Data Collection, Cleaning, and Analysis of Gamers at Kutztown and their social habits

Connor Ellis

- **4.2.a What is the source of your data? Include any links or references to the data source.**
  - The source for my data is from my Independent Study advised by Patrick Earl. The data comes from a google form that my 'gamer' subjects filled out during the 10th week of this semester.
- **4.2.b What is your intended goal in analyzing this data set? Are you extending previous analysis or starting new analysis?**
  - I am starting new analysis. To my knowledge, Gamers at KU have never had data collected about their social habits. I intend to use this data in my independent study to find if there are profound and predictable connections between different aspects of gamers' lives.
- **4.2.c What steps have you taken so far to get the dataset to its state for 4.1 above? What problems did you encounter?**
  - Apart from spending hours creating fair survey questions and ordering them correctly, I have converted the file from an .xls to a .csv and imported it into WEKA. I have some data cleaning to do before I can properly do some classifying.
  - After using RemoveUseless to remove my two disclaimer questions. I used MergeManyValues to merge many of my 'Major' and 'Minor' instances together into standardized groups. I went from ~40 unique majors to 18, and ~25 unique minors to 13.
- **4.2.d How could the results of the analysis be used in a commercial or research setting?**
  - These results could be used to study student retention at Kutztown. When students are involved in clubs, they tend to stay at KU longer. The school could use this data to see how being a gamer affects a students social life, academic life, and health. The same questions could also be given to traditional athletes to compare against gamers.
- **4.2.e What machine learning / modeling techniques do you anticipate using? Nominal classification, numeric estimation, other?**
  - How do(es) the planned modeling technique(s) relate to 4.2.d?

        - I plan on performing numeric estimation. I am interested in how well regressive models can make predictions on a few target attributes from data that I collected.
    - Identify the modeling tool or tools you plan to use.
        - I plan on using some of the following regression models in WEKA as my main regression tool. ZeroR, REPTree, SMOreg, MulitlayerPerceptron, and GaussianProcesses.
- **4.2.f Document any other aspect of the project that you feel is important to communicate.**
    - This data and some of the work is also part of my independent study with Partick Earl.
- **5.2.a What additional data did you collect during analysis, if any? Include any links or references to the data source.**
    - You can include compressed data files to D2L if it accepts them. A URL is good enough.
        - There was no additional data collected for this project.
- **5.2.b Did you achieve your intended goal in analyzing this data set? Explain how analysis shows goal achievement or refutation.**
    - Include classification results and explain how they achieve, refute, or otherwise relate to your goals.
        - Using the Apriori Associator, I was able to determine some interesting/profound rules about my data. My intent with collecting this data is to discover patterns in gamers
        - Preliminary Associations created some obvious answers. I feel some of them are still worth mentioning as they were a good starting place to learn about how association in my data set works. Below are some of the most obvious associations and what they mean in english.
            - Class=Freshman Gender=Male 26 ==> Semesters Completed at KU=0 - (Have not yet completed a semester at Kutztown) 26 <conf:(1)> lift:(2.78) lev:(0.16) [16] conv:(16.66)
                - If you are a Freshman Male, you also have zero semesters completed at KU.
            - Class=Freshman Current GPA=Not yet assigned a GPA 30 ==> Semesters Completed at KU=0 - (Have not yet completed a semester at Kutztown) 30 <conf:(1)> lift:(2.78) lev:(0.19) [19] conv:(19.22)
                - If you are a freshman without a GPA, you have completed zero semesters at KU

        - After removing attributes that create obvious rules (gender was almost all male, major was almost all CS/IT, setup was almost all yes) I was able to create some interesting rules. Here are some of those rules. Below each rule I have the english translation of the rule in the context of my survey questions.
            - Current GPA=3.50 - 4.00 20 ==> On average, EVERY WEEK, how many DAYS do you party?=0 20 <conf:(1)> lift:(1.3) lev:(0.05) [4] conv:(4.66)
                - Every person that has a GPA of 3.50 - 4.00, also parties zero days a week on average.
            - Simulation & Sports Games=1 20 ==> On average, EVERY WEEK, how many DAYS do you party?=0 20 <conf:(1)> lift:(1.3) lev:(0.05) [4] conv:(4.66)
                - Every person that enjoys Sports/Simulation games at a 1/10, also parties zero days a week on average.
            - Sleep on NON-School Night=1:00:00 AM Action-Adventure Games=10 11 ==> Do you have a job?=Yes 10 <conf:(0.91)> lift:(1.91) lev:(0.05) [4] conv:(2.88)
                - At a very high rate, if you go to sleep at 1AM when you do not have class the next day, and you like Action-Adventure games at a 10/10, you also have a job
            - HOW do you play those games?=4 41 ==> Gender=Male 39 <conf:(0.95)> lift:(1.14) lev:(0.05) [4] conv:(2.26)
                - At a 95% rate, if you prefer to play games in a competitive manner at a 4 / 5, your gender is Male
- **5.2.c What machine learning/modeling steps have you taken?**
    - Show classification / regression results. Show filtering used. Explain it using the detail that I use in assignments 1-3 solutions.
        - Filtering
            - RemoveUseless: Removed both of my IRB-required questions about consent for the study.
            - StringToNominal: Took all my string data, which was quite a lot, and merged them into nominal categories
            - NumericToNominal: This takes all the remaining non-nominal data and standardizes it.
            - SortLabels: Helps me visualize the data in the preprocess tab.

