



2019-5-21

PDS REPORT

STUDENT PERFORMANCE DATASET

CHEN QIUYU S3739209 HU ZIDI S3695699

RMIT UNIVERSITY
qiuyucc@yahoo.com




Table of content

Executive Summary	2
Introduction	2
Methodology	2
Result.....	3
Single columns	3
Column relationship	4
Data Modelling.....	8
Discussion	10
Conclusion.....	11
Reference.....	11

Executive Summary

The purpose of this report was to analyse different factors' impacts on student performances of final grade, such as attendance, parent's cohabitation status and so on. A survey of student achievement in secondary education was conducted by using school reports and questionnaires. Overall, the result indicates that there are many attributes can affect students' performance in secondary schools. Parents can take some considerations based on this research and make improvements, like preventing their child from romantic relationship. The report concludes that student's final grade can be predicted based on data attributes which contain strong correlation. It's recommended that parents should consider many factors while sending their children to school and student should have a positive attitude on studying for their career.

Introduction

Students' grade has been a big concern for every family as the education quality that student received has influences on their future. The property in school district has getting more and more popular and it sells at higher price than other areas. Live closer to school may reduce time spent on travelling, student is able to concentrate on study with less interruption. Meanwhile, the tuition fee of early childhood education is even more expensive than university. We can image that building up children originates from family, in terms of parent's education, especially for mother's education level. The story of Three Moves by Mencius's Mother tells us these concerns were real. This report will discuss different factors' impacts on student performances, and student's final grade can be predicted by these factors.

Methodology

The data used for this research provided taken from Machine Learning Repository, provided by Paulo Cortez from University of Minho. The data approach student achievement in secondary education of two Portuguese schools which was collected by survey and school reports. (Cortez, 2014) The data attributes contain student final scores, parents' cohabitation status, social life and forth on. The student applied data science knowledge to dig the factor affect student's grade. First of all, data preparation has been conducted to load and cleaning the data. To simplify data exploration processing, the grade was divided into three categories, such as distinction, pass and fail.

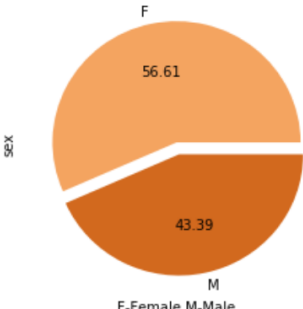
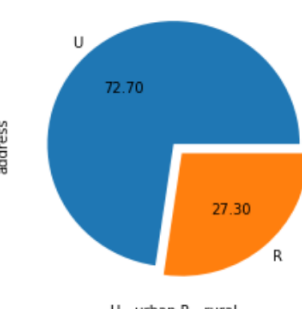
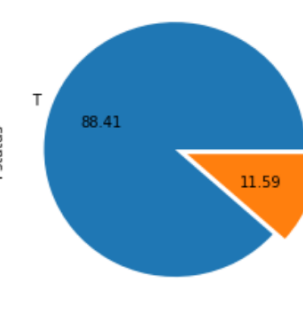
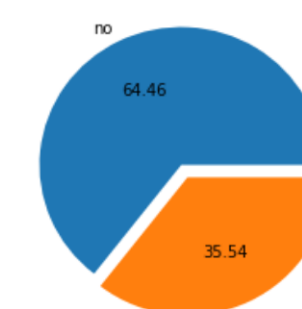
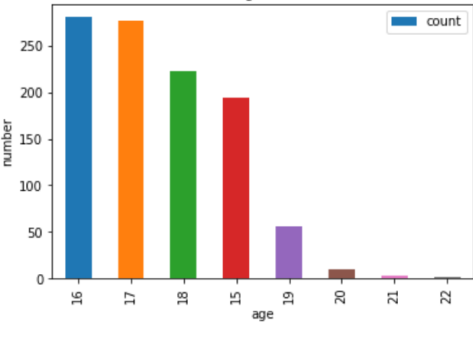
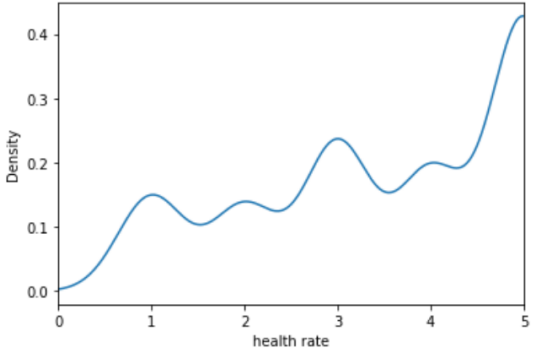
```
#convert final grade which is G3 to categorical: Distinction, pass and fail  
df['Grade'] = 'NaN'  
df.loc[(df.G3 >= 15) & (df.G3 <= 20), 'Grade'] = 'Distinction'  
df.loc[(df.G3 >= 10) & (df.G3 <= 14), 'Grade'] = 'Pass'  
df.loc[(df.G3 >= 0) & (df.G3 <= 9), 'Grade'] = 'Fail'
```

