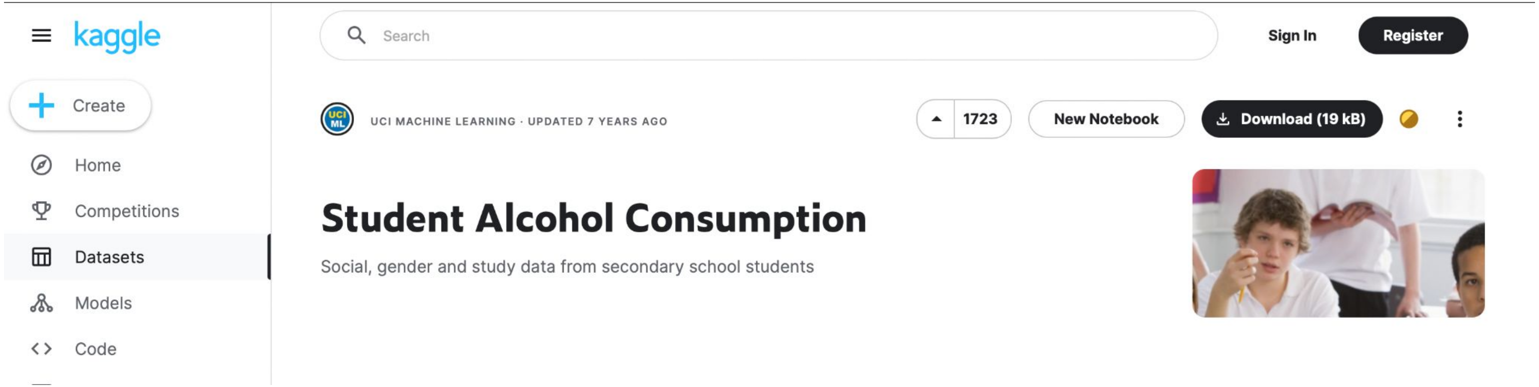
SC1015 MINI PROJECT

A139 Team 5
Ev Chua Yee Suan (U2222559E)
Fan Jin Min (U2220358C)
Boonkitrungpaisarn Potala (U2222312E)



Student Alcohol Consumption

Social, gender and study data from secondary school students

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

Problem Definition

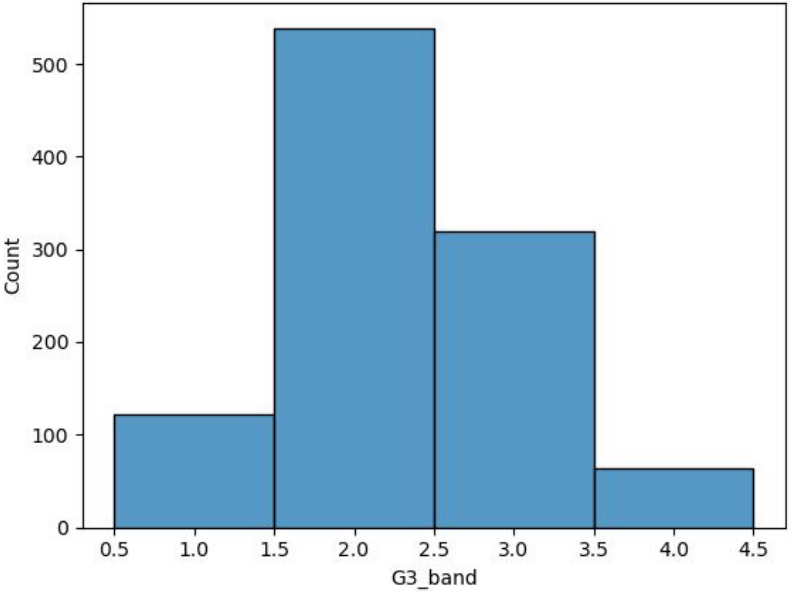
Factors that could affect a student's grade and predicting a student's grades