        - MergeManyValues: This was very important to the string data that users had entered for the survey. The most important attributes that I merged were, Major, and Minor.
    - Classification/Regression Results
        - **ZeroR**

```
ZeroR predicts class value: 3.00 - 3.49

Correctly Classified Instances              30                   29.4118 %
Incorrectly Classified Instances            72                   70.5882 %
Kappa statistic                    -0.0286
Mean absolute error                0.2511
Root mean squared error                     0.3534
Relative absolute error            100      %
Root relative squared error                 100      %
Total Number of Instances                   102
```

        - **REPTree**

```
Class = Senior
|  Sleep on School Night = 12:00:00 AM : 1.50 - 1.99 (3/2) [0.23/0]
|  Sleep on School Night = 12:20:00 AM : 3.00 - 3.49 (0/0) [1/0]
|  Sleep on School Night = 12:30:00 AM : 3.50 - 4.00 (1/0) [0.08/0.08]
|  Sleep on School Night = 1:00:00 AM : 3.00 - 3.49 (4/0) [2.31/0.31]
|  Sleep on School Night = 1:30:00 AM : 3.50 - 4.00 (0/0) [1/0]
|  Sleep on School Night = 2:00:00 AM : 3.00 - 3.49 (1/0) [0.08/0.08]
|  Sleep on School Night = 2:30:00 AM : 3.00 - 3.49 (1/0) [0.08/0.08]
|  Sleep on School Night = 3:00:00 AM : 3.00 - 3.49 (0/0) [1/0]
|  Sleep on School Night = 3:30:00 AM : 3.00 - 3.49 (0/0) [0/0]
|  Sleep on School Night = 4:00:00 AM : 3.00 - 3.49 (0/0) [0/0]
|  Sleep on School Night = 9:30:00 PM : 3.00 - 3.49 (0/0) [0/0]
|  Sleep on School Night = 10:00:00 PM : 3.50 - 4.00 (1/0) [0.08/0.08]
|  Sleep on School Night = 10:30:00 PM : 3.00 - 3.49 (0/0) [0/0]
|  Sleep on School Night = 11:00:00 PM : 3.00 - 3.49 (2/1) [1.15/0.15]
|  Sleep on School Night = 11:30:00 PM : 2.50 - 2.99 (0/0) [1/0]
|  Sleep on School Night = 11:59:00 PM : 3.00 - 3.49 (0/0) [0/0]
|  Sleep on School Night = 11:15:00 PM : 3.00 - 3.49 (0/0) [0/0]
Class = Junior
|  Major = General Business  : 2.50 - 2.99 (0/0) [0/0]
|  Major = Arts Administration : 2.50 - 2.99 (0/0) [0/0]
|  Major = History : 2.50 - 2.99 (1/0) [0/0]
|  Major = Accounting : 2.50 - 2.99 (0/0) [0/0]
|  Major = Social Media Theory and Strategy : 2.50 - 2.99 (0/0) [0/0]
|  Major = Professional Writing : 3.50 - 4.00 (1/0) [0/0]
|  Major = Computer Science and Physics : 3.50 - 4.00 (1/0) [0/0]
|  Major = Computer Science : 3.00 - 3.49 (6/4) [8/4]
|  Major = Undeclared : 2.50 - 2.99 (1/0) [0/0]
|  Major = Criminal Justice : 2.50 - 2.99 (1/0) [0/0]
|  Major = Cinema, Television, and Media : 2.50 - 2.99 (1/0) [0/0]
|  Major = Applied Digital Arts : 3.50 - 4.00 (1/0) [1/0]
|  Major = Psychology : 2.50 - 2.99 (2/1) [0/0]
|  Major = Secondary Education : Not yet assigned a GPA (1/0) [0/0]
|  Major = Geography : 2.50 - 2.99 (0/0) [0/0]
|  Major = Communication : 2.50 - 2.99 (0/0) [0/0]
|  Major = Political Science and German : 2.50 - 2.99 (0/0) [0/0]
|  Major = Sports Management : 2.50 - 2.99 (0/0) [0/0]
Class = Freshman : Not yet assigned a GPA (20/0) [11/1]
Class = Graduate Student : 3.50 - 4.00 (7/3) [2/1]
Class = Sophomore : 3.00 - 3.49 (12/4) [4/3]
```

```
Correctly Classified Instances          61               59.8039 %
Incorrectly Classified Instances        41               40.1961 %
Kappa statistic                  0.4422
Mean absolute error              0.1734
Root mean squared error                 0.3211
Relative absolute error          69.0641 %
Root relative squared error             90.8704 %
Total Number of Instances               102
```

