Academic year: 2019/2020
Computer Science Department

Big Data Analytics

Assignment: Student Performance case study (clustering/classification)

Nafaa Jabeur / Hana Gharrad

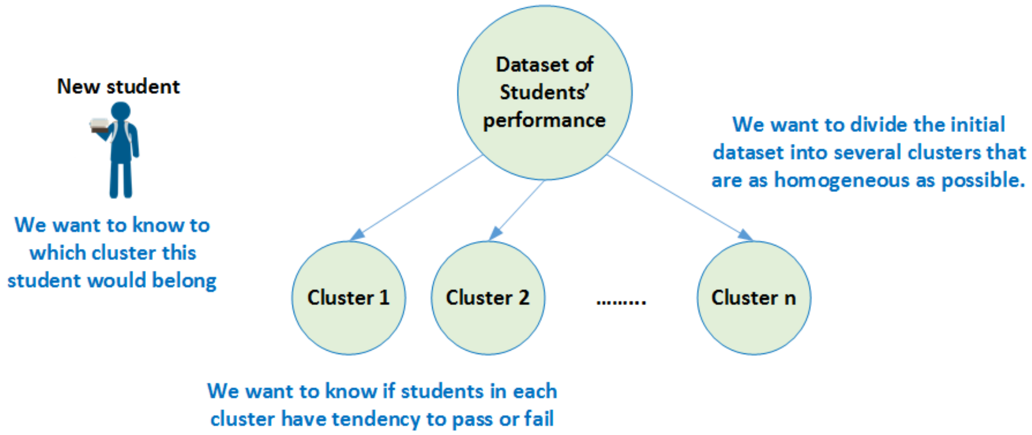
Instructions

- Assignment to be done in group of 3 students maximum.
- Dataset is available on Moodle.
- Send your assignments to nafaa.jabeur@gutech.edu.om and hana.gharrad@gutech.edu.om
- Due date: **January 26, 2020**.

1. Introduction

In this assignment, you are asked to analyze the performance of students based on a given dataset. The dataset, which is composed of data from two schools, measures the performance of students in Math. The observations describe each student according to multiple variables (31 in total), including the school of the student, the family size, the father's job and the weekly study time.

As a first step, we want to organize the students into clusters in order to identify similar students (students belonging to the same cluster should be reasonably more similar than those belonging to other clusters). As a second step, we want to classify new observations (i.e. new students) by predicting to which clusters they belong. Ultimately, we want to know if students in each cluster have tendency to pass or fail. The figure below is drafted to show some thoughts related to the case study in hands. Your task is to answer the questions in Section 3 of this assignment.


New student
We want to know to which cluster this student would belong
Dataset of Students' performance
We want to divide the initial dataset into several clusters that are as homogeneous as possible.
Cluster 1
Cluster 2
....
Cluster n
We want to know if students in each cluster have tendency to pass or fail

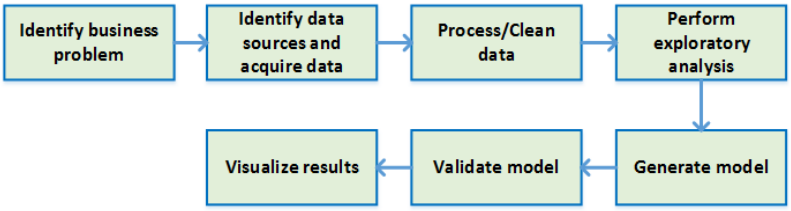
To achieve your tasks, you are provided with a dataset of 31 features and 400 observations. The table below describes the features of the dataset.

