Sip to Success: A Machine Learning Approach to Forecasting Students' Final Grade Based on Alcohol Consumption Patterns

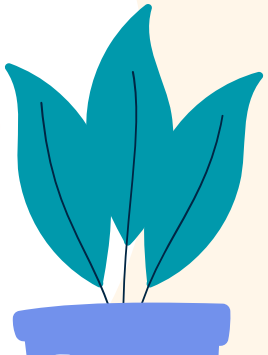
Team 7

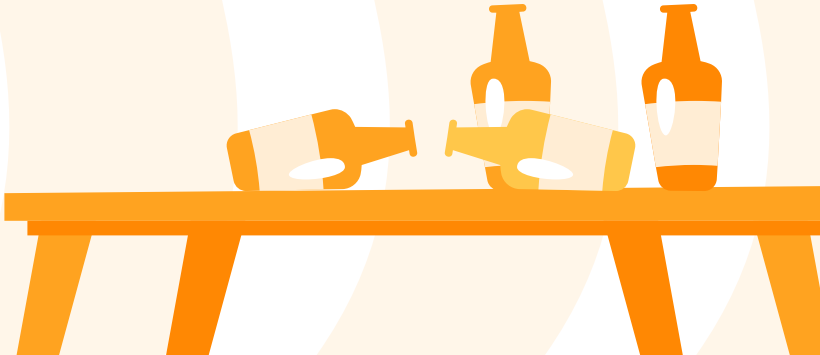
Jose Sanchez Jr, Alexis Flores, Nayeli Ramirez

About the Project

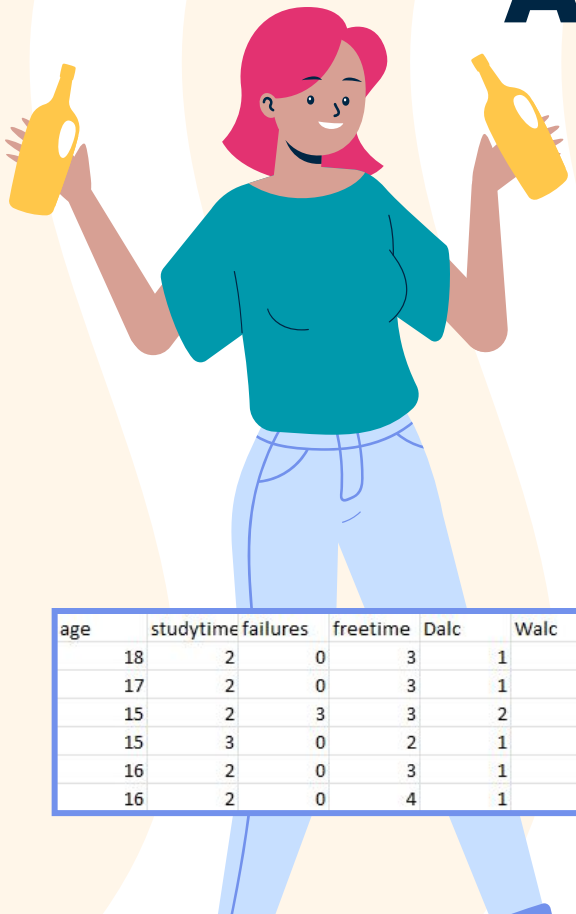
Scope: Predict the final grade a student will receive depending on their alcohol consumption. This will shed light on the importance of healthy consumption of alcohol. We will be able to conclude how much alcohol affects us even in our student life.

Having to just know or find information about how drinking affects students academically compare to some that don't or rarely drink was a very interesting research and result.
The data is all coming from a math course not all the subject together.





About the Data



Obtained from kaggle.

The data was obtained in a survey of students math courses in secondary school found in Portugal.

It contains a lot of interesting social, gender and study information about students.

Our data preprocessing phase involved discerning and retaining only the features integral to the model's learning process.

age	studytime	failures	freetime	Dalc	Walc	Finalgrade	sex_F	sex_M	schoolsup	schoolsup	famsup_n	famsup_y	paid_no	paid_yes	higher_nc	higher_ye	internet_	internet_yes
18	2	0	3	1	1	0	1	0	0	1	1	0	1	0	0	1	1	0
17	2	0	3	1	1	0	1	0	1	0	0	1	1	0	0	1	0	1
15	2	3	3	2	3	0	1	0	0	1	1	0	0	1	0	1	0	1
15	3	0	2	1	1	1	1	0	1	0	0	1	0	1	0	1	0	1
16	2	0	3	1	2	0	1	0	1	0	0	1	0	1	0	1	1	0
16	2	0	4	1	2	1	0	1	1	0	0	1	0	1	0	1	0	1

