Students' Exam Score Based on Multiple Factors

Group members:
Leonardo Langgeng, Lily Hou

Can we predict exam scores from external factors?

(individual traits, parents, lifestyle, etc.)

Dataset

https://archive.ics.uci.edu/ml/datasets/student+performance

- 395 data points
- 30 features
- 6 class (Math and Portuguese; G1, G2, G3) — we focus on Math G1

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltime	studytime	failures
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	2	2	0
1	GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1	2	0
2	GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	1	2	3
3	GP	F	15	U	GT3	T	4	2	health	services	home	mother	1	3	0
4	GP	F	16	U	GT3	T	3	3	other	other	home	father	1	2	0
..	...	..	...	...	...	...	...	...	...	...	...	...	...	...	...
390	MS	M	20	U	LE3	A	2	2	services	services	course	other	1	2	2
391	MS	M	17	U	LE3	T	3	1	services	services	course	mother	2	1	0
392	MS	M	21	R	GT3	T	1	1	other	other	course	other	1	1	3
393	MS	M	18	R	LE3	T	3	2	services	other	course	mother	3	1	0
394	MS	M	19	U	LE3	T	1	1	other	at_home	course	father	1	1	0

schoolsup	famsup	paid	activities	nursery	higher	internet	romantic	famrel	freetime	goout	Dalc	Walc	health	absences	G1
yes	no	no	no	yes	yes	no	no	4	3	4	1	1	3	6	5
no	yes	no	no	no	yes	yes	no	5	3	3	1	1	3	4	5
yes	no	yes	no	yes	yes	yes	no	4	3	2	2	3	3	10	7
no	yes	yes	yes	yes	yes	yes	yes	3	2	2	1	1	5	2	15
no	yes	yes	no	yes	yes	no	no	4	3	2	1	2	5	4	6
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	..
no	yes	yes	no	yes	yes	no	no	5	5	4	4	5	4	11	9
no	no	no	no	no	yes	yes	no	2	4	5	3	4	2	3	14
no	no	no	no	no	yes	no	no	5	5	3	3	3	3	3	10
no	no	no	no	no	yes	yes	no	4	4	1	3	4	5	0	11
no	no	no	no	yes	yes	yes	no	3	2	3	3	3	5	5	8

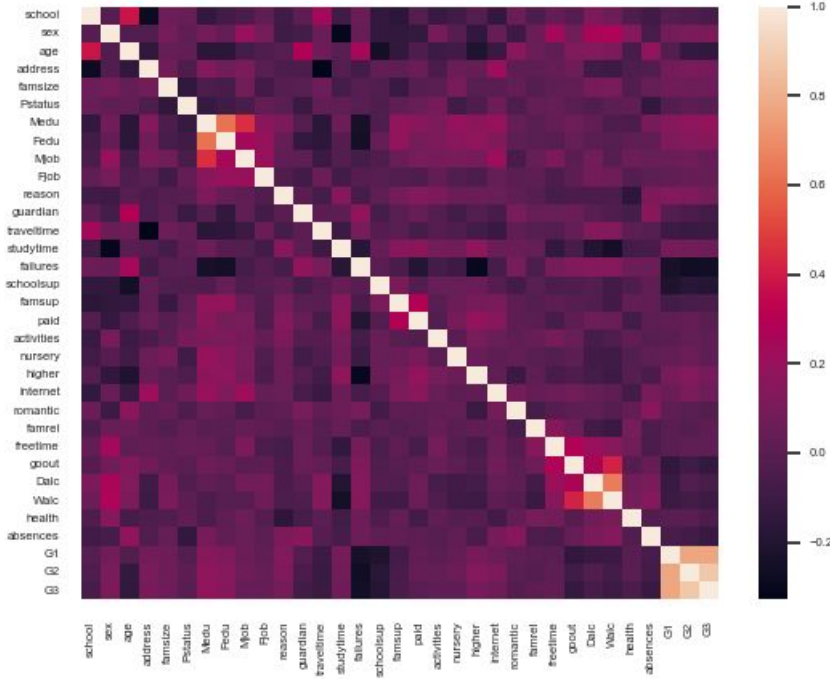
school - student's school (binary: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira)
sex - student's sex (binary: "F" - female or "M" - male)
age - student's age (numeric: from 15 to 22)
address - student's home address type (binary: "U" - urban or "R" - rural)
famsize - family size (binary: "LE3" - 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)

