Exploring the role of alcohol and other factors in determining student achievement :D

Group 6 member:
Lei Zhu, Ruiyuan Ming, Minghao Mo, Yiming Xu, Xinrui Zhong

DATA: https://www.kaggle.com/datasets/uciml/student-alcohol-consumption/data

```
Data Shape: (1044, 32)
First Few Records:
   Unnamed: 0 school sex  age address famsize Pstatus  Medu  Fedu     Mjob  \
0           0     GP   F   18       U     GT3       A     4     4  at_home
1           1     GP   F   17       U     GT3       T     1     1  at_home
2           2     GP   F   15       U     LE3       T     1     1  at_home
3           3     GP   F   15       U     GT3       T     4     2   health
4           4     GP   F   16       U     GT3       T     3     3    other

   ... internet romantic famrel  freetime  goout  Dalc Walc health absences  \
0  ...       no       no      4         3      4     1    1      3         6
1  ...      yes       no      5         3      3     1    1      3         4
2  ...      yes       no      4         3      2     2    3      3        10
3  ...      yes      yes      3         2      2     1    1      5         2
4  ...       no       no      4         3      2     1    2      5         4

   G3
0   6
1   6
2  10
3  15
4  10

[5 rows x 32 columns]
```

Variables:

school, sex, age, address,

famsize(family size)

Pstatus(parent's cohabitation status)

Medu(mother's education)

Fedu(father's education)

Mjob(mother’s job)

Fjob(father's job)

guardian(student's guardian)

traveltime(home to school travel time)

studytime(weekly study time)

failures(number of past class failures)

schoolsup(extra educational support)

famsup (family educational support)

paid (extra paid classes within the course subject activities)