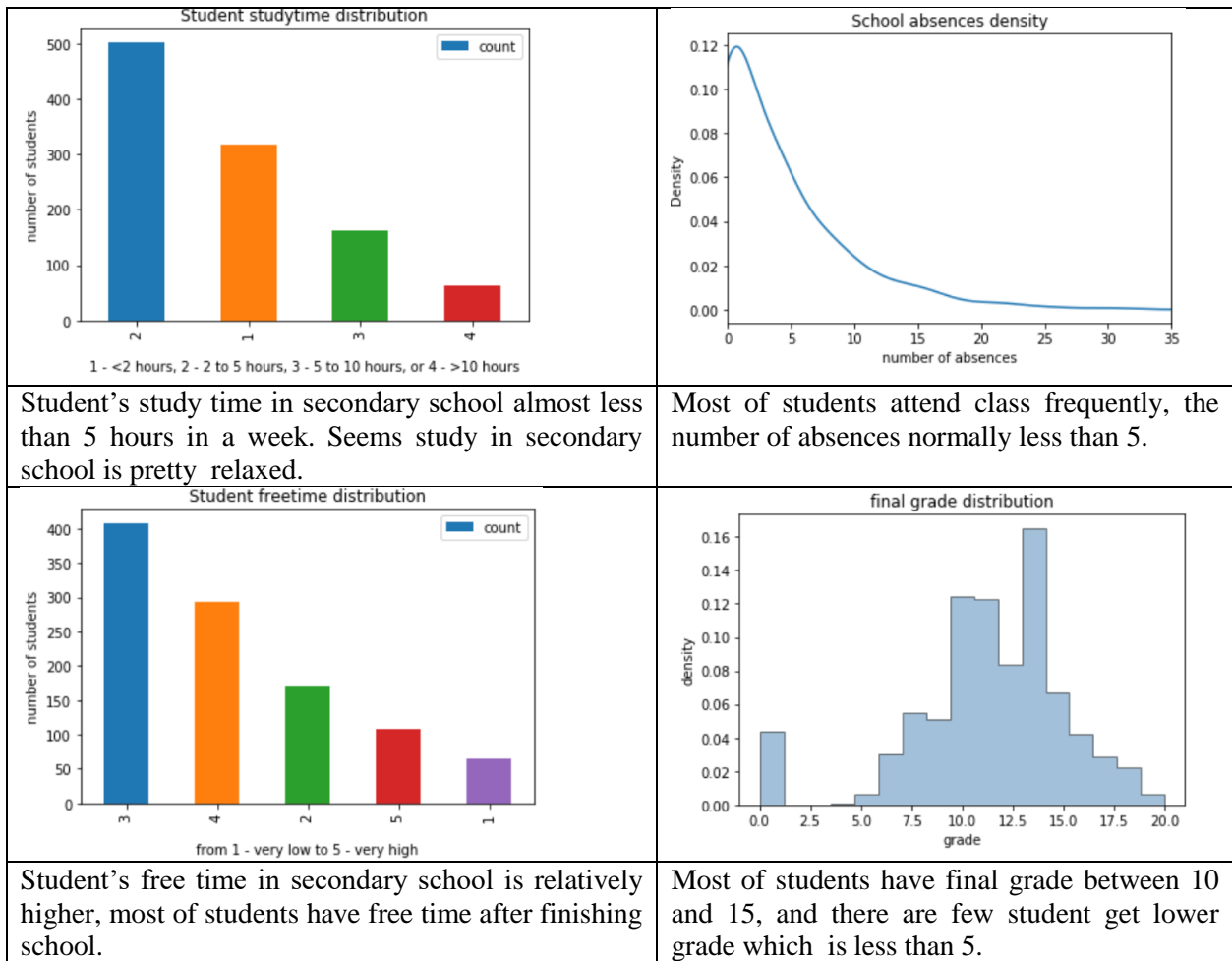
Secondly, data exploration was performed to visualized data by using appropriate descriptive statistics and graph. Like numerical attribute can be measured by distribution as the proportion of categorical attribute. Meanwhile, column of dataset will be paired and address a plausible hypothesis as a result. The relationship between other attribute with student grade, it tells that factors affect student's grade which can let student and his family pay more attentions.