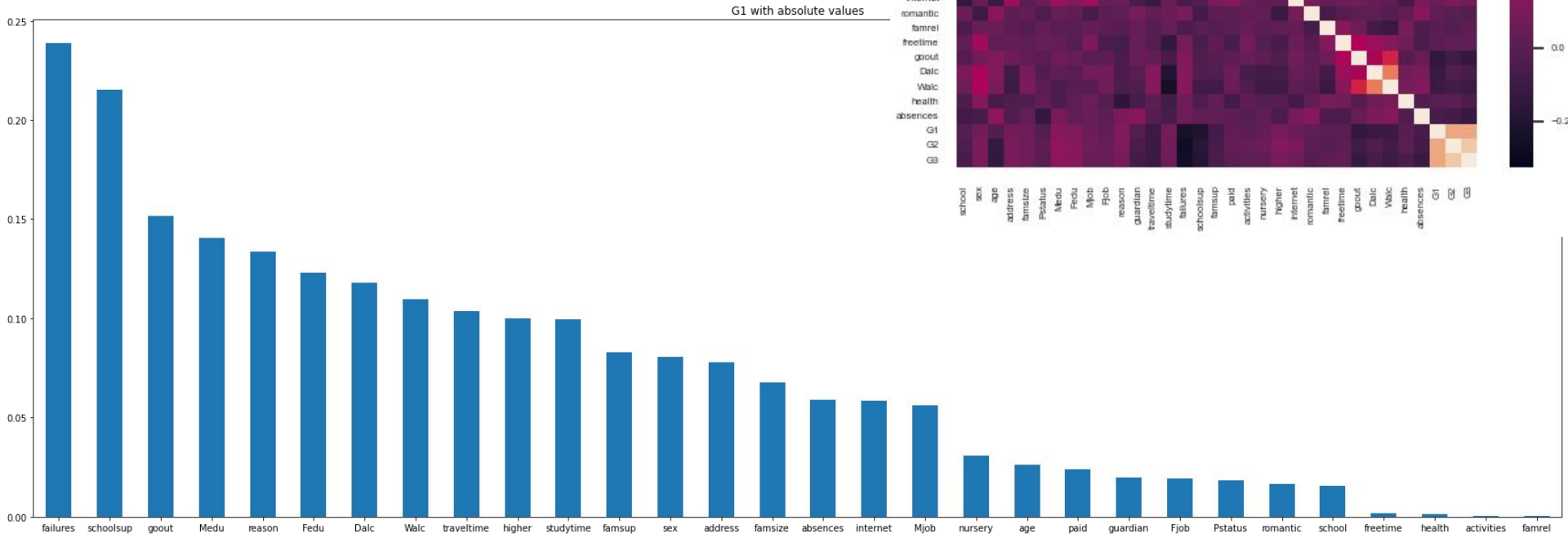
Platform Analysis

- Analysis via correlation plot
  - Feature vs grade plot

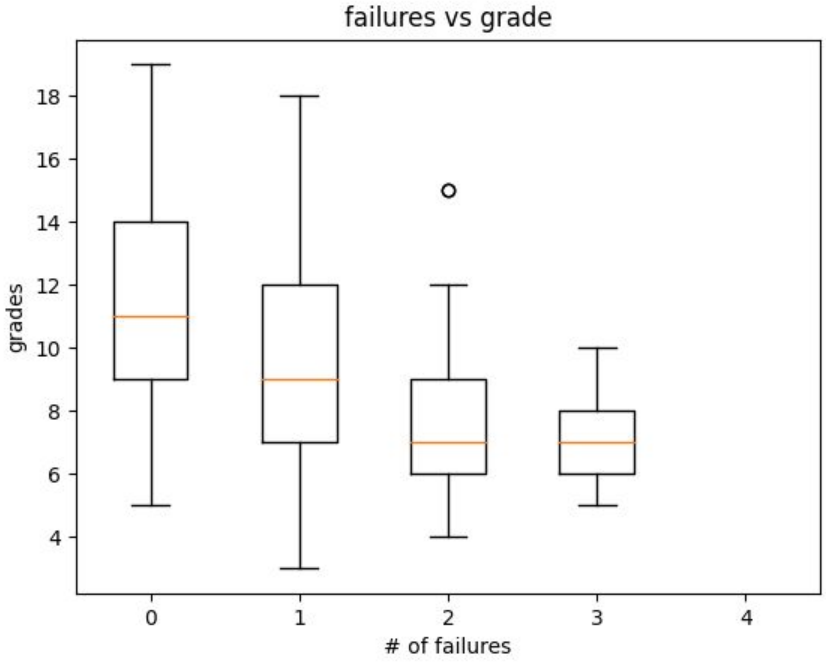
- Gaussian Naive Bayes
  - Probability based ML
  - Simple and fast training
  - Able to handle large features
  - Classification, passing score >= 10
    Data split 80/20

