Technical University of Denmark

Introduction to Machine Learning and data mining

Data: Feature extraction, and visualization

Khushboo Nyman (s182328)
Enrico Tolotto (s190057)
Alexandros Spyropoulous (s182346)

October 1 2019

Contents

1 Contributions

Table 1: **Contributions**

Section Number	Section name	Student responsible
2	Introduction	Alexandros
3	Description of data set	Khushboo
4.1	Classification of attributes	Alexandros
4.2 and 4.3	Data issues and basic summary statistics	Khushboo
5.1	Data distribution and outliers	Enrico
5.2	Feasibility of Machine Learning Modelling	All
5.3	PCA analysis for the report	All
6	Discussion	Enrico

2 Introduction

Students' performance is a topic that interests many people around the globe. Our interest in this report is to check, based on given data, whether there is an association between a student's score and his gender. As the data that is gathered in this area keeps increasing, with the help of today's computer power we are able to extract, analyze more data and conclude to interesting results.

In this first report we are going to analyse and obtain an overview of the data including potential issues such as outliers and missing data. Furthermore, we are going to check whether the attributes appear to be normal distributed and highly correlated. These training data will be later on used as an input to an appropriate machine-learning method in order to produce our first model.

A principal component analysis (PCA) is going to be performed on the training data. With that way we will check whether it's possible to reduce the number of attributes without losing information. Eventually we will achieve better visualization and less complexity.

Finally we are going to address whether the primary machine learning modeling aim appears to be feasible based on our visualizations.

3 Description of data set

Our data set contains information about student performances in the exams and it consists of 8 attributes and 1000 measurements obtained from students. The data set has been downloaded from this link.

Our primary machine learning modelling aim is to classify the gender of a student based on the exam scores that he/she achieved in different courses. The attributes that seem the most interesting for the classification and regression techniques are the gender, math score, reading score, writing score.

By using these techniques on the data, the goal of this project is to predict the gender class label (male, female) based on the exam scores on three different courses (math, reading and writing).

Our data set doesn't seem to have any spurious attributes or missing data. The only thing that we need to perform is the one-out-of-K encoding that will help us convert some of our categorical variables into a numerical format so that we can draw the various plots and also feed our ML algorithm so to do a better job in prediction.

4 Detailed explanation of attributes of data

4.1 Classification of attributes

Table 2: **Attributes classification**

Attribute name	Discrete/Continuous	Nominal/Ordinal/Interval/Ratio
gender	Discrete	Nominal
race/ethnicity	Discrete	Nominal
parental level of education	Discrete	Ordinal
lunch	Discrete	Ordinal
test preparation course	Discrete	Ordinal
math score	Discrete	Ordinal/Interval/Ratio
reading score	Discrete	Ordinal/Interval/Ratio
writing score	Discrete	Ordinal/Interval/Ratio

- **Gender:** is described as discrete because there are only two possibilities 'male' and 'female' and one cannot rank male and female, so it is also or dinal.
- **Race/ethinicity:** is discrete because there are fixed number of possibilities. They also cannot be ranked, hence ordinal.
- **Parental level of education** : is also discrete because it can only take certain values. However, it is classified as ordinal because we can rank the level of education in this order (from high to low) Masters degree, Bachelor degree, Associate degree, some college, high school, some high school.
- **Lunch:** is discrete as it also takes up only two values. Our assumption is that the students who get free/reduced lunch, could belong to a lower income group family compared to the students who pay regular price. Hence, we have considered lunch as ordinal.
- **Test preparation course:** is discrete, again due to the same reason that it takes only two values. It is ordinal because it describes the level of preparation. So, if it is 'none' then the student's level of preparation is lower than if it is 'completed'.
- **Math score, Reading score, Writing score:** are discrete because they take certain values between 0 and 100. They cannot be continuous because there are no scores which have a fractional value. They are interval as they have values of equal intervals that mean something. Here, these scores have intervals of one. A ratio variable, has all the properties of an interval variable, and also has a clear definition of 0. When the score equals 0, there is none of the score. It is ordinal because we can rank the scores as higher or lower.

