UNIVERSITY OF ONTARIO INSTITUTE OF TECHNOLOGY

Cluster Techniques and Prediction Models for a Digital Media Learning Environment

by

Arturo Fernandez Espinosa

A thesis submitted in partial fulfillment for the
degree of Master of Science
in
Computer Science

in the
Faculty of Business and Information Technology

August 2012

Declaration of Authorship

I, Arturo Fernandez Espinosa , declare that this thesis titled, Cluster Techniques and Prediction Models for a Digital Media Learning Environment, and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: ________________________________

Date: ________________________________

“During my life I have seen too many shiny stars losing their light... I would say that it is never too late to bright again... I am trying to take my own advice.”

Arturo Fernandez

Abstract

The present work applies well-known data mining techniques in a digital learning media environment in order to identify groups of students based on their profile. We generate identifiable clusters where some interesting patterns and rules are observed. We generate a neural network predictive model intended to predict the success of the students in the digital media learning environment. One of the goals of this study is to identify a subset of variables that have the biggest impact in student performance with respect to the learning assessments of the digital media learning environment. Three approaches are used to perform the dimensionality reduction of our dataset.

The experiments were conducted with over 69 students of health science courses who used the digital media learning environment.

Acknowledgements

I would like to acknowledge my supervisor Dr. Miguel Vargas Martin for his guidance and patience during this process. There is no words to express my gratitude for the given opportunity to materialize this document that 2 years ago was no more than a dream.

Dr. Jayshiro Tashiro, thanks for sharing your experience and insight in the educational aspects of this work. Without your contribution this work would have not been possible.

Meaghen Regts, thanks for your big effort recruiting participants and collecting the information that is one of the main parts of this work.

I want to acknowledge Dr. Miguel Vargas Martin and Dr. Jayshiro Tashiro for their brilliant work on the design of the IPSims which is the main source of information and the main motivation of this thesis work.

I also want to thank my team mates, coaches from the Ridgebacks soccer team and the good friends I have made here in Canada. Thanks for your support in the hard moments and for making the best of your efforts to make this Mexican guy feel like at home.

Thanks to my parents and family for the unconditional support in every moment. This would not be possible without all the support and the words of encouragement. Muchas gracias por todo...

This work was supported in part by the Social Sciences and Humanities Research Council of Canada (SSHRC).

Contents

Declaration of Authorship **i**

Abstract **iii**

Acknowledgements **iv**

List of Figures **vii**

List of Tables **x**

1 Introduction **1**
- 1.1 Introduction . . . 1
- 1.2 General Description of the Problem . . . 2
- 1.3 Research Questions . . . 4
- 1.4 Scope, Assumptions and Limitations . . . 5
- 1.5 Approach . . . 6
- 1.6 Summary of Characteristics of the Dataset and Data Mining Process . . . 8
- 1.7 IPSims General Description . . . 11
  - 1.7.1 IPSims First Time Users . . . 12
- 1.8 Why IPSims? . . . 14
  - 1.8.1 IPSims Tracking Engine Instead of Commercial Tools . . . 15
- 1.9 Contributions and Organization . . . 16

2 Related Literature **20**
- 2.1 Related Literature . . . 20
  - 2.1.1 Educational Data Mining . . . 20
  - 2.1.2 Cluster Analysis . . . 26

3 Data Mining Process in IPSims **30**
- 3.1 Data Mining Process on IPSims . . . 30
  - 3.1.1 Data Acquisition . . . 30
    - 3.1.1.1 IPSims Original version . . . 31
    - 3.1.1.2 Selected Variables . . . 32
    - 3.1.1.3 First Time Users and Storing of User Profile and User's activity . . . 33

3.1.2 Data Integration 41
3.1.3 Data Pre-Processing 43
3.1.3.1 Qualitative Data into Quantitative Data 44
3.1.3.2 Time Variables for lsp_tracks Database 46
3.1.3.3 Outliers Treatment 48
3.1.4 Data Mining Techniques on IPSims 50
3.1.5 Data Evaluation and Interpretation 52

4 Experiments 54
4.1 Methods 54
4.2 Dimensionality Reduction 55
4.2.1 Factor Analysis 56
4.2.2 Standard Deviation Analysis 59
4.3 Cluster Analysis Experiment 61
4.4 Prediction Model Experiment 72
4.4.1 Training Algorithms 77

5 Conclusions 84
5.1 Summary 84
5.2 Discussions 85
5.2.1 Overfitting 86
5.2.2 Clustering 94
5.3 Contributions 109
5.4 Closing Remarks 111
5.5 Future Work 112

A 113

B 144

C 148

D 152

E 155

Bibliography 158

List of Figures

1.1 Data mining workflow Han et al. [1] 7
1.2 IPSims login screen 12
1.3 IPSimsMainMenuScreen 13
1.4 IPSimsScenarioTemplate 14

2.1 Results of the predictions for the Sembiring et al. paper 23

3.1 IPSims original datasources merging diagram 34
3.2 IPSims architecture and persistence flow 35
3.3 lspl_users figure 38
3.4 Table Generation 40
3.5 New database figure 42
3.6 Data Integration Process 44
3.7 Outliers first approach 49
3.8 Outliers second approach 50
3.9 Linear Separability with the use of NN 52
3.10 Linear Separability with the use of SVM 52

4.1 Experimentation process workflow 56
4.2 Scree plot for our factor analysis 58
4.3 Distribution of the variables age 61
4.4 Distribution of the variable graduation year from high school 61
4.5 Distribution of the variable play hour per week 62
4.6 Distribution of the variable surf hour per week 62
4.7 Vector class definition 64
4.8 Cluster class definition 64
4.9 High school average vs grade by cluster 67
4.10 Intention to continue post graduate studies after UOIT vs grade 68
4.11 Grade vs surf hour vs gender 3D Plot 69
4.12 Grade vs high school average vs age 3D plot 69
4.13 Surface grade vs computer literacy 70
4.14 Surface grade vs surf hour peer week 70
4.15 Surface high school average Vs grade 71
4.16 A diagram of a Perceptron 75
4.17 Feed forward network representation 75
4.18 Plot generated for the conjugate gradient algorithm 81
4.19 Plot generated for the conjugate gradient algorithm 82
4.20 Plot generated for the genetic training algorithm 82

4.21 Plot generated for the distributed training algorithm 83

5.1 Training plot for the genetic training algorithm using standard deviation analysis dataset 86
5.2 Training plot for the conjugate gradient training algorithm using standard deviation analysis dataset 86
5.3 Training plot for the distributed training algorithm using standard deviation analysis dataset 87
5.4 Testing plot for the genetic training algorithm using standard deviation analysis dataset 88
5.5 Testing plot for the conjugate gradient training algorithm using standard deviation analysis dataset 88
5.6 Testing plot for the distributed training algorithm using standard deviation analysis dataset 89
5.7 Training plot for the genetic training algorithm using the components from factor analysis 90
5.8 Training plot for the distributed training algorithm using the components from factor analysis 90
5.9 Testing plot for the distributed training algorithm using the components from factor analysis 91
5.10 Training plot for the conjugate gradient training algorithm using the components from factor analysis 91
5.11 Testing plot for the conjugate gradient training algorithm using the components from factor analysis 92
5.12 Training plot for the genetic training algorithm using correlation, significance and variance analysis dataset 92
5.13 Training plot for the conjugate gradient training algorithm using correlation, significance and variance analysis dataset 93
5.14 Training plot for the distributed training algorithm using correlation, significance and variance analysis dataset 93
5.15 Testing plot for the genetic training algorithm using correlation, significance and variance analysis dataset 94
5.16 Testing plot for the conjugate gradient training algorithm using correlation, significance and variance analysis dataset 94
5.17 Testing plot for the distributed training algorithm using correlation, significance and variance analysis dataset 95
5.18 College or university education before UOIT vs. grade by Cluster 97
5.19 Computer literacy vs. grade by Cluster 98
5.20 Expected grade in the course vs. grade by Cluster 98
5.21 Experience vs. grade by Cluster 99
5.22 High school average vs. grade by Cluster 99
5.23 Interest in course material vs. grade by Cluster 100
5.24 Likelihood for choosing a career in health sciences vs. grade by Cluster . . 100
5.25 Perceived value of computer based simulations vs. grade by Cluster 101
5.26 Play hour peer week vs. grade by Cluster 101
5.27 Intention to pursue graduate education after UOIT vs. grade by Cluster . 102
5.28 Expected grade for the course vs. grade by Cluster 102
5.29 Interest in course material vs. grade by Cluster 103

5.30 Play hour peer week vs. grade by Cluster 103
5.31 Surf hour peer week vs. grade by Cluster 104
5.32 College or university education before UOIT vs. grade by Cluster 104
5.33 3-D plot for the age, cluster and surfhour 105
5.34 3-D plot for the age, cluster and high school average 105
5.35 3-D plot for cluster, expected grade in the course and grade 106
5.36 3-D plot for cluster, perceived value for computer based simulations and experience with computer based simulations 106
5.37 3-D plot for cluster, computer literacy and likelihood in choosing a career in health sciences 107
5.38 Distribution by cluster and high school average 107
5.39 Surf hour peer week distribution by cluster 108
5.40 Computer literacy distribution by cluster 108
5.41 Age distribution by cluster 108

List of Tables

3.1 First refinement of lspl_tracks 47
3.2 last refinement for time stamps 48

4.1 Variables analyzed in factor analysis 59
4.2 Standard Deviation Analysis Table 60
4.3 Cluster Analysis Table for a Generated Cluster 67
4.4 Grading rank table 79
4.5 R-squared and MSE for each training algorithm using our correlation, significance and variance analysis dataset 80

Dedicated to my parents and sisters for being my inspiration to keep going in the pursuit of my dreams... You guys are the key of this achievement. The sacrifice of being far from you is paying off, this document is the materialization of our team work. Step by step we are reaching our goals. Your teachings and love have make me become on what I am now. There is not enough words in this language or any other to express all my gratitude and love for you. Los amo. . . .

Chapter 1

Introduction

1.1 Introduction

Data mining is the emerging science created as a necessity for our understanding of information. With the advent of the digital era we are surrounded by huge amounts of data stored in digital media units known as databases. Unfortunately in most of the cases these data are useless given the fact that humans are not able to process and extract knowledge from them using conventional methods. With the advances in technology and the growing speed of computational power, new knowledge areas in data understanding have been opened. A clear example of this is our area of study known as educational data mining, which is the discipline, generated to extract knowledge from educational datasets.

We use in this work the set of words: knowledge extraction and knowledge discovery as the final result of applying data mining techniques and acquire valuable, usable and easy-to-understand information to present to the final user.

Data is present everywhere and in all the imaginable forms. From satellite images to genetic chains, from simple numeric variables to binary strings representing a natural phenomenon. Data is diverse and sometimes untreatable for non-digital methods. The intention of this study is to apply data mining techniques in a specific data environment called IPSims in order to discover valuable knowledge from the available information in our digital media learning environment IPSims. Applying data mining techniques to our digital learning environment we can have an understanding of the variables that may have the biggest impact in the health sciences student's performance in the IPSims activities. This will help the instructors to have an idea of which students are more likely to be successful and give an adequate input to the ones that are not.

Data is present in multiple forms, some of them digital and some other in non-digital formats the diversity and nature of all these datasets our computer systems cannot process them in most of the cases without giving special computational and statistical treatments to them. The generated process to select, treat computationally, statistically and the discovery of patterns is known as data mining. Then it is the obligation of the data mining specialist to create and apply techniques for data pre-processing in order to make the input data understandable and high-quality for the computer. In this data mining study one of our purposes is to show how we can apply data mining techniques in order to generate usable and high quality data for data mining algorithms in digital media learning environments of a similar nature.

The present work focuses on what we call educational data; this makes the present study a study in educational data mining. There are multiple works in educational data mining, some of them try to generate effective predictive models to understand why some students are successful. Another kind of works show a comparison between techniques to classify students according to some criteria and some others make the comparison of predictive models in educational environments. The importance of these kinds of studies can be visualized as the opportunity to provide the student with adequate learning feedback in their courses, as well as to give the instructor the opportunity to take action before the undesirable result of a student failing a given course happens. We can find diverse research topics related to the educational data mining all of them with remarkable importance in the educational area as well as in the computer sciences.

One of our main goals is to generate an accurate neural network prediction model capable of predicting the final grade of a student within the IPSims activities based on a user profile generated for our digital media learning environment. Diverse statistical and computational techniques are applied in order to accomplish this goal. Before we can predict the final grade of the user within the IPSims activities we want to identify which variables are the ones that have a more direct correlation with the student performance in the activities. This might allow us to identify the user profiles that are more likely to display a poor performance in the IPSims activities.

1.2 General Description of the Problem

Digital databases provide huge amounts of information on diverse areas and disciplines. This information is hard-to-use unless we can apply techniques that provide us with useful and easy to understand information. This retrieved information is commonly called knowledge. This knowledge is retrieved by data mining techniques. Data mining is an emergent discipline of computer science that includes the extraction of patterns from large datasets.

Before we can accomplish with knowledge discovery we need to generate a data mining framework based in a structured process that allows us to extract the required information to be processed for our data mining algorithms, then it is one of our main problems to generate a data mining framework in order to extract patterns that are potential knowledge. We can discover diverse patterns but just some of them will become knowledge, these selections of knowledge depend in our interpretation and the visualization techniques used in the discovered patterns.

IPSims is a media learning environment applied in a real life scenario with students of health science with the original intention to study and understand learning misconceptions. Misconceptions can be defined as the misunderstanding of the lectures in classroom, the difficulty to identify patterns or take appropriate decisions based in acquired knowledge. Misconceptions can be formed by past experiences, incorrect past teaching or other possible factors.

We believe that there are diverse factors that contribute to the success of a student in a course or in their academic career. Regts et.al. [2] cite some of the believed gaps in human learning process and misconceptions is one of these gaps. In order to understand and fill the learning gaps presented during the learning process in digital media learning environments we need to understand which factors are affecting the student performance in our digital media learning environments activities. We need to identify patterns in the information stored in our system with the objective to identify a group of variables that may have the biggest impact in the student's grade in the IPSims activities. The identification of these factors may result in a better understanding of why the students with certain characteristics are being successful in these activities.