nursery - attended nursery school

higher - wants to take higher education

internet - Internet access at home

romantic - with a romantic relationship

famrel - quality of family relationships

freetime - free time after school

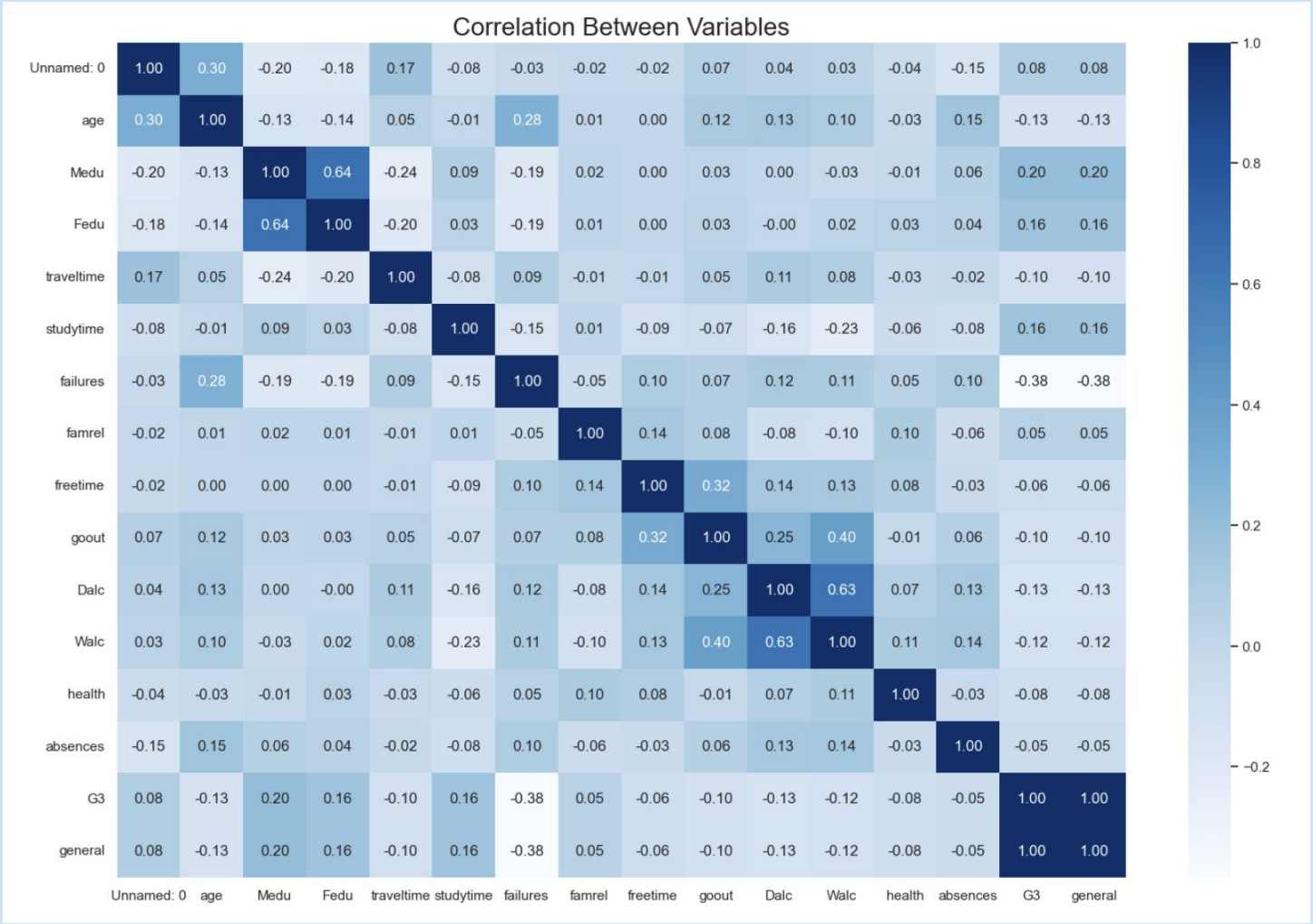
goout - going out with friends

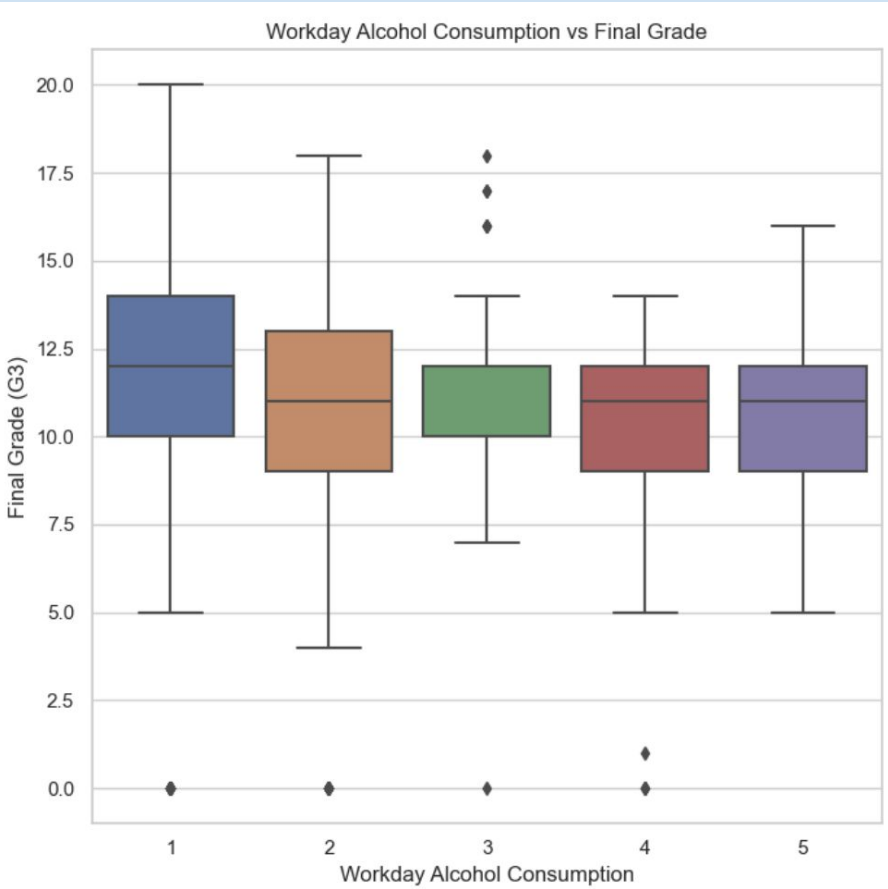
Dalc - workday alcohol consumption

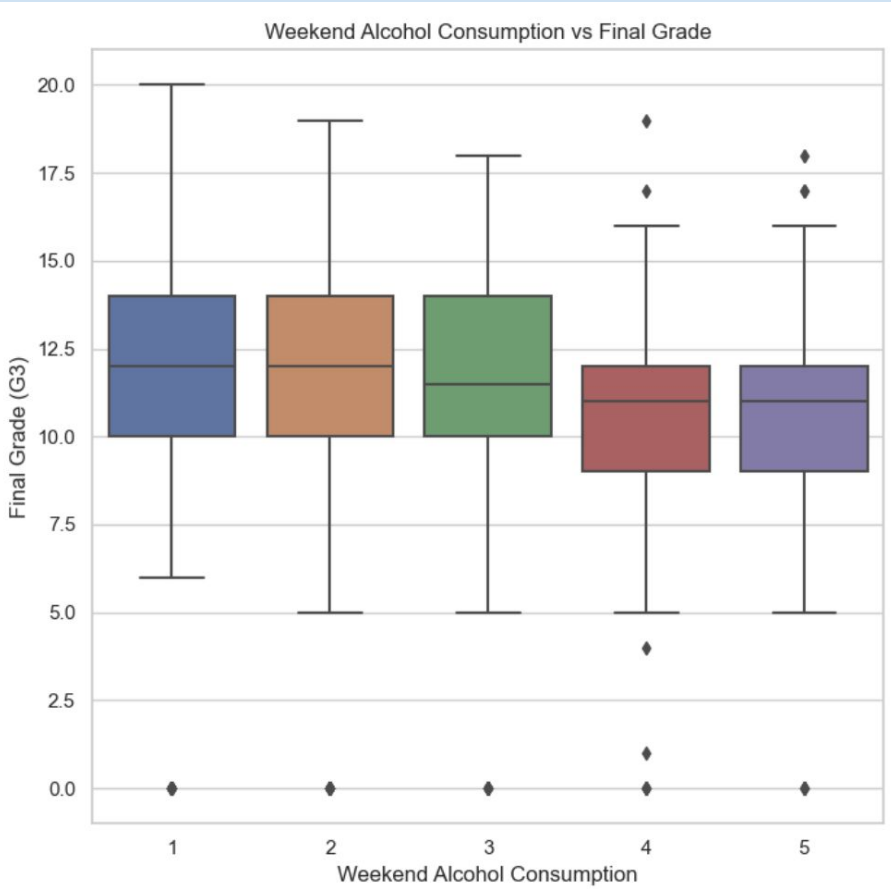
Walc - weekend alcohol consumption

health - current health status

absences - number of school absences


Correlation Between Variables
Unnamed: 0 1.00 0.30 -0.20 -0.18 0.17 -0.08 -0.03 -0.02 -0.02 0.07 0.04 0.03 -0.04 -0.15 0.08 0.08
age 0.30 1.00 -0.13 -0.14 0.05 -0.01 0.28 0.01 0.00 0.12 0.13 0.10 -0.03 0.15 -0.13 -0.13
Medu -0.20 -0.13 1.00 0.64 -0.24 0.09 -0.19 0.02 0.00 0.03 0.00 -0.03 -0.01 0.06 0.20 0.20
Fedu -0.18 -0.14 0.64 1.00 -0.20 0.03 -0.19 0.01 0.00 0.03 -0.00 0.02 0.03 0.04 0.16 0.16
traveltime 0.17 0.05 -0.24 -0.20 1.00 -0.08 0.09 -0.01 -0.01 0.05 0.11 0.08 -0.03 -0.02 -0.10 -0.10
studytime -0.08 -0.01 0.09 0.03 -0.08 1.00 -0.15 0.01 -0.09 -0.07 -0.16 -0.23 -0.06 -0.08 0.16 0.16
failures -0.03 0.28 -0.19 -0.19 0.09 -0.15 1.00 -0.05 0.10 0.07 0.12 0.11 0.05 0.10 -0.38 -0.38
famrel -0.02 0.01 0.02 0.01 -0.01 0.01 -0.05 1.00 0.14 0.08 -0.08 -0.10 0.10 -0.06 0.05 0.05
freetime -0.02 0.00 0.00 0.00 -0.01 -0.09 0.10 0.14 1.00 0.32 0.14 0.13 0.08 -0.03 -0.06 -0.06
goout 0.07 0.12 0.03 0.03 0.05 -0.07 0.07 0.08 0.32 1.00 0.25 0.40 -0.01 0.06 -0.10 -0.10
Dalc 0.04 0.13 0.00 -0.00 0.11 -0.16 0.12 -0.08 0.14 0.25 1.00 0.63 0.07 0.13 -0.13 -0.13
Walc 0.03 0.10 -0.03 0.02 0.08 -0.23 0.11 -0.10 0.13 0.40 0.63 1.00 0.11 0.14 -0.12 -0.12