Exploratory Data Analysis

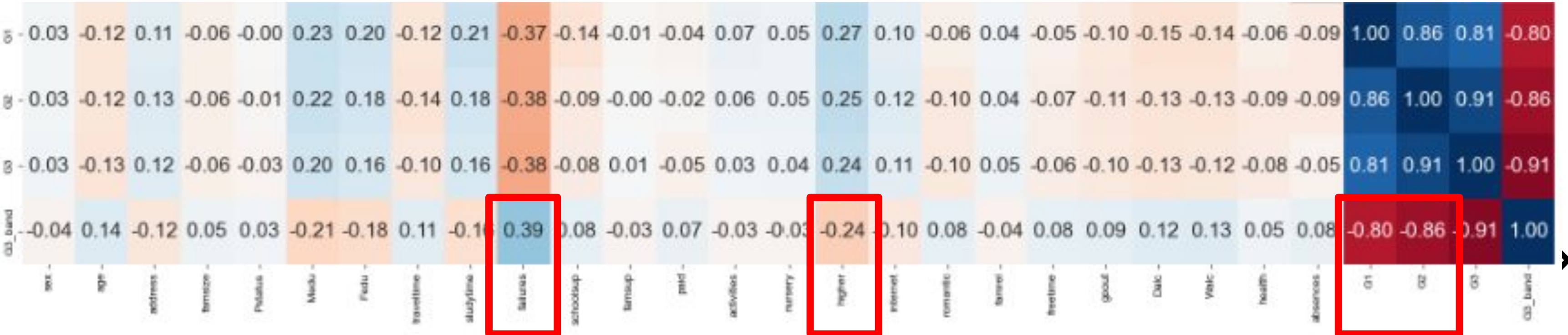
Grade Bandings

G3_band

- Band 1 (best): G3 > 15
- Band 2: 10 < G3 <= 15
- Band 3: 5 < G3 <= 10
- Band 4: G3 <= 5



Grade Bandings



	sex	age	address	famsize	Pstatus	Medu	Fedu	traveltime	studytime	failures	schoolsup	famsup	paid	activities	nursery	higher	internet	romantic	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3	G3_band
G1	0.03	-0.12	0.11	-0.06	-0.00	0.23	0.20	-0.12	0.21	-0.37	-0.14	-0.01	-0.04	0.07	0.05	0.27	0.10	-0.06	0.04	-0.05	-0.10	-0.15	-0.14	-0.06	-0.09	1.00	0.86	0.81	-0.80
G2	0.03	-0.12	0.13	-0.06	-0.01	0.22	0.18	-0.14	0.18	-0.38	-0.09	-0.00	-0.02	0.06	0.05	0.25	0.12	-0.10	0.04	-0.07	-0.11	-0.13	-0.13	-0.09	-0.09	0.86	1.00	0.91	-0.86
G3	0.03	-0.13	0.12	-0.06	-0.03	0.20	0.16	-0.10	0.16	-0.38	-0.08	0.01	-0.05	0.03	0.04	0.24	0.11	-0.10	0.05	-0.06	-0.10	-0.13	-0.12	-0.08	-0.05	0.81	0.91	1.00	-0.91
G3_band	-0.04	0.14	-0.12	0.05	0.03	-0.21	-0.18	0.11	-0.1	0.39	0.08	-0.03	0.07	-0.03	-0.0	-0.24	0.10	0.08	-0.04	0.08	0.09	0.12	0.13	0.05	0.08	-0.80	-0.86	-0.91	1.00