4.2 Data issues

It doesn't seem that there are any data issues except that 'test preparation course' is filled with 'none' and 'completed'. It has been assumed that 'none' means test preparation course was neither taken nor completed.

4.3 Basic summary statistics

Table 3 shows the summary statistics of continuous interval/ratio attributes.
The various statistics fields are count, mean, standard deviation (std), minimum score (min), maximum score (max) and the different quartiles. The quartile 25% shows the cut off value for the first

Table 3: Summary statistics

	math score	reading score	writing score
count	1000.00000	1000.00000	1000.00000
mean	66.08900	69.169000	68.054000
std	15.16308	14.600192	15.195657
min	0.00000	17.000000	10.000000
25%	57.00000	59.000000	57.750000
50%	66.00000	70.000000	69.000000
75%	77.00000	79.000000	79.000000
max	100.00000	100.000000	100.000000

25% of the data. Similarly, 50% and 75% quartiles can be understood as the cut off value for first 50% and 75% of the data.

Figure 1 shows the correlation between the attributes math score, reading score and writing score. We can see that all the scores are highly positively correlated as they are very close to 1. But, reading and writing score are better correlated as they have the highest correlation coefficient of 0.95 among the three coefficients. Then comes reading score and math score correlation and the last one is writing score and math score correlation.
It means that if a student performs better in one course, he/she also performs better in other courses. Similarly, if a student performs bad in one course, he/she also performs bad in other courses.
Figures 2 , 3 and 4 also show the correlations of different scores separating the genders.

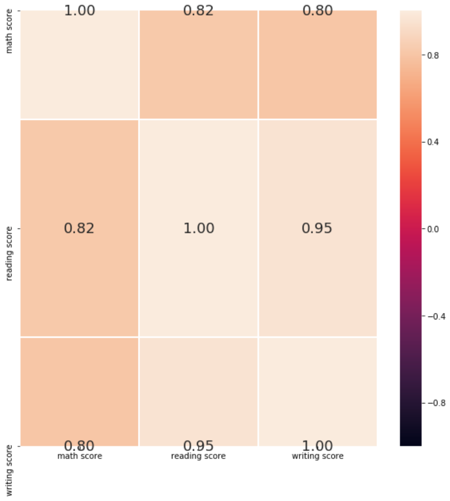

Figure 1: Correlation Matrix

Table 4: Covariance

	math score	reading score	writing score
math score	229.9	181.0	184.9
reading score	181.0	213.16	211.79
writing score	184.93	211.79	230.9

Table 4 shows the co-variance between the attributes math score, reading score and writing score. Co-variance determines the direction of linear relationship between two variables. From the table, we can see that all coefficients are positive which means the values increase or decrease together.