The design of a data mining framework required address carefully diverse stages, these stages can be classified as part of our problem in the process of knowledge discovery.

Then we will say that given the nature of our data, which fits into the educational data, it is our intention to generate a complete data mining framework that will serve to two specific purposes of knowledge discovery. These two specific purposes are to provide knowledge related with the user profile and the final grade in the IPSims activities.

The generation of a complete data mining framework for our educational digital media learning environment arise diverse issues that have to be considered such as: The nature of the collected educational data, the problems that could raise when collecting the data, the fact that educational datasets are usually small in size, etc. In our present scenario for a digital media learning environment inappropriate data treatment and interpretation leads to trivial and inconclusive results.

The nature of the media learning environment provide us with a large amount of information related to user profile and description of user activities in the system; then this capability makes attractive the possibility to exploit the information in order to retrieve valuable knowledge that might support different assumptions related with educational aspects.

We understand for the term “effective exploit of the information” the discovery of useful patterns in our IPSims database. Useful patterns means patterns that serves to a specific purpose. They accomplish with specific requirements and standards, they are easy-to-understand for the final user and overall they are a valid and trustable source of information for further research works related with educational user behavioural aspects.

Even when it is our intention to apply our data mining framework to two specific educational data mining applications, we want to provide knowledge that can be used by the final user in diverse forms according to their needs and requirements. Based in the two specific educational data mining applications we generate our research questions, without forgetting that the main intention of this research work is to provide a data mining framework for knowledge discovery within the IPSims.

Based in the previous paragraphs we can say that the motivation of this research work is to extract knowledge from our digital media learning environment IPSims. The IPSims system is a framework that offers a complete platform for diverse studies and analysis including the generation of a complete data mining system for knowledge discovery and applications in the educational data mining.

1.3 Research Questions

Based on the past paragraphs we know that our main intention is to generate a data mining framework for the IPSims and based on this data mining framework we want to provide knowledge for two specific educational data mining applications that might be answer for the next research questions:

- Is there an identifiable subset of user profile obtained from IPSims variables that may have the biggest impact in the user final grade?
- Is it possible to generate an accurate neural network prediction model for grade prediction in the IPSims activities?

1.4 Scope, Assumptions and Limitations

In this research work it is not our intention to make a contribution in educational aspects. Even when the environment that provides the data for this research is an educational digital environment we are more focused in the computational processes and techniques that are required to extract patterns and knowledge from the educational databases of our system in study IPSims.

We want to test our data mining framework in two specific applications of educational data mining, as a computer science thesis work the educational area is not the main topic of this research, then we draw conclusions and perform our analysis based on the output of our data mining algorithms. The applications where our data mining frameworks is applied are predictions and classifications, however the number of study subjects is small to provide conclusive results for our applications in the educational data mining, but the amount of study subjects is enough to set the starting point for future studies related with IPSims and may reinforce the results presented in this first stage of our data mining framework.

The processes and transformations applied to the data establish a consolidated source of information for IPSims that can be used later on to apply diverse machine learning algorithms in order to discover diverse patterns, or simply compare techniques and results. The present study gives two applications:

- Prediction of the final grade of the student in the activities of IPSims based in the student profile.
- Identification of clusters of students according to their user profile.

These two applications retrieve valuable knowledge that the instructors and students can use in diverse ways according to their needs. As mentioned before, the educational area is not one of our main topics of research. Our contribution is a set of processes to generate a suitable data mining framework from where we can apply well-known techniques to produce valuable knowledge to the educational field. Our data mining framework is created specifically for the context of IPSims and the analysis of the targeted information contained in it. While IPSims is a comprehensive digital learning environment the present work only uses certain portions of the IPSims database. Nevertheless we were able to prepare the entire database for future work not only including follow ups of the present one but any other study involving the IPSims database such as: students behaviour, social network analysis, correlation between student profile and navigation preferences.

1.5 Approach

The solution to the task of "effective exploitation of the information" raises multiple problems. These problems are classified and addressed as follows:

- Testing of the elements of our digital learning media environment (speed, bugs,).
- Error prevention through the User Interface (user interface elements validation).
- Error prevention in the data storage media (database).
- Error prevention in user and system communications (networking and stress tests).
- Appropriate selection of the data source (students of a specific course).
- Definition of an appropriate data model (database architecture).
- Collect the potential input data for the proposed data model (recruitment).
- Clean and heal the dataset (data pre-processing).
- Adequate data extraction from our database (extraction scripts).
- Normalization and formatting of the dataset (data transformation).
- Patterns discovery through data mining techniques (machine learning algorithms).
- Interpretation and validation of the results (plots and conclusions).

The previous steps are addressed in order to reach our goal of discovering useful patterns in an educational data environment.

Our data exploitation process does not mean just extracting information from digital media. Exploitation involve an entire process from validation of user interfaces, verification of the system communications with the user (networking response tests), bug detection, definition of our data model, selection of the data source and the adequate formatting and normalization techniques, application of machine learning algorithms and finally the correct interpretation and visualization of our results. This entire process is what we call data exploitation.

With the completion of our data exploitation process we are opening a window to other research studies that can involve:

- Generation of predictive models.

- Identify specific groups of students according to their user profile or behaviour (clustering).
- Generation of reduced datasets which may have a direct correlation with specific user features and the IPSims.
- Generation of association rules.
- Identify tendencies in users behaviours within the IPSims.

The efficient exploitation of information involves a well-defined, justified, and organized process. The proposed approach to solve our tasks is to follow this organized structured process for information exploitation and in order to fully exploit the capabilities of our system it is necessary to create appropriate mechanisms that allows information retrieval and knowledge discovery in the IPSims.

In order to meet our objectives data mining techniques are applied following the methodologies presented in the literature. These methodologies are already proven and are the definition of a formal process known as knowledge discovery in databases. Figure 1.1 show a common data mining process.

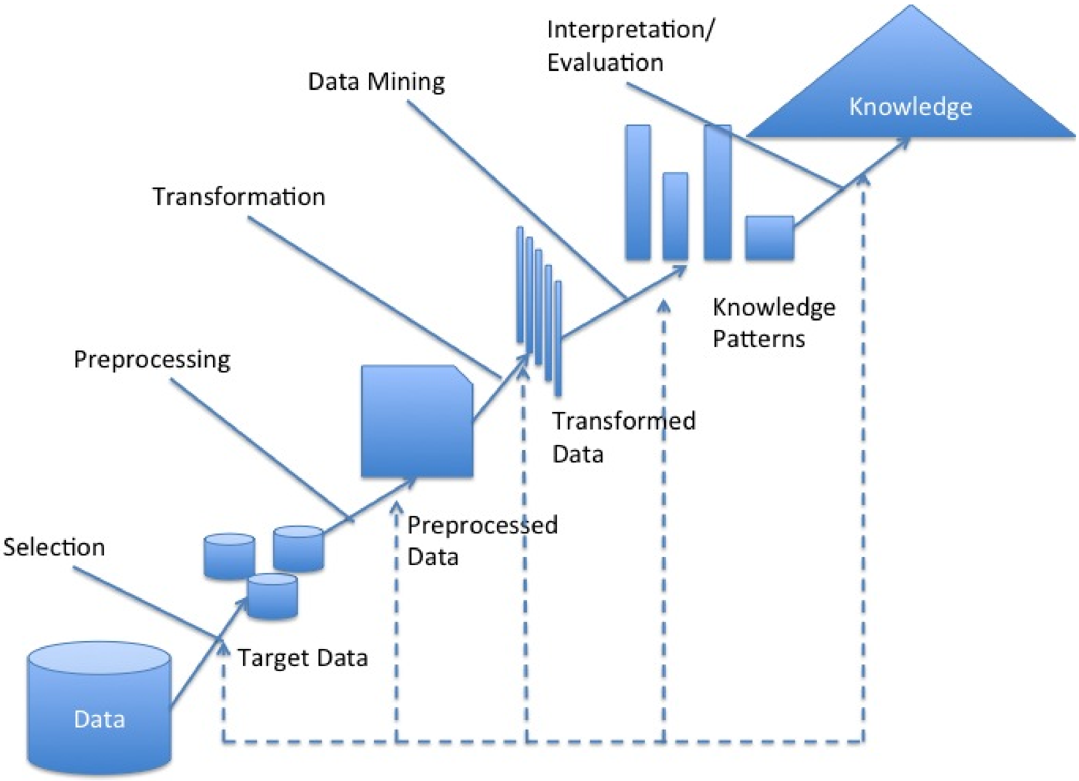


FIGURE 1.1: Data mining workflow Han et al. [1]

We adopt this approach and modify it according to our needs in order to discover interesting patterns that eventually we conceptualize in knowledge. This knowledge intends to be an easy to understand source of information for the instructor and students.

1.6 Summary of Characteristics of the Dataset and Data Mining Process

Our information repository is basically made of three sources: 1)the user profile stored in a digital database, 2) users choices and preferences as they use IPSims environment, and 3) student surveys collected after completion of all the learning activities performed in IPSims. The student surveys were designed for Meaghen Regts one of the members of this research team. Without her outstanding effort the consolidation of our database would not have been possible. The student surveys include nine different sets of questions, namely 1) learning assessment based on assigned learning activities, 2) demographic information, 3) preferences for learning resources and educational scaffolding, 4) rating of web-base course work, 5) rating of IPSims learning environment, 6) satisfaction with educational simulations and serious games, 7) disposition to engage effortful cognitive endeauvor (need for cognition and ambiguity tolerance), 8) expectancy-value questionnaire, and 9) performance evaluation in inter-professional learning activities. These three sources of data were collected after applying the system in two different health science nutrition and health related courses in the University of Ontario Institute of Technology (UOIT). 69 students were recruited for the study.

The possibility to start applying our system as an educational supportive tool in the selected courses was planned and approved with the professors of each class, and according to University of Ontario Institute of Technology Research Ethics Board. Dr. Miguel Vargas Martin and Dr. Jayshiro Tashiro were the authors of the Research Ethics Board document. Their expertise was critical in the realization of this thesis work.

Once the data were collected in IPSims database and a paper questionnaire. See Appendix A. The process of the knowledge discovery was applied step by step.

The steps applied to the information collected for the two courses in health sciences (including the data acquisition process mentioned in Chapter II) were:

- System validation(Testing)
- Data acquisition
- Data integration

- Data pre-processing
- Data mining techniques on IPSims
- Data evaluation and interpretation

These five steps summarize a standard data mining process for our specific study case. Each stage has to be accomplished carefully. We want to eliminate inconsistencies or bias in our data but at the same time we are trying to keep as much as possible the integrity of our original dataset. After we collect the data it is necessary to store all the information in a single digital database (Data warehousing) in order to be able to extract the required information for our study in an efficient way. Multiple data treatments are required over our original databases and the information stored in paper questionnaires has to be dumped into a digital database that eventually needs to be merged with the information stored in the IPSims database.

The merging and integration process of our sources of information is developed with the use of*Python* and *SQL* queries. The database management system storing the data is *Postgres* and the reason of this choice is that the original IPSims database was developed on *Postgres* and it makes easier to us to integrate the information over the same technology which IPSims was originally developed. The *Python* language is selected given its simplicity and it also provides easy integration with *Postgres* through modules like *psycopg2*. The *psycopg2* module allows us to manipulate *Postgres* databases in an easy way with few lines of code.

Once we consolidate our data warehouse model (Data integration) and we populate it with the appropriate information we can proceed to pre-process our data (Data pre-processing). This stage of our knowledge discovery process is accomplished with the use of diverse techniques to eliminate outliers, inconsistencies and bias in our dataset. An important part in this stage is the data reduction analysis.

Our original dataset is not proper to conduct an analysis (or at least a one that does not retrieve trivial results). The original data model including the paper questionnaire questions contains more than 100 variables and the study subjects are just 69, eventually this 100 variables would be truncated for our expert in education. Under this condition any type of predictive analysis are inconclusive and trivial given that the ratio between variables and study cases is not adequate. Studies about prediction models recommend around 20 study cases for each variable ideally or at least five study cases for each variable. Unfortunately these studies are dependent of the environment and the type of data. A different group of studies related with study subject vs variable ratio indicates that is a better approach to base the quality of the predictions using the size of the sample as an indicator. They established that less than 100 will retrieve really poor prediction results. Our dataset presents 69 study cases and originally more than 20 variables for our user profile. This means in order to keep an adequate study cases - variable ratio we need at least 100 study cases to keep just an acceptable ratio of five:one this means five study cases for each variable. Fortunately in our study we can generate different reduced datasets based on standard deviation analysis, correlation, significance, variance analysis and factor analysis the results of the test for dimensionality reduction are satisfactory keeping the validity of our research but in counter case by now we are not able to generate more study cases, then if we based the assumptions of obtaining good predictions just based in the approach of the sample size, since now we could make the assumption that our predictions will not be satisfactory. But as mentioned before these studies are highly related to the nature of the data where they are applied.

Once we obtained a reduced dataset of acceptable quality we can apply diverse machine learning algorithms over our dataset (data mining techniques on IPSims). Diverse techniques are applied such as cluster techniques and neural network predictive models. The intention of applying machine learning algorithms is to discover interesting patterns for our research. Those patterns are evaluated and interpreted in order to prove the validity of our results and our assumptions.

Diverse machine learning algorithms and statistical techniques could be apply at this point in order to identify patterns: Multiple regression, decision trees for predictions, support vector machine, neural networks are all of them techniques for prediction purposes. In the case of clustering we could apply diverse well-known algorithms as well. The decision to use neural networks and K-means is based on their simplicity and effectiveness.

For the case K-means we know that we can guarantee a fast convergence to a local optimum while we can evaluate easily and effectively the performance of our algorithm given its nature based on Euclidean distance optimization. For this specific case where our assumptions are based in groups of grades related to the Grade Point Average the nature and a specific user profile for each possible grade K-means allows us to select the number of clusters to be generated this is an ideal situation for our purposes.

In the case of our prediction model we selected neural networks over other possible prediction techniques based on the ability of neural networks to describe non-linear correlations, to handle large number of features, fast application over our educational environment, they are easy to implement, they are non probabilistic and they have the learning ability through the training process.

