Assignment 2

Erin Sutton
s3707294@student.rmit.edu.au

Abdul Raheman Nasir Gazi
s3705448@student.rmit.edu.au

27/05/2019

Table of Contents:

Abstract 1

Introduction 1
Methodology 1
Results and Discussion 2

Data Exploration: Selected Columns 2
Data Exploration: Pairs of Attributes 3
Tree Decision 5

K Nearest Neighbors 6
Refences 8

Abstract

The aim of this report was to investigate whether a student's circumstances would affect their final score in portuguese. The data attributes where collected through a questionnaire that was taken in two schools in addition to school reports. Overall the results show that students circumstances have a significant impact on their final grade. The report concludes that while other other factors do contribute to a student's final grade their circumstances have a substantial effect.

Introduction

There has been a lot of discussion in the media and wider community over the past few years about what affect a student's circumstances have on their education and what needs to happen to "even the playing field". This report investigates whether a student's circumstances can be used to predict their final score and thus if it truly plays a considerable role in a student's success in their education.

Methodology

This study was conducted via questionnaires and school reports, and covered a range demographic and social features as well as student grades, they contained only closed questions. It investigated aspects of a student's life in reference to their grades in Portuguese. A total of 649

students answered the questionnaires from two secondary schools in the Alentejo region of Portugal during the 2005-2006 school year.

Results and Discussion

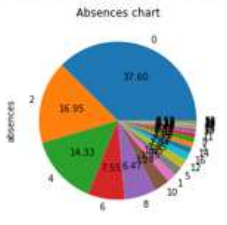
Data Exploration: Selected Columns:

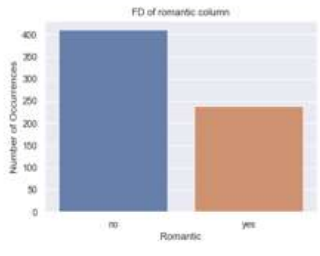
```
count    649.000000
mean       1.568567
std        0.748660
min        1.000000
25%        1.000000
50%        1.000000
75%        2.000000
max        4.000000
Name: traveltime, dtype: float64
```

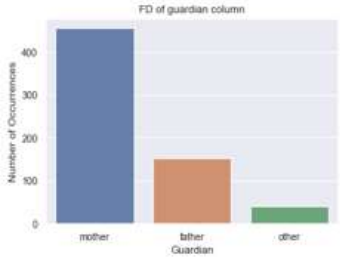
```
count    649.000000
mean      11.906009
std        3.230656
min        0.000000
25%       10.000000
50%       12.000000
75%       14.000000
max       19.000000
Name: finalgrade, dtype: float64
```

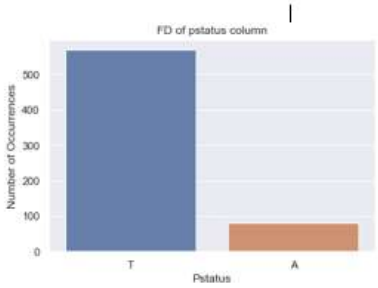
```
count    649.000000
mean       2.280431
std        1.284380
min        1.000000
25%        1.000000
50%        2.000000
75%        3.000000
max        5.000000
Name: walc, dtype: float64
```

```
count    649.000000
mean       3.930663
std        0.955717
min        1.000000
25%        4.000000
50%        4.000000
75%        5.000000
max        5.000000
Name: famrel, dtype: float64
```










```
U    452
R    197
Name: address, dtype: int64
```


```
U    452
R    197
Name: address, dtype: int64
```


traveltme:- Descriptive statistics was done for traveltime to find the average number of hours one needs to travel.

finalgrade:- Descriptive statistics was done for finalgrade to find the average of grades and also to fid the minimum and maximum marks.

heath:- Health chart shows a good number of students with adequate health.

absences:- Due to large number of students being fit and healthy, it was found that there was a big number of students who didn't skip even one class.