health -0.04 -0.03 -0.01 0.03 -0.03 -0.06 0.05 0.10 0.08 -0.01 0.07 0.11 1.00 -0.03 -0.08 -0.08
absences -0.15 0.15 0.06 0.04 -0.02 -0.08 0.10 -0.06 -0.03 0.06 0.13 0.14 -0.03 1.00 -0.05 -0.05
G3 0.08 -0.13 0.20 0.16 -0.10 0.16 -0.38 0.05 -0.06 -0.10 -0.13 -0.12 -0.08 -0.05 1.00 1.00
general 0.08 -0.13 0.20 0.16 -0.10 0.16 -0.38 0.05 -0.06 -0.10 -0.13 -0.12 -0.08 -0.05 1.00 1.00
Unnamed: 0 age Medu Fedu traveltime studytime failures famrel freetime goout Dalc Walc health absences G3 general
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


Workday Alcohol Consumption vs Final Grade
Final Grade (G3)
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
1
2
3
4
5
Workday Alcohol Consumption


Weekend Alcohol Consumption vs Final Grade
Final Grade (G3)
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
1
2
3
4
5
Weekend Alcohol Consumption

Feature engineering

1. Check for missing values – no missing values
2. Split data: 80% for training and 20% for testing
3. One-hot encoding for qualitative variables
4. Min-max normalization for quantitative variables

Feature selection

Method: VarianceThreshold
Parameter: threshold=0.24 vs 0.25
Number of selected variables: 14 vs 7

Threshold=0.24

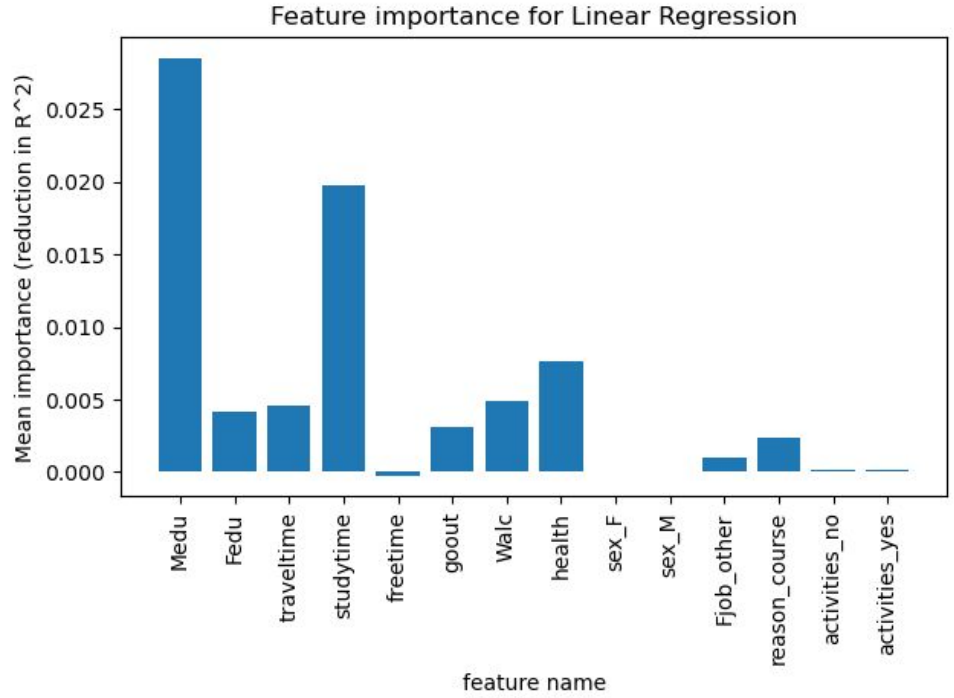
	Medu	Fedu	traveltime	studytime	freetime	goout	Walc	health
0	-0.5	-0.5	-1.000000	-0.333333	0.5	-0.5	-0.5	1.0
1	0.5	0.5	-1.000000	-0.333333	-0.5	0.0	-0.5	0.0
2	1.0	0.0	-1.000000	0.333333	0.0	0.0	0.0	-1.0
3	0.5	0.5	0.333333	-0.333333	0.0	0.0	0.0	-0.5
4	-0.5	-0.5	-1.000000	0.333333	0.0	0.5	-1.0	1.0

sex_F	sex_M	Fjob_other	reason_course	activities_no	activities_yes
1.0	0.0	0.0	1.0	1.0	0.0
0.0	1.0	1.0	0.0	0.0	1.0
1.0	0.0	1.0	0.0	0.0	1.0
0.0	1.0	1.0	0.0	1.0	0.0
1.0	0.0	1.0	0.0	0.0	1.0