There are a number of studies where neural networks are compared against multiple regression analysis and neural networks usually presents the best results. Multiple linear regression based the predictions in generated variables that describe relationships between the existing variables. The linear regression fits a straight line to a set of data values, which makes linear regression incapable to describe complex non-linear relationship between variables.

The data in multiple linear regressions is modeled by linear prediction functions the same kind of functions used in the perceptron. But the neural network approach differs from multiple linear regression for the fact that neural networks are able to describe non linear relationships. A neural network can be seen more as a black box that delivers results without a clear explanation in how these results were computed but what we know is that neural networks solve the linear separability problem which fits more our research purposes.

While we require discarding some variables, based in diverse methods the versatility of neural networks to describe more complex relationships is what we required for this study. These characteristic added to the fact of an easy implementation was the facts that determine the selection of a neural network as a prediction tool for our experiments.

With linear regression we try to explain relationships between the variables, our intention in this part of the study is to predict based in observed cases. A neural network approach is a more powerful and simplistic tool in this environment.

Finally our study concludes with the evaluation and interpretation of the results obtained after all the previous stages. The evaluation and interpretation are accomplished with the use of visualization techniques such as plots also we make use of some statistics measures for our prediction model evaluation and cluster generation quality.

1.7 IPSims General Description

IPSims is a digital learning environment designed to support and enhance students learning experience. IPSims is an interactive environment supported by multimedia materials like videos, medical profiles, local and external information sources used in an effective way in order to facilitate students learning approach. IPSims has different functionalities to support the instructor and faculty members in the under- standing of students academic performance. IPSims tracking algorithm keep track of the activity of the students within IPSims, as well IPSims persists a user profile. Then we can summarize the IPSims digital media learning environment as follows:

- IPSims is a digital learning environment designed to support and enhance students learning experience.

- IPSims is an interactive environment supported by multimedia materials including videos, medical profiles, local and external information resources, that are presented to the user in an effective way with the objective to facilitate students learning approach.

- IPSims is a complete educational suite which supports the instructor and faculty members in the understanding of students academic performance.

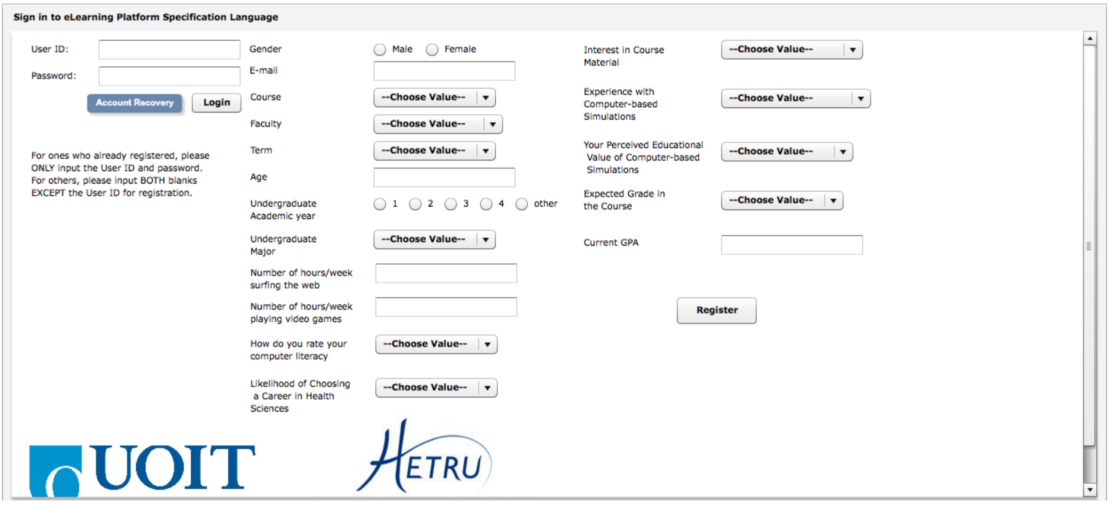
- IPSims is a tracking tool that apply an advanced tracking algorithm which trace the activity of the students in the system.

- IPSims is a data warehouse for educational data.

- IPSims is a search engine for user related information in the system.

1.7.1 IPSims First Time Users

IPSims first-time users get registered in the system with a user ID that is generated automatically after they fill the user profile and select a password. The information required by the system is obtained through the use of a web form generated by FLEX technology [7]. See Figure 1.2 shows the login screen for IPSims.



Sign in to eLearning Platform Specification Language

User ID:

Password:

For ones who already registered, please ONLY input the User ID and password. For others, please input BOTH blanks EXCEPT the User ID for registration.

Gender ○ Male ○ Female

E-mail

Course --Choose Value--

Faculty --Choose Value--

Term --Choose Value--

Age

Undergraduate Academic year ○ 1 ○ 2 ○ 3 ○ 4 ○ other

Undergraduate Major --Choose Value--

Number of hours/week surfing the web

Number of hours/week playing video games

How do you rate your computer literacy --Choose Value--

Likelihood of Choosing a Career in Health Sciences --Choose Value--

Interest in Course Material --Choose Value--

Experience with Computer-based Simulations --Choose Value--

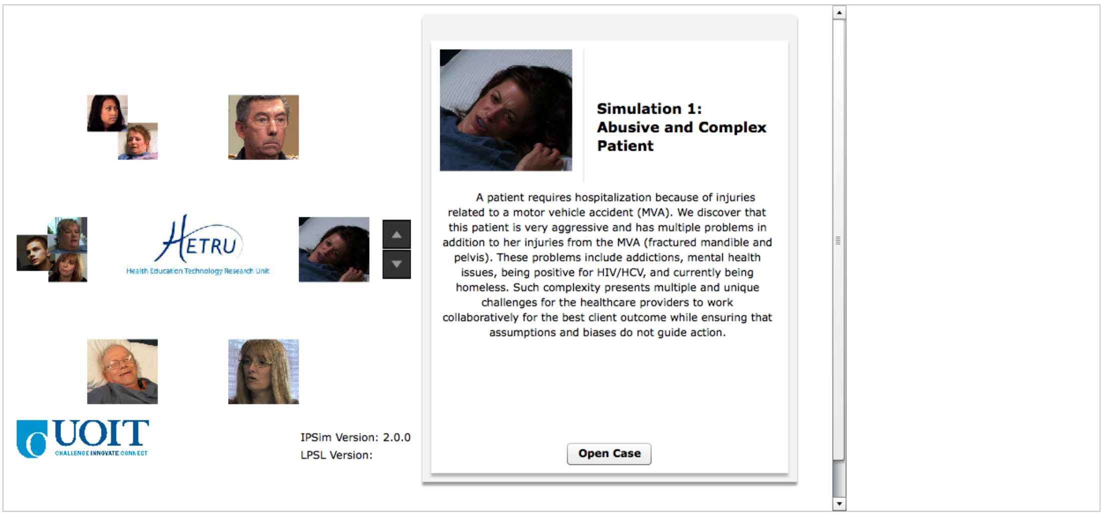
Your Perceived Educational Value of Computer-based Simulations --Choose Value--

Expected Grade in the Course --Choose Value--

Current GPA

Figure 1.2: IPSims login screen generated by Adobe FLEX

The user profile, user ID, password and navigation preferences are stored in a secure database just for login purposes. Inside this same database, which is the main source of information in the knowledge discovery process we can find another important group of variables related with the user profile and navigation behaviours this database is used by the authors to discover patterns and relationships between the elements contained on it. After the registration process the system presents a menu screen to the users where they can select one of the six simulation scenarios according to the instructors instructions. Each simulation contains three scenarios, a library with web links and scholarly journals, scopes of practice, inter-professional competencies, inter-professional perspectives, case records, case encounter (video), main menu, logout, and bookmark links. Figure 1.3.


HETRU
Health Education Technology Research Unit
UOIT
CHALLENGE INNOVATE CONNECT
Open Case

IPSim Version: 2.0.0
LPSL Version:

Simulation 1: Abusive and Complex Patient

A patient requires hospitalization because of injuries related to a motor vehicle accident (MVA). We discover that this patient is very aggressive and has multiple problems in addition to her injuries from the MVA (fractured mandible and pelvis). These problems include addictions, mental health issues, being positive for HIV/HCV, and currently being homeless. Such complexity presents multiple and unique challenges for the healthcare providers to work collaboratively for the best client outcome while ensuring that assumptions and biases do not guide action.

FIGURE 1.3: IPSims main menu screen

Each simulation screen presents an aesthetic user interface that allows the participant to navigate through the multimedia resources. Figure 1.4 displays one of the simulations where the template can be observed.

IPSims has the capability to trace the user activities. For activities we mean: The time spent in every resource and the name of this resource stored in chronological order of visit. We want to mention that there is a strong relation between the user interface presented through FLEX technology and the IPSims track functionality. The relation between these elements and the description of the system control layer is described in the next chapter.

In conclusion IPSims is an educational platform that opens a wide branch for research in the educational and computer science area. It is one of the objectives of this work to exploit some of those capabilities and resources that IPSims provide to the researchers.

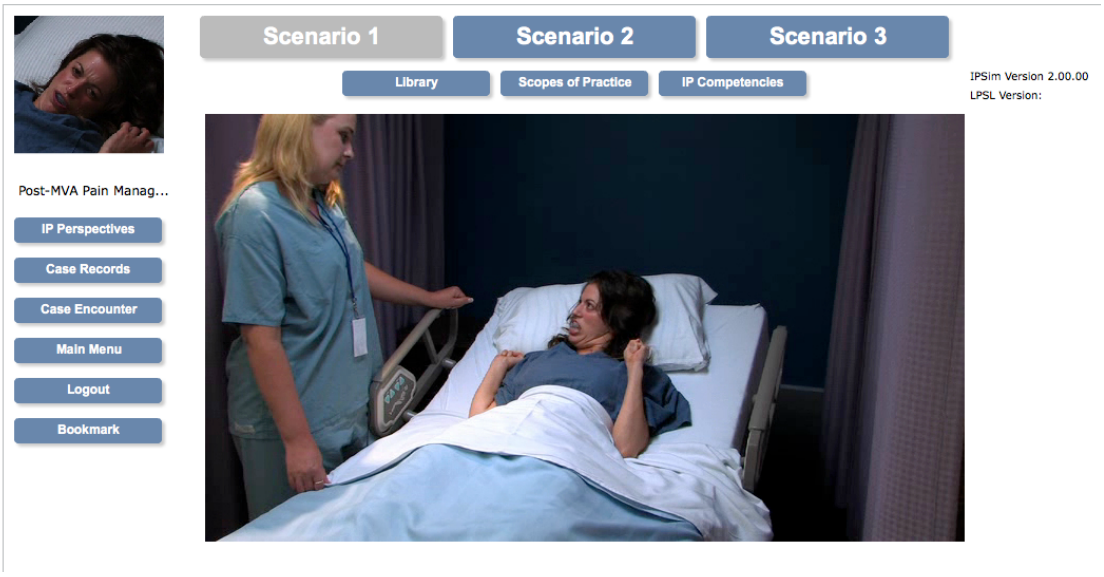


FIGURE 1.4: IPSims scenario template

1.8 Why IPSims?

IPSims is a tool that provides a multimedia learning experience for the students and at the same time give the instructor a powerful platform for the understanding of the students performance and preferences. The capability of IPSims to provide participants related information makes of IPSims an ideal platform for research within the educational data mining discipline. The research alternatives in IPSims are diverse and this work will present some hypotheses and the justifications for the selection of IPSims as a research tool.

IPSims give us different options for research studies such as usability analysis, user navigation preferences and the valuable option to be able to modify our source code to identify specific aspects of our interest. The current study is focused on data mining techniques and it could be applied to some other data sets with more consolidated information.

IPSims is populated with information related to real students from the University of Ontario Institute of Technology. This characteristic makes our study applicable to a real life situation providing solutions for the instructors and the students of the mentioned institution. The results of this study provide with interesting knowledge that can be used to improve teaching techniques and support in a more efficient ways the learning process of students with a profile that reflects a tendency to fail within the activities in IPSims.

As a conclusion we can say that IPSims was selected for the study because it is a customizable educational platform that as show in further chapters will allow us in an efficient and reliable way to:

- Provide valuable knowledge to our institution letting know the instructor which students are in high risk to fail the activities in the IPSims.
- Identify the variables in the user profile that may have a direct relation in the student's performance.
- Analyze which values of the detected variables are the ones that may reflect a positive or negative impact within the student performance.
- Generate association rules for the student behaviour prediction and performance in the IPSims.
- Generate customizable groups of students based in the information provided by IPSims.
- Generation of student clusters based on their characteristics and behaviours.

These are some of the possibilities that IPSims offer as a educational research platform as well as all the educational benefits for the student learning process reinforcement. The potential use of IPSims as a tool for the understanding of the students performance set a starting point for further research also IPSims represents a novel research platform for different areas like education, computer sciences, risk analysis and others.

1.8.1 IPSims Tracking Engine Instead of Commercial Tools

IPSims tracking engine mentioned by Tashiro et al. [3] is used instead of commercial tools such as Google Analytics or Web Trends. These commercial tools offer a lot of capabilities including statistics, measuring of diverse activities in the selected websites, counting of bounces, origin of connections and some other web metrics. Having these tools why would we choose IPSims tracking engine instead? The answer is simple and it is because we can have direct control and access to IPSims database also we are the only responsible for the integrity and security of our students information. Google Analytics and Web Trends can retrieve advanced and specific statistics from websites but we do not have a direct free access to the databases located in the corresponding company's cloud. Everything is retrieved to us through a user interface with already defined capabilities. We are not able to modify content inside the databases and we are restricted to the use of their own API's.

Web Trends restrict the use of the collected information; the more we want to know the more we need to pay depending on the sensitivity of our information. Google Analytics as a free tool does not give us any guarantee that the information recollected by the system will not be used for commercial purposes or in benefit of Google. IPSims tracking engine offer the liberty to modify not just the interface and the control layer, also we are able to modify the data model and the tracking features according to our requirements. We can manage how to trigger and manage our events, how to store the information for the trace of students activities and we have complete knowledge of the activities performed in the IPSims. Google Analytics and Web Trends are ideal tools for web activity metrics while in IPSims we can code specific functionalities related with our specific research purposes like the counter for the number of sessions of an specific user. The capability to keep track of users and their actions is what makes the IPSims tracking engine the prefered tool.