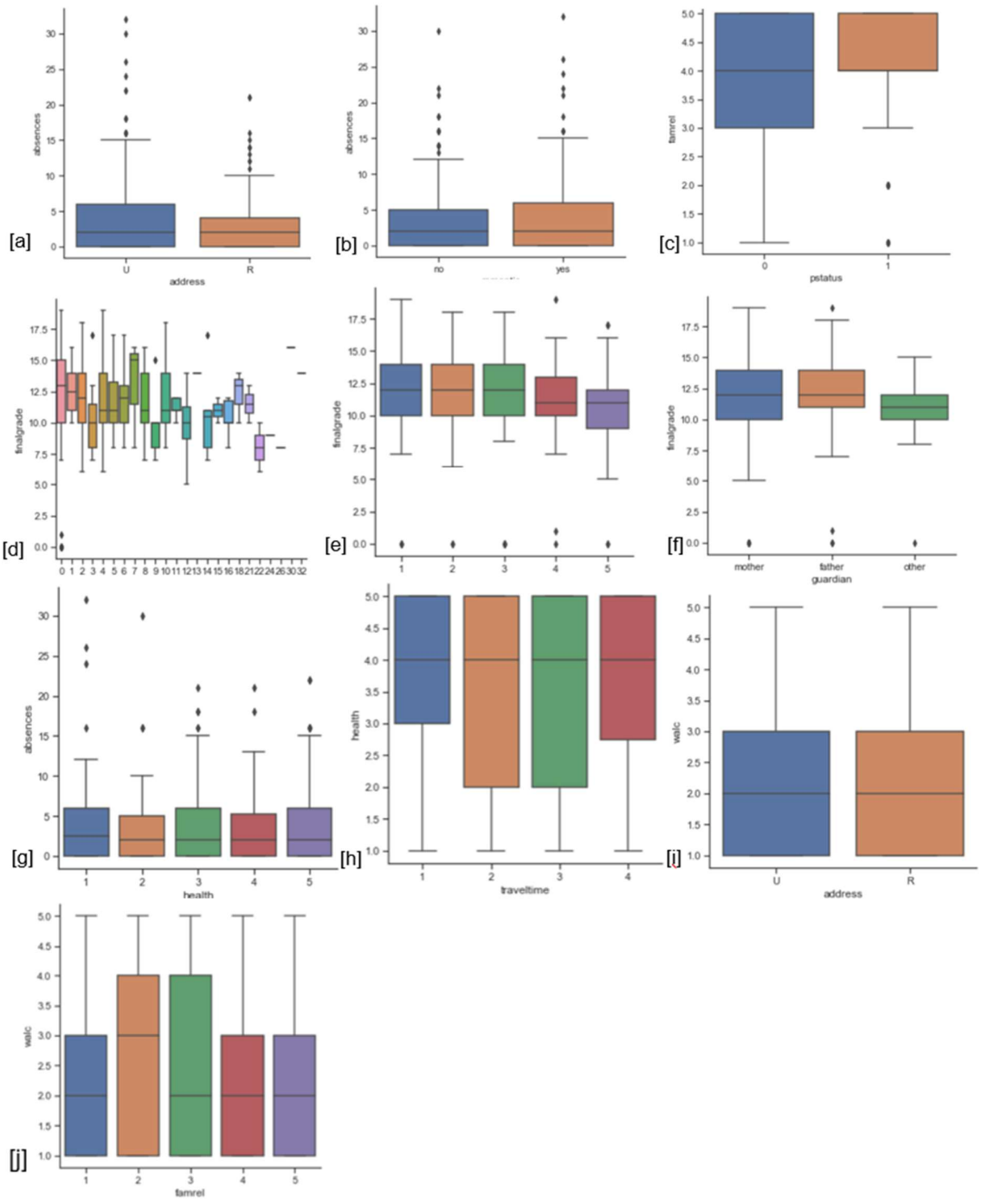
romantic:- Large number of students answered 'no' when they were asked if they were romantically involved with someone.

address:- It was found that most of the students came from urban areas.

guardian:- Big number of students replied with ‘mother’ when asked about their guardian situation. A few of them also replied with other.

(For walc, famrel, and pstatus nothing explanatory was found which needed to be written down)

Data Exploration: Pairs of Attributes:



Address vs. Absences: *Students who live in regional areas are absent more often*

Graph [a] shows that there is a relationship between absences and address, the hypothesis is false as students in urban areas have a greater number of absences than those in rural areas.

Romantic vs. Absences: *Students who are in a romantic relationship have more absences*
Graph [b] shows that a relationship between romantic and absences as students who are in a romantic relationship do have more absences from school supporting a the hypothesis.

Parent status vs. Family Relationship: *Students whose parents are together will a better family relationship*
Graph [c] shows that a relationship between parent status and family relationships exist, such that student's whose parents are together tended to score their family relationship higher than whose parents are apart which supports the hypothesis.

Absences vs. Final grade: *Students with more absences will have a lower score*
Graph [d] shows that a relationship exists between a student's absences and final grade, such that the less absences a student has the higher their final score, supporting the hypothesis.