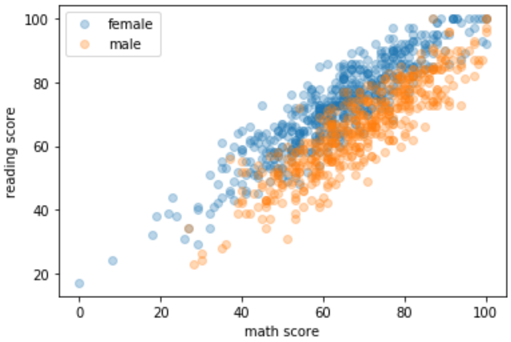

Figure 2: Math Score vs Reading Score

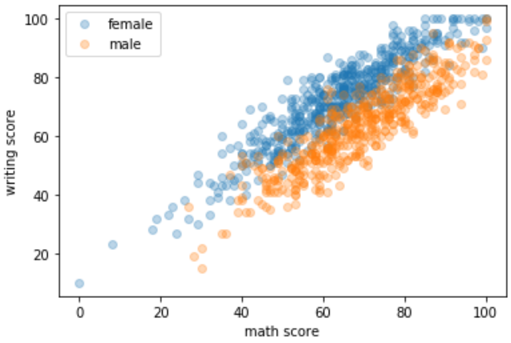

Figure 3: Math Score vs Writing Score

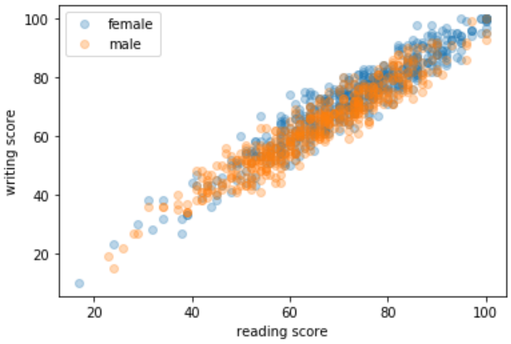

Figure 4: Reading Score vs Writing Score

Covariance is similar to correlation but when the co-variance is calculated, the data are not standardized. Because the data are not standardized, we cannot use the co variance statistic to assess the strength of a linear relationship. On the other hand, data is standardized when calculating correlation. So, correlation can be used to assess the strength of a linear relationship.

In the covariance matrix in the output, the off-diagonal elements contain the co variances of each pair of variables. The diagonal elements of the co variance matrix contain the variances of each variable. The variance measures how much the data are scattered about the mean. The variance is equal to the square of the standard deviation.

5 Data Visualization

5.1 Data distribution and outliers

To have a general understanding of the chosen dataset, it is important to visualize and understand how the data is distributed in the set. An optimal way for doing so is to plot a histogram of each relevant feature for our machine learning model.

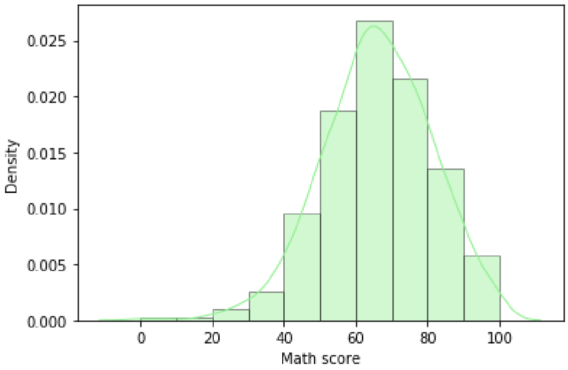

Figure 5: Math score

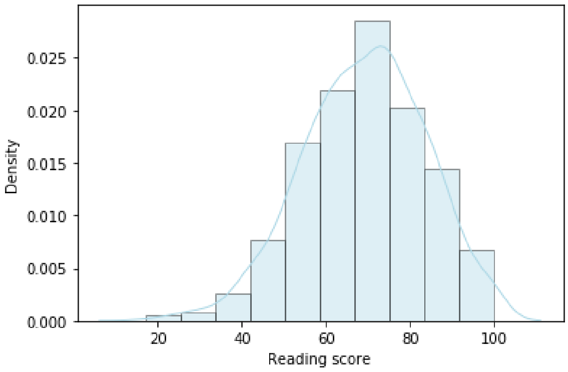

Figure 6: Reading score

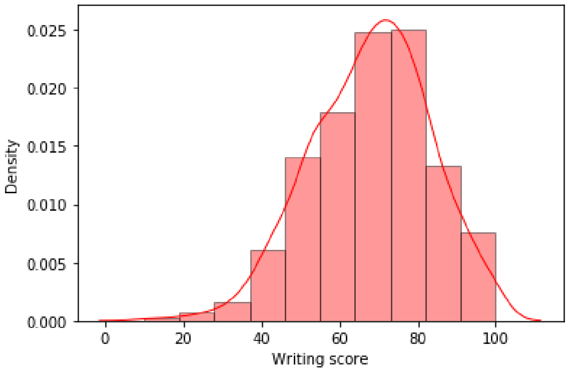

Figure 7: Writing score

The figures 5, 6 and 7 are the density distributions of the three main attributes in our dataset, displayed with 10 intervals or bins. From the density plot we can see that our histograms are roughly symmetric and bell-shaped, which is an indication that all three feature are normally distributed.