Last but not least, student chooses classification to perform data modelling to predict student grade by other attributes of student. The two classification model used in the report are K-nearest neighbour algorithm and decision tree. KNN algorithm is one of the supervised algorithm that calculates the distance between new data points and all training data to distinguish its class. (robinson, 2018) Rather than KNN algorithm, decision tree identify the most significant attribute and break the dataset into smaller subsets. (Seif, 2018) Both of algorithm worked well on predict student grade in this case, the accuracy score compared in the report.

Result

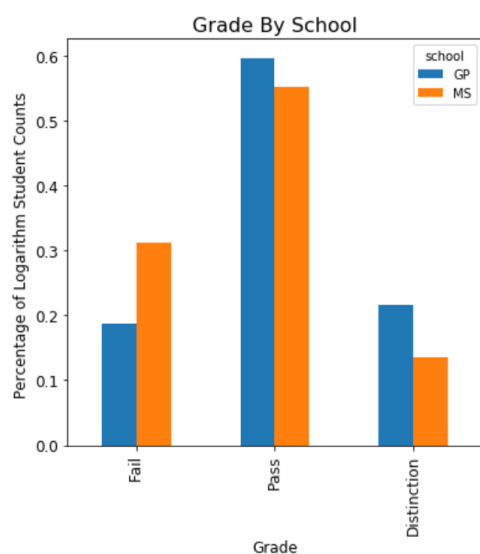
Single columns

<p>Gender Proportion</p>  <p>sex</p> <p>F-Female M-Male</p>	<p>Address Proportion</p>  <p>address</p> <p>U - urban R - rural</p>
<p>Female take 56.61% of population in this dataset, and male takes 43.39% of total students.</p>	<p>72.70% of students live in urban area and 27.30% of students live in rural area.</p>
<p>Parent conhabitation status Proportion</p>  <p>Pstatus</p> <p>T - together A - apart</p>	<p>Student in relationship Proportion</p>  <p>romantic</p> <p>yes</p>
<p>88.41% of student's parents live together, and 11.59% of student's parents live apart.</p>	<p>36.54% of students fall in relationship, and 64.46% of students are single.</p>
<p>Student Age distribution</p>  <p>number</p> <p>age</p>	<p>Student health density</p>  <p>Density</p> <p>health rate</p> <p>from 1 - very bad to 5 - very good</p>
<p>Most of students attend in secondary at age of 15 -18 years old.</p>	<p>Good news that most of students are relatively health in the dataset.</p>



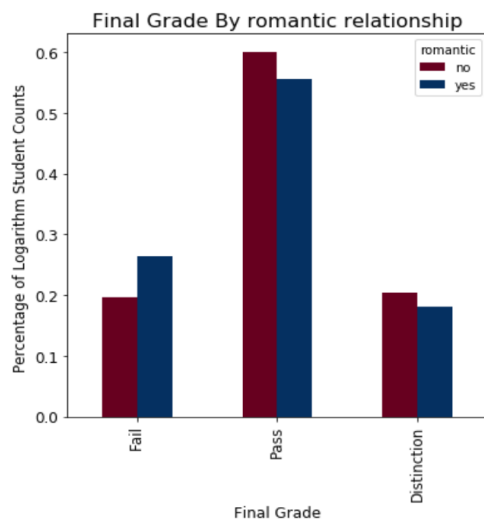
Column relationship

Assumption 1: Do you think the quality of school's education quality affect student's grade?