Weekend drinking vs. Final grade: *Students who drink more on the weekend score less*
Graph [e] shows a relationship between weekend drinking and a student's final grade exists, such as the lower a student scored their weekend drinking the higher their final grade, thus the hypothesis is proven.

Guardian vs. Final Grade: *Students whose guardian is 'other' score less*
Graph [f] shows a relationship between guardian and final grade, student's whose guardians are neither their mother or father score lower on their final grade, supporting the hypothesis.

Health vs. Absences: *Students with poor health have more absences*
Graph [g] shows that there is no relationship between a student's health and their absences. Thus, the hypothesis is false.

Travel Time vs. Health: *Students who have less travel time are more healthy*
Graph [h] shows that students whose travel time is more than 15 minutes and less than hour are less healthy than student's who must travel less than 15 minutes or greater than an hour. The hypothesis is thus false.

Address vs. Weekend drinking: *Students who live in urban areas drink more*
Graph [i] shows that there is no relationship between address and weekend drinking, thus the hypothesis is false.

Family relationship vs. Weekend drinking: *Students who score their family relationship lower drink more on weekends*
Graph [j] shows that there is no relationship between family relations and weekend drinking. The hypothesis is thus false.

Tree Decision:

Final Grade: **0** $= 0 \leq grade < 5$ **1** $= 5 \leq grade < 10$ **2** $= 10 \leq grade < 15$ **3** $= 15 \leq grade \leq 20$

---

	precision	recall	f1-score	support
0	0.00	0.00	0.00	11
1	0.19	0.18	0.18	45
2	0.61	0.59	0.60	207
3	0.22	0.26	0.24	62
micro avg	0.45	0.45	0.45	325
macro avg	0.25	0.26	0.25	325
weighted avg	0.45	0.45	0.45	325

```
[[  0   1   7   3]
 [  0   8  33   4]
 [  6  28 122  51]
 [  1   6  39  16]]

0.4492307692307692
```

Model: 50% training, 50% testing

In this model 50% of the data was used for training and the other 50% for testing, the classification report shows that ratio of correctly predicted positives against all predicted positives in very low for targets 1 and 3 (0.19 and 0.22 respectively) whilst target 2 has a good ratio (0.61). The same could be said for the ratio of correctly predicted positives against all observations, only target 3 ratio increases, targets 1 and 2 decrease. The classification error rate (approx. 0.449) is the second lowest of the three tree decision models however, it has the highest number of correctly predicted targets according to the confusion matrix. Notably, target 2 has a much higher number of correct predictions then the other 3 targets combined.

---

	precision	recall	f1-score	support
0	0.00	0.00	0.00	8
1	0.12	0.14	0.13	35
2	0.60	0.54	0.57	167
3	0.20	0.24	0.22	50
micro avg	0.42	0.42	0.42	260
macro avg	0.23	0.23	0.23	260
weighted avg	0.44	0.42	0.43	260

```
[[ 0  2  4  2]
 [ 1  5 26  3]
 [ 3 30 91 43]
 [ 1  6 31 12]]

0.4153846153846154
```

Model: 60% training, 40% testing

In this model 60% of the data was used for training and 40% for testing, the classification report shows that the ratio of correctly predicted positives against all predicted positives is even lower than the model above, the recall, confusion matrix and the classification error rate support this, the number of correctly predicted targets is less in each target. The classification error is the lowest of the model. Despite a greater percentage of data used training the model it produced a less accurate model. Most of the false positives were predicted to be target 2 which indicates that a student's circumstances alone can not predict their final grade.

---

	precision	recall	f1-score	support
0	0.00	0.00	0.00	1
1	0.13	0.29	0.18	14
2	0.74	0.54	0.62	92
3	0.28	0.35	0.31	23
micro avg	0.48	0.48	0.48	130
macro avg	0.29	0.29	0.28	130
weighted avg	0.58	0.48	0.52	130

```
[[ 0  1  0  0]
 [ 0  4  8  2]
 [ 1 22 50 19]
 [ 2  3 10  8]]
```

0.47692307692307695

Model: 80% training, 20% testing

In this model 80% of the data was used for training and the other 20%f for testing, the classification report shows that the ratio of correctly predicted positives against all predicted positives for target 2 is quite good however, the targets remain low. The ratio of correctly predicted positives against all observations remains constant with the previous two model expect for target 3 which is high. Thus, the f1-score for all targets are the best of the tree decision models. The classification error rate (approx. 0.4769) is also the highest of the models and the confusion matrix supports this.