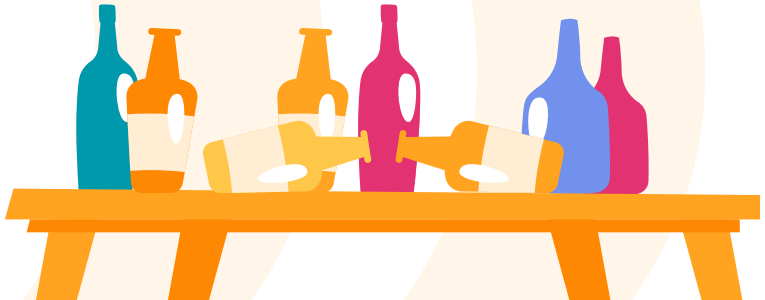
Features

sex - student's sex (binary: 'F' - female or 'M' - male)
age - student's age (numeric: from 15 to 22)
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) (binary: yes or no)
higher - wants to take higher education (binary: yes or no)
internet - Internet access at home (binary: yes or no)
freetime - free time after school (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)

	A	B	C	D	E	F	G	H	I	J	K	L
1	sex	age	studytime	failures	schoolsup	famsup	paid	higher	internet	freetime	Dalc	Walc
2	F	18	2	0	yes	no	no	yes	no	3	1	1
3	F	17	2	0	no	yes	no	yes	yes	3	1	1
4	F	15	2	3	yes	no	yes	yes	yes	3	2	3
5	F	15	3	0	no	yes	yes	yes	yes	2	1	1
6	F	16	2	0	no	yes	yes	yes	no	3	1	2
7	M	16	2	0	no	yes	yes	yes	yes	4	1	2
8	M	16	2	0	no	no	no	yes	yes	4	1	1
9	F	17	2	0	yes	yes	no	yes	no	1	1	1
10	M	15	2	0	no	yes	yes	yes	yes	2	1	1

Portugal’s Grading Criteria

Grade	Qualification
20 - 17.5	Excellent
17.4 - 15.5	Very Good
15.4 - 13.5	Good
13.4 - 9.5	Sufficient
9.4 - 3.5	Weak
3.4 - 0	Poor