Attribute	Description	Type	Sample values
School	Student's School	Factor	'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira
Sex	Student's Sex	Factor	'F' - female or 'M' - male
Age	Student's Age	Integer	numeric: from 15 to 22
Address	Student's Home Address Type	String	'U' - urban or 'R' - rural
FamSize	Family Size	Factor	'LE3' - less or equal to 3 or 'GT3' - greater than 3
Pstatus	Parent's Cohabitation Status	String	'T' - living together or 'A' - apart
Medu	Mother's Education	Factor	0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education
Fedu	Father's Education	Factor	0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education
Mjob	Mother's Job	Factor	'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other'
Fjob	Father's Job	Factor	'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other'
Reason	Reason To Choose This School	Factor	close to 'home', school 'reputation', 'course' preference or 'other'
Guardian	Student's Guardian	String	'mother', 'father' or 'other'
TravelTime	Home To School Travel Time	Integer	1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour
StudyTime	Weekly Study Time	Integer	1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours
Failures	Number Of Past Class Failures	Integer	n if 1<=n<3, else 4
Schoolsup	Extra Educational Support	Boolean	yes or no
Famsup	Family Educational Support	Boolean	yes or no
Paid	Extra Paid Classes Within The Course Subject (Math Or Portuguese)	Boolean	yes or no
Activities	Extra-Curricular Activities	Boolean	yes or no
Nursery	Attended Nursery School	Boolean	yes or no
Higher	Wants To Take Higher Education	Boolean	yes or no
Internet	Internet Access At Home	Boolean	yes or no
Romantic	With A Romantic Relationship	Boolean	yes or no
Famrel	Quality Of Family Relationships	Factor	from 1 - very bad to 5 - excellent
Freetime	Free Time After School	Factor	from 1 - very low to 5 - very high

Goout	Going Out With Friends	Factor	from 1 - very low to 5 - very high
Health	Current Health Status	Factor	from 1 - very bad to 5 - very good
Absences	Number Of School Absences	Integer	from 0 to 93
G1	First Period Grade	Integer	from 0 to 20
G2	Second Period Grade	Integer	from 0 to 20
G3	Final Grade	Integer	from 0 to 20, output target

2. Quick recall on Data Analysis Steps

The figure below captures the different steps that we generally needed to solve a data analysis problem.


Identify business problem
Identify data sources and acquire data
Process/Clean data
Perform exploratory analysis
Generate model
Validate model
Visualize results

The required tasks in each step are summarized as follows:

Step 1: **Identify business problem**

In this step, you must understand the problem that you are trying to solve. You should also identify the central objectives of your project by identifying the variables that need to be predicted.

Step 2: Identify data sources and acquire data

In this step, you need to identify the needed data and gather your data from different sources.

Step 3: **Process/Clean data**

Check if some data are missing, evaluate the missing data and identify how to deal with any instances that have missing values.

Step4: **Perform exploratory analysis**

The data exploration stage is like the brainstorming of data analysis. This is where you understand the patterns and bias in your data. It could involve pulling up and analyzing a data to see the general trend, or even creating an interactive visualization that lets you dive down into each data point. Visualizing the relationship between result and studying hours is an example of actions that you can do during the data exploration stage.

Step5: Generate the model

In this step, you need to extract relevant features and thing about additional possible features that could be useful for the model. Which is also named feature engineering (is the process of selecting and transforming variables when creating a predictive model). Typically perform two types of tasks in feature engineering: feature selection and construction. For example, if you were predicting student scores and had features for the number of hours of sleep on each night, you might want to create a feature that denote the average sleep that the student had instead. After selecting and generating the features, you need to split the data if it is needed into train and test sets. Then, you need to run the model.