Google Analytics and Web Trends are not able to track the user by the field user ID by itself contained in IPSims database. We need to define Javascript functions and variables in order to store the user ID when the user log into the system, but this action compromises the information integrity of our users. The privacy of our users is an important aspect and when using Google Analytics or Web Trends we are involving a third party in the management of our information. Storing our users information in the cloud, making visible their navigation preferences is something that we can not allow.

The user tracking capability of IPSims opens a wide window for future studies and research; unfortunately by now the 69 study cases is a small number to think in a study related with navigation decisions or user behaviour. Once the project increase the number of study cases a research based on this valuable information stored by IPSims may result in an interesting area of study and contribution to the educational sciences.

IPSims offer different capabilities besides the tracking engine. For example users are immerse in a real life environment simulated by videos, electronic resources and medical documents. The presence of an educational digital environment that was planned originally for the study of learning gaps. Tashiro et al. [3] gives a deeper explanation in the theoretical framework behind IPSims, as well as the original purpose of the platform, which allows us to have a clearer idea of why choosing IPSims over other tools.

1.9 Contributions and Organization

It is our main contribution the generation of a data mining system for knowledge discovery in the IPSims system.

It is our intention contribute to the understanding of the students using our proposed data mining framework for knowledge discovery in the IPSims as well as provide a neat explained process in how to apply well-known data mining techniques in a digital learning media environment exploring the research possibilities in educational data mining, this with the objective to provide a source of knowledge for further improvements in diverse educational areas.

One of the first premises of this job is to identify a valid subset of variables that is part of a user profile, this subset of variables may reflect a big impact in the final grade of our user in the activities on IPSims (Fernandez et al. [4, 5]). This first objective leads us to generate other ideas and contributions that are mandatory to attend in order to accomplish with one of our initial goals.

There was a number of contributions in our work whose purpose was to support other major contributions these supporting contributions are:

- Improvement and validation of IPSims login screen.
- Normalization of IPSims database.
- Application of knowledge discovery process over IPSims.
- Monitoring of the network performance for IPSims.
- Generation of an accurate neural network prediction model.
- Management of IPSims database.
- Reduction of the IPSims dataset based on factor analysis and standard deviation analysis.
- Identification of the possible high impact variables in students' performance in IPSims dataset with the use of cluster analysis.

The improvement and validation of IPSims login screen is a task required by the education experts. The addition of new variables as well as the automatic validation in order to avoid the input of invalid data that could contaminate our database is a requirement that has to be achieved in order to improve the presentation of our system to the students. This requirement is addressed with the use of FLEX user interface elements which prevent the user of making some common mistakes while inputting the information of their user profile. Fernandez et al. [5] study this elements and gives a general description of how are they validated.

The normalization (techniques to constraint the values of our variables into a valid range for our specific machine learning algorithms) scripts for the IPSims database is another contribution in this work. Without a normalized database we would not be able to apply efficiently our data mining algorithms for knowledge discovery. Before get into the normalization process, cleaning, healing and transformation processes were required in order to consolidate a database that were understandable for the selected machine learning algorithms for patterns detection. (Fernandez et al. [4]).

Data mining techniques were applied over IPSims in order to reach what we call knowledge discovery in IPSims. These knowledge is represented as: The prediction model for the grade in the IPSims activities based in the user profile and in the IPSims (Fernandez et al. [4, 6]). Networking monitoring is vital component that we need to address in order to guarantee the integrity of the data stored in the system database. As a distributed system with client- server architecture our task is to guarantee a fast responsive user interface that enhance the student in the activities trying to avoid any possible distraction to the user generated by the system as well as prevent the lose or inconsistencies in the recorded information.

We extracted a meaningful subset of variables by diverse techniques with different objectives. On one hand the objective is to generate a reduced dataset that keep the properties of our original dataset in order to reduce the difference in the ratio for the number of study cases against the number of variables. In the other side we apply cluster techniques to determine the variables that may have the biggest impact in the student's performance in the IPSims activities.

Being IPSims a system used at University of Ontario Institute of Technology the results of a prediction model and the identification of the user profile variables that may affect the student's performance in the IPSims will lead to the improvement of teaching techniques, preventive actions in order to avoid undesired results for the student, and some other solutions in the area. Our contribution are not just to prove the implementation of well-known computational techniques into a system used in a real scenario, we also provide valid information and results based in data mining techniques. This will open a number of possible educational and user analysis research topics within the IPSims.

IPSims is a research platform where multiple approaches and researches can be conducted. At the same time the present thesis work is developed other research within the IPSims is in process and one of our contributions is to provide the necessary IPSims database information for the other research work purposes. Special data treatments are required to satisfy the other research needs and these data treatments consists in different tasks such as: Transformations from one domain to another, transformations in the structure of the data, transposition of the information, and some other cleaning and healing data treatments.

We face challenging tasks when the experiments are performed in IPSims. Our study cases are few in comparison with the variables for the required analysis. We apply dimensionality reduction techniques trying to keep the significance of the data as well as a cluster technique analysis to identify the relevant variables to our study. A well-known cluster technique is required in order to try to identify meaningful variables that may have the biggest impact in the final grade of the user within the IPSims activity. Our contributions are diverse and it is one of our priorities to substantiate the validity of our results.

The rest of the thesis is presented as follows. Chapter II is the related literature. The description of how the data mining process is accomplished is described in Chapter III. In Chapter IV the experiments made in IPSims are reviewed. Chapter V presents the conclusions, contributions and future work.

Chapter 2

Related Literature

2.1 Related Literature

Nowadays data mining is a very useful tool to retrieve knowledge, but "what does it mean to retrieve knowledge?" according to our data mining definition as Vreken [7] mentioned data mining retrieving knowledge process is the action of finding useful patterns in a specific set of data. We are mentioning the words "useful patterns" because according to Vekren [7] and Han et al. [1] finding patterns in a data is easy, but not all the patterns retrieve knowledge. Find useful patterns are the real task, and the real objective of data mining.

How can we know if a pattern is useful? According to Han et al. [1] a pattern is useful if its easily understood by humans, valid on new or tested data with some degree of certainty, potentially useful and novel. There is something that we have to consider every time we want to apply data mining in order to retrieve knowledge. The art of finding patterns in data is not enough; we have to find the useful patterns that allow us to retrieve real knowledge. In this chapter we review some of the techniques and applications of data mining in order to make clearer and concise the points that we mentioned in the past paragraphs. We want to highlight the importance and relevance of data mining in our actual information system IPSims. It is our intention to mention literature in the educational data mining area in order to have a starting point to compare our results and methodology.

2.1.1 Educational Data Mining

An educational digital learning environment like IPSims has large amount of educational and user behaviour data. These data can be related with different aspects and subjects in the educational area, subjects and aspects such as: Students, professors, educational methodologies, scheduling and so on. The task of finding useful patterns in this educational context is defined as educational data mining.

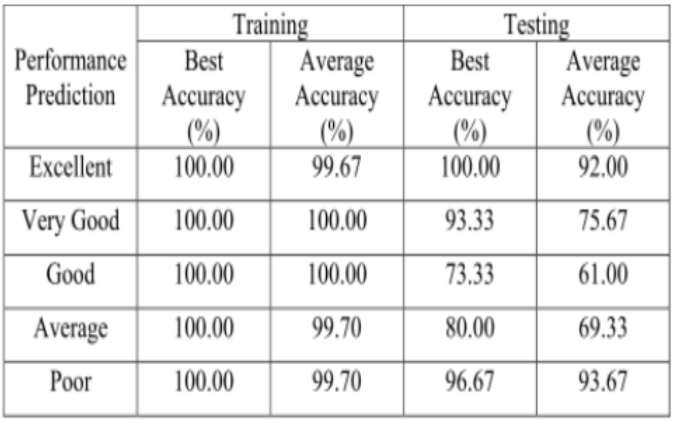
IPSims provides us with student data and student behaviour-navigation data. The intention of this work is to exploit these data in an efficient way providing a complete framework for knowledge discovery in a digital learning media environment. Most of the works in the educational data mining analyze diverse aspects and knowledge areas related with the education and the educational data providing interesting and relevant information for diverse educational applications.

Hamalainen et al. [8] presents a complete study of multiple techniques that can be applied to educational data mining. In the paper diverse techniques are mentioned as well as advantages and disadvantages of each of them. The work mentions: "The feed-forward neural networks architecture consist of layers of nodes: one for input nodes, one for output node, and at least one layer of hidden nodes. On each hidden layer the nodes are connected to the previous and next layer nodes and the edges are associated with individual weights. The most general model contains just one hidden layer. This is usually sufficient, because in principle any function can be represented by a three-layer network". Our proposed architecture contains just one hidden layer that is enough to represent any function according to the cited lines. Another part of this article mentions:"Unfortunately, there are no foolproof instructions and the parameters have to be defined by **trial-and-error**. However there are some general rules of thumb, which restrict the number of trials needed. For example Duda et al. [9] suggest the use of a three-layer network as a default number of layers and add layers only for serious reasons". This is a justification for the approach used on this thesis work. The well known rules of thumb are applied and the rest of the parameters for our neural network are generated based in trial-and-error approach.

Hamalainen et al. [8] mentioned the disadvantages of applying neural networks in educational data mining."The main disadvantage is that feed-forward neural networks need a lot of data much more than typical educational data sets contain. They are very sensitive to over-fitting and the problem is even more critical with small training sets. Knowing that in our work we make use of techniques such as factor analysis for dimensionality reduction, this with the intention to minimize the susceptibility of our model to the problems carried by a small dataset. Hamalainen points, "The neural network model is a black box and it is hard for people to understand the explanations for the outcomes". That is why we present our results as plots, and at the end as a numbers reflecting the predicted final grade for the students.

Kabra [10] proposes a framework to predict engineering students performance. The prediction model proposed by this author is a decision tree algorithm applied over 346 engineering first year students in the periods from 2009-2010 and 2010-2011. The data used for this study is obtained from the administrative department of the school where the experiments are applied. From the information collected Kabra [10] selects a set of 17 variables. From this set of variables collected by the author gender is the only coincidence with the variables selected in our work. Kabra [10] Records previous academic performance. We have in our work the variable high school average as the general point average obtained in the high school education, Kabra decompose this variable in more sub variables that specify some skills for the student such as: Math and science. Kabra [10] makes use of WEKA mining software to generate the prediction model based on a decision tree using the J48 algorithm. The author generates a prediction model with an accuracy of 60.46% that means that 209 elements out of 346 are correctly classified based in the five possible classes for the Grade Point Average scale. We present a model with better effectiveness; unfortunately our number of objects is dramatically lower just 69. The work of Kabra [10] does not contain a formal justification for the selected variables, they just mention their interest for variables that display student's previous academic performance.

Sembiring et al. [11] presents a work where the intention is to find a good prediction model for the student success in higher learning institutions. The use of clustering techniques and supervised learning algorithms (support vector machines) call our attention given the similarity of the techniques used in our research. Sembiring et al. [11] mentions that there is multiple works which deals with prediction models over educational data. But there is no certainty if there are any predictors that accurately predict the success of a student. By the use of psychometric factors Sembiring et al. [11] aims to find the relationship between the student's behaviour and their success in a higher education institution. The experiments were applied in the University Malaysia Pahang and the authors use a questionnaire to collect the relationship between the psychometric factors and their final academic performance. The variables collected by the author are: Interest, study behaviour, engage time, believe and family support. After collecting the data in multiple files, they merged all of them into a single table. The authors describe briefly the pre-processing and transformation stage. They store the grades as: Excellent, very good, good, average. The author modified the original values of the variables and adjusted them to a normal distribution. They categorized each variable presented in the questionnaire as: High, medium, low. This means that they modify the real input of the users to adjust their distribution to a normal one. In our work we can adjust the distribution of our variables to a normal distribution, this in our data pre-processing stage, but the modification of our data to shape it into a normal distribution means to modify our real input given by the students. We make use of different techniques to eliminate the outliers and possible bias in our information trying to keep the data the closest possible to the original input given by the users; these techniques will be explained in the Chapter III. Sembiring et al. [11] proposed a similar clustering technique to the one used in our work. It is curious to see that the selection of the parameter K (number of clusters) is also similar to the one used in this study. Trial and error and this is not a surprise. Most of the literature about K-means proposes the selection of the parameter K based on trial and error testing. For the prediction (classification) experiment the authors used 300 students, each student is described by ten variables. The supervised learning algorithm used for this task was a smooth support vector machine having as a training set 90% of the cases and just 10% were used as testing elements. The author does not give a clarification for the selection of these values. Most of the literature and rules of thumb for these kinds of systems indicates that 75% of the cases are used for testing and 25% for testing. Sembiring et al. [11] finds prediction accuracy values of 61% to 93.67%. See Figurer̃effig:tableSembiring based on these results the author claims to found a valid and accurate enough predictive model for student success rate based on psychometric factors.



Performance Prediction	Training		Testing	
	Best Accuracy (%)	Average Accuracy (%)	Best Accuracy (%)	Average Accuracy (%)
Excellent	100.00	99.67	100.00	92.00
Very Good	100.00	100.00	93.33	75.67
Good	100.00	100.00	73.33	61.00
Average	100.00	99.70	80.00	69.33
Poor	100.00	99.70	96.67	93.67

Figure 2.1: Results of the predictions for the Sembiring et al. paper. Taken verbatim from [11]

Ramaswami [12] presents another prediction model for students performance of higher secondary school education in India. The author highlights that there are many factors that influence the student performance. The data for this study is collected from two different sources: The main source is collected directly from the students while the secondary data is obtained from the school office. A total of 1000 datasets from five different schools in three different districts of a Hindu region is collected. After the preprocessing and transformation stage 772 student records are obtained. These records are used as an input for the CHAID based prediction model. The generated questionnaire is constructed for educational experts just as the questionnaire used for data recollection in our work. 35 variables are the dataset, from these set of variables we found: Sex and previous level marks as the most similar variables to the ones in our study. It is interesting to notice how some of the selected variables for Ramaswami [12] are related with physical features of the students. The prediction model for that Ramaswami [12] proposes is a classification tree, a similar algorithm to the one proposed in Kabra [10]. The use of feature selection in the framework is present and a generated table for Chi-Square values for each variable displays the high potential variables that have a high impact in the student performance. This is from our interest, instead of a statistic technique we based our decision of the possible high impact variables based on our cluster analysis. The possibilities and tools in order to detect patterns are multiple having each of them different pros and contras. Kabra [10] selects the prediction model input variables with values of Chi-Square greater than 100. For the prediction a data mining component of the software STATISTICA 7 was used. The overall accuracy of the prediction model was 44.69% that means that 345 students out of 772 can be classified correctly. The predicted value is the GPA in the form of: A, B, C, D and F. In the conclusions mentions that generalization of the outcomes could not be made due to the limited samples of students. The author has 772 study cases while we just have 69 this leads to think that our prediction model is unable to react to unexpected inputs given the short solution space explored for our small training set while training the neural network. With such a small dataset for training we are susceptible to face the overfitting phenomenon for over training or as mentioned the reduced number of study subjects.