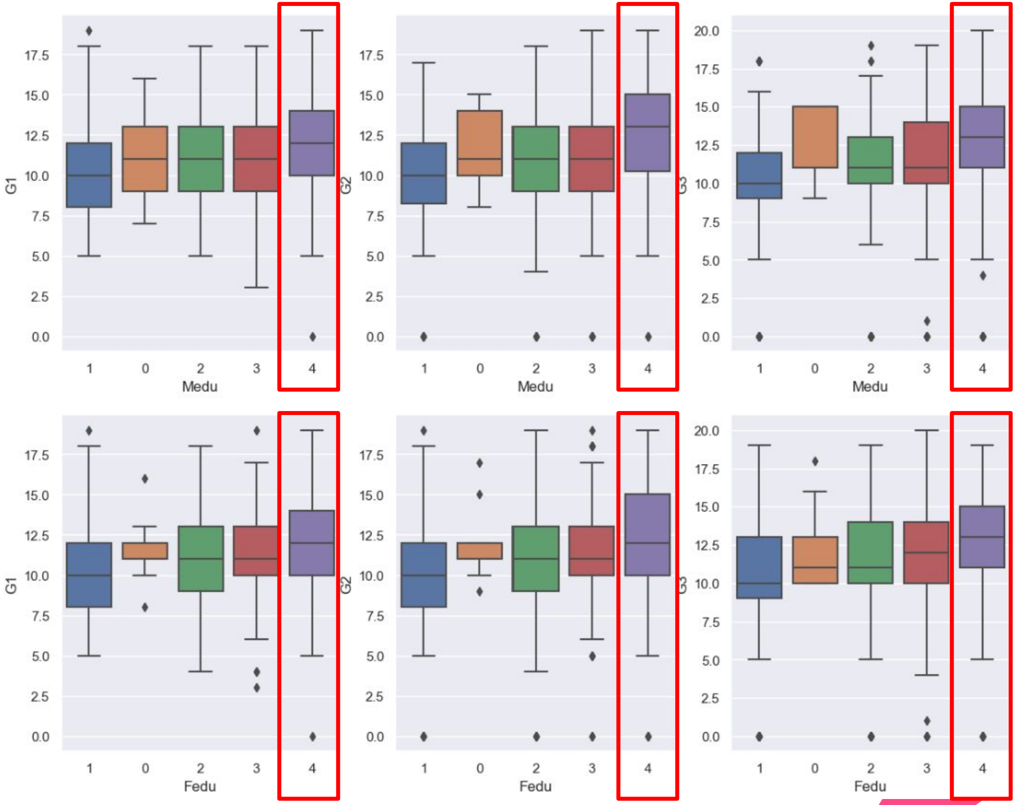
Failures

Higher

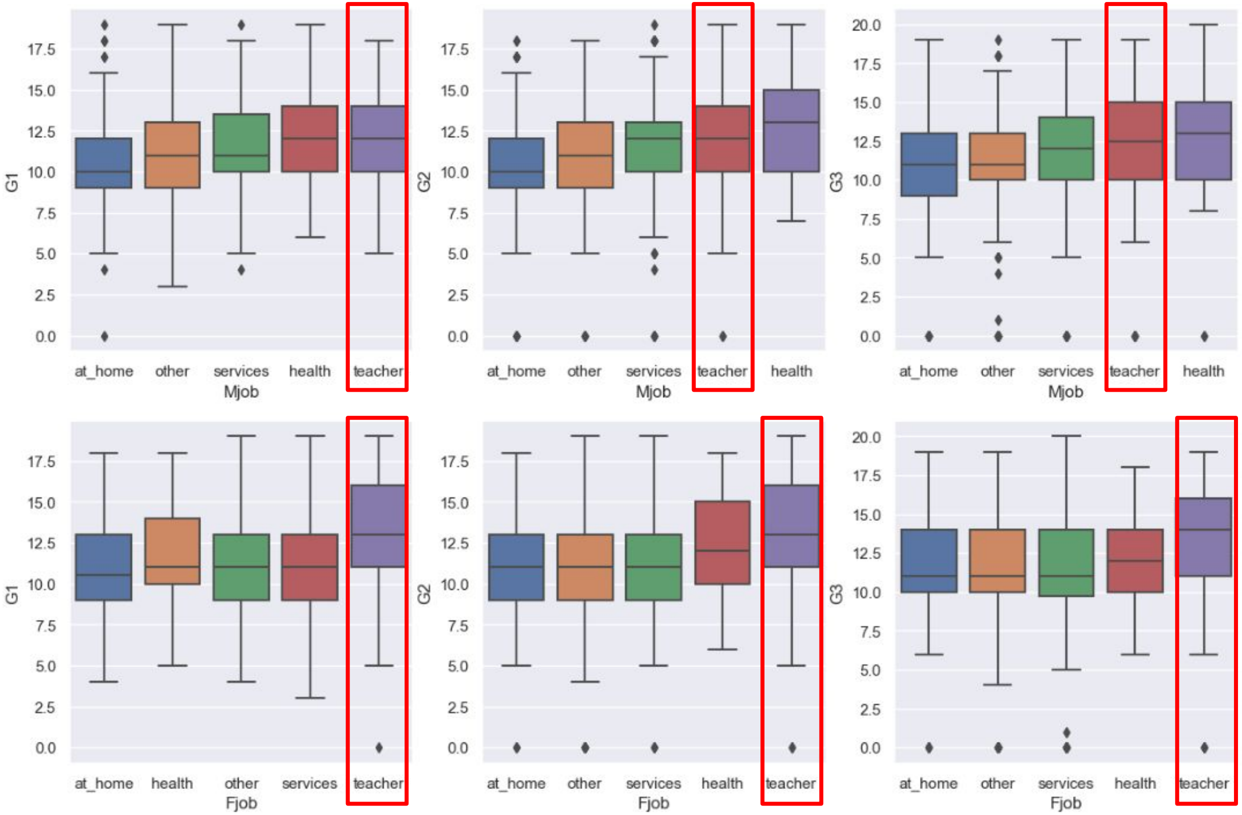
G1 and G2

Have relatively stronger correlations with G3_band

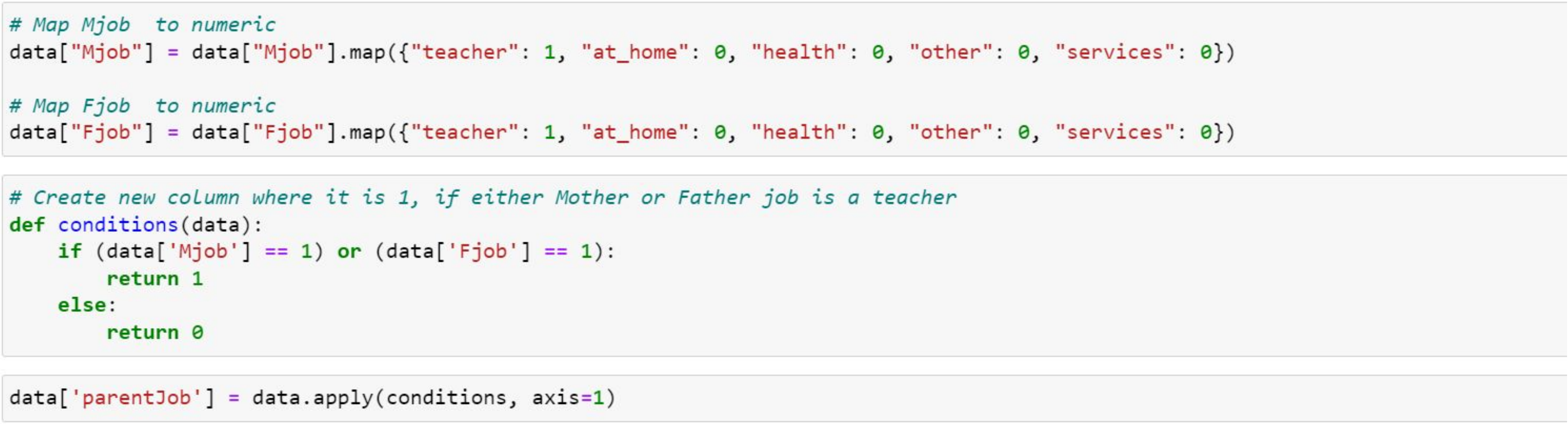
Parents’ Education



Parents’ Jobs



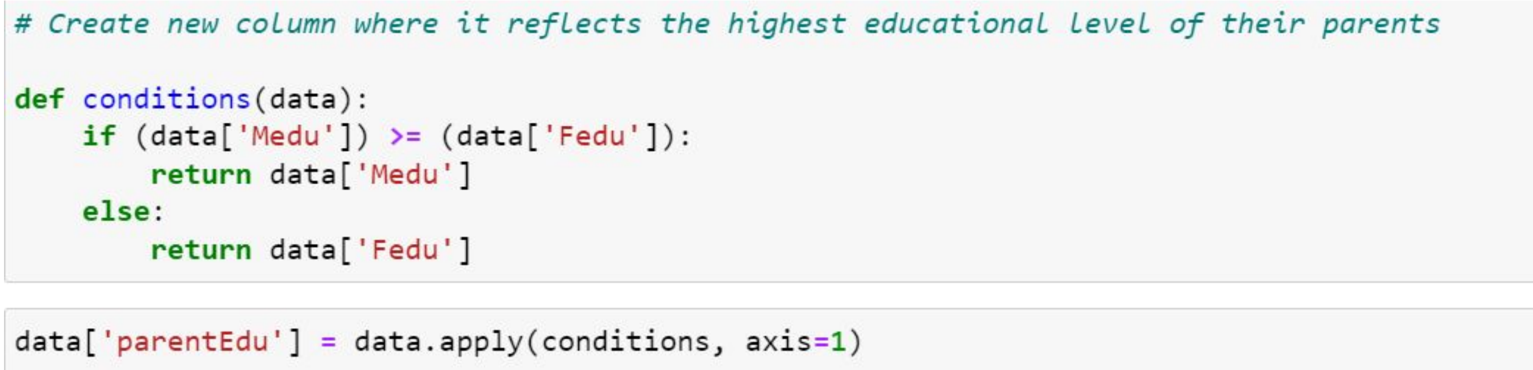
Data Preparation