Step6: Validate the model

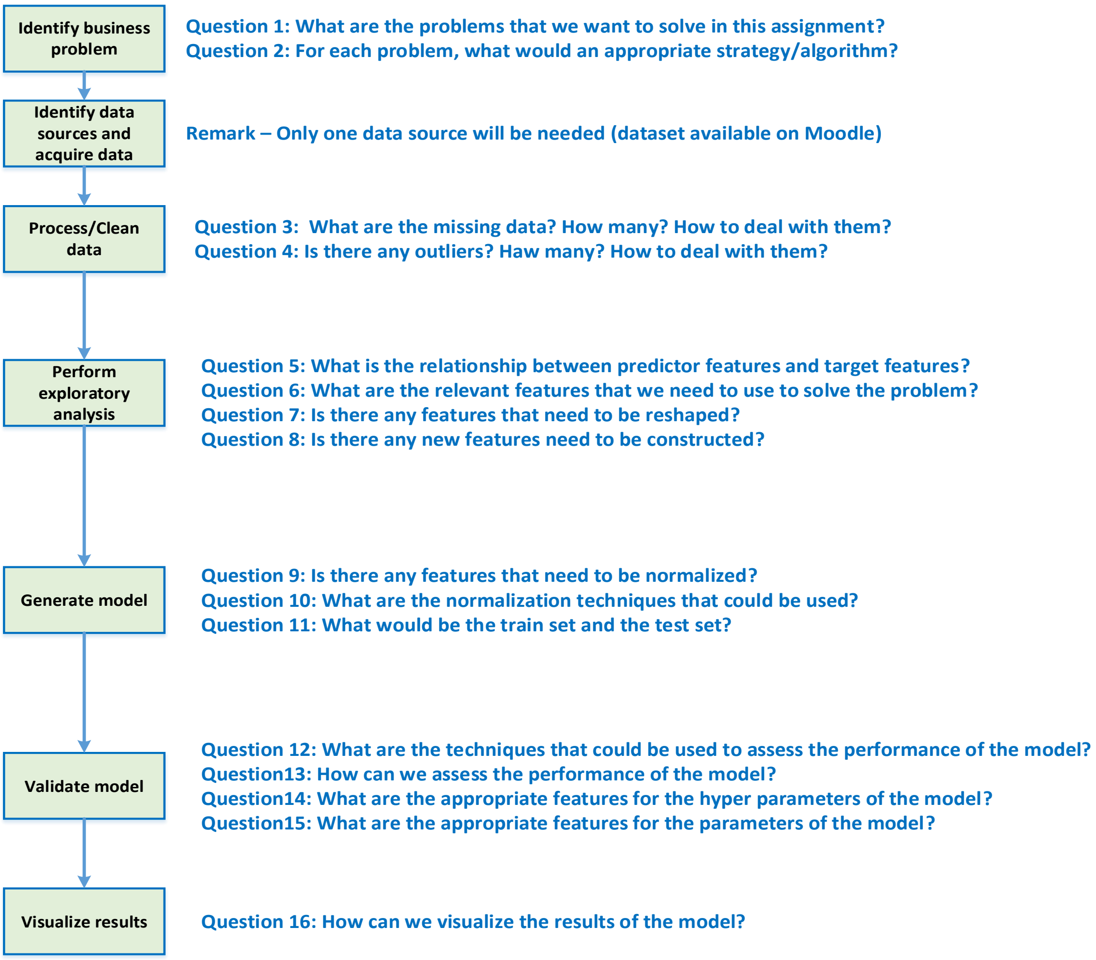
After generating the mode, it is critical that you evaluate its success. Some processes (like elbow, confusion matrix, and k-fold cross validation) are commonly used to measure the accuracy of a model. Plots (such as ROC curves, which is the true positive rate plotted against the false positive rate), are also used to benchmark the success of a model.

Step7: Visualize results

Once you have derived the intended insights from your model, you must represent them in an easy way that the different key stakeholders in the project can understand.

3. Case study: Student performance clustering and classification

In order to make sure that you are going to follow the steps for solving a data analysis problem, you will be asked to answer questions related to each of these steps. Examples of questions are highlighted in the figure below.


Identify business problem
Identify data sources and acquire data
Process/Clean data
Perform exploratory analysis
Generate model
Validate model
Visualize results

Question 1: What are the problems that we want to solve in this assignment?
Question 2: For each problem, what would an appropriate strategy/algorithm?

Remark – Only one data source will be needed (dataset available on Moodle)

Question 3: What are the missing data? How many? How to deal with them?
Question 4: Is there any outliers? Haw many? How to deal with them?

Question 5: What is the relationship between predictor features and target features?
Question 6: What are the relevant features that we need to use to solve the problem?
Question 7: Is there any features that need to be reshaped?
Question 8: Is there any new features need to be constructed?

Question 9: Is there any features that need to be normalized?
Question 10: What are the normalization techniques that could be used?
Question 11: What would be the train set and the test set?

Question 12: What are the techniques that could be used to assess the performance of the model?
Question13: How can we assess the performance of the model?
Question14: What are the appropriate features for the hyper parameters of the model?
Question15: What are the appropriate features for the parameters of the model?