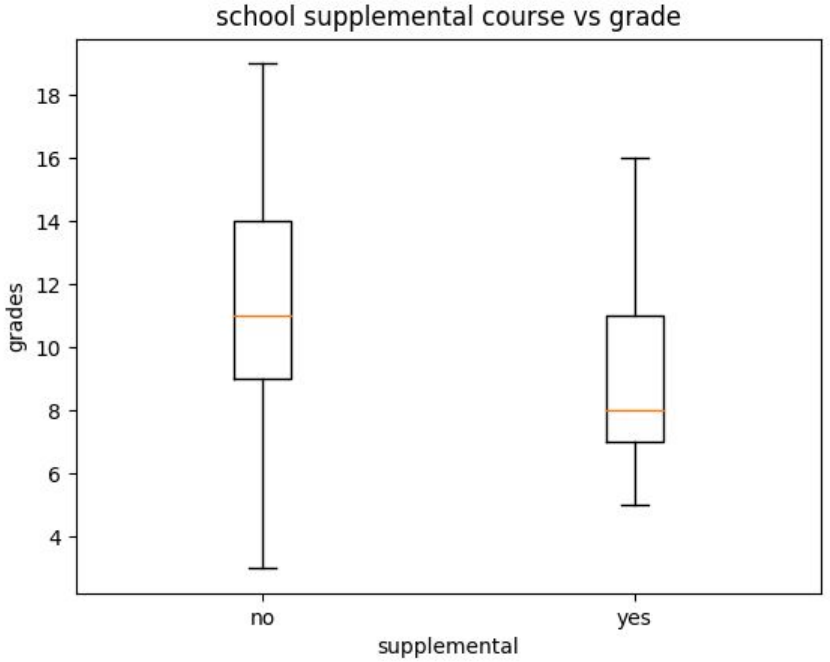
Correlations

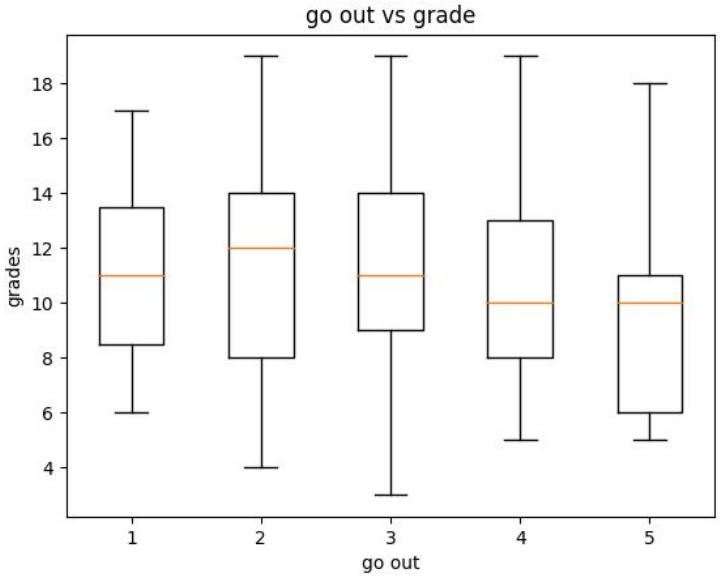


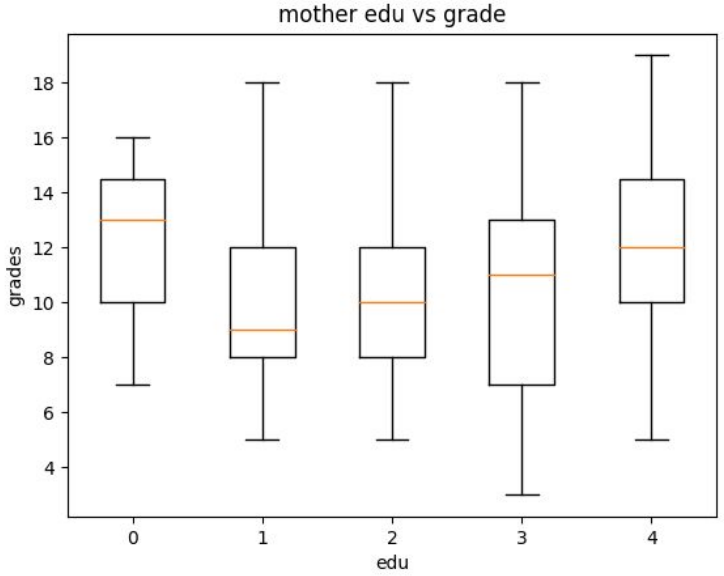


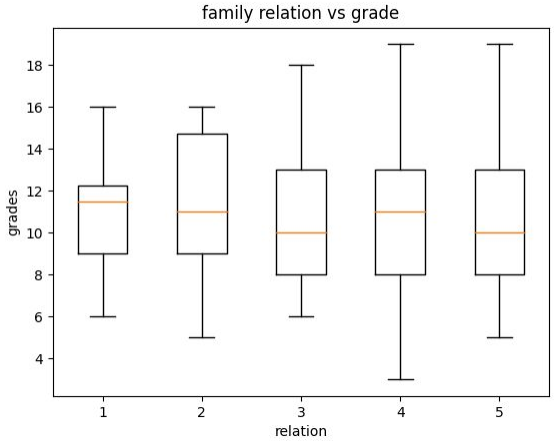
Some Plots


failures vs grade
grades
18
16
14
12
10
8
6
4
0
1
2
3
4
# of failures


school supplemental course vs grade
grades
18
16
14
12
10
8
6
4
no
yes
supplemental