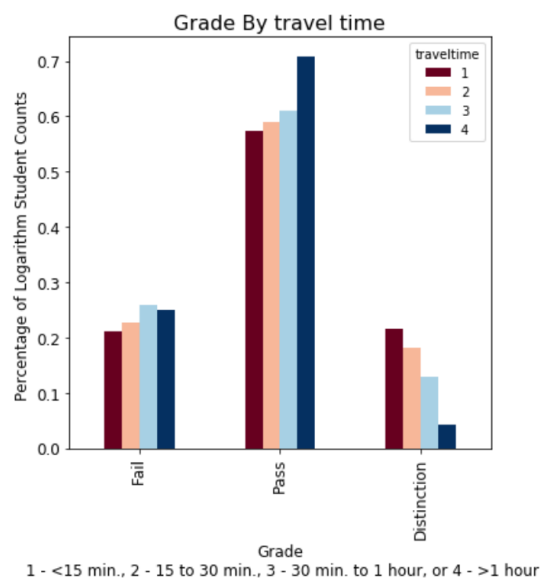
Take a look in this figure, grades of student have been categorized into fail, pass and distinction. School from GP (Gabriel Pereira) has lower failure density and relative higher final grade than MS (Mousinho da Silverira). It tells us the importance to choose school for parents.

Assumption 2: Would parents allow their children have romantic relationship in school?



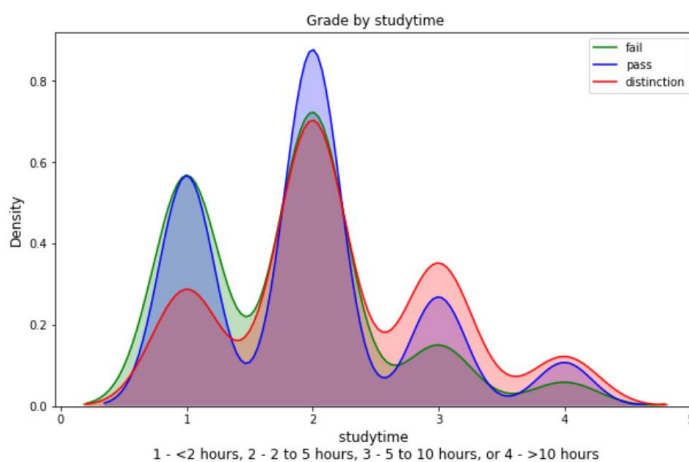
Take a look at this figure, students who have romantic relationship may not perform well on academic. The statistics show that having romantic relationship has significant influence on academic performance. As parents, having romantic relationship is prohibition for their children at young age.

Assumption 3: Should you choose to live in school districts to save travel time for children?



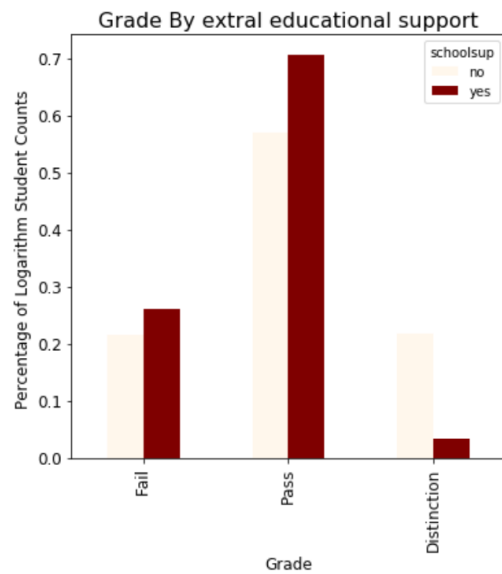
Take a look at this figure, there is no doubt that student who lives closer to school has higher academic performance. This factor may not lead student to fail, but student who lives closer to school has higher possibility to get distinction. I suggest that student can live near school.

Assumption 4: Should you restrict children to spend more time on studying?