Project Timeline

Found Data

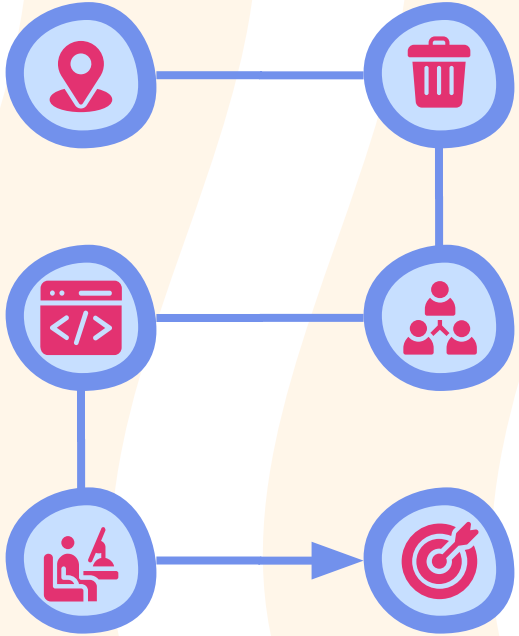
Got dataset from kaggle

Implemented

Finally putting the models to work

Tested Models

Improving on the implementation for better results



Cleaned Data

Took out any features that didn't seem redeeming

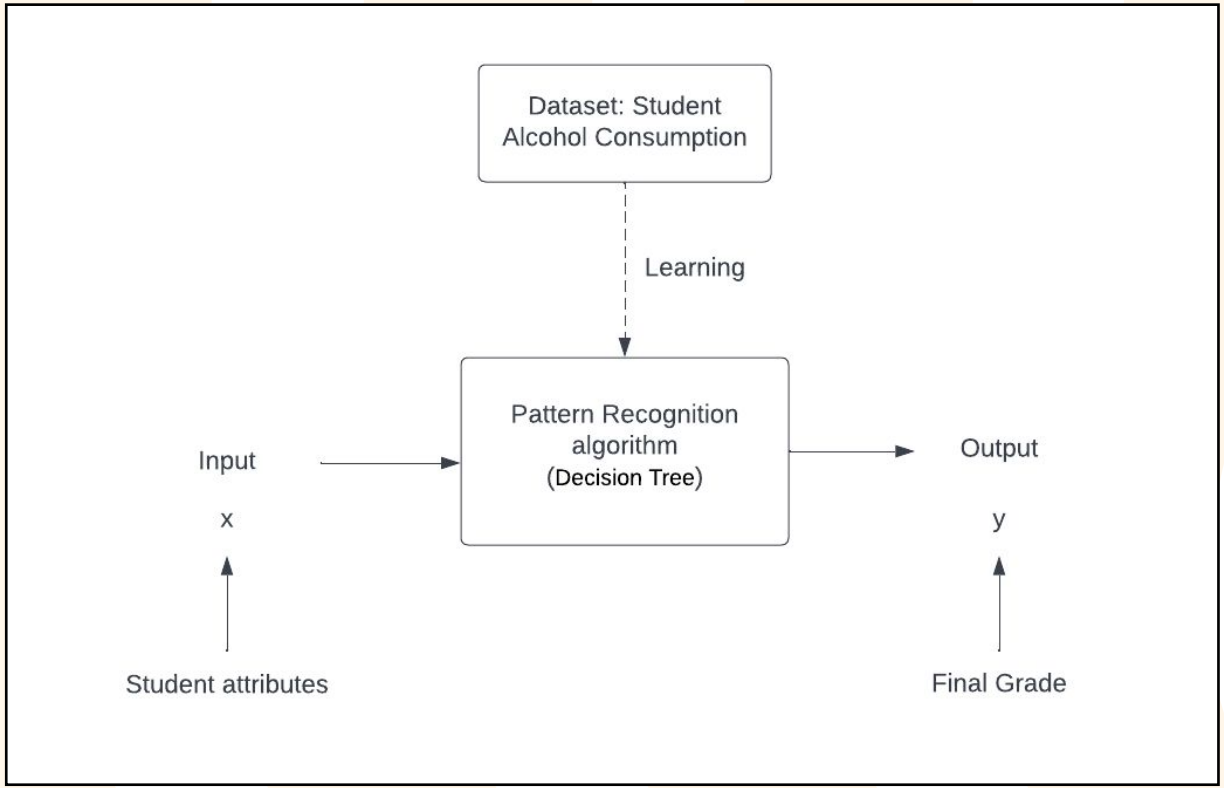
Determined models to use

Decision Tree and kNN

Results

Write a satisfied result

Decision Tree



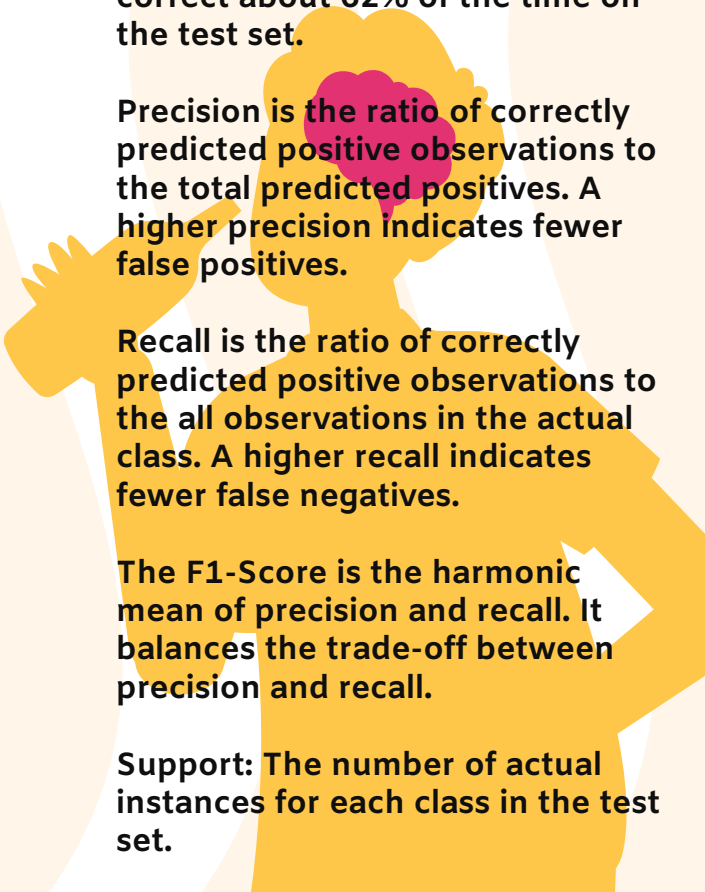
Decision Tree

Our initial thought was that a decision tree would be a great fit for predicting final grades based on alcohol consumption because it can understand complex connections in the data, making it useful for this problem. Decision trees are easy to explain, handling different types of information well, like numbers and categories.

```
Accuracy: 0.6075949367088608
Classification Report:
              precision    recall  f1-score   support

           0       0.70      0.74      0.72        53
           1       0.39      0.35      0.37        26

    accuracy                           0.61        79
   macro avg       0.54      0.54      0.54        79
weighted avg       0.60      0.61      0.60        79
```



Accuracy is the ratio of correctly predicted instances to the total instances. In this case, the model is correct about 62% of the time on the test set.