Suthan et al. [13] propose the multidimensional student assessment (MUSTAS) to measure the student's performance through dimensional attributes. By multidimensional the authors refer to a suit of variables that cover: demographic factors, academic performance of the student and dimensional factors. These dimensional factors are classified as self assessment, institutional assessment and external assessment. The author uses the Chi-Square based prediction model CHAID and classification tree just as Kabra [10]. Suthan et al. [13] presents the structure of their MUSTAS framework and mentioned: "We believe the academic performances of the students are not always depending on their own effort. Our investigation shows that other factors have got significant influence over student's performance". This work does not show experimental results it lacks of information related to the environment where the experiment is applied such as: number of students, conditions and data recollection instruments.

Raychaudhuri et al. [14] present a work in detection of factors that affects student's academic performance. The premise of Raychaudhuri et al. [14] is similar to the hypothesis of this thesis work. Given the fact that we propose that there is an identifiable subset of variables that may have the biggest impact in the students final grade in the activities within the digital media learning environment. Raychaudhuri et al. [14] recollects 332 student's information to apply a regression model. Raychaudhuri et al. [14] work is more rudimentary but the importance of this work is that using an statistic approach the authors wants to prove a similar hypothesis to the one proposed in our thesis work. Raychaudhuri et al. [14] presents a regression model with an accuracy of 52% based on the study of 332 students described by eight variables. One of the interesting results on this work is the discovery of a direct relation between the education level of the mother and the student performance. Students with mothers, which have a higher education level, reflect better academic performance results in the study.

We find in our work interesting relationships based in our cluster analysis. The discovery of these patterns can lead to diverse branches of data mining such as generation of association rules. Based on the observation of our generated plots we can conclude which subset of variables is the one that may have a direct impact in the student performance in virtual scenarios.

Barbour et al. [15] mentions: "There are likely many factors that influence the student-participants online learning experiences, but clearly there are concrete measures that could be taken to prevent difficulties for many virtual school students". Educational Success Prediction Instruments (ESPRI) as the author calls them are powerful tools that help to determine the specific factors that can affect the student's performance. In order to prove this the author mentions: "Our results, although promising, should be viewed as preliminary for this population. The next step to confirm the validity of the ESPRI and to test the predictive model is to use it with additional groups of VHS [Virtual High School] students to determine if the instrument discriminates as well between pass/fail groups in other populations as it did this one". The same applies to our work it is mandatory to use our predictive model in next courses to prove the validity of our work.

There is diverse works that shown a similar approach to the one presented on this thesis El Moucary et al. [16] presents a study where the authors propose a framework based on neural networks and clustering techniques using K-Means the authors found a prediction model for the students GPA for courses in a foreign-language. The authors collects 200 records that after the pre-processing stage down to just 73 records almost the same number we are using in our study. The author mentions: "As previously mentioned all records in the data set were used for training and testing since data set is small". This approach is useful to know if our neural network received an adequate training for the"known cases" but if we do not have a specific testing set we do not know how the neural network behaves when unexpected cases are presented. In our case if we train and test the neural network with the entire dataset we obtain results for R-squared above of 99% and values for the maximum square error less than 0.05. These results reflect an adequate training for our neural network but there is no more cases to test the behaviour of our model against unexpected inputs. In El Moucary et al. [16] there is no certainty that the neural network can predict the results properly when an unexpected input is presented. Lets understand for unexpected input a case that is not presented in the training set. For the paper presented by El Moucary et al. [16] the selected number of cluster K was three this based on experimentations for values of K equal to two, three and four. They mention: "The three-cluster approach clearly identifies that students with high GPA in English courses, are most likely to obtain a high GPA for the core and major courses pursued in their specialty... We based our selection not just in the intention to have a division of five groups representing the possible obtained GPA in the scale A, B, C, D, F but we also measure the quality in our clusters keeping the one that reflect the best results for our quality measure.

It is worth to mention that there is not enough variation in our study cases grade to observe a clear grouping of elements according to their characteristics by grade. It is expected when the number of study cases increase to observe average grade of: A, B, C, D and F each of them corresponding to one of the five possible clusters.

El Moucary et al. [16] does not have a measure for the quality of their generated clusters then there is not certainty that the generated clusters are adequate to present the assumed results. In all the analyzed paper there is never a descriptive mention in which training algorithm is being used in the neural network as well as we never found a complete description in how the training datasets were selected. Our work presents the methodologies for the training set collection we also mention the training algorithms used for the neural network and we report the performance of these training algorithms.

2.1.2 Cluster Analysis

Cluster analysis is an easy to implement and powerful tool for pattern recognition. The ability to of these algorithms to group elements by their characteristics make this unsupervised learning technique an ideal tool for fast identification of potential relevant features in a dataset. In order to succeed with the use of cluster techniques it is required to have an ideal visualization and adequate representation techniques of our results.

Our work presents a module where clustering techniques are applied with the intention to discover strong relations between the user profile variables and the student performance in the IPSims activities. These strong relations let us know which are the variables that may have the biggest impact in the student's performance in the IPSims activities.

Erdogan et al. [17] present a study where the results in the university student entrance examination and the academic success of the students is analyzed with the use of clustering techniques. The data is collected from Maltepe University for the 2003 academic period. 722 student records are analyzed. The K-Means algorithm is applied to the collected data with the use of Matlab software. Erdogan et al. [17] data mining process is divided in four stages: Data preparation, data selection and transformation, data mining and presentation. For the data preparation the authors joined multiple tables into a single one. This is a similar approach to the one presented in our work where we merge all the variables, which described our user profile. After the data preparation, the data selection and transformation stage is a simple transformation of qualitative variables into quantitative. Erdogan et al. [17] does not make mention in why the transformed variables take the new values. In the data mining stage the authors inputed the transformed data into the algorithm K-means and the decision of the parameter K is described as follows: "The number of clusters was determined as an external parameter. Different cluster numbers were tried, and a successful partitioning was achieved with five clusters". This paragraph is relevant to us because we adopt a similar technique where we experimented with different values of the parameter K (clusters), being number five the one that displayed the best cluster quality. For the stage of presentation plots are presented and analyzed. The conclusions of the author reflect expected results such as the fact that students with high percentage in the entrance examination are most successful usually because they got scholarships and they need to work hard and obtain high grades to keep their scholarships. The analysis of Erdogan et al. [17] work is a reference point for our cluster analysis approach. We consider this work just as a basic reference because some of the important aspects in data mining process are not described. It is the intention of this thesis work to establish a neat process in how the knowledge is obtained from cluster analysis.

Ayesha et al. [18] proposed a framework for the analysis of the student learning behaviour based on the educational environment factors (class quiz, mid and final test). K-means was the clustering algorithm used on this work. The data of 120 students is collected from the department of computer science in the University of Agriculture Faisalabad. Ayesha et al. [18] as the past works mention just the transformation process but they do not explain why did they apply transformations and which is the criteria for the selected values for the transformation. It seems to be a common denominator in most of the educational data mining works analyzed until now. This is not a surprise given the inability of the mining software to extract and process information from multiple data sources. In our case we implement our own algorithm K-means. We extract the information from multiple sources storing temporarily just the used variables in a list data structure. After the algorithm finishes its iterations the list is released from memory.

The results of Ayesha et al. [18] are presented graphically and in tables but the work does not explain how the parameter K was selected so there is not a justification for the nine presented clusters neither a measure in how valid are the classifications for a value of nine for the parameter K. In our thesis work we explain how the parameter K is selected and we also justify the validity of our selection based in the calculation of the average distance from all the elements in a cluster with respect to their centroid. Being K-means an algorithm based in Euclidean distances calculations the best way to test the quality of our clusters is with the minimum average distance value of the points to the corresponding centroid.

Tair et al. [19] propose another educational mining work based on the K-means algorithm in this work the authors used a dataset compose of six variables. 3360 student records are collected, 51 of these records have missing information and the authors decided to ignore them keeping just 3314. In our work we find missing values but given the low number of cases was unacceptable for us to eliminate these records from our analysis. A healing technique is applied to overcome this issue. Tair et al. [19] claims the discovery of association rules which will help to the improvement of graduate students performance. This work is a complete mining system where classification, cluster techniques, association and outliers detection methods are applied. We consider that in order to find association rules more efficiently faster techniques can be used such as classification trees where we can avoid the use of cluster analysis, which is more an intuitive tool, based on the researchers criteria and experience. Tair et al. [19] discriminate variables from the original dataset just by the use of common sense.

Multiple clustering works have been analyzed but there is not a clear description in how the patterns or the knowledge is being provided to the users after the clustering process is completed. El-Halees [20] is one of the works that show how to manage the information after the construction of the clusters. The approach is similar to ours: El-Halees [20] generates a table where each column represent one of the five generated clusters and each row represents one of the variables: attendance, GPA, hours, e-resources, e-excercise, e-homeworks, midterm, lab, final, grade. The justification for this generated table is that the author assumes that with the stored information they can divide the students according to their performance and they will be able to guide the students based on their presented behaviour. This is accomplished through the matching of the student characteristics with the row in the table that displays the more adequate values for the given user profile. Our work presents a table with the generated mean for each variable by cluster. This calculated mean by variable has the intention to organize the clusters and extract the variables that may have the direct impact in the final grade of the user. The generation of association rules could be straight-forward with the use of this approach but the generation of these association rules is out of the scope of this study.

Chapter 3

Data Mining Process in IPSims

3.1 Data Mining Process on IPSims

It is the objective of the previous chapter to show some of the works related to the educational data mining in order to have a reference to compare our work. The previous chapter presents the last advances in data mining techniques applied to educational area and it is important to compare our work to the works showed in Chapter II in order to provide complementary results and ideas for the educational data mining discipline.

3.1.1 Data Acquisition

The data acquisition stage for the present study case is built with three different data sources. Two digital databases and a paper questionnaire filled by the users. IPSims as media learning environment is the system that records the information in the digital sources (databases).

In order to populate this databases and paper questionnaires the participation of students in the health science area is required and a consent form is generated in order to be fulfilled by the users this with the intention to accomplish with UOIT's requirements. We also required the authorization from the instructors of two health sciences courses at the University of Ontario Institute of Technology in order to be able to start the data acquisition process.

The data gathered from the students of University of Ontario Institute of Technology is used in this study. The collected data corresponds to the academic period of March 2011 and includes records of 69 students.

Further aspects related to this process are explained in detail in Chapter IV.

3.1.1.1 IPSims Original version

The original version of IPSims is an incomplete Flex template. This template displays labels of type VariableX where X would be a letter from the alphabet ordered in ascending order. These labels match with a corresponding combo box. The combo boxes contains the values Variable1, ... VariableN. This original interface lacks of meaning for our research purposes.

IPSims research environment is an interesting theoretical concept that offers a lot of capabilities for the understanding of diverse educational aspects. A design based in educational studies and theories by a multidisciplinary research team ended in a final educational map with numerous possibilities for research purposes. In order to exploit all this capabilities we generate our data mining framework that make use of the potential and conceptual design of the IPSims with the intention to provide knowledge to the instructors and students.

The research team on this study is conformed by specialists in education and computer science areas. The specialist in education selected a group of disaggregation variables that will be part of the user profile dataset for our analysis. Some of these variables are displayed in the actual user interface other ones are gathered form the paper questionnaire. All the meaningless labels are substituted for meaningful ones. This new meaningful labels corresponds to the set of names that the experts in education assign to each variable. We also need to substitute the original combo boxes and place the best suited validated user interface elements instead.

The variables contained in the main screen of IPSims are: Gender, e-mail, course, faculty, term, age, undergraduate academic year, undergraduate major, number of hours/week surfing the web, number of hours/week playing video games, how do you rate your computer literacy, likelihood of choosing a career in health sciences, interest in course material, experience with computer based simulations, your perceived educational value of computer based simulations, expected grade in the course, current GPA.

In data mining is a common issue the inconsistency and bad representation of the data provided by the user. Data treatments are required in the stage of data preprocessing this with the objective to heal and correct the inconsistencies and bad representations in our data.

It is a good strategy to generate an adequate and validated user interface that prevents the input of inconsistent or noisy data into the database. We present a modified user interface to substitute the previous IPSims login screen. We generate combo box and validated text box that restrict the input of the user to the system, allowing just valid and consistent values in order to avoid excessive data treatments. An example of this validation is the user interface element for the variable grade point average. This variable is captured using a text box that allows the input of just numeric values according to the UOIT GPA scale that goes from 0.0 to 4.3. The user cannot type alphabetic characters or values that go out of range.

Simple preventive measures like the mentioned in the previous paragraph work as an error prevention method for the user inputs in the IPSims login screen. With a consistent and validated user interface we can prevent and reduce a lot of noise and inconsistencies in our data analysis.

After validating all the text box and populate our combo box with adequate values, the connections with the database have to be modified in order to store correctly the values that are provided through the user interface. We need to generate dummy user accounts in order to test that the values are being stored correctly into the database.

After these modifications the system is visually and internally functional to start with the experimentation but we need stress tests in order to guarantee an acceptable response from the system to the user actions. In order to test the response of the system diverse metrics are used: Time response, peak response times, latency for given hours of the day. All these measures give us an idea in how well our system behaves when the users work with it.