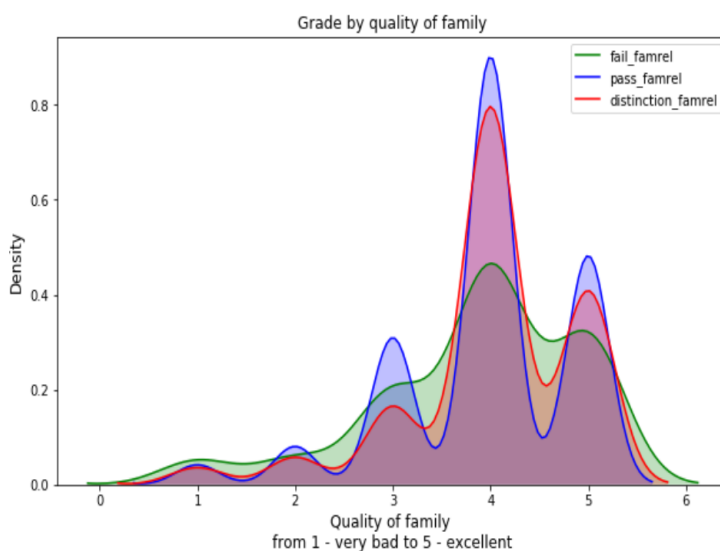
From this figure, some students spent more time on studying, but still get failed, and some students only have few study time, but they have higher grade. Learning efficiency is more important on academic performance, rather than time.

Assumption 5: Would extra educational support really helpful to improve grade?



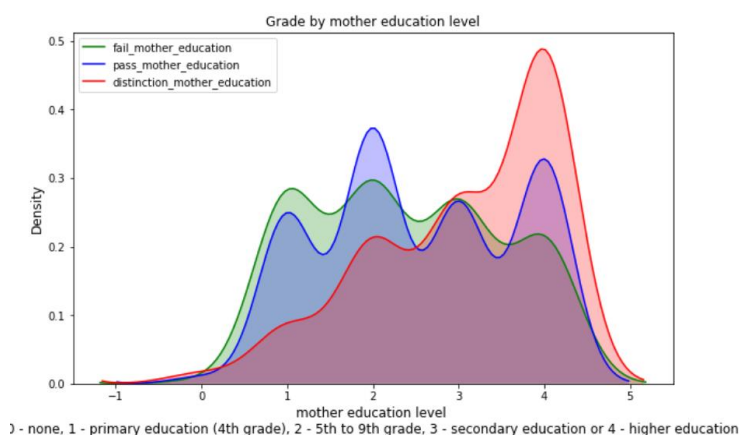
From the figure, it's not hard to tell that extra educational support don't help student to get distinction, self-learning seems more effective. Surprisingly, student who have extra educational support fails more than student who doesn't have. Extra educational support is unnecessary.

Assumption 6: Should parent maintain their relationship quality for student's grade?



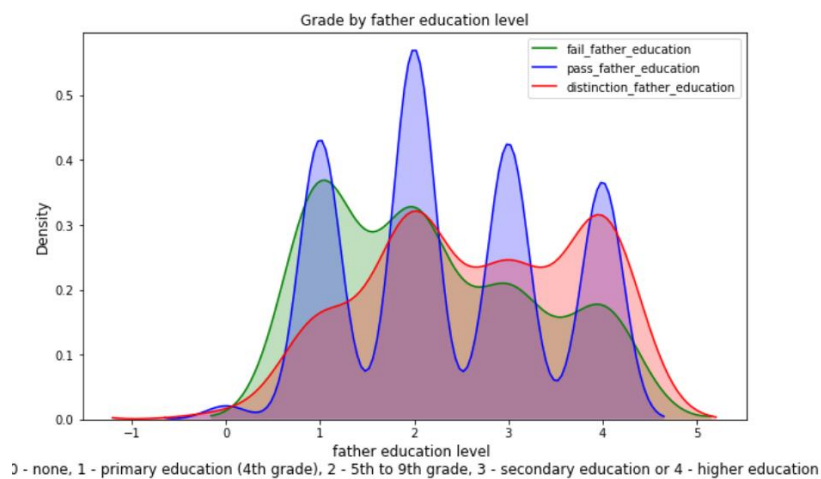
The quality of family relationship has impacts on student's grade. Bad quality of family relationship may not provide good environment for student to study well. However, the excellent family relationship cannot ensure the distinction, maybe they spoil too much. Parent need to maintain their relationship, but shouldn't spoil children.

