

# Report of practical work of data mining

Realized by : Supervised by:

Youssef Faraby M.Ali IdrI

University year 2017 – 2018

1-Dataset description

# **Students' Academic Performance Dataset (xAPI-Edu-Data)**

Data Set Characteristics: Multivariate

Number of Instances: 480

Area: E-learning, Education, Predictive models, Educational Data Mining

Attribute Characteristics: Integer/Categorical

Number of Attributes: 16

Date: 2016-11-8

Associated Tasks: Classification

Missing Values? No

File formats: xAPI-Edu-Data.csv

# **Source:**

Elaf Abu Amrieh, Thair Hamtini, and Ibrahim Aljarah, The University of Jordan, Amman, Jordan, [http://www.Ibrahimaljarah.com](http://www.ibrahimaljarah.com/)

# **Dataset Information:**

This is an educational data set which is collected from learning management system (LMS) called Kalboard 360. Kalboard 360 is a multi-agent LMS, which has been designed to facilitate learning through the use of leading-edge technology. Such system provides users with a synchronous access to educational resources from any device with Internet connection.

The data is collected using a learner activity tracker tool, which called experience API (xAPI). The xAPI is a component of the training and learning architecture (TLA) that enables to monitor learning progress and learner’s actions like reading an article or watching a training video. The experience API helps the learning activity providers to determine the learner, activity and objects that describe a learning experience. The dataset consists of 480 student records and 16 features. The features are classified into three major categories: (1) Demographic features such as gender and nationality. (2) Academic background features such as educational stage, grade Level and section. (3) Behavioral features such as raised hand on class, opening resources, answering survey by parents, and school satisfaction.

The dataset consists of 305 males and 175 females. The students come from different origins such as 179 students are from Kuwait, 172 students are from Jordan, 28 students from Palestine, 22 students are from Iraq, 17 students from Lebanon, 12 students from Tunis, 11 students from Saudi Arabia, 9 students from Egypt, 7 students from Syria, 6 students from USA, Iran and Libya, 4 students from Morocco and one student from Venezuela.

The dataset is collected through two educational semesters: 245 student records are collected during the first semester and 235 student records are collected during the second semester.

The data set includes also the school attendance feature such as the students are classified into two categories based on their absence days: 191 students exceed 7 absence days and 289 students their absence days under 7.

This dataset includes also a new category of features; this feature is parent parturition in the educational process. Parent participation feature have two sub features: Parent Answering Survey and Parent School Satisfaction. There are 270 of the parents answered survey and 210 are not, 292 of the parents are satisfied from the school and 188 are not.

(See the related papers for more details).

# **Attributes**

**1 Gender** - student's gender (nominal: 'Male' or 'Female’)

**2 Nationality-** student's nationality (nominal:’ Kuwait’,’ Lebanon’,’ Egypt’,’ SaudiArabia’,’ USA’,’ Jordan’,’ Venezuela’,’ Iran’,’ Tunis’,’ Morocco’,’ Syria’,’ Palestine’,’ Iraq’,’ Lybia’)

**3 Place of birth**- student's Place of birth (nominal:’ Kuwait’,’ Lebanon’,’ Egypt’,’ SaudiArabia’,’ USA’,’ Jordan’,’ Venezuela’,’ Iran’,’ Tunis’,’ Morocco’,’ Syria’,’ Palestine’,’ Iraq’,’ Lybia’)

**4 Educational Stages-** educational level student belongs (nominal: ‘lowerlevel’,’MiddleSchool’,’HighSchool’)

**5 Grade Levels**- grade student belongs (nominal: ‘G-01’, ‘G-02’, ‘G-03’, ‘G-04’, ‘G-05’, ‘G-06’, ‘G-07’, ‘G-08’, ‘G-09’, ‘G-10’, ‘G-11’, ‘G-12 ‘)

**6 Section ID-** classroom student belongs (nominal:’A’,’B’,’C’)

**7 Topic-** course topic (nominal:’ English’,’ Spanish’, ‘French’,’ Arabic’,’ IT’,’ Math’,’ Chemistry’, ‘Biology’, ‘Science’,’ History’,’ Quran’,’ Geology’)