go out vs grade
grades
go out
1
2
3
4
5
4
6
8
10
12
14
16
18


mother edu vs grade
grades
edu
0
1
2
3
4
4
6
8
10
12
14
16
18


family relation vs grade
grades
relation
1
2
3
4
5
4
6
8
10
12
14
16
18

Naive Bayes in a nutshell

	school	sex	age	P/NP
student1	A	M	15	P
student2	B	F	15	P
student3	B	M	15	NP
student4	A	F	12	NP
student5	B	F	17	NP

newStudent	A	F	15	????

P(class) * P(features show up on class)

P(Pass) = (2/5) * (1/2) * (1/2) * (2/2) = 0.1 → 0.692

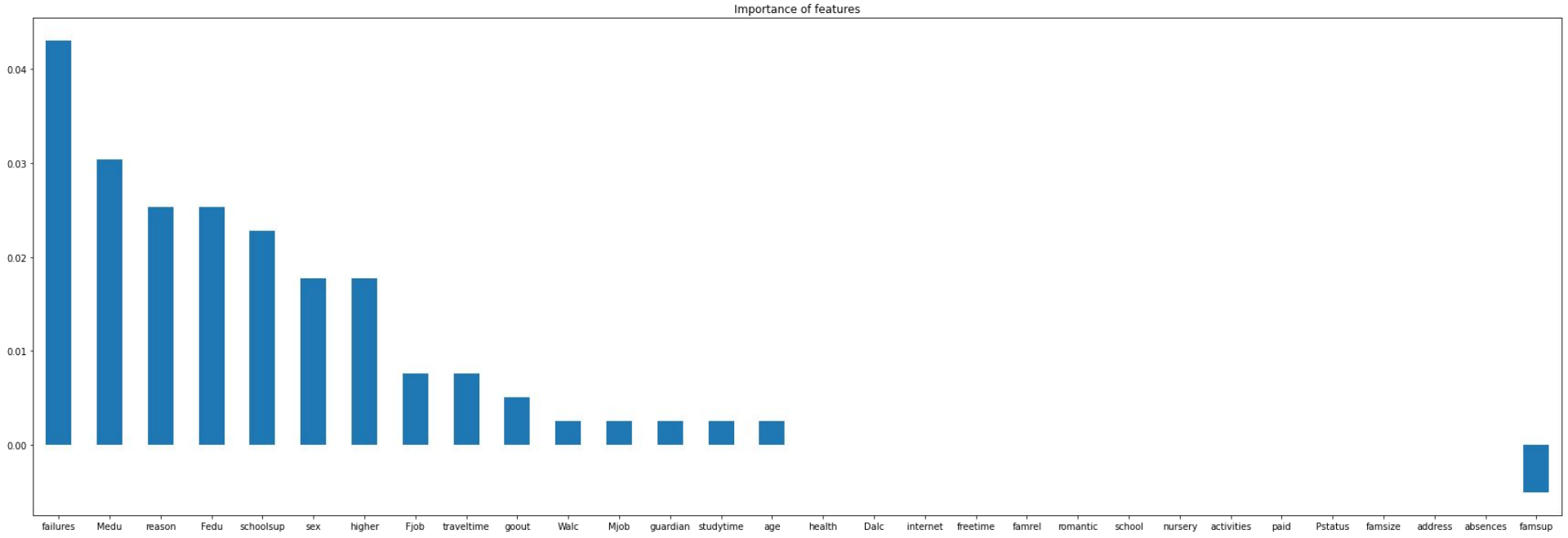
P(NoPass) = (3/5) * (1/3) * (2/3) * (1/3) = 0.044 → 0.308

Model result

Training data 316
Testing data 79

Accuracy ~73%

	Data P	Data NP
Pred P	47	12
Pred NP	9	11



Classifier demo

Recap

Can we predict exam scores from external factors?

Yes, but not that accurate(0.7).

GIGO? We tried different dataset and got similar results.

What can we learn from this dataset?

Students that failed before have a tendency to fail again.

If any attention is given, it should be for this group.

The extra educational support results are underwhelming,

but again, the successful students won't need it.

Can this project be useful in a real-world scenario?

It can be a tool for HS teachers/counselors to keep an eye on students that may need help.