Precision is the ratio of correctly predicted positive observations to the total predicted positives. A higher precision indicates fewer false positives.

Recall is the ratio of correctly predicted positive observations to the all observations in the actual class. A higher recall indicates fewer false negatives.

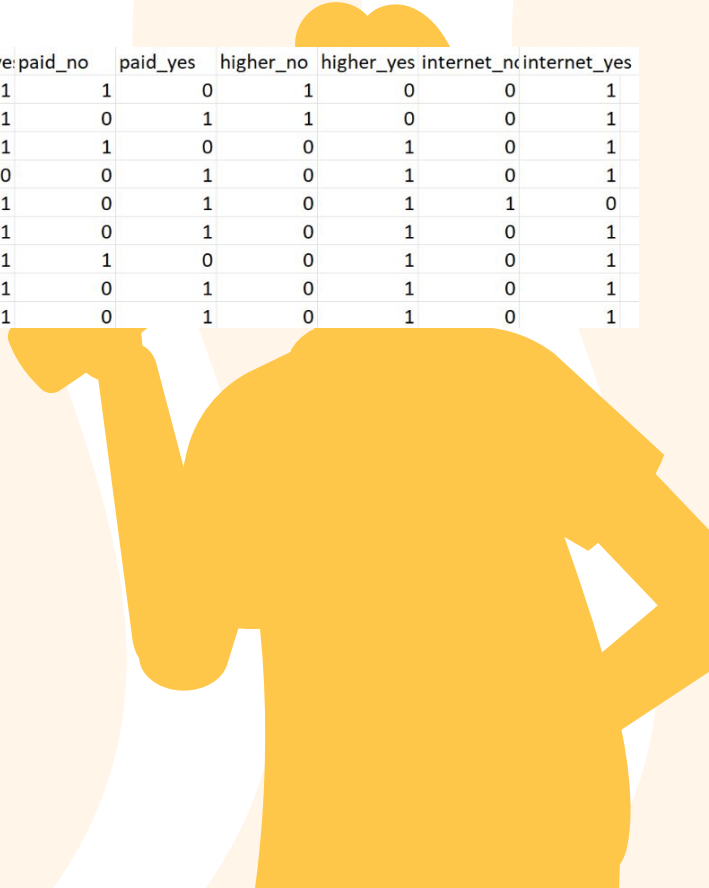
The F1-Score is the harmonic mean of precision and recall. It balances the trade-off between precision and recall.

Support: The number of actual instances for each class in the test set.

Results

age	studytime	failures	freetime	Dalc	Walc	sex_F	sex_M	schoolsup_	schoolsup_	famsup_no	famsup_ye	paid_no	paid_yes	higher_no	higher_yes	internet_n	internet_yes
17	1	3	5	1	1	0	1	0	1	0	1	1	0	1	0	0	1
18	1	0	3	2	3	0	1	1	0	0	1	0	1	1	0	0	1
18	2	1	3	1	3	0	1	1	0	0	1	1	0	0	1	0	1
16	2	0	3	1	1	1	0	1	0	1	0	0	1	0	1	0	1
20	2	2	5	4	5	0	1	1	0	0	1	0	1	0	1	1	0
18	2	0	3	5	5	0	1	1	0	0	1	0	1	0	1	0	1
15	2	0	3	1	1	0	1	1	0	0	1	1	0	0	1	0	1
16	2	0	4	1	1	0	1	1	0	0	1	0	1	0	1	0	1
18	2	0	3	1	2	1	0	1	0	0	1	0	1	0	1	0	1

Finalgrade
0
0
0
0
0
0
1
0
0
1



k-Nearest Neighbor

A k-Nearest Neighbor model was a viable option for predicting a student's final grade based on alcohol consumption because it can classify data points based on their neighbors' classification.

```
Training set size: 197
Test set size: 198
---------  ------
Predicted  Actual
8          10
8          12
0          5
13         10
7          9
9          13
0          18
9          6
19         0
0          14
16         15
14         7
11         15
12         10
11         14
9          8
13         8
9          11
15         15
---------  ------
Accuracy: 15.15%
```

Comparison between the models

Decision Tree

Accuracy: 62%

k-NN

Accuracy: 15.15%

Similarities

Non-parametric

Differences

Supervised vs unsupervised learning algorithms

The decision tree model emerged as the better choice for predicting final grades based on alcohol consumption.

Improvements

We could have implemented another model like Logistic Regression and Support Vectors Machines.

We would also try to use the decision tree with different features to try to improve the accuracy.

More optimized and easier to understand data set.

Another model to compare our results to.

Sources

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

Thank you!