```python
# Map Mjob  to numeric
data["Mjob"] = data["Mjob"].map({"teacher": 1, "at_home": 0, "health": 0, "other": 0, "services": 0})

# Map Fjob  to numeric
data["Fjob"] = data["Fjob"].map({"teacher": 1, "at_home": 0, "health": 0, "other": 0, "services": 0})
```

```python
# Create new column where it is 1, if either Mother or Father job is a teacher
def conditions(data):
    if (data['Mjob'] == 1) or (data['Fjob'] == 1):
        return 1
    else:
        return 0
```

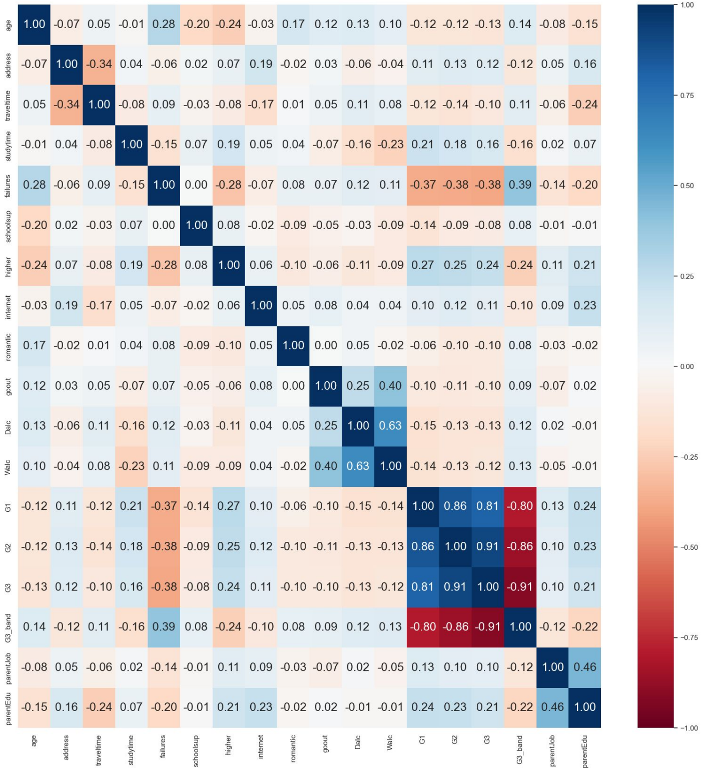
```python
data['parentJob'] = data.apply(conditions, axis=1)
```