The stress tests using 100 concurrent users displayed acceptable results. The only detected issue at that time is the bandwidth of the university connection that implies some delay in the system response especially during examination dates at UOIT. The server where IPSims is hosted used to pass through UOIT's student residence. In hours peak the excessive use of bandwidth by the students living in the residence affected IPSims connection.

The issue is resolved by having IT services isolating the hosting server from the common connection bridge within the UOIT's firewall.

Once the user interface is validated and the stress tests are passed satisfactorily the experiments in the two health science courses are started.

3.1.1.2 Selected Variables

It is not the intention of this thesis work to make a deep study of educational aspects but is important to point out the source from where the variables used in this work are selected. Then is important to mention that the group of variables to analyze in this work is selected for our group of experts in education based on their experience and knowledge with learning gaps and misconceptions. These selected variables are selected with the intention to identify learning misconceptions in a digital media learning environment. It is important to mention that the original suit of variables contains more than 100 elements, while the final set to analyze contains only 22 variables.

Regts et. al. [2] Give a deeper overview of the educational research intentions in IPSims. The original goal of IPSims is to reach a better understanding of misconceptions generated in students of health science programs although the present work focuses more in the application of computational techniques in order to retrieve knowledge from the available data in IPSims we consider important to give at least one reference that provides a deeper understanding of the information presented by IPSims.

The variables that are part of the user profile are obtained from the IPSims data sources these sources are: Two digital databases and a paper questionnaire. The data sources are explained with more detail in the following paragraphs. Figure 3.1 depicts IPSims original data sources and their relationships.

3.1.1.3 First Time Users and Storing of User Profile and User's activity

The first time the user access the platform they have to fill the variables presented by the web form called login screen of IPSims. Figure 1.2 displays the login screen of IPSims. These displayed variables in the login screen are a portion of the user profile. The users have to select a password and subsequently click the button register; eventually the system retrieves automatically a user ID (uid).

The integrity of the fields user ID (uid) and password are granted with the use of the MD5 algorithm. The secured variables presented on the login screen: user ID, and password are stored into a first database named lspl_users. This database is described in detail in the next sections. After the user registration, the user can access the system anytime with the use of a valid user ID and password.

Part of the user profile is obtained from the main screen and the other variables are collected using a paper questionnaire applied to the participant students. After the participants answer to the paper questionnaire the information is captured into CVS (comma-separated values) files which eventually are transformed into *Postgres* tables using SQL commands.

IPSims is a platform capable of storing the paths followed by the users inside the system. Every time a user clicks in a hyperlink IPSims triggers an event that unchains a persistence process, a more detailed description of IPSims architecture is mentioned by Tashiro et al. [3].

It is important to give a brief description in how IPSims handle the events in order to store the time stamps, and page identifiers. These events have the intention to keep track of the information representing the decision sequela of the users. Figure 3.1 presents the original data architecture of IPSims data sources, here we can observe how the databases are independent from each other. These databases will be merged in order to obtain a consistent and consolidated source of information to extract data related with IPSims users.

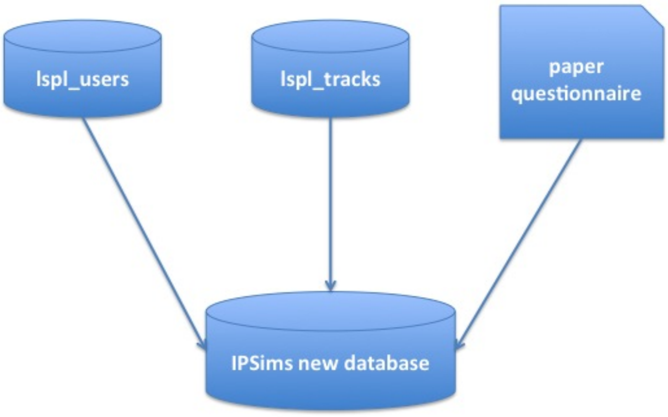


FIGURE 3.1: IPSims original datasources merging diagram

IPSims Architecture and Persistence Engine

IPSims technology stack is conformed by three different technologies. *Adobe FLEX*, *PHP* and *Postgres* as showed in Figure 3.2.

IPSims works under MVC (Model, View, Controller) architecture. The view is presented in *Adobe FLEX* technology and the control layers are constructed under *FLEX* and *PHP*. The data model layer is supported by *Postgres*.

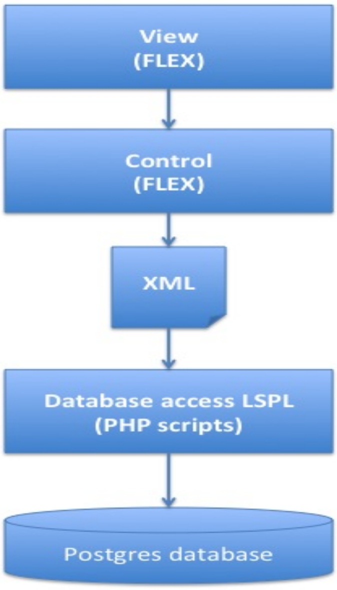


Figure 3.2: IPSims architecture and persistence flow

The system loads a template for each scenario presented in the IPSims this loaded template is described under an XML document. XML for his hierarchical structure allow us to have a tree data structure definition for our media-learning environment. The hierarchical structure of our media learning environment described by a XML document allows to visualize the user decision sequela as a tree data structure, where each node correspond to a web document or resource within the IPSims environment and the traverse of this tree represents the navigation performed by the user. The generated tree structures are helpful to compare the decision sequela of each user. This could be a point of comparison between successful and unsuccessful students we can also analyze user navigation preferences.

The decision sequela provides another unit of information called time stamps. Every time the user click on a user interface element the action triggers an event. Eventually a time stamp is recorded into the database called *lspl_tracks*. The time stamp is persisted with the corresponding page identifier (*pid*). Page identifier is a unique string to identify a web document. The mentioned *pid* and time stamp are stored under an auto-generated table.

lspl_tracks database keeps record of all the activities of the users organized in different auto-generated tables. This database is explained on detail later on this chapter.

Another important part of IPSims digital learning environment is the LSPL engine. LSPL is a group of *PHP* scripts that are in charge of formatting the input coming from our XML document and storing the users data into the *Postgres* IPSims database.

LSPL has some modules that execute a number of simple queries to the stored data within the IPSims databases. These simple queries unfortunately lack from an understandable presentation format and the results of the queries cannot be shown on the web browser. If the information retrieved is relatively numerous the retrieved information overflows and lacks of organization.

We require to modify our *PHP* script with the intention to make the system capable to show large amounts information retrieved by the SQL queries but the modification is useless given the completely lost of shape when displaying the information in our browser. For lost of shape we can understand tables without alignment as well as data being displayed completely disorganized in the web browser window.

The *PHP* layer (LSPL engine) seems to be an add-on that does not accomplish properly with some of the required tasks such as knowledge retrieving, it is just an intermediate layer that provides with simple queries and execute persistence to the database. Figure 3.2 displays the location of the *PHP* layer in the architecture. This module of the system could be discarded and the persistence process of the users decision sequela could be made using a different technique that does not involve another layer in a different language such as *PHP*.

The stored data is presented in an undesirable format this leads to a disrupted database, which by itself is disorganized. The data stored in it does not retrieve any useful information by the original version of the system. The treatment required to make these data readable for the data mining tools is an exhaustive process. In summary IPSims is build under different technologies in order to perform the next main tasks:

- Present an interactive enhanced system where the students can reinforce their learning.
- Persist user profile into a database.
- Persist user decision sequela into a database.
- Retrieve information using simple queries.

IPSims data sources

IPsims original data model consist of three databases, an exhaustive treatment over this data model is required in order to generate a normalized database this with the intention retrieve the information in a more efficient and organized process.

The original databases for the IPSims are explained in this section.

IPSims data source include a paper questionnaire generated for the experts in the educational area.

LSPL_USERS

LSPL_USERS is the first database it consists of three tables: *mgtlogin*, *user_info* and *user_attr*. Each table is designed as follows:

- mgtlogin: name, pass, salt, rank.
- user_info: uid, course, faculty, term, age, year, major, surfhour, playhour, comskill, variablei, variablej, variablek, variablel, variablem, variablen, variableo, variablep.
- user_attr: uid, password, gender, contactemail, academicbg, incomeperyr, race, prevsession.

In the first version of *lspl_users* database we can observe that a lot of the columns in *user_info* lack of meaning for the research like: variablei, variablek, variablel. In next stages of data pre-processing these fields are renamed and populated with the adequate information. Figure 3.3 depicts the structure of *lspl_users* table.

Unfortunately this database is designed in an incorrect way for consequence is not a normalized database. While *uid* is a primary key in *user_attr*. *user_info* presents *uid* just as a field, is not even a foreign key.

In the *mgtlogin* we do not have a primary key defined. And these kind of issues are attempting against the first normal form. In consequence there are some sections of the database that are incongruent with our user interface. For example *user_attr* table is filled with some of the variables that the user interface captures in our login screen but the user interface does not show fields for academicbg, incomeperyr (we have to get rid of these fields in next stages).The previous researchers decide to leave these variables out of their investigation and the variables are not being eliminated from the table just from the user interface. Another incongruence is that gender clearly fits better in the table *user_info* but it is defined on *user_attr*. The original database for *lspl_users* has multiple issues when talking about data normalization and good data modeling practices.

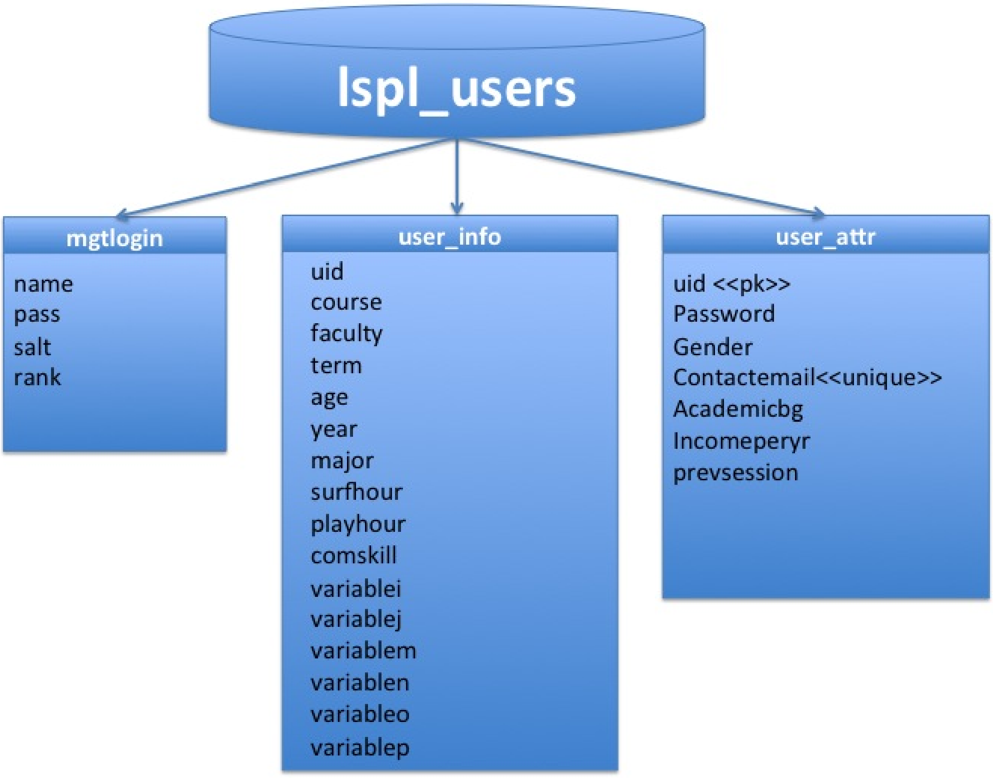


FIGURE 3.3: lspl_users database

It is desired to create tables based in some kind of congruency. Some variables are grouped naturally by their characteristics. *user_info* contains all the information related to the user profile. *user_attr* contains data related with the login information of the user like password and user id. Then the common sense indicates that gender should be defined in *user_info*. We can observe that some variable names are not descriptive of the values that are being stored on them. One of our tasks is to create descriptive field names for the new variables using the names provided by the education expert researchers. Adequate user interface elements are required for these new names.

The non descriptive column names in the original database are not modified in order to avoid possible problems with the names in the IPSims interface, the control tier, the XML documents and *PHP* scripts. The non descriptive names are intact in the original database but scripts in *Python* are generated to move this information into a new database with similar characteristics. The new database has descriptive column names for each variable and a normalized architecture that accomplish with the third normal form. The new database is described on the next sections.

As we can see *user_info* and *user_attr* tables are in charge to persist the information provide by the user when they are first time users. It is clear that even if *uid* is not declared as a foreign key in *user_info* we can use this field as a relationship between the two tables.

mgtlogin is a table that persists the name, password, encrypted email and rank of the users that have access to the management section of IPSims. IPSims management system allows some basic queries to the databases. The table *mgtlogin* is not from our interest for this document.

LSPL_TRACKS

LSPL_TRACKS is our second digital data source and it goes on scene once the user login for the first time in the system. As the name indicates this database is in charge of storing the paths (decision sequela) that every user takes in the system.

The database *lspl_tracks* has a peculiar design. This design is neither common nor normalized. The design consists in something that we call dynamic table generation (DTG). That means that the database generates tables automatically according to a given action triggered for our UI. This action which determines the creation of a new table in our *lspl_tracks* is the login in the system for a given user. That means that every time a user clicks the login button with a correct *uid* and password the event triggers an action in the control layer. This action creates a new table in our *lspl_tracks* database. Every generated table has the next columns: *pid* and time. Where *pid* is a web document identifier (page identifier) and time is a time stamp that indicates the exactly moment when the user went into the web document.