Assumption 7: As a male, do you need to get married with woman with higher education level?



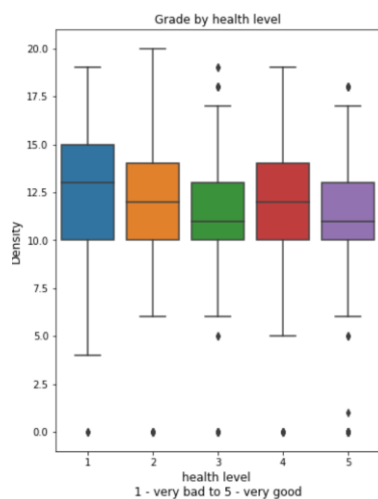
The figure shows that mother's education level has impacts on student's grade. As higher education that mother has, the percentage of getting distinction for their children are higher. I suggest that guy should marry someone with higher education level.

Assumption 8: As a female, do you need to get married with male with higher education level?



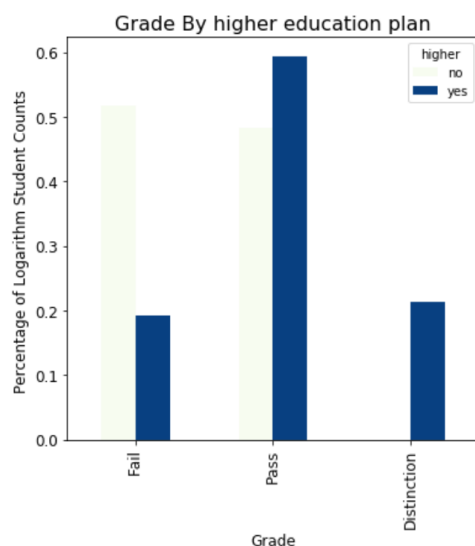
From the figure, father's education level has less influence on student's grade than woman. But father's higher education may help children reduce percentage of fails in final. I suggest that female needs to study harder, rather than thinking to get married with a smart man.

Assumption 9: Does health matter for student to score higher in final?



From the figure, we can see the health level doesn't have huge impact on student's grade. Surprisingly, student who has bad health level has better grades slightly. I think that student who score lower doesn't need to use unhealthy as excuse. Health is import for lifestyle, but it's not an excuse to face realistic.

Assumption 10: Should parent tell their children to have a well-designed plan for future?



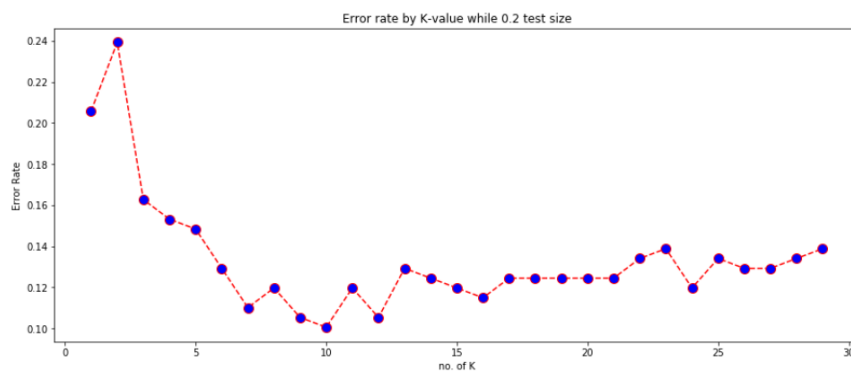
From the figure, we can tell the student who has plan to further study will definitely have higher scores than others student. The ambition and education plan is important to measure whether a student has positive attitude for studying. I think parent may tell children early to prepare future.