Question 16: How can we visualize the results of the model?

Step 1: **Identify business problem**

Answer the following questions.

<u>Question 1 –</u> What are the problems that we are trying to solve in this assignment?
<u>Question 2 –</u> What are the algorithms that could be used to solve each problem?

Step 2: Identify data sources and acquire data

The data used in this assignment are provided from one source only In this regard, we will not be interested in aggregating data from different sources. To carry out your assignment, please refer to Moodle to download the dataset.

Step 3: **Process/Clean data**

<u>Question 3 –</u> Find the number of missing data for each feature.

Question 4 – What is the best way to deal with missing data (delete, calculate the average or the median, etc.)?

Question 5 – Using an outlier detection technique (such as scatter, Z-score, or Box Plot), find if there are any outliers in the data (give these outliers).

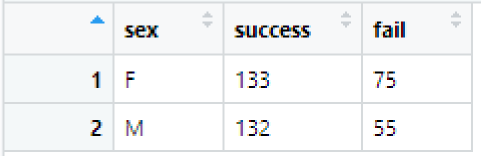
- Remark – Remove any existing outliers.

Step 4: **Perform exploratory analysis**

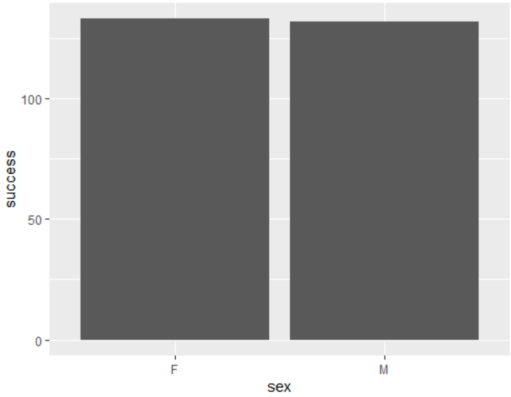
In this step, we aim to perform an exploratory analysis in order to understand the data and identify the important features to use. For example, we want to see if the gender of a student has any relation with his/her results. To this end, we will answer the question "**How many males and females did pass?**" by calculating the number of male students who succeeded and the number of female students who succeeded. To this end, we create in the code below a pipeline. The input of this pipeline is the data that we import from the dataset. We create two additional columns named "success" and "fail". Success will be equal to 1 if the value of the feature G3 (final grade) is greater or equal to 10 (otherwise success will be equal to 0). We give this data as input to be filtered and then grouped by gender. In the last step of the pipeline we calculate the sum of success and fail for each group (M/F). The result will be stored in a new data frame. The table below shows the content of the new data frame created. The result is plotted using `ggplot` function from the library `ggplot2`.

```
result2Sex<-data%>%
   mutate(success=ifelse(G3>=10,1,0), fail= ifelse(G3<10,1,0))%>%
   filter(sex=="F"|sex=="M")%>%
   group_by(sex)%>%
   summarise(success=sum(success), fail=sum(fail))
View(result2Sex)
result2Sex%>%
  ggplot(aes(x=sex,y=success))+
  geom_bar(stat="identity")
```

Output:



	sex	success	fail
1	F	133	75
2	M	132	55



In order to evaluate the importance of others features, answer the following questions:

How many male and female did pass? (already answered above)

Question 6 – Discover the relationship between going out and the grades of students (in other words, does, for example, going out so often affect the results of the student?)

Question 7 – Is there a relationship between family size and the result of students?

Question 8 – Does the quality of the family relationship affect students' results? Explain your answer.

Question 9 – Does the school travel time affect students result? Explain your answer.

Question 10 – What type of relationship does exist between the previous failures and students' results? (in other words, if a student previously failed does this mean he/she will pass?!)

Question 11 – Do extra activities help to improve student performance?

Question 12 – Does internet home access help to improve student performance?

Question 13 – What is the distribution of students who succeeded over the levels of free time?

Question 14 – What is the relationship between age and student result? (Hint: Generate a graph and say if the relationship is proportional or inversional proportional)

Question 15 – What is the relationship between address and student result?

Question 16 – What is the relationship between study time and student result?

Question 17 – Does absence affect student performance?