There is no primary key on these tables in consequence there is not an obvious relationship between the different tables on this database and there is not a field that indicates user belonging for each table. The only clue that we have about the ownership of the tables is the name of each table. DTG generates each table name under the next convention:

$uidx_0x_1x_2x_3x_4sny_0y_1\ldots y_n$

Where:

$\forall\, x_i \in$ "*tablename*"; $i = 0,\ 1\ ,\ 2,\ 3,\ 4;\ 0\ \leq\ x_i\ \leq 9;$

$\forall\, y_j \in$ "*tablename*"; $1\ \leq\ y_0\ \leq\ 9\, and\, for\, j\ =\ 1,\ 2,\ldots,\ n;\ 0\ \leq\ y_j\ \leq\ 9;$

Here the set of x's correspond to an specific *uid*. In the cases where the *uid* is build by one, two, three or four digits, the previous digits are filled with 0's in order to complete the five spaces.

The y's correspond to the number of session. Every time a user login in the system this counter represented by the y's increase being the number one the first y_0 value corresponding to the first generated table for a given user. Every time we need to increase the number of digits a y_j is generated

The regular expression representing our definition for the DTG is:

$uid[0-9][0-9][0-9][0-9][0-9]sn[1-9][0-9]*$

An example of the first auto generated table for a user with an uid 104 is shown in Figure 3.4.

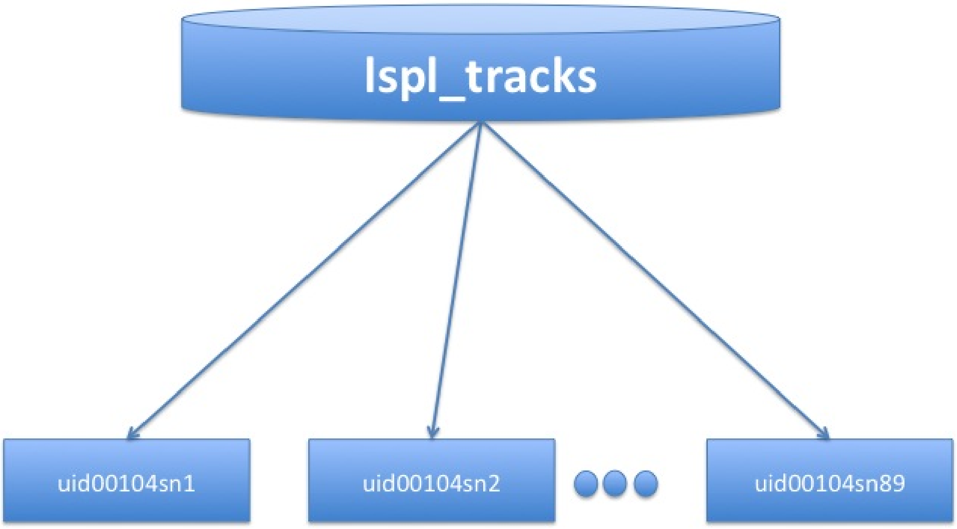


FIGURE 3.4: Generation of 89 different tables for user 104

With all these information we have a general idea in how the information is collected for our media learning environment this give us a clue which problems we need to address in order to retrieve the stored information in an efficient way. This with the intention to generate a complete data mining system.

It is important to make reference to previous paragraphs and highlight that there is an extra source of data that is different from the two digital data sources that we already describe. The third source of data is the paper questionnaire. This paper questionnaire was designed for an expert in education. The objective of this questionnaire is to add valuable information that supports different studies about learning processes, student's outcome predictions and student learning misconceptions. The questionnaire is subdivided in groups of questions namely 1) Learning assessment based on assigned learning activities, 2) demographic information, 3) rating of web-based course work, 4) rating of IPSims learning environment, 5) satisfaction with educational simulations and serious games, 6) disposition to engage in effortful cognitive endeavor (need for cognition and ambiguity tolerance), 7) expectancy-value questionnaire, 8) performance evaluation in inter-professional learning activities. A deeper description can be found in Regts et al. [2].

The educational fern is not the main topic of this thesis work and is not one of our principal concerns the educational background to generate the questionnaire. We trust in the selection of our variables that our expert in education selected . We just want to mention that this questionnaire is another valid source of data for our research.

Summarizing we could say that data acquisition stage is accomplished through three different well-defined processes. User profile in the main page of the system, decision sequela in the system and the paper questionnaire. Focusing in the two digital processes described above we can see at first sight some inconsistencies and design problems arise. The next section of this document talks about the integration of the three sources and simple solutions to all the inconsistencies and problems within the data sources.

3.1.2 Data Integration

The stage of data integration refers as the name says to the integration of our data sources or data sets.

In this particular case the IPSims database is generated by three data sources: Two electronic databases (*lspl_users* and *lspl_tracks*) and a paper questionnaire. Before we can start the integration of these sources we need to apply isolated treatments to each of them in order to generate datasets that accomplish with the type of information required for our analysis.

It is convenient for the system to have all the information located in a single normalized database which should accomplish at least with the third normal form for databases normalization this with the objective extract and analyze the information more efficiently in the IPSims system. The proposed database will consist of tables and columns as shown in Figure 3.5:

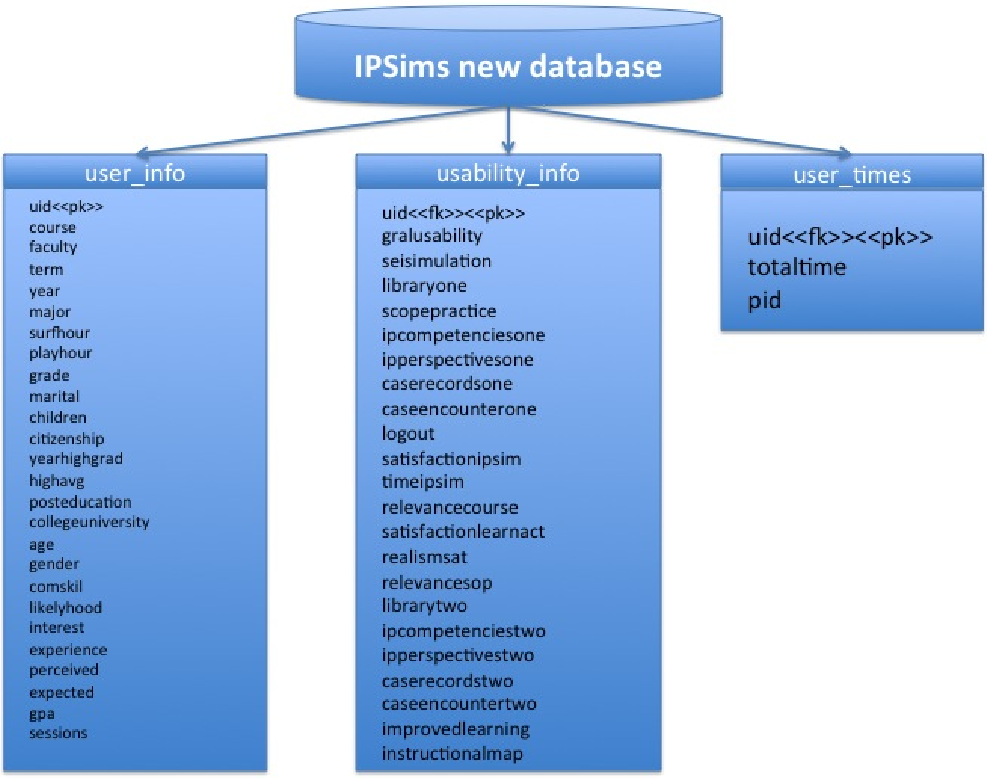


Figure 3.5: New generated single database for IPSims

The table *user_info* shown in the Figure 3.5 is the resulting table from the form located in the login page of the system and the addition of the total number of sessions per user calculated within a *Python* script. The number of sessions is displayed within the DTG $uidx_0x_1x_2x_3x_4sny_0y_1$ where $0 \leq y_i \leq 9$ with i=0,1. The concatenation of these y_i values give us the number of sessions. This process describes the extraction of the number of sessions from the name of the auto-generated tables contained in *INFORMATION_SCHEMA.table_names*. We have to take the last token corresponding to the number of sessions and transform it into an integer then we take the greatest value for each user. This greatest value is the total number of sessions.

Note: The sessions start with a value of 1 for y_0 and this value grows incrementally for each new session. For example:

$$uidx_0x_1x_2x_3x_4sn_1$$

$$uidx_0x_1x_2x_3x_4sn_2$$

$$uidx_0x_1x_2x_3x_4sn_{99}$$

We have to take the greatest value for each different user $x_0x_1x_2x_3x_4$ in order to fill the total number of sessions for a given *uid*.

usability_info is the result of our paper questionnaire. The paper questionnaire is in fact an evaluation of different aspects related with the user experience in IPSims and some user profile variables. The variables from the questionnaire were truncated for our expert in order to reduce the size of the generated table for *usability_info*.

user_times is the table containing the total time spent for every user in a given *pid* where *pid* is the acronym for page ID. Every web document in IPSims has associated a specific *pid* that differentiates it from other web documents related to IPSims inside or outside the system.

As a conclusion we can notice that IPSims principal dataset is structured by three different sources: Two digital databases and one paper questionnaire. The original IPSims databases present normalization issues also the information in the questionnaire paper is not in a digital format; the process described in this section proposes an architecture where a normalized database is presented. The proposed database complies with the third normal form making easier and more efficient the access to the required data.

The data integration process can be summarized in the Figure 3.6.

In essence the objective of this process is to generate the input dataset for the cluster and prediction algorithms.

3.1.3 Data Pre-Processing

Before we can integrate all our data into a single desirable source of information some parts of the data have to receive a special treatment for their nature inside the database. After applying these processes we can generate data structures containing the desired formatted data that fills the required tables into the unique single desirable source of information.

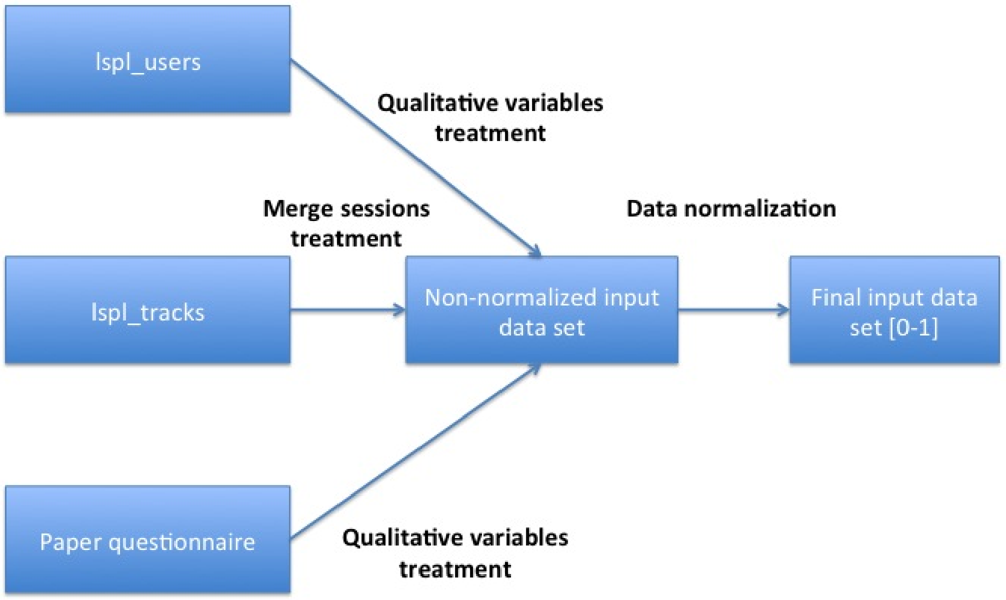

lspl_users
Qualitative variables treatment
Merge sessions treatment
Data normalization
lspl_tracks
Non-normalized input data set
Final input data set [0-1]
Paper questionnaire
Qualitative variables treatment

FIGURE 3.6: Data integration process

3.1.3.1 Qualitative Data into Quantitative Data

The paper questionnaire and the web form included in the login screen of IPSims contain variables that belong to a qualitative nature; the qualitative nature of these variables give us a rich description of an event but unfortunately does not provide an adequate input dataset for data mining techniques. We require quantitative or normalized and scaled values as an input for our data mining techniques.

There is diverse literature that talks about the treatment that could be given to qualitative variables such as the techniques mentioned by Driscoll et al. [21] . In our case we defined an incremental scale from 1 to n from the less desirable to the most desirable value according to a criteria proposed by our educational experts. Eventually a normalization process is required in order to present values between zero to one to provide the adequate input to our prediction model experiment.

Our *Python* API *ffnet* [22] counts with a module for normalization and feature scaling that automatically normalizes and scale all our variables based in the formula:

$$X_i = \frac{X_i - \mu}{S}$$

Where μ represents the mean over all the values of X and S represents the standard deviation for all the *X's*. A description of some of the variables and the source from they are extracted are shown in the next lines:

- Surf hours per week: Will be shown in the web form as a text box where the user can record any positive number or zero according to the number of hours peer week the participant surf on internet. The persistence in the database will be using the same format.
- High school average: This variable is taken from the questionnaire and the participant will write her average from high school based on 5 intervals between 0-100. The variable is persisted as an integer from one to five where one indicates the lowest interval and five the highest one.
- Pursue of post secondary education: This variable indicates if the student is interested on keep with her studies after finishing the university. The variable is taken from the survey paper and is persisted as one if they want to pursue studies after the university and two if they don't want.
- Age: The variable will keep the age of the user, this variable is taken from the survey paper and will be stored in the database as an integer positive number.
- Computer literacy: The variable is stored from the web form and presented to the participant as a combo box containing the values: Terrible, Poor, Average, Good, Excellent. The variable is persisted as a number from one to five one corresponding to Terrible, two for Poor and so on.
- Likelihood of choosing a career in health sciences: The variables is stored from the web form as well and presented in a combo box displaying the values: Not At All Likely, Not So Likely, Likely, Very Likely. The values are persisted in the database as integers going from one to four starting from one mapping to Not At All Likely.
- College or university before the current university: This variable stores a value of one or two, one if the student took college or university before the current one. And it uses two in the case of have not course any of these mentioned before.

- Gender: This variable stores the gender of the participant. Two for women and one for men.

We also presented this in Fernandez et al.[6].

3.1.3.2 Time Variables for lsp_tracks Database

We showed how *lspl_tracks* store the information using DTG.

