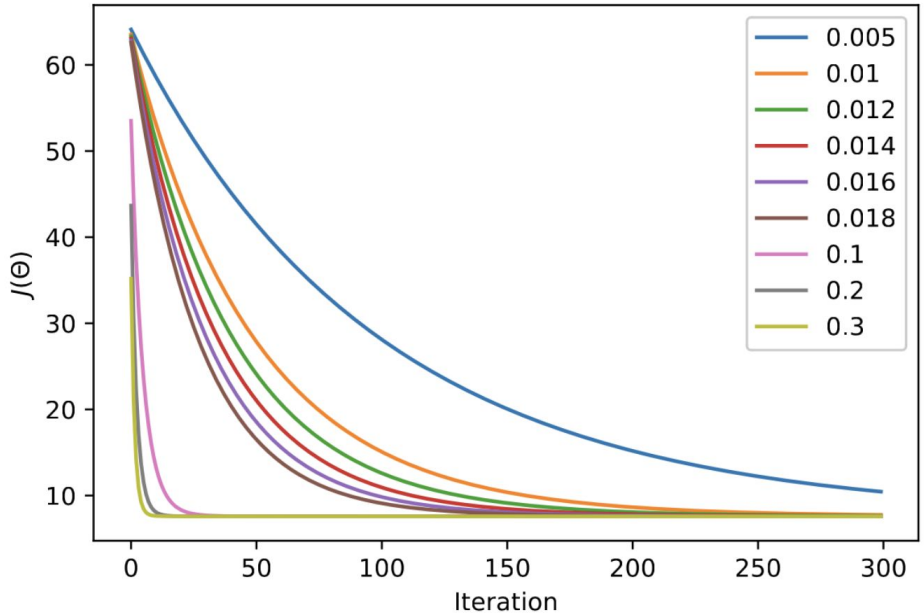
Learning Rate

- Implement gradient descent
- Compute MSE using K-fold validation function
- Repeat the process for all the learning rate values:

$$0.005, 0.01, 0.012, 0.014, 0.016, 0.018, 0.1, 0.2, 0.3.$$

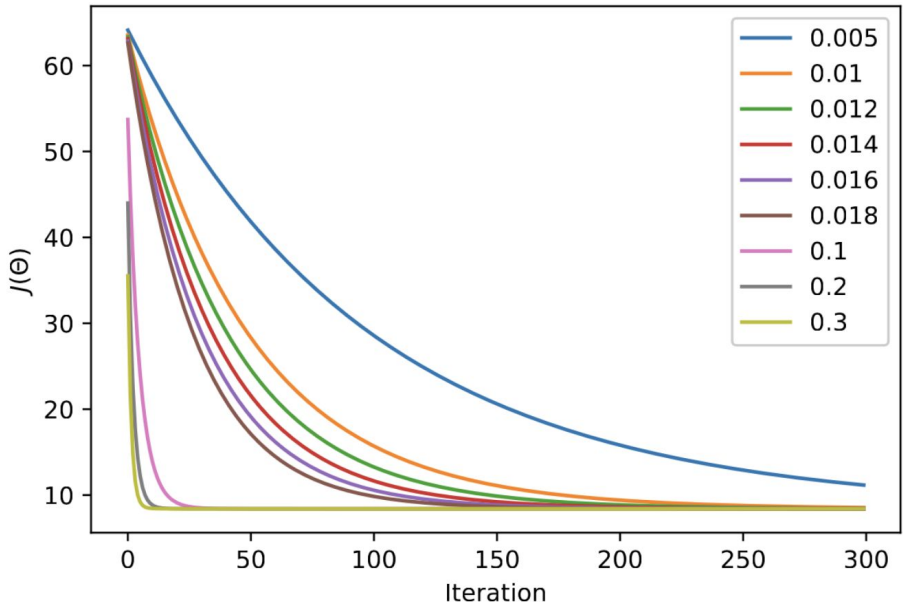
Results of regression G3 for the Math data

- Using all data



α	$J(\theta)$
0.005	18.1586
0.01	12.5738
0.012	11.9432
0.014	11.5469
0.016	11.2694
0.018	11.0595
0.1	10.0492
0.2	10.0667
0.3	10.0673

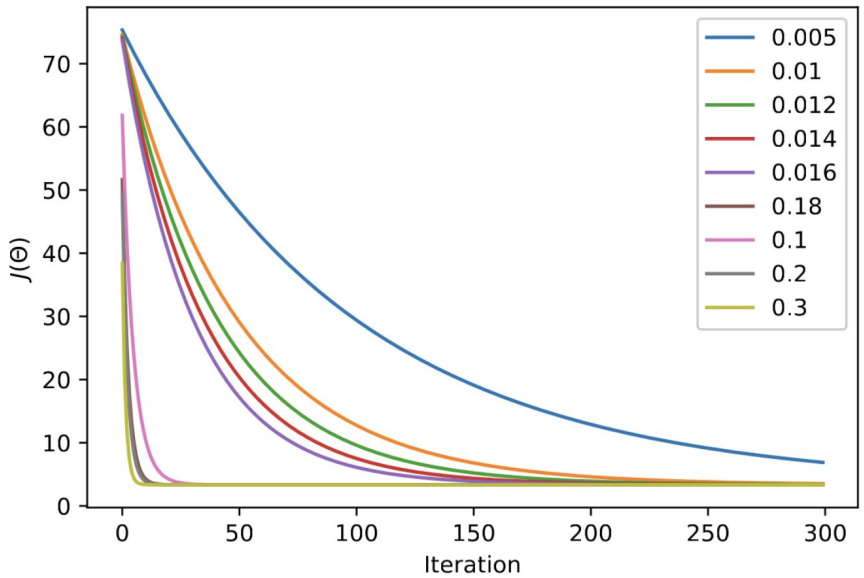
- Using 10 best data



α	$J(\theta)$
0.005	12.9262
0.01	9.4852
0.012	9.3323
0.014	9.2892
0.016	9.2824
0.018	9.2865
0.1	9.3213
0.2	9.3213
0.3	9.3213