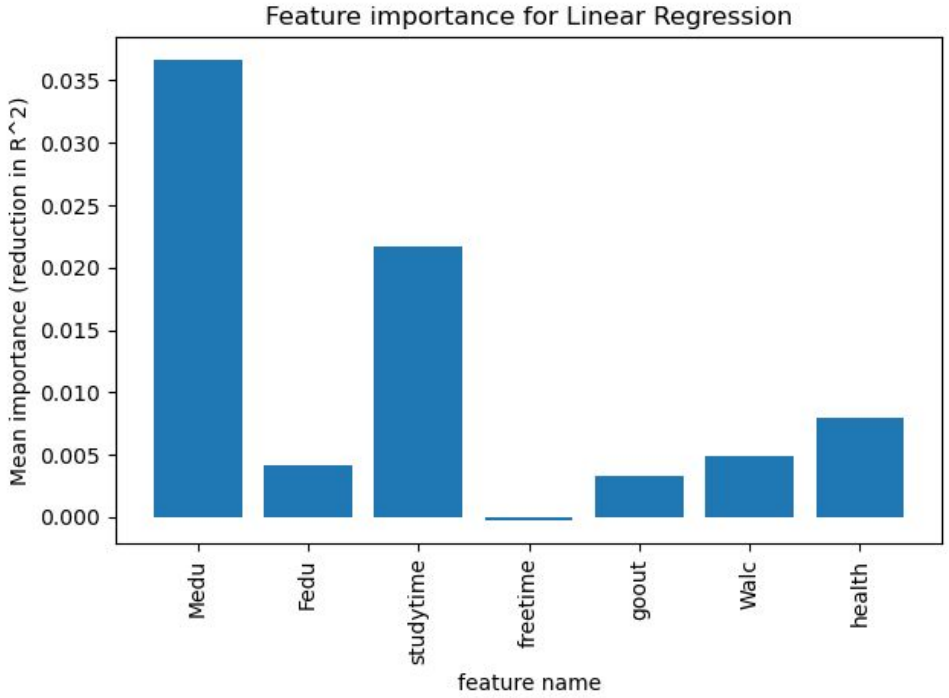
Threshold=0.25

	Medu	Fedu	studytime	freetime	goout	Walc	health
0	-0.5	-0.5	-0.333333	0.5	-0.5	-0.5	1.0
1	0.5	0.5	-0.333333	-0.5	0.0	-0.5	0.0
2	1.0	0.0	0.333333	0.0	0.0	0.0	-1.0
3	0.5	0.5	-0.333333	0.0	0.0	0.0	-0.5
4	-0.5	-0.5	0.333333	0.0	0.5	-1.0	1.0