Data Modelling

K Nearest Neighbour model

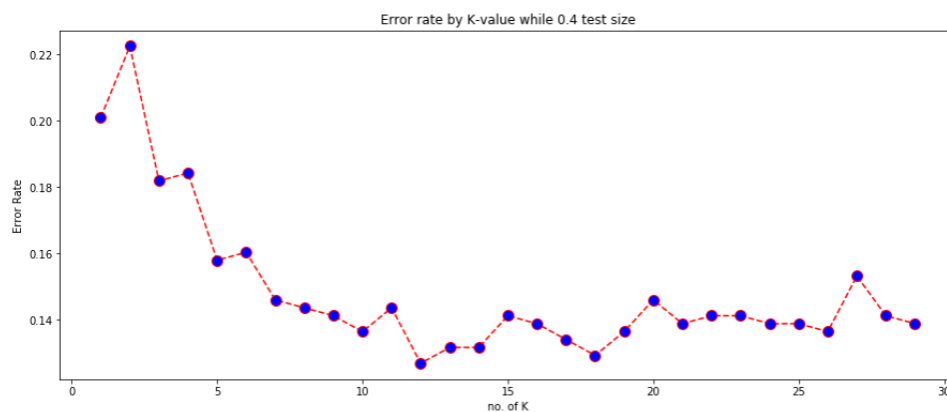
The KNN model is one of the data modelling methods used for classify student's final grade based on student's information. First of all, encode final grade label into numeric data, like 0 for distinction, 2 for pass and 3 for fail. Secondly, not matter what is test size, the K value needs to be selected seriously. The student use elbow method to create a for loop that trains with different K values, the low error rate K will be selected, and it usually comes up with an odd value. After select K-value, KNN model applied and get the result below.

80% for training and 20% for testing



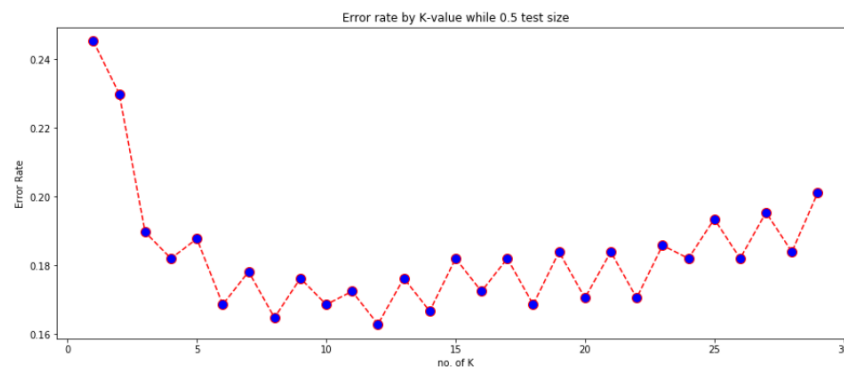
Confusion Matrix		precision	recall	f1-score	support
[[34 0 11] [0 26 13] [3 4 118]]	0	0.92	0.76	0.83	45
	1	0.87	0.67	0.75	39
	2	0.83	0.94	0.88	125
micro avg		0.85	0.85	0.85	209
macro avg		0.87	0.79	0.82	209
weighted avg		0.86	0.85	0.85	209

60% for training and 40% for testing



Confusion Matrix		precision	recall	f1-score	support
[[64 0 14] [0 59 26] [8 7 240]]	0	0.89	0.82	0.85	78
	1	0.89	0.69	0.78	85
	2	0.86	0.94	0.90	255
micro avg		0.87	0.87	0.87	418
macro avg		0.88	0.82	0.84	418
weighted avg		0.87	0.87	0.87	418

50% for training and 50% for testing



Confusion Matrix		precision	recall	f1-score	support	
[[78	0 25]	0	0.94	0.76	0.84	103
[0	77 50]	1	0.89	0.61	0.72	127
[5	10 277]]	2	0.79	0.95	0.86	292
micro avg		0.83	0.83	0.83	522	
macro avg		0.87	0.77	0.81	522	
weighted avg		0.84	0.83	0.82	522	

Decision Tree

The Decision tree model is one of the data modelling methods used for classify student's final grade based on student's information. It calculates gini value to decide which feature has more influences on student's grade. Student apply len function for X_train list, there are totally 58 features to distinguish in decision tree. The for loop will be applied to get optimal minimum sample leaf, see which one has more accurate score in data modelling performance.

```
from sklearn.tree import DecisionTreeClassifier
msl=[]
for i in range(1,58):
    tree = DecisionTreeClassifier(min_samples_leaf=i)
    t= tree.fit(X_train, y_train)
    ts=t.score(X_test, y_test)
    msl.append(ts)
msl = pd.Series(msl)
msl.where(msl==msl.max()).dropna()
```

```
12    0.873206
13    0.873206
dtype: float64
```