The main problem with DTG is that each generated database just contains the two fields *pid* and *time*. None of them is useful as a primary key or foreign key to relate each table with the correspondent user. The only way to accomplish with the relation $U_i \rightarrow T_{i,j}$; where U_i is the i-th user and $T_{i,j}$; is the j-th table for the i-th user is using the name of the table and the metadata contained in the *Postgres* table *INFORMATION_SCHEMA*.

The abstraction of the solution at a higher level can be visualized as follows

- 1.- For each element in INFORMATION_SCHEMA.TABLE_NAME decompose the string with format: $uidx_0x_1x_2x_3x_4sny_0y_1 \ldots y_n$ into: $x_0x_1x_2x_3x_4$
- 2.- Eliminate the zeros before the first digit distinct from zero in $x_0x_1x_2x_3x_4$
- 3.- The resulting set of $x_0x_1x_2x_3x_4$ with no zeros in the left side before the first non zero digit where each x_i represents a value from zero to nine will be the correspondence between the table and a user id.

That is how we can accomplish the mapping: $U_i \rightarrow T_{i,j}$

We decide to use *Python* as a programming language to accomplish with this pre-processing method given the well-known simplicity and integration of *Python* with *Postgres*. The module *psycopg2* has high-level procedures that facilitates the interaction *Python-Postgres* allowing the adequate treatment and storage of the data into the desirable places and formats.

Once we accomplished the mapping $U_i \rightarrow T_{i,j}$ the generation of a data structure is required in order to store our information related with time; the time is an interesting source of information for different research areas; at the end of the treatment process with time variables in *lspl_tracks* the total time spent for each user in a given *pid* is the final value to be stored in the database.

uid	pid	time stamp
1598	dasdasda	$12\backslash01\backslash1997\,13:34,23$
1598	reyehdss	$12\backslash01\backslash1997\,13:45,36$
1598	ratatga	$12\backslash01\backslash1997\,13:48,22$
1598	parufy	$13\backslash01\backslash1997\,10:25,02$

Table 3.1: First refinement for lspl_tracks

For future works and required information from the specialist in education. We have to apply specific treatments to the time stamps this with the intention to retrieve not the time stamps but instead to show and persist the total time spent for every user in a given web document. It is necessarily to apply an extra refinement process in our database to store the times. The process is briefly explained in the next paragraphs.

Psycopg2 [23] allows the extraction of the information stored in *postgres* databases. This extracted information can be temporarily saved in a list data structure where each sublist corresponds to a row in the database. This is convenient and allows to try a simple technique to determine the total time per user spent in a given web document (pid).

All the DTG corresponding to one *uid* are arranged in ascending order by the number of session and grouped by *uid*. We can find all the pid times for a given user in ascendant order that help us to avoid the process of ordering and grouping the data. This process would not be hard to accomplish commands like GROUP BY and ASC in SQL realize the grouping and ordering in ascendant order respectively. We traverse the list data structure retrieved from the database using *Psycopg2* to calculate the time spent in seconds for each row. The traverse of the list is make by pairs of elements *uid* and *pid* keeping in a temporal variable the previous pair of elements; in order to calculate the times it is enough to make a subtraction using the *Time* class in *Python* between the actual element and the stored in our temporal variable (corresponding to next and previous time stamp respectively); the subtraction using the *Time* class retrieves the difference between the two time stamps in seconds. This operation is realized while the *uid* column value for the *uid* in the actual position over the list is the same one stored in the temporal variable; when these *uid* are not equal we can assume that there is no more sessions and activities recorded for this user. To conclude this process we have to search tuples of corresponding *pid's* and *uid's* in the list data structure retrieved for *psycopg2* and a simple sum of all the times results in the final time in seconds for a given *pid* and *uid*.

Note: This final time will be stored in a new table.

There is an special *pid* called *logout* this *pid* is the one that indicates that a session has been terminated. This *pid* represents a flag that let us know when the session is finished. This is a useful aspect that helps to calculate the total time spent by session unfortunately not all the users click the logout button, some of them simply close the browser window.

The new table will look as follows:

uid	pid	total time
1598	Dsadas	60.32
1598	Gryty	70.52
236	Jkykkj	80.12
236	gryty	188.45

Table 3.2: Last refinement of time stamps

We are basically transforming the format of *time* from a time stamp into a total time in seconds and merging the rows with the same *pid* and *uid*. Reducing the dimensions of the table and eliminating redundancy.

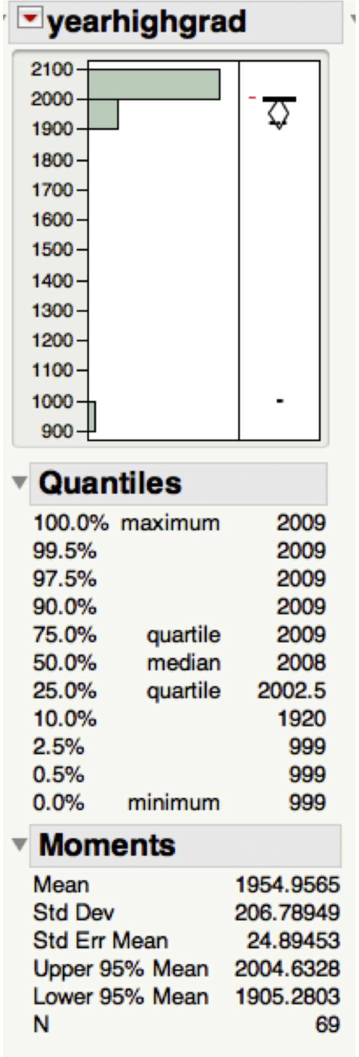
3.1.3.3 Outliers Treatment

In previous sections of this document we mentioned the validation of the user interface in order to reduce or avoid the inconsistencies and noise in the data. Unfortunately these factors are uncontrollable in the paper questionnaire.

The expert in education decided to give values of 999 for the missing information and for the fact of being a paper questionnaire there is no validation of the data inputted into it for the users of IPSims. This situation leads to outliers and bias in our persisted information. Once the data from the paper questionnaire is digitalized we proceed to treat the outliers. Two common techniques are applied and the one that shows better results is the one to be used in the treatment of outliers.

The first technique consists in the substitution of the 5% top and bottom values of a variable for the calculated mean, taking in account the entire column that represent this variable. It means that we take all the registers for the variables, we calculate the mean and then with the resulting mean we substitute the 5% top and bottom values for that column. The second technique is similar to the first one with the difference that instead of using the mean calculated for all the values in the column, firstly we discard the 5% top and bottom values for the column after this we calculate the mean without using in the calculation the discarded values and we obtain a truncated mean that substitutes the discarded values.

The second approach achieves better results for our analysis. This was an expected result given the fact that the elimination of the top and bottom values before the calculation of the average gives a more consistent value for the substitution of our outliers Figure 3.7 shows the results for the variable year of graduation from high school applying the first approach, and Figure 3.8 shows the results of the second approach.

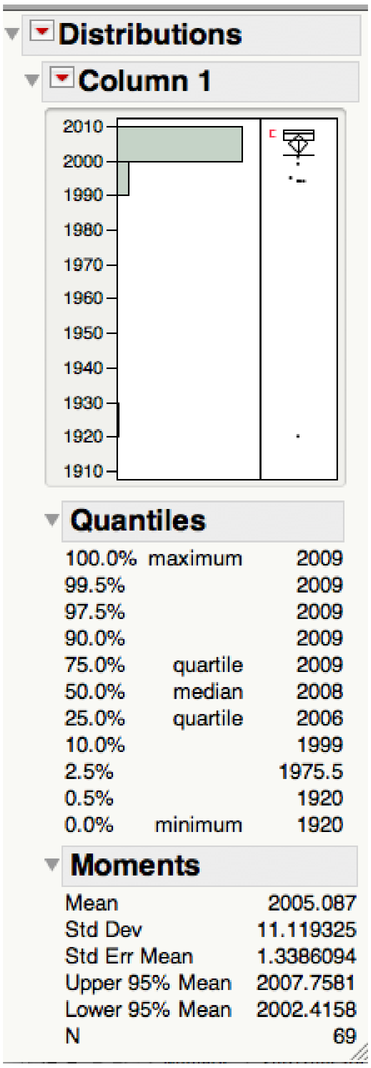


```
Quantiles
100.0% maximum       2009
99.5%                2009
97.5%                2009
90.0%                2009
75.0%     quartile   2009
50.0%     median     2008
25.0%     quartile 2002.5
10.0%                1920
2.5%                  999
0.5%                  999
0.0%      minimum     999

Moments
Mean             1954.9565
Std Dev          206.78949
Std Err Mean      24.89453
Upper 95% Mean   2004.6328
Lower 95% Mean   1905.2803
N                       69
```

FIGURE 3.7: First technique applied to one of our variables

The results for the second approach applied to this variable are better than the results show by the first one. These improved results for the second approach are observed in the other variables as well.



Quantiles

100.0%	maximum	2009
99.5%		2009
97.5%		2009
90.0%		2009
75.0%	quartile	2009
50.0%	median	2008
25.0%	quartile	2006
10.0%		1999
2.5%		1975.5
0.5%		1920
0.0%	minimum	1920

Moments

Mean	2005.087
Std Dev	11.119325
Std Err Mean	1.3386094
Upper 95% Mean	2007.7581
Lower 95% Mean	2002.4158
N	69

Figure 3.8: Second technique applied to one of our variables

3.1.4 Data Mining Techniques on IPSims

Once we have integrated and pre-processed the data we need to select the data mining techniques that fit better on our objectives for this thesis work. In order to accomplish this we need to analyze our research questions:

- There is an identifiable subset of user profile variables that may have the biggest impact in the user final grade?
- It is possible to generate an accurate neural network prediction model for grade prediction?

For the first objective we decided to apply cluster techniques in order to generate groups of students that could be analyzed according to their characteristics. The related literature shows some works about cluster techniques.

Clustering is the technique into the machine learning algorithms that group objects into clusters being the objects contained in the same cluster more similar and the ones from different clusters more different. These clusters are unknown and the objects as mentioned are grouped according to their similarities. An object can belong just to one cluster and no more.

We can find diverse cluster algorithms based on distinct techniques some of the most popular are: Hierarchical clustering, k-Means, EM-clustering, density-based clustering to mention some. For this study that counts just with 69 study cases a simple technique with fast convergence to at least a local minimum is adopted. The limited number of cases is the main reason to take the decision of choosing k-Means as a cluster technique.

A deeper description of this process and how cluster techniques are applied to the IPSim is shown in next chapters.

For the second objective there are multiple options such as neural networks, support vector machines, decision trees. The first abstraction of this problem is to deal with what we call linear separability problem. Where every object to be classified is represented by a high dimensional vector in a high dimensional space. The intention is to separate these high dimensional vectors based on an input-target approach. The input is made by the variables of the high dimensional vectors and the target is the variable used to classify these vectors in a class. Figure 3.9 displays how neural networks trace lines over the plane to solve the linear separability problem.

There are different techniques to solve the linear separability problem and those techniques can provide diverse solutions. The more simplistic tools use straight lines to solve this problem; this can lead to the issue that a straight line cannot separate some of the objects. For these kind of situations we have more sophisticated tools such as smooth SVM that are able to use curves in order to classify elements in complex arrangements.

Sometimes sophisticated not necessarily means better especially for cases where the number of objects to separate is low. Figure 3.10 shows a smooth line that solves the linear separability problem with the use of support vector machines.

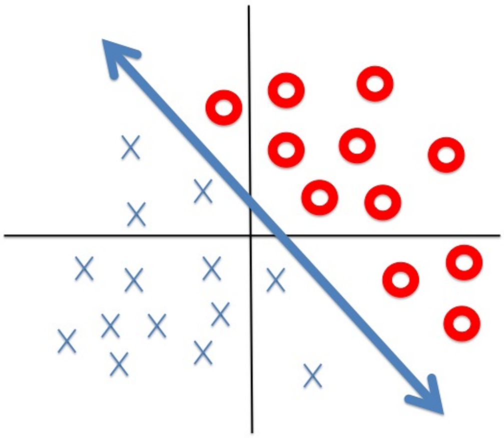


Figure 3.9: Linear Separability with the use of NN

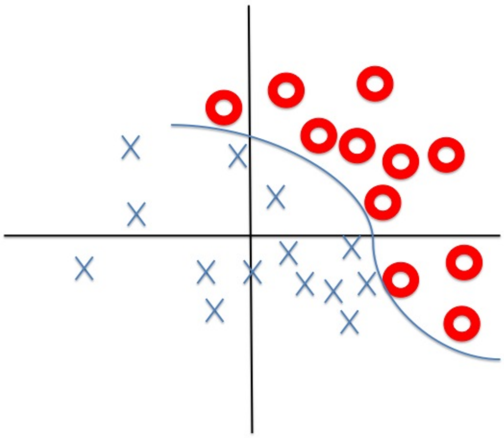


Figure 3.10: Linear Separability with the use of SVM

Given the low number of study cases we decided to use a powerful but at the same time a simple technique for our prediction model. The selected technique is neural network. It will be shown in next chapters that the selection of this tool might be adequate for this work.

3.1.5 Data Evaluation and Interpretation

For this last stage of our process of knowledge discovery within IPSims multiple images and plots are displayed showing the obtained knowledge and the validity of the results.

In some cases a well designed plot or image can be more meaningful that an equation or a paragraph. It is mandatory to select the adequate information-representation tools in order to make our results valuable for the final user.

Chapter 4

Experiments

4.1 Methods

The strategy selected to recruit the participants for our experiments was a complete and structured process.

In order to start with our experimentations we requested permission to instructors of two courses of health sciences at University of Ontario Institute of Technology that showed interest in the use of IPSims. The content of the two courses were related to the nutrition and health area. We suggested them to make IPSims part of the evaluation of their students enrolled in the course. After we got authorization from the instructors the recruitment of participants was performed. The recruitment was realized through a letter of invitation and a telephone script. (Refer to the Appendix for further details.)

This letters of invitation were sent to the students enrolled in the health sciences courses related with the nutrition and health area where IPSims use was approved. We did not offer or provide any kind of compensation to the participants.

In our research experiments there was the possibility of coercing feeling and although any risks were anticipated