Linear Regression

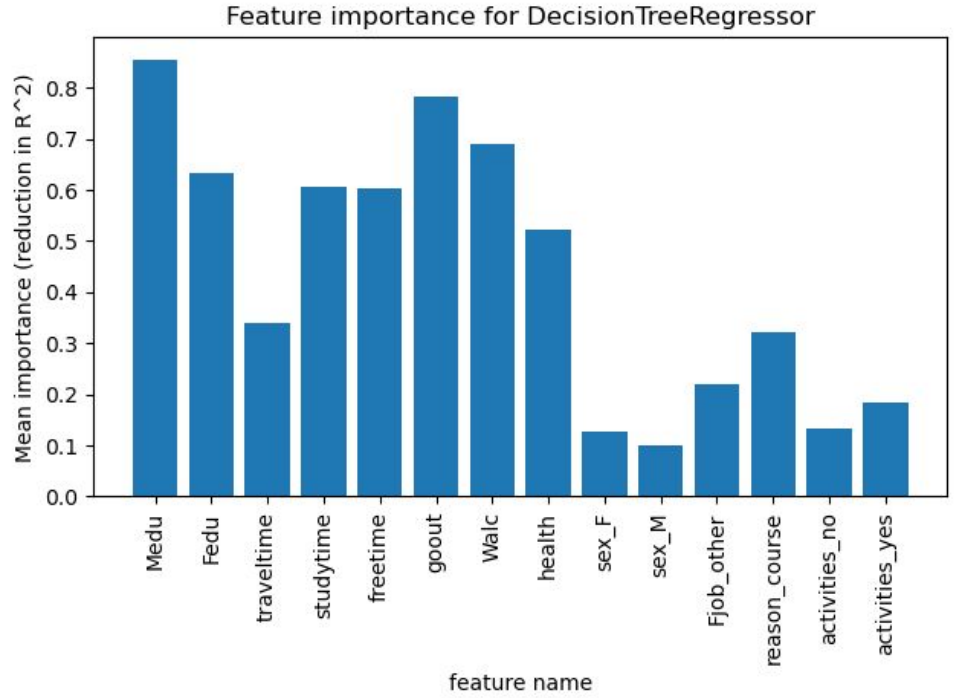


threshold = 0.24
MSE = 7.956

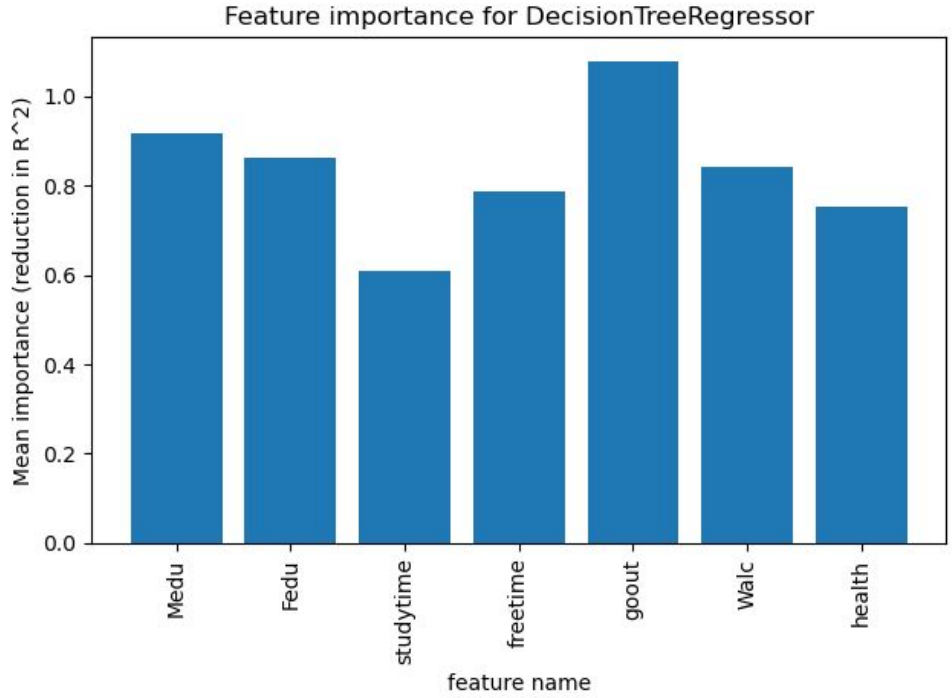


threshold = 0.25
MSE = 7.972

DecisionTree Regressor

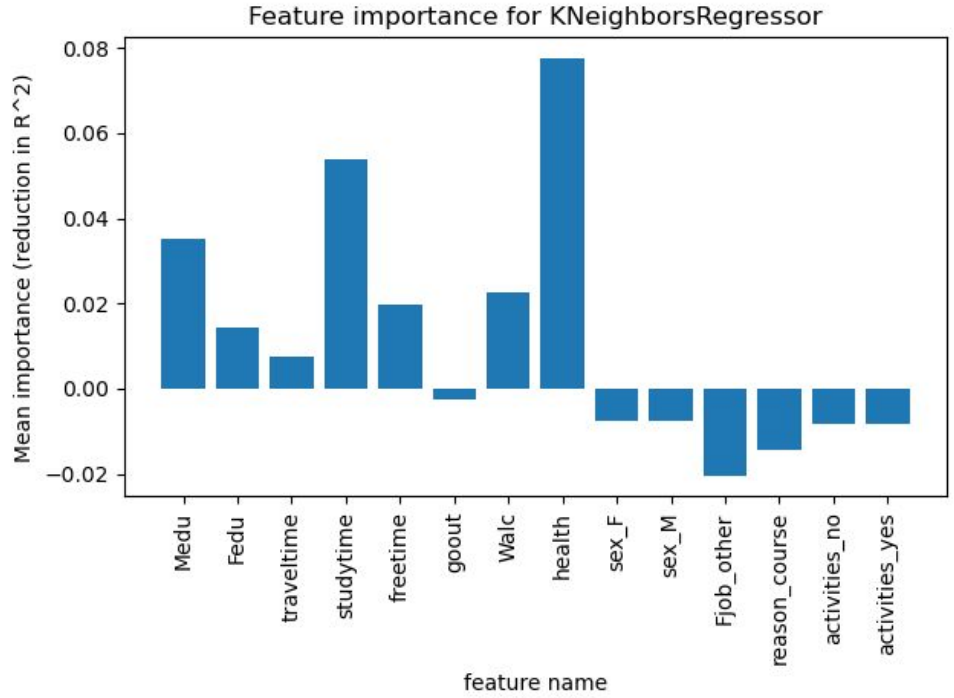


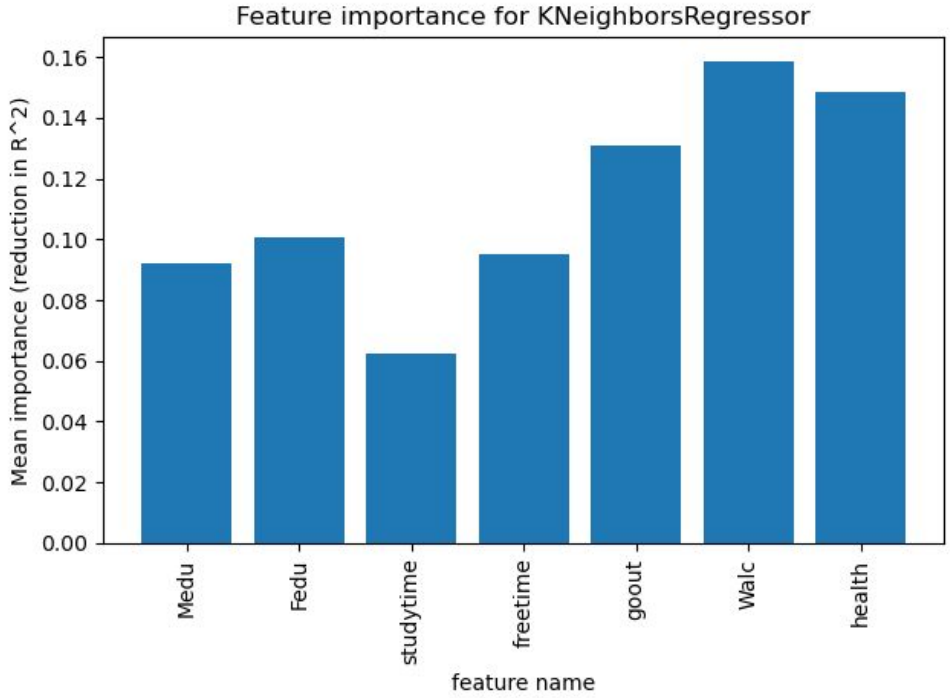
threshold = 0.24
MSE = 19.228



threshold = 0.25
MSE = 17.650

KNeighbors Regressor