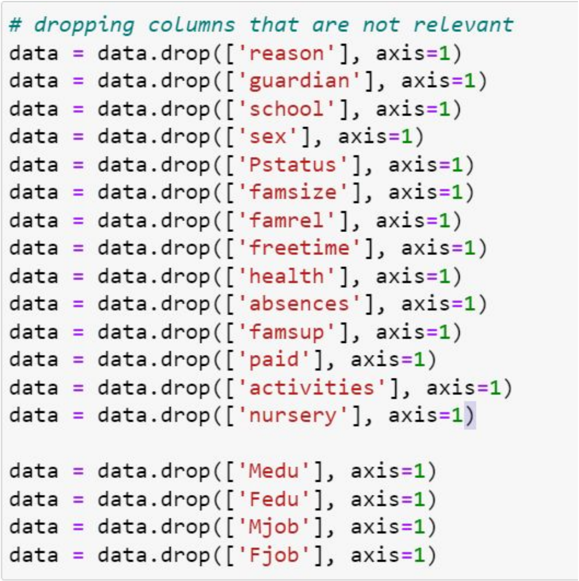
```
# Create new column where it reflects the highest educational level of their parents

def conditions(data):
    if (data['Medu']) >= (data['Fedu']):
        return data['Medu']
    else:
        return data['Fedu']
```

```
data['parentEdu'] = data.apply(conditions, axis=1)
```