- **MulitlayerPerceptron**

Model too large for this document

```
Correctly Classified Instances          58               56.8627 %
Incorrectly Classified Instances        44               43.1373 %
Kappa statistic                  0.417
Mean absolute error              0.1494
Root mean squared error                 0.3435
Relative absolute error          59.4991 %
Root relative squared error             97.193  %
Total Number of Instances               102
```

- **OneR**

```
Class:
   Senior    -> 3.00 - 3.49
   Junior    -> 2.50 - 2.99
   Freshman    -> Not yet assigned a GPA
   Graduate Student    -> 3.50 - 4.00
   Sophomore    -> 3.00 - 3.49
(66/102 instances correct)
```

```
Correctly Classified Instances          66               64.7059 %
Incorrectly Classified Instances        36               35.2941 %
Kappa statistic                  0.5245
```

```
Mean absolute error                  0.1176
Root mean squared error                        0.343
Relative absolute error              46.8485 %
Root relative squared error                    97.0538 %
Total Number of Instances                      102
```

- What problems did you encounter?
    - Some of the Attribute Evaluators that I used with my data were
        - CorrelationAttributeEval w/ the Ranker Method
        - OneRAttributeEval w/ the Ranker Method
        - PrincipalComponents w/ the Ranker Method
        - ReliefFAtrtributeEval w/ the Ranker Method
    - All other Evaluators and Methods either return no information or information that adds nothing of immediate value to the project. Some of the Evals and Methods cannot operate on my discrete data.
    - For my purposes I will be using Current GPA as my target attribute and…
        - CorrelationAttributeEval w/ the Ranker Method
            - Ranked attributes:
            - 0.2942 1 Class
            - 0.2824 6 Semesters Completed at KU
            - 0.1444   11 Setup
            - 0.1354   30 Do you have a job?
            - 0.1289   23 What percentage of your gaming setup was purchased with money that you earned?
        - OneRAttributeEval w/ the Ranker Method
            - Ranked attributes:
            - 64.706 1 Class
            - 58.824 6 Semesters Completed at KU
            - 41.176   30 Do you have a job?
            - 36.275   18 Puzzle & Party Games
            - 36.275   23 What percentage of your gaming setup was purchased with money that you earned?
        - ReliefFAtrtributeEval w/ the Ranker Method
            - 0.399771   1 Class
            - 0.319816   6 Semesters Completed at KU
            - 0.05229   10 Sleep on NON-School Night
            - 0.049002   24 On average, EVERY DAY, how many hours do you spend GAMING ALONE?
            - 0.045032   8 HOW do you play those games?
    - …as my Attribute Evals

- **5.2.d Use SMO, or SMOreg, or MultiLayerPerceptron, or clustering, OR at least one other technique not used in assignments 1-3.**
  - Give results and explain how this step relates to earlier steps.
    - MultiLayerPercerptron was used above
- **5.2.e Revise explaining how could the results of the analysis be used in a commercial or research setting?**
  - The results can be used to discover what aspects of a gamer's life at KU most affects their GPA.
  - The results could be compared against traditional student athletes to see where their lives affect their GPA.
  - The school could leverage this data to determine where to fund student events to help student GPA numbers.
- **5.2.f Document any other aspect of the project that you feel is important to communicate.**
  - Here are some of the interesting discoveries.
    - **Every person that has a GPA of 3.50 - 4.00, also parties zero days a week on average.**
      - While this may seem like an obvious thing to predict, it is really cool to see it using data.
    - Every person that enjoys Sports/Simulation games at a 1/10, also parties zero days a week on average.
      - People who tend to party also tend to play Sports/Simulation games. Those games are usually advertised to the 'party' crowd. Again, seeing it in data is neat.
    - At a very high rate, if you go to sleep at 1AM when you do not have class the next day, and you like Action-Adventure games at a 10/10, you also have a job
    - At a 95% rate, if you prefer to play games in a competitive manner at a 4 out of 5, your gender is Male
    - The easiest way to guess the GPA of a gamer at KU is to use their class (freshman, sophomore, junior, senior, graduate)
      - `Graduate Student    -> 3.50 - 4.00`
      - `Senior              -> 3.00 - 3.49`
      - `Junior              -> 2.50 - 2.99`
      - `Sophomore           -> 3.00 - 3.49`
      - `Freshman            -> Not yet assigned a GPA`