**8 Semester-** school year semester (nominal:’ First’,’ Second’)

**9 Parent responsible for student** (nominal:’mom’,’father’)

**10 Raised hand-** how many times the student raises his/her hand on classroom (numeric:0-100)

**11- Visited resources-** how many times the student visits a course content(numeric:0-100)

**12 Viewing announcements-**how many times the student checks the new announcements(numeric:0-100)

**13 Discussion groups-** how many times the student participate on discussion groups (numeric:0-100)

**14 Parent Answering Survey-** parent answered the surveys which are provided from school or not (nominal:’Yes’,’No’)

**15 Parent School Satisfaction-** the Degree of parent satisfaction from school(nominal:’Yes’,’No’)

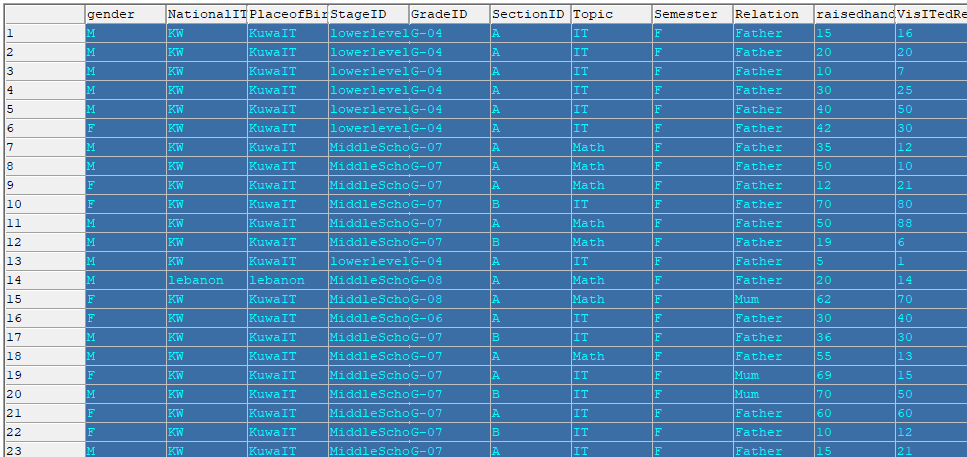
**16 Student Absence Days-**the number of absence days for each student (nominal: above-7, under-7)

# **The students are classified into three numerical intervals based on their total grade/mark:**

**Low-Level:** interval includes values from 0 to 69,

**Middle-Level:** interval includes values from 70 to 89,

**High-Level:** interval includes values from 90-100.

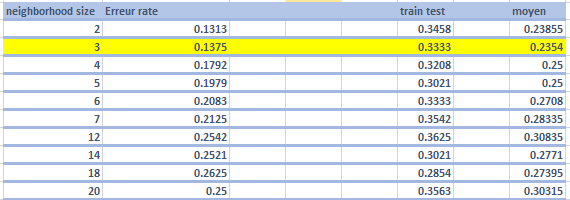


**Figure 1-Dataset**

2-K-NN Algorithm

KNN is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure. The purpose of the k Nearest Neighbors (kNN) algorithm is to use a database in which the data points are separated into several separate classes to predict the classification of a new sample point. This sort of situation is best motivated through examples.

We first start by identifying different values of error rate from learning and testing database that match different values of K.



*Tableau 2 : Learning and testing error rate vs K*

The objective is to find the suitable value of K by defending the minimum of testing and learning errors rate.

To choose the most suitable K, we should calculate the average between the two curves, in our case, we found a minimum error rate lower that 20%, which led us to choose the smaller error rate that coincide with K = 3.

We find the classifier performance above of the same K:

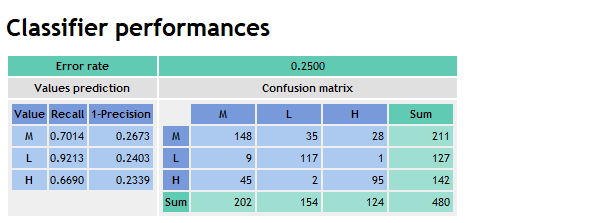


Figure 4 – K-NN confusion matrix

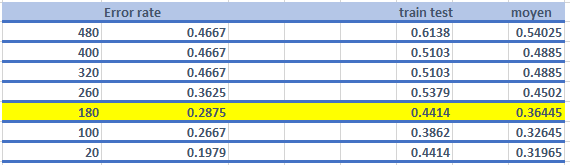
The confusion matrix show typically a good result, we rarely find that the class representing reality is not confused with the one representing the model, which brought us a very small error rate equal to 0.25, and this is the best case for us.

ID3 Algorithm

In decision tree learning, ID3 (Iterative Dichotomiser 3) is an algorithm used to generate a decision tree from a dataset. ID3 is typically used in the machine learning and natural language processing domains.

The ID3 algorithm begins with the original set S as the root node. On each iteration of the algorithm, it iterates through every unused attribute of the set S and calculates the entropy H(S) (or information gain IG(A)) of that attribute. It then selects the attribute which has the smallest entropy (or largest information gain) value. The set S is then split by the selected attribute to produce subsets of the data. The algorithm continues to recurse on each subset, considering only attributes never selected before.

We start by calculating learning and testing error rate in function of the min size of split, the table below show this result:

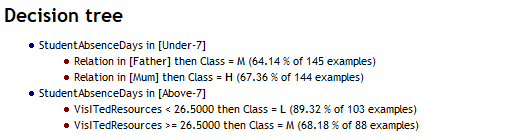


*Tableau 3 : Learning and testing error rate vs split*

Our objective is to determinate the minimum value of error rate in both curves.

We can see that the learning test keep the same value of error rate which is 0.46 with large number of minimum size of split.

we are looking for the maximum value of the «minimum size for split »with minimum error rate which led us to choose the value 180.



Conclusion: Comparison between K-NN and ID3:

The error rate of K-NN algorithm is 0.25 in the testing database, while in ID3 we find 0.46, which mean that ID3 generate better result than K-NN.

The error rate of K-NN in learning test is the same for the ID3 all over the curve.

The both algorithm typically gives very satisfying results.