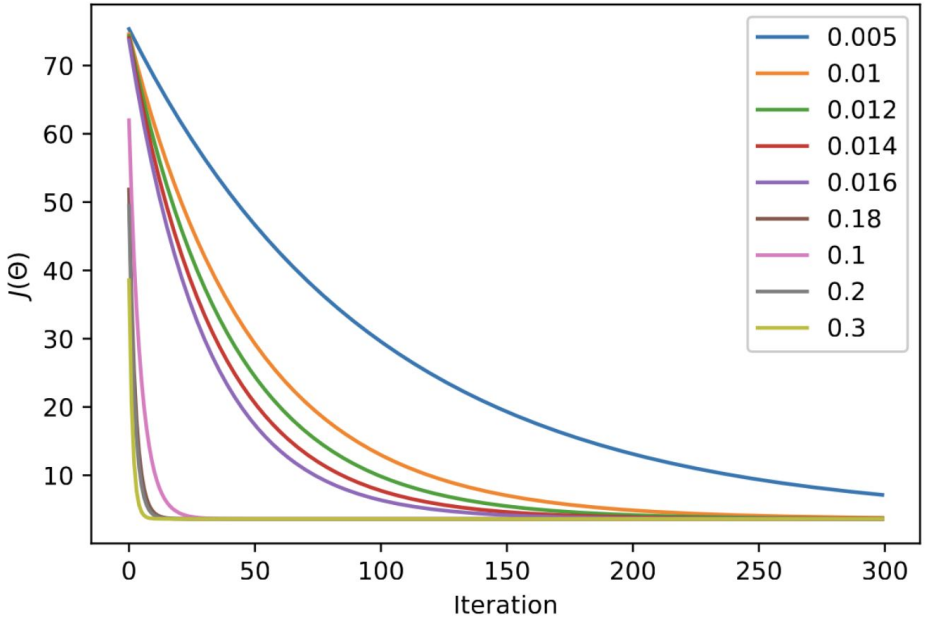
Results of regression G3 for the Portuguese data

- Using all data



α	$J(\theta)$
0.005	10.6939
0.01	4.5815
0.012	4.1968
0.014	4.0423
0.016	3.9788
0.018	3.9413
0.1	3.9347
0.2	3.9347
0.3	3.9347

- Using 10 best data



α	$J(\theta)$
0.005	8.9664
0.01	4.2415
0.012	4.0093
0.014	3.9267
0.016	3.8961
0.018	3.8789
0.1	3.8715
0.2	3.8715
0.3	3.8715

Best models for math dataset

$$h(x) = 10.4 + 20.23 \cdot x_1 + 0.63 \cdot x_2 + -0.48 \cdot x_3 + 0.23 \cdot x_4 + 0.32 \cdot x_5 + -0.1 \cdot x_6 + 0.5 \cdot x_7 + -0.11 \cdot x_8 + -0.17 \cdot x_9 + 0.46 \cdot x_{10} + -1.28 \cdot x_{11} + -0.45 \cdot x_{12} + -0.42 \cdot x_{13} + 0.17 \cdot x_{14} + -0.16 \cdot x_{15} + -0.07 \cdot x_{16} + 0.3 \cdot x_{17} + 0.19 \cdot x_{18} + -0.52 \cdot x_{19} + 0.21 \cdot x_{20} + 0.3 \cdot x_{21} + -0.66 \cdot x_{22} + -0.24 \cdot x_{23} + 0.34 \cdot x_{24} + -0.25 \cdot x_{25} + 0.45 \cdot x_{26} + 0.02 \cdot x_{27} + 0.29 \cdot x_{28} + -0.15 \cdot x_{29} + 0.31 \cdot x_{30} + -0.42 \cdot x_{31} + 0.05 \cdot x_{32} + 0.12 \cdot x_{33} + -0.19 \cdot x_{34} + -0.1 \cdot x_{35} + 0.41 \cdot x_{36} + -0.14 \cdot x_{37} + -0.09 \cdot x_{38} + 0.14 \cdot x_{39} + 0.15 \cdot x_{40} + -0.06 \cdot x_{41} + -0.04 \cdot x_{42}$$

$$h(x) = 10.42 + -1.34 \cdot x_1 + 0.5 \cdot x_2 + 0.23 \cdot x_3 + -0.09 \cdot x_4 + -0.05 \cdot x_5 + -0.49 \cdot x_6 + -0.43 \cdot x_7 + -0.15 \cdot x_8 + 0.2 \cdot x_9 + -0.17 \cdot x_{10} + 0.31 \cdot x_{11} + -0.2 \cdot x_{12} + -0.42 \cdot x_{13}$$

Thank you for your attention