Considering the precision, ratio, f1-score and classification error rate values along with the confusion matrix, the 50% testing and training model is the most accurate of the tree decision models. It should be noted that most student's final grade falls within the range of target 2, which may explain why the ratio's for target 2 has been consistently higher than the other targets. This may lead us to believe that the models are more accurate than they truly are.

K Nearest Neighbors:

Final Grade: $\mathbf{0} = 0 \leq grade < 5$ $\mathbf{1} = 5 \leq grade < 10$ $\mathbf{2} = 10 \leq grade < 15$ $\mathbf{3} = 15 \leq grade \leq 20$

	precision	recall	f1-score	support
0	0.00	0.00	0.00	11
1	0.17	0.04	0.07	45
2	0.64	0.83	0.72	207
3	0.21	0.15	0.17	62
micro avg	0.56	0.56	0.56	325
macro avg	0.25	0.25	0.24	325
weighted avg	0.47	0.56	0.50	325

```
[[  0   0   8   3]
 [  0   2  39   4]
 [  0   9 171  27]
 [  1   1  51   9]]
```

In this model 50% of the data was used for training and the other 50% for testing, the classification report shows that ratio of correctly predicted positives against all predicted positives in very low for targets 1 and 3 (0.17 and 0.21 respectively) whilst target 2 has a good ratio (0.64). The same could be said for the ratio of correctly predicted positives against all observations, only target 2 ratio increases, targets 1 and 3 decrease. The KNN score (approx. 0.56) is the second lowest of the three KNN

models however, it has the highest number of correctly predicted targets according to the confusion matrix. Notably, target 2 has a much higher number of correct predictions then the other 3 targets combined.

---

	precision	recall	f1-score	support
0	0.00	0.00	0.00	8
1	0.15	0.06	0.08	35
2	0.64	0.84	0.73	167
3	0.22	0.12	0.16	50
micro avg	0.57	0.57	0.57	260
macro avg	0.25	0.26	0.24	260
weighted avg	0.48	0.57	0.51	260

```
[[  0   0   6   2]
 [  0   2  30   3]
 [  0  10 141  16]
 [  1   1  42   6]]
```

In this model 60% of the data was used for training and 40% for testing, the classification report shows that the ratio of correctly predicted positives against target 2 predicted is higher than the model above, the recall, confusion matrix and the classification error rate support this, the number of correctly predicted targets is high in only target 2. Despite a greater percentage of data used training the model it produced a less accurate model. Most of the false positives were predicted to be target 2 which indicates that a student's circumstances alone cannot predict their final grade.

---

	precision	recall	f1-score	support
0	0.00	0.00	0.00	8
1	0.15	0.06	0.08	35
2	0.64	0.84	0.73	167
3	0.22	0.12	0.16	50
micro avg	0.57	0.57	0.57	260
macro avg	0.25	0.26	0.24	260
weighted avg	0.48	0.57	0.51	260

```
[[  0   0   6   2]
 [  0   2  30   3]
 [  0  10 141  16]
 [  1   1  42   6]]
```

In this model 80% of the data was used for training and the other 20%f for testing, the classification report shows that the ratio of correctly predicted positives against all predicted positives for target 2 is quite good however, the targets remain low. The ratio of correctly predicted positives against all observations remains constant with the above model but is high if we compare it to the very first model. Thus, the f1-score is similar to the above model which also means it is higher than the first model. The KNN score is 0.573 which is also quite similar to the model above.

After much consideration we have concluded that the best model is 80% training and 20% testing using the k Nearest Neighbours method. Based on the precision, recall and f1 – score in conjunction with the confusion matrix to give the most accurate model.

Conclusion

Our best model 80% training and 20% testing using the k Nearest Neighbours method shows that only class 2 of the final grade is predicted well, this indicates that we cannot predict a student's final grade from only their circumstances that other factors play a part in their final grade. Due to the high number of students that scored between 10 and 15 to is hard to say exactly how much of a role that such attributes contribute to the student's success and whether its importance is overstated in the wider community.

Refences

Navlani, A. (2019, May 20). *Decision Tree Classification in Python*. Retrieved from Data Camp: https://www.datacamp.com/community/tutorials/decision-tree-classification-python

Ren, D. Y. (2019, March). Practical Data Science: Introduction.

Ren, D. Y. (2019, April). Tutorial 7.

Tony_tiger. (2019, May 20). *Find our error rate using sklearn*. Retrieved from Stack Overflow: https://stackoverflow.com/questions/10318884/find-out-error-rate-using-sklearn

Waskom, M. (2019, May 20). *Categorical scatterplots*. Retrieved from seaborn: https://seaborn.pydata.org/tutorial/categorical.html

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.