80% for training and 20% for testing

Confusion Matrix		precision	recall	f1-score	support
[[37 0 8] [0 28 11] [3 8 114]]	0	0.93	0.82	0.87	45
	1	0.78	0.72	0.75	39
	2	0.86	0.91	0.88	125
	micro avg	0.86	0.86	0.86	209
	macro avg	0.85	0.82	0.83	209
	weighted avg	0.86	0.86	0.86	209

60% for training and 40% for testing

Confusion Matrix		precision	recall	f1-score	support
[[68 0 10]	0	0.86	0.87	0.87	78
[0 63 22]	1	0.86	0.74	0.80	85
[11 10 234]]	2	0.88	0.92	0.90	255
	micro avg	0.87	0.87	0.87	418
	macro avg	0.87	0.84	0.85	418
	weighted avg	0.87	0.87	0.87	418

50% for training and 50% for testing

Confusion Matrix		precision	recall	f1-score	support
[[95 0 8]	0	0.90	0.92	0.91	103
[0 106 21]	1	0.83	0.83	0.83	127
[11 21 260]]	2	0.90	0.89	0.90	292
	micro avg	0.88	0.88	0.88	522
	macro avg	0.88	0.88	0.88	522
	weighted avg	0.88	0.88	0.88	522

KNN & Decision Tree Comparison

	KNN			Decision Tree		
Test Size	0.2	0.4	0.5	0.2	0.4	0.5
Model Score	0.873563	0.865900	0.873563	0.883141	0.896551	0.902298
Cross validation	0.827586	0.823754	0.827586	0.858237	0.888888	0.883141

By comparing the score of modelling score and cross validation score, Decision tree is better than KNN model in this case.

Discussion

The main arguments for this report are two, one is data exploration result which address that how different factors affect student's final grade. How to grow and education is big concerns for family, as they put all the efforts on them. As parents, they are worried about their children's future.

Another argument is more on data modelling method, in terms of using KNN and decision tree on data classification. It hard to make a conclude that decision tree is better than KNN to some certain extend. Both of them server different purposes, as KNN determines neighbourhoods, which indicate all the feature should be numeric since it was measured by distance. However, decision tree can measure nominal and numeric data and it is supervised learning. It doesn't require frequently look up data set as it has in-memory classification model ready. Decision tree is easier algorithm comparing with KNN, which provides more accurate result as well. (Mohanapriya, 2018)

Conclusion

In a nutshell, the student has found there are many factors have influences on final grade and these factors should be paid more attention by parents and students themselves. For example, student should improve learning efficiency towards studying, a positive attitude is necessary to score a higher grade. Besides, parents should have a plan and some limitation for their children, like restricting them from going out, taking alcohol or failing relationship. I suggest that parent should consider their children while purchasing a property, living near to school provide a nature advantage for students. Besides, after using data modelling method, the student has found decision tree is easier and more accurate than KNN model based on result. There are still many classification methods and data science knowledge, student will keep learning data science to enrich skill sets.

Reference

- Boschetti, A., 2016. *Python Data Science Essentials*. 2 ed. Birmingham: Packt Publishing.
- Cortez, P., 2014. *Student Performance Data Set*. [Online] Available at: <http://archive.ics.uci.edu/ml/datasets/Student+Performance#> [Accessed 10 5 2019].
- Mohanapriya, M., 2018. *Comparative study between decision tree and knn of data mining*, s.l.: IOP Publishing.
- robinson, s., 2018. *K-Nearest Neighbors Algorithm in Python and Scikit-Learn*. [Online] Available at: <https://stackabuse.com/k-nearest-neighbors-algorithm-in-python-and-scikit-learn/> [Accessed 10 5 2019].
- Seif, G., 2018. *A guide to decision trees for machine learning and data science*. [Online] Available at: <https://towardsdatascience.com/a-guide-to-decision-trees-for-machine-learning-and-data-science-fe2607241956> [Accessed 16 5 2019].