Removing variables with weak correlations



	age	address	traveltime	studytime	failures	schoolsup	higher	internet	romantic	goout	Dalc	Walc	G1	G2	G3	G3_band	parentJob	parentEdu
age	1.00	-0.07	0.05	-0.01	0.28	-0.20	-0.24	-0.03	0.17	0.12	0.13	0.10	-0.12	-0.12	-0.13	0.14	-0.08	-0.15
address	-0.07	1.00	-0.34	0.04	-0.06	0.02	0.07	0.19	-0.02	0.03	-0.06	-0.04	0.11	0.13	0.12	-0.12	0.05	0.16
traveltime	0.05	-0.34	1.00	-0.08	0.09	-0.03	-0.08	-0.17	0.01	0.05	0.11	0.08	-0.12	-0.14	-0.10	0.11	-0.06	-0.24
studytime	-0.01	0.04	-0.08	1.00	-0.15	0.07	0.19	0.05	0.04	-0.07	-0.16	-0.23	0.21	0.18	0.16	-0.16	0.02	0.07
failures	0.28	-0.06	0.09	-0.15	1.00	0.00	-0.28	-0.07	0.08	0.07	0.12	0.11	-0.37	-0.38	-0.38	0.39	-0.14	-0.20
schoolsup	-0.20	0.02	-0.03	0.07	0.00	1.00	0.08	-0.02	-0.09	-0.05	-0.03	-0.09	-0.14	-0.09	-0.08	0.08	-0.01	-0.01
higher	-0.24	0.07	-0.08	0.19	-0.28	0.08	1.00	0.06	-0.10	-0.06	-0.11	-0.09	0.27	0.25	0.24	-0.24	0.11	0.21
internet	-0.03	0.19	-0.17	0.05	-0.07	-0.02	0.06	1.00	0.05	0.08	0.04	0.04	0.10	0.12	0.11	-0.10	0.09	0.23
romantic	0.17	-0.02	0.01	0.04	0.08	-0.09	-0.10	0.05	1.00	0.00	0.05	-0.02	-0.06	-0.10	-0.10	0.08	-0.03	-0.02
goout	0.12	0.03	0.05	-0.07	0.07	-0.05	-0.06	0.08	0.00	1.00	0.25	0.40	-0.10	-0.11	-0.10	0.09	-0.07	0.02
Dalc	0.13	-0.06	0.11	-0.16	0.12	-0.03	-0.11	0.04	0.05	0.25	1.00	0.63	-0.15	-0.13	-0.13	0.12	0.02	-0.01
Walc	0.10	-0.04	0.08	-0.23	0.11	-0.09	-0.09	0.04	-0.02	0.40	0.63	1.00	-0.14	-0.13	-0.12	0.13	-0.05	-0.01
G1	-0.12	0.11	-0.12	0.21	-0.37	-0.14	0.27	0.10	-0.06	-0.10	-0.15	-0.14	1.00	0.86	0.81	-0.80	0.13	0.24
G2	-0.12	0.13	-0.14	0.18	-0.38	-0.09	0.25	0.12	-0.10	-0.11	-0.13	-0.13	0.86	1.00	0.91	-0.86	0.10	0.23
G3	-0.13	0.12	-0.10	0.16	-0.38	-0.08	0.24	0.11	-0.10	-0.10	-0.13	-0.12	0.81	0.91	1.00	-0.91	0.10	0.21
G3_band	0.14	-0.12	0.11	-0.16	0.39	0.08	-0.24	-0.10	0.08	0.09	0.12	0.13	-0.80	-0.86	-0.91	1.00	-0.12	-0.22
parentJob	-0.08	0.05	-0.06	0.02	-0.14	-0.01	0.11	0.09	-0.03	-0.07	0.02	-0.05	0.13	0.10	0.10	-0.12	1.00	0.46
parentEdu	-0.15	0.16	-0.24	0.07	-0.20	-0.01	0.21	0.23	-0.02	0.02	-0.01	-0.01	0.24	0.23	0.21	-0.22	0.46	1.00



```
# dropping columns that are not relevant
data = data.drop(['reason'], axis=1)
data = data.drop(['guardian'], axis=1)
data = data.drop(['school'], axis=1)
data = data.drop(['sex'], axis=1)
data = data.drop(['Pstatus'], axis=1)
data = data.drop(['famsize'], axis=1)
data = data.drop(['famrel'], axis=1)
data = data.drop(['freetime'], axis=1)
data = data.drop(['health'], axis=1)
data = data.drop(['absences'], axis=1)
data = data.drop(['famsup'], axis=1)
data = data.drop(['paid'], axis=1)
data = data.drop(['activities'], axis=1)
data = data.drop(['nursery'], axis=1)

data = data.drop(['Medu'], axis=1)
data = data.drop(['Fedu'], axis=1)
data = data.drop(['Mjob'], axis=1)
data = data.drop(['Fjob'], axis=1)
```