Another insightful visualization of the distribution of the three features can be obtained with a boxplot. Figure 8 is the comparison of the three-box plots of math, reading, and writing score. Here, the light grey line that divides each box into two parts corresponds to the 0.5 percentile of the corresponding feature. From this graph, it can be easily seen that 50% of the students got a score above 60 for all the three subjects. On the opposite side, considering the upper line and bottom line that form the box, which correspond respectively to the 0.75 and 0.25 percentile of the data, it is clear that less than 25% of the students scored less than 55 and less than 25% of the students more than 75.

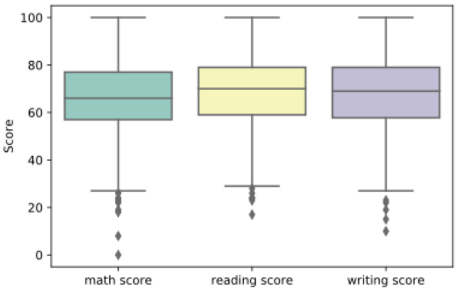

Figure 8: Box plot of Math, Reading and Writing scores

The black lines, known as the whiskers, outline how ample the distribution of the data is. Observations that fall outside the whiskers bounds are displayed as grey diamond and they are defined as outliers.

In the current dataset the outliers are found only in the lower part of the graph and are all the observation which have a value less then the value present in the table 5 The total number of outliers per attribute is listed in the table 6

Table 5: Lower whisker value for reading, writing and math score

	math score	reading score	writing score
Lower Whisker Value	27	29	25.8

Table 6: Outliers of reading, writing and math score

	math score	reading score	writing score
Outliers	8	6	5

5.2 Feasibility of Machine Learning Modelling

The main objective of this report is to analyze the ”students-performance-in-exams” dataset and verify the feasibility of a machine learning modeling, that finds an association between gender and scores. Looking at the visualization of the dataset, it is safe to assume that the data is normal distributed for the three attributes considered, and the number of outliers per feature is less then 0.1%. A second consideration to make in terms of feasibility is the size of the dataset. The current set is formed by 1000 observations, which is a discrete figure for a machine learning algorithm. Based on this observations and the data visualization, it seems feasible to predict the gender of a student based on its scores in math, reading and writing.

5.3 PCA analysis for the report :

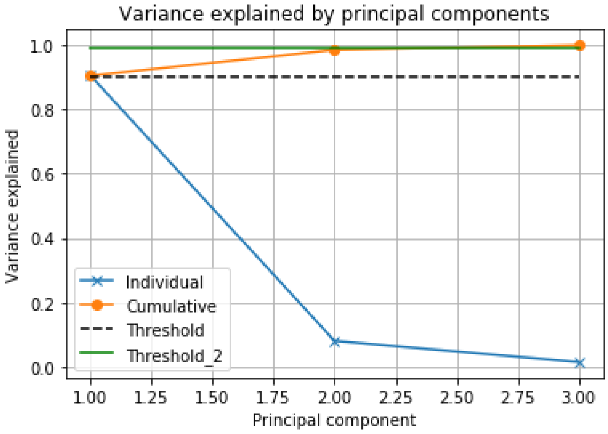

Figure 9: Variation explained as a function of the number of principal components

Figure 9 shows that PC1 alone accounts for 90% of the variance and PC1 and PC2 account for 99% of the variance. So, we can consider PC1 and PC2 mainly to analyse our data using PCA.

