Seth Ball
Keller Sedillo-Garrido
MD Zaman

CS 488
Team Project

The dataset that we have chosen, involves student achievement in secondary education of two Portuguese schools. Some data points that are available include general information of the students including age, sex, DOB, etc. There is also information on their parents including parents education, habitual status, current employment status, etc. There are two datasets in this problem, one for Math and the other for Portuguese. The Problem that we are looking to solve with this dataset is: are there any observable correlations that we can find between any combination of attributes and the grades of the students? Also can we create a decision tree to predict the grades students will get based on these attributes?

Our motivation for the project is to further expand our knowledge of data mining by using real world datasets to make predictions on future data presented. Student performance is something that is constantly being tracked to inform teachers on how well they are teaching their students, but there are many factors that can contribute to students' grades other than their aptitude for school. Using this dataset we hope to link possible non-school attributes to find whether or not these attributes play large roles in how well students do in school and if there are correlations, which ones have the biggest impact on students performance in school.

Other people have looked into finding a coalition between grades and different attributes. One study looks into how Assessment Grades and Online Activity Data can affect their performance and even predict their potential grade. According to the research paper, **predicting a student's performance** is not an easy task due to the complexity of the problem of predicting grades. There is so much in a student's life that can affect the way that they perform.

Now that we understand the background of our dataset, lets look at the processes we plan to take. There is a few steps we want to follow:

1) Data preparation:

Data transformation operations change the data to make it useful in data mining. Following transformation can be applied. For instance, the name of the student is different in different tables.

2) Data transformation:

There is a few techniques we can apply to improve our dataset:

Smoothing: Helps to remove noise from the data.

Aggregation: Summary or aggregation operations are applied to the data. Meaning that bi-yearly data can be calculated to find a pattern.

Generalization: In this step, Low-level data is replaced by higher-level concepts with the help of concept hierarchies. For example, the city is replaced by the county.

Normalization: Normalization performed when the attribute data are scaled up o scaled down. Example: Data should fall in the range -2.0 to 2.0 post-normalization.

Attribute construction: the attributes are constructed and include the given set of attributes helpful for data mining.

3) Data Modeling

For predictive models, this could be an objective function, for dataset augmentation a success criteria **Classification** and **Regression** analysis are standard methods to predict student behavior, automate the sorting of students into demographic groups, forecast revenue, and more. Classification and regression are similar because analysts use each to predict outcomes. Classification method takes the form of a **Decision tree**. We will discuss the type of model we will be using later in the report.

5) Visualization:

With python programming language we will reflect our analysis of work on the visual graphs. We will find the pit-falls and accuracy rate of our predictive models.

4) Evaluation

Gaining success in rates(error or good predictive values) and understanding is an iterative process. In fact, new model requirements may be raised because of data mining process we will be discovering as we analysis the data set . A go or no-go decision is taken to move the model in the deployment phase.

6) Deployment

The knowledge or information discovered during the data mining process should be made easy to understand for non-technical stakeholders. A detailed deployment plan, for shipping, maintenance, and monitoring of data mining discoveries is created

So, each member wants to be involved in all steps of our selective data.

Reference:
https://www.researchgate.net/publication/341100298_Predict_Students'_Academic_Performance_based_on_their_Assessment_Grades_and_Online_Activity_Data

Short Description of the Data Set:

These statistics look at secondary student achievement at two Portuguese schools. The data was gathered utilizing school reports and surveys, and the attributes include student grades, demographic, social, and educational factors. Regarding the performance in two different areas, Portuguese language and mathematics (mat), two datasets are offered (por). The two datasets were modeled using binary/five-level classification and **regression** tasks in [Cortez and Silva, 2008]. The target attribute G3 shows a substantial association with the attributes G2 and G1, which is significant. This is due to the fact that G3 represents the final year grade and is assigned during the third period, whereas G1 and G2 represent the first and second period grades. Without G2 and G1, it is harder to forecast G3, yet such a prediction is significantly more accurate.

Data preparation

Check for NULL values: there are no NULL values, let's see the data preview

There are 33 columns:

school: student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)

sex: student's sex (binary: 'F' - female or 'M' - male)

age: student's age (numeric: from 15 to 22)

address: student's home address type (binary: 'U' - urban or 'R' - rural)

famsize: family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)

Pstatus: parent's cohabitation status (binary: 'T' - living together or 'A' - apart)

Medu: mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 5th to 9th grade, 3 secondary education or 4 higher education)

Fedu: father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 5th to 9th grade, 3 secondary education or 4 higher education)

Mjob: mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')

Fjob: father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')

reason: reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')

guardian: student's guardian (nominal: 'mother', 'father' or 'other')

traveltime: home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)

studytime: weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

failures: number of past class failures (numeric: n if 1<=n<3, else 4)

schoolsup: extra educational support (binary: yes or no)

famsup: family educational support (binary: yes or no)

paid: extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)

activities: extra-curricular activities (binary: yes or no)

nursery: attended nursery school (binary: yes or no)

higher: wants to take higher education (binary: yes or no)

internet: Internet access at home (binary: yes or no)

romantic: with a romantic relationship (binary: yes or no)

famrel: quality of family relationships (numeric: from 1 - very bad to 5 - excellent)

freetime: free time after school (numeric: from 1 - very low to 5 - very high)

goout: going out with friends (numeric: from 1 - very low to 5 - very high)

Dalc: workday alcohol consumption (numeric: from 1 - very low to 5 - very high)

Walc: weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)

health: current health status (numeric: from 1 - very bad to 5 - very good)

absences: number of school absences (numeric: from 0 to 93)

G1: first period grade (numeric: from 0 to 20)

G2: second period grade (numeric: from 0 to 20)

G3: final grade (numeric: from 0 to 20)

Five main phases include data collection, data pre-processing, sub-datasets generation, classification algorithms application, and evaluation. However, assessment and activity data together work better to enhance the accuracy of the prediction model.

Now lets create a grading based on its G Average:

Above 90% = Grade A

Between 70% & 90% = Grade B

Below 70% = Grade C

From the data above, let's create one more column to get the average grade from G1 to G3 (3 years average):

Proposed Solutions:

Feature selection is applied to select a subset of features that have a greater impact on student academic performance. Moreover, the subset produced by feature selection allows classifiers to reach optimal performance and can be a helpful solution for imbalanced class distribution in the dataset. Therefore, six different filter and wrapper methods are applied on the student dataset. Three filter methods we will apply to include Correlation Attribute Evaluation, Information Gain Attribute Evaluation, Decision Tree , Naive Bayes, and K-Nearest Neighbor are used to implement the wrapper method. The results of these feature selection algorithms show that assessment grades are the most important features that affect student academic performance. To look into possible solutions further, we implemented a K nearest neighbors model, Gaussian Mixture model, K-means clustering model, SVM and K-fold Model.

Initial analysis:

Our initial analysis is based on the general analysis that we did during the 3rd stage where we looked for correlations between individual attributes and their relation to their scores in G1, G2, and G3. We found that the school GP had better overall average scores throughout each grading period for both Math and Portuguese compared to the school MS. Men did better in Math, while women did better at portuguese. There was a pretty strong correlation between the number of failed classes and the final grades, when the students had failed one or more classes they had a high likelihood of receiving a low final grade. The higher the level of education the father had attained led to the students having a higher grade in portuguese regardless of student sex. There was a strong correlation between study time and final grades with portuguese. There was no correlation between the number of absences and the final grade of students.