Question 18 – Does health affect student performance?

Question 19 – Does mother and father education/job affect student result?

Step 5: Generate the model

Based on the answers of the previous questions:

Question 20 – Give the useful variables to perform the clustering and classification. Remove the non-useful variables if there are any.

Question 21 – Are there any features that we need to add to the dataset? Identify and add these features to the dataset?

Question 22 – Is there any data needed to be normalized? What techniques of normalization could be used?

Hint: You can display correlation matrix to evaluate the hidden relationships between features.

Step 6: Validate the model

- **Part 1**: We want to find a good clustering of students based on the most important features that do affect the results of student. A good clustering will provide a less **within-cluster variation** . A good number of clusters will be the smallest number of clusters with the lowest values of **within-cluster variation**.
    - Let us assume that we will use the features: sex, age, address, Pstatus, Medu, Fedu, Mjob, Fjob, studytime, absences.

      You are requested to:
        - Question 23 – Do the clustering based on these features with k=7. Showcase your results (plot of clusters).
        - Question 24 – Calculate the within-cluster variation.
    - Let us assume that we will use the features: age, address, Fjob, traveltime, studytime, failures, and absences.

      You are requested to:
        - Question 25 – Do the clustering based on these features with k=7. Showcase your results (plot of clusters).
        - Question 26 – Calculate the within-cluster variation
    - Let us assume that we will use the features: age, address, famsize, Pstatus, Medu, Fedu, Mjob, Fjob,studytime, traveltime ,failures, internet, famrel, freetime, goout,health, absences.

      You are requested to:
        - Question 27 – Do the clustering based on these features for k=7.
        - Question 28 – Calculate the within-cluster variation.

        - Question 29 – Is this clustering better or worse than the previous ones? Explain your answer.
    - Question 30 – Find the appropriate number of clusters (k) for the student performance data.

Hint: Refer to Lab 1 to calculate within-cluster variation (within-cluster sum of square (WSS)) or Average silhouette for different values of K (number of cluster).

- **Part 2** In this part, we want to use classification algorithm in order to be able to classify new observations. In order to make sure that we will be using a convenient classification algorithm, we commonly test more than one algorithm and assess their results. In this assignment, we will use SVM and K-Nearest Neighbor (KNN).

  Let us assume that we will use the features: age, address, famsize, Pstatus, Medu, Fedu, Mjob, Fjob, studytime, traveltime, studytime, failures, internet, famrel, freetime, goout, health, absences,G1, G2, G3, Cluster. Where Cluster is the target feature.

You are requested to:

- Question 31 – Do the classification using KNN algorithm.
- Question 32 – Assess the performance of the model for different values of K (number of neighbors) using confusion-matrix and accuracy.
- Question 33 – Find the best value of K.

You are requested to:

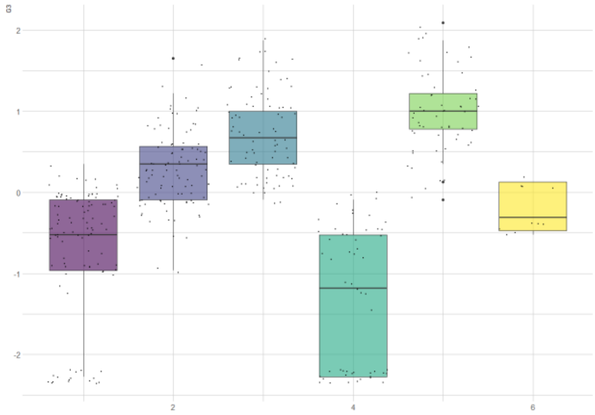
- Question 34 – Do the classification using SVM algorithm.
- Question 35 – Assess the performance of the model using confusion-matrix and accuracy.

Question 36 – Compare the performance of SVM to the accuracy of KNN.

- **Part 3** In this part we want to find if students in the same cluster have similar results.

  Question 37 – In order to answer this question, show the distribution of the results for each cluster. Explain the results.

As an example, if we already found 6 clusters, the following graph shows the distribution of the results over different clusters.



Step 7 : Step7: Visualize results

Question 38 – Create graphs to present important results concluded from the previous step. Comment on those graphs.