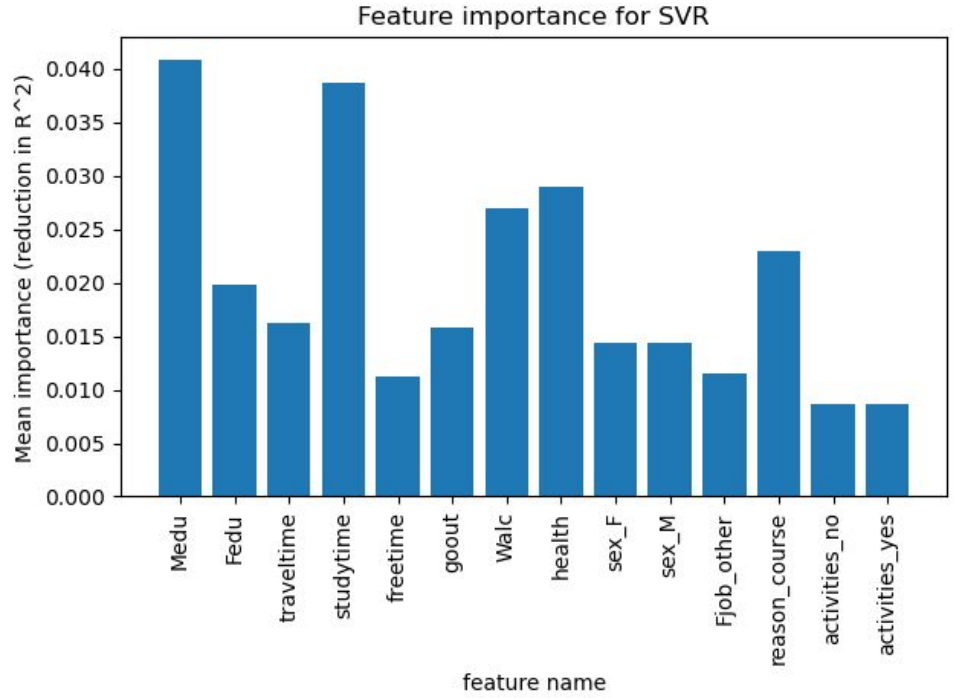




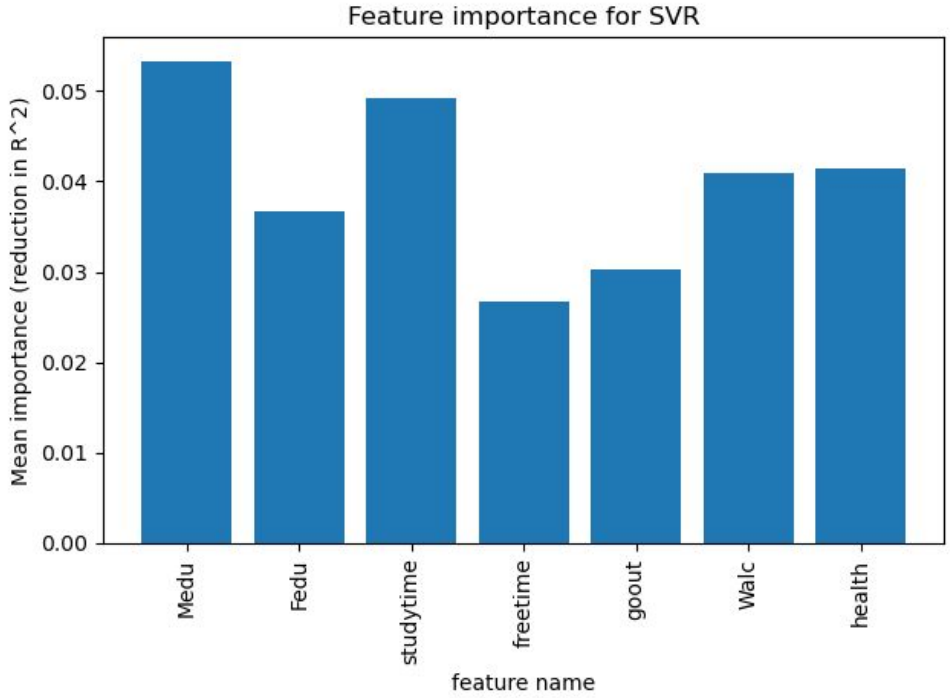
threshold = 0.24
MSE = 9.775

threshold = 0.25
MSE = 10.523

SVR

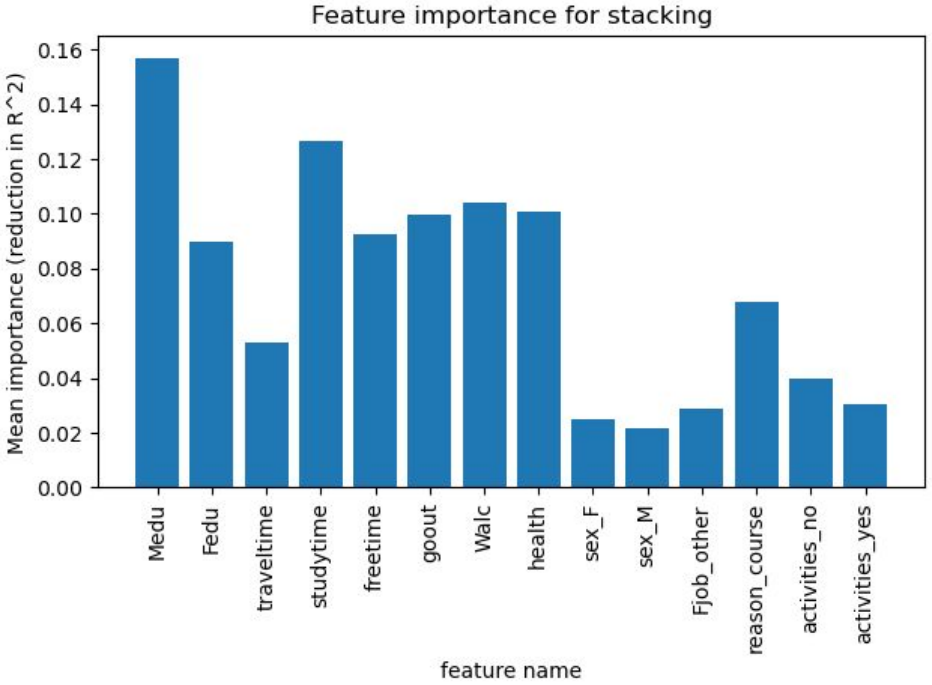


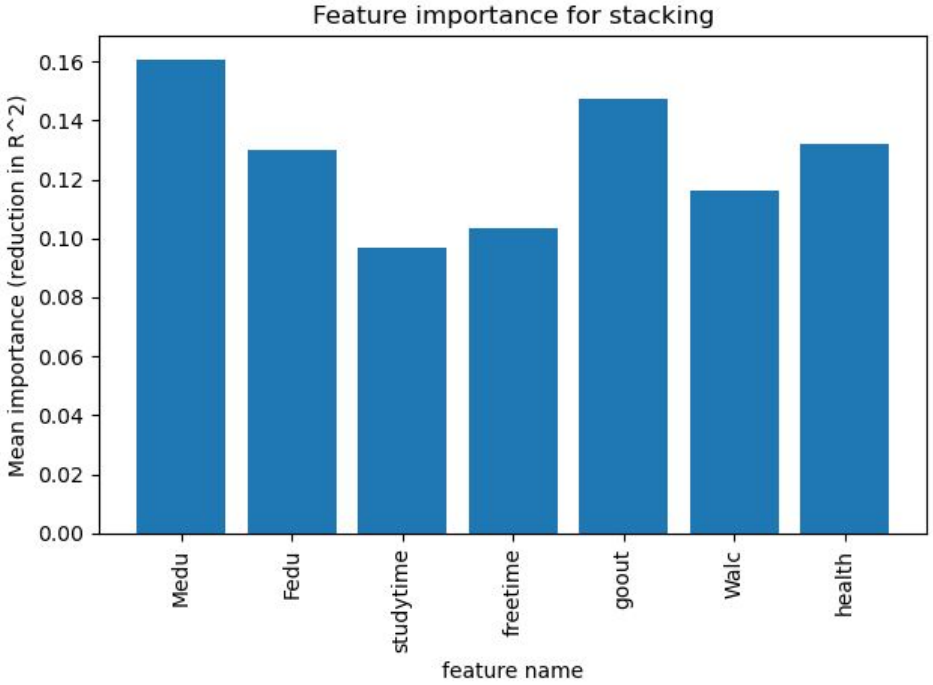
threshold = 0.24
MSE = 7.696



threshold = 0.25
MSE = 8.044

Stacking



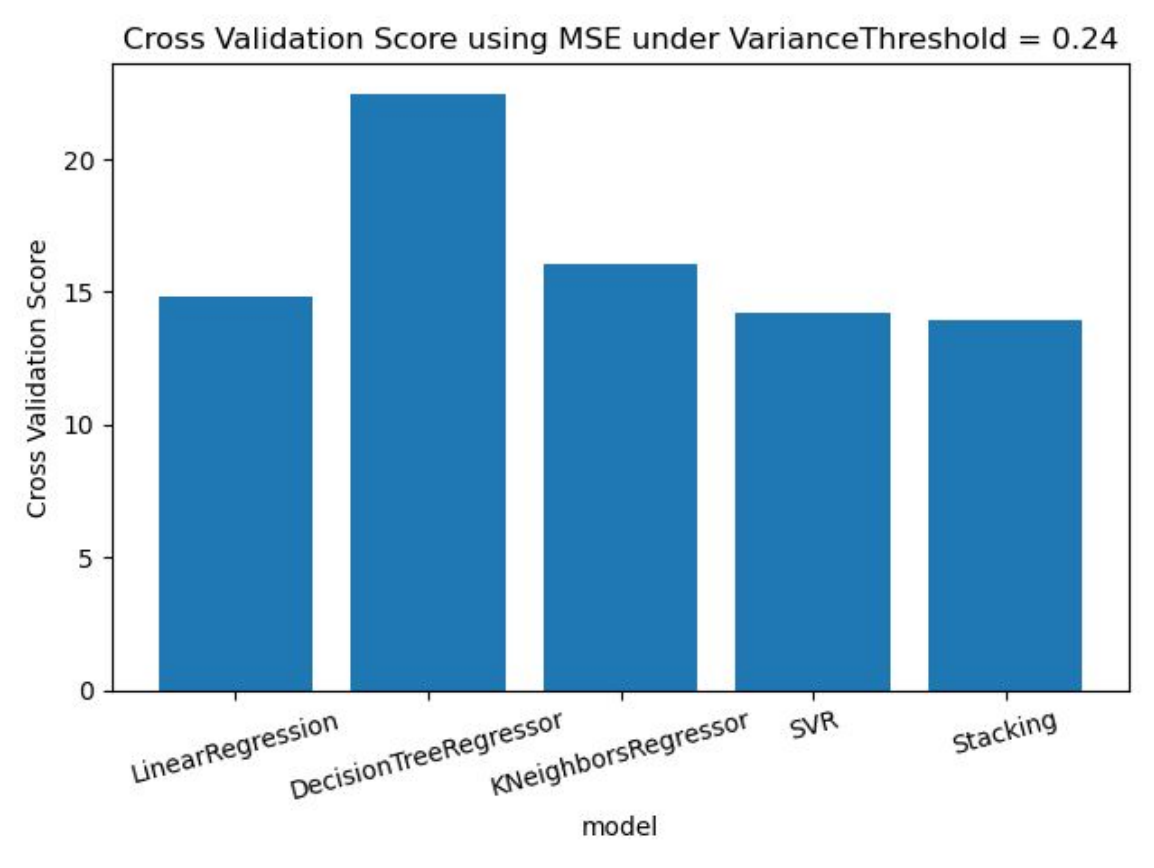