Core Analysis: Machine Learning

Machine Learning Techniques

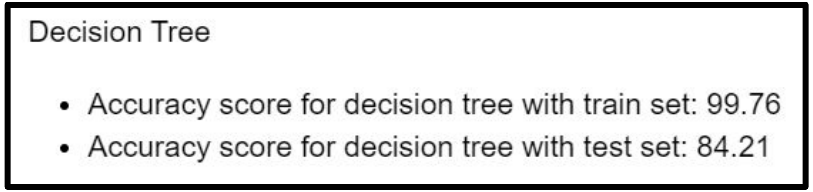
Decision Tree

Support Vector Machines

k-Nearest Neighbours

Decision Tree

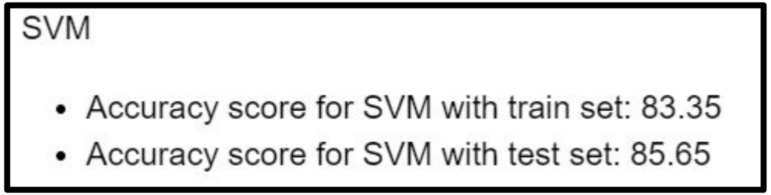
- Issue of overfitting due to huge difference in accuracy scores between train and test set



Decision Tree

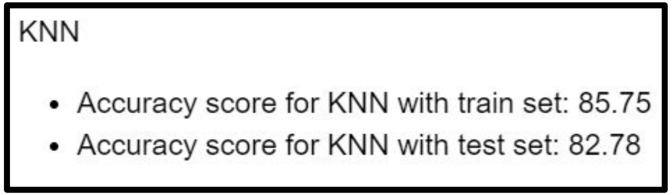
- Accuracy score for decision tree with train set: 99.76
- Accuracy score for decision tree with test set: 84.21

Support Vector Machine & k-Nearest Neighbours



SVM

- Accuracy score for SVM with train set: 83.35
- Accuracy score for SVM with test set: 85.65

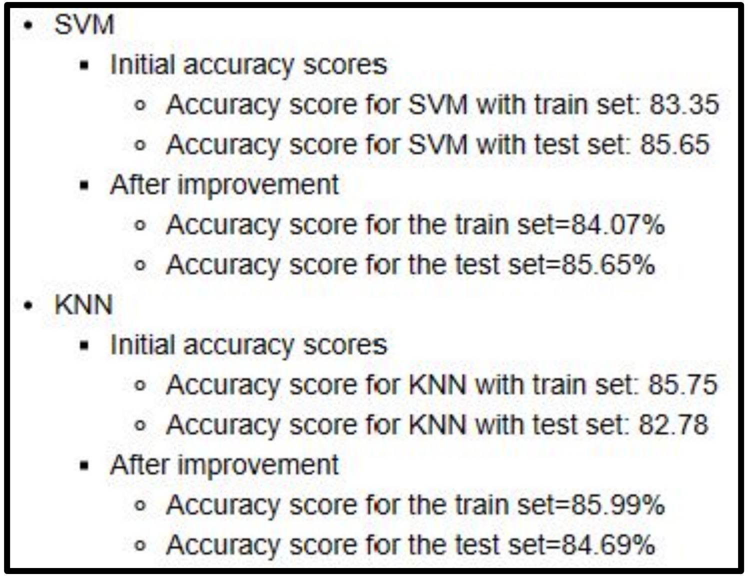


KNN

- Accuracy score for KNN with train set: 85.75
- Accuracy score for KNN with test set: 82.78

Improving Accuracy

- Using only columns: failures, higher, G1, G2, parentJob and parentEdu



- SVM
  - Initial accuracy scores
    - Accuracy score for SVM with train set: 83.35
    - Accuracy score for SVM with test set: 85.65
  - After improvement
    - Accuracy score for the train set=84.07%
    - Accuracy score for the test set=85.65%
- KNN
  - Initial accuracy scores
    - Accuracy score for KNN with train set: 85.75
    - Accuracy score for KNN with test set: 82.78
  - After improvement
    - Accuracy score for the train set=85.99%
    - Accuracy score for the test set=84.69%

Conclusion & Data-Driven Insights

DATA-DRIVEN INSIGHTS

Failures

Previous fail rates could negatively impact grades

Higher

Interest in higher education may make one more motivated to score better

G1, G2

Past results may mean one is likely to score similarly in latest tests (G3)

Parents' Jobs

When a parent is a teacher, it could impact student's grade positively

Parent's Education

Students with parents with higher education level likely to score better

THANK YOU