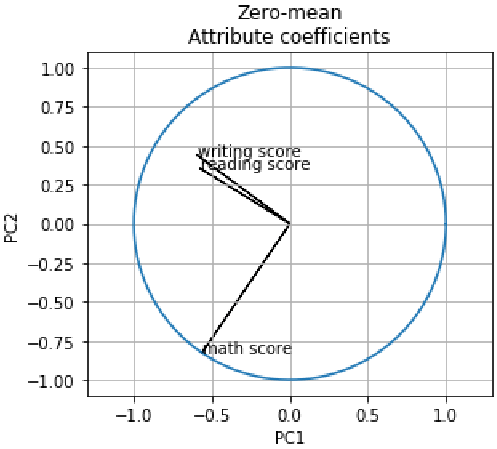

Figure 10: Attribute coefficients displaying the principal directions of considered PCA components

When we don’t divide the data by standard deviation after multiplying with the mean, then we achieve the figure 10. It shows that the math score points in the negative direction of both PC1 and PC2 whereas reading score and writing score point in negative direction of PC1 and positive direction of PC2.

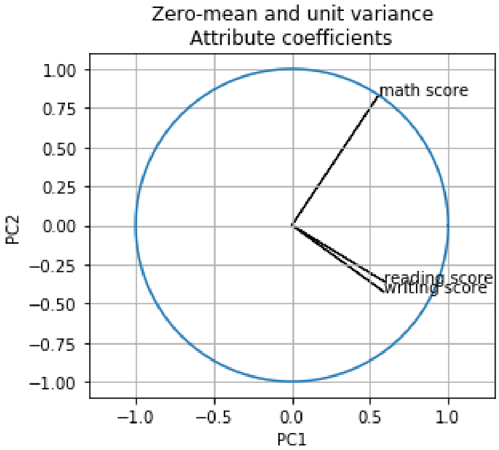

Figure 11: Attribute coefficients displaying the principal directions of considered PCA components (after further standardisation of data)

When we further standardize the data by giving it unit variance, we get figure 11. This, however shows a different picture from the earlier one. Here, the math score points in positive direction of both PC1 and PC2. Reading score and writing score point in positive direction of PC1 and negative direction of PC2. It clearly shows that the direction of the attributes reversed on PC1 PC2 plane, after standardizing the data.

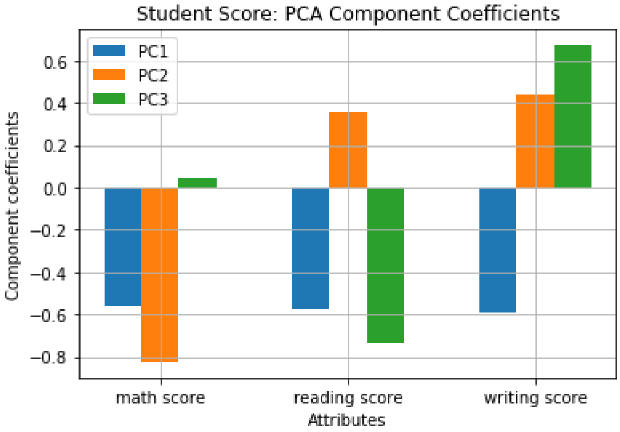

Figure 12: Data projected onto considered principal components

Figure 12 shows the data projected onto all of the principal components PC1, PC2 and PC3. PC1 has moderate negative coefficients for math score and reading score and little bit higher coefficient for writing score. PC2 has a high negative coefficient for math score, low positive coefficient for reading score and moderate coefficient for writing score. PC3 has a very small positive coefficient for math score, high negative coefficient for reading score and high positive coefficient for writing score. In general, the magnitude of the coefficients matter and not the direction (positive or negative). The direction only implies how the components affect the attribute with respect to increase or decrease.

6 Discussion

The aim of this report was to show whether it was feasible to implement a machine-learning model to our dataset. We started by checking for issues in our data set and then selected the important attributes (math score, reading score, writing score and gender) for our machine learning aim. Another important observation is that the data is normalized and the percentage of outliers is very low that is why we don't consider to remove them as we expect that they will not affect our model. We showed there is a high correlation between these attributes and also that there is a clear distinction between male and female scores. Lastly we implemented the PCA and we were able to identify the two major principal components which together account for 99 % of the data. So we conclude that our machine learning aim is feasible.