threshold = 0.24
MSE = 7.572

threshold = 0.25
MSE = 7.850

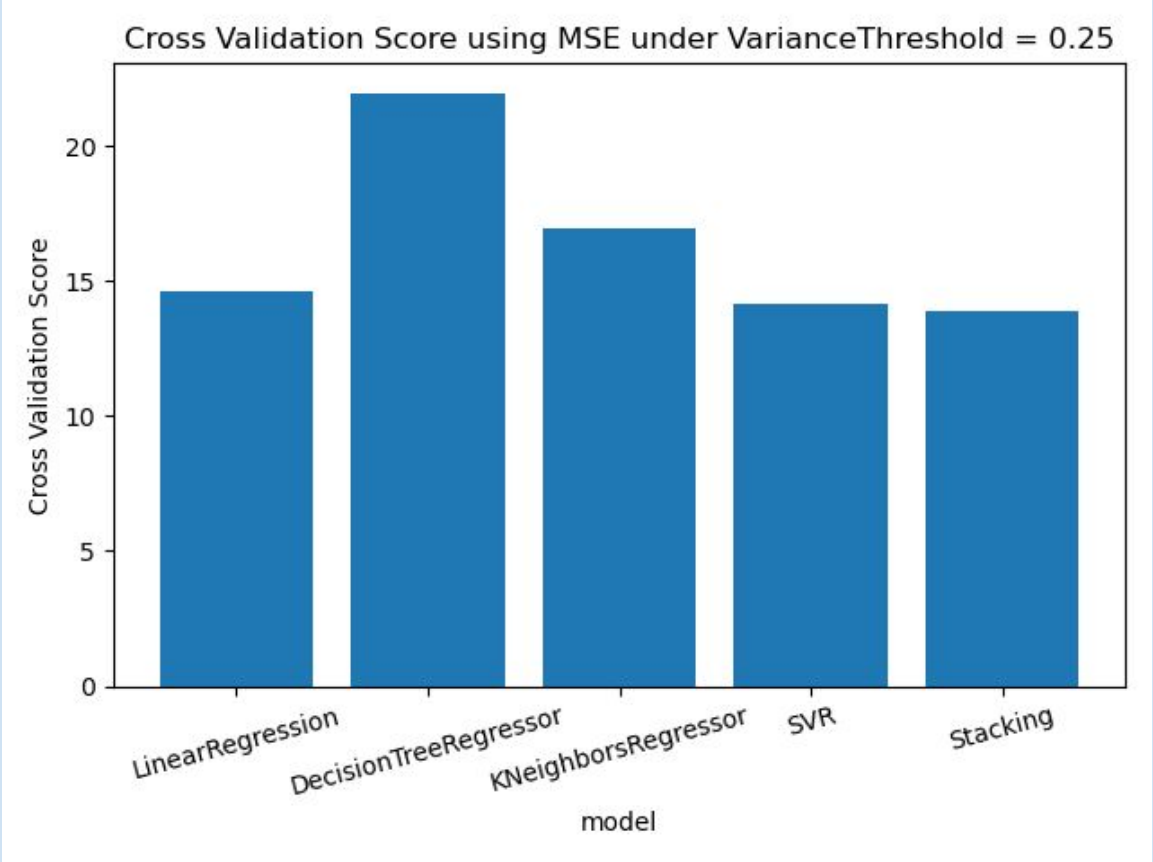
Cross Validation Score using MSE under VarianceThreshold = 0.24

```
CV score for LinearRegression: 14.8
CV score for DecisionTreeRegressor: 22.5
CV score for KNeighborsRegressor: 16.0
CV score for SVR: 14.2
CV score for Stacking: 13.9
```

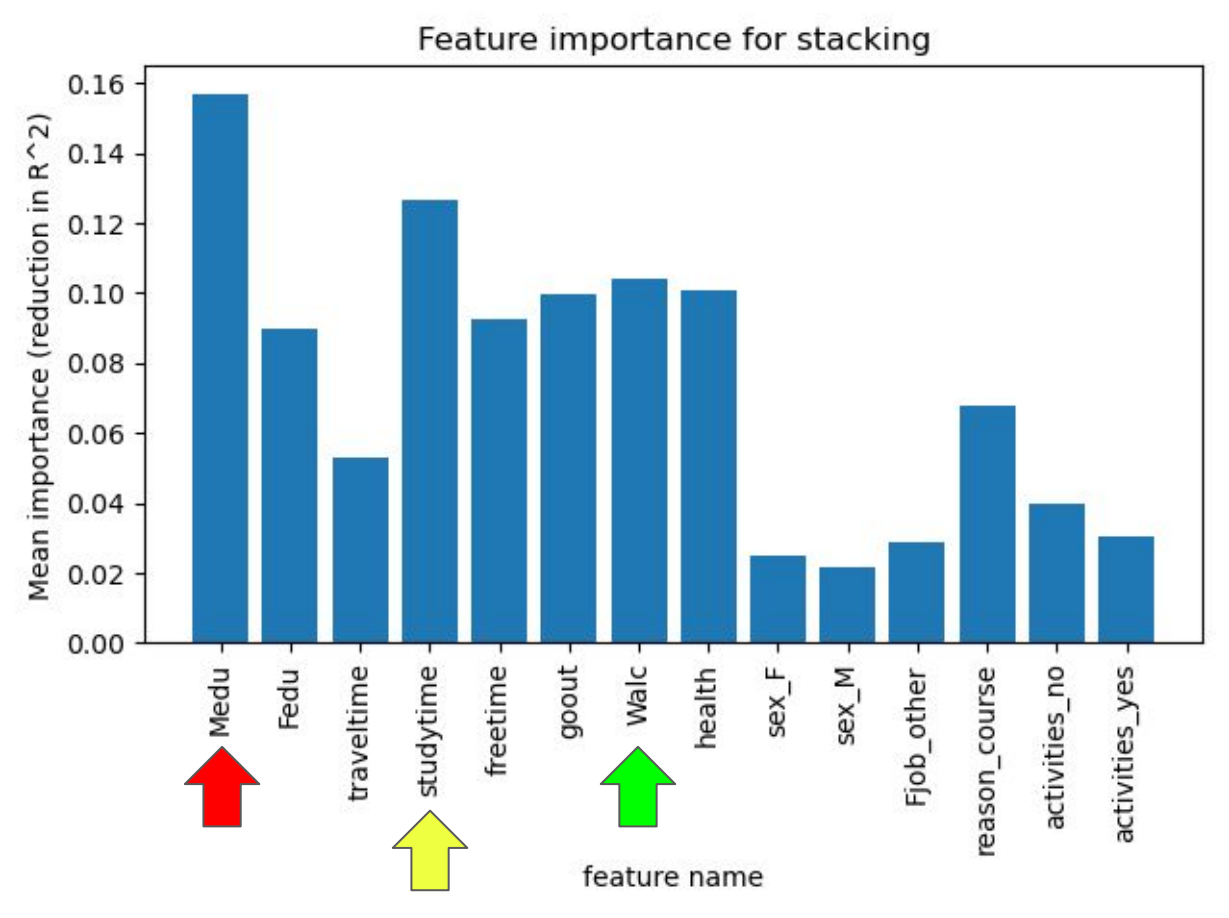


Cross Validation Score using MSE under VarianceThreshold = 0.25

```
CV score for LinearRegression: 14.6
CV score for DecisionTreeRegressor: 21.9
CV score for KNeighborsRegressor: 16.9
CV score for SVR: 14.2
CV score for Stacking: 13.9
```



Best model: Stacking under VarianceThreshold = 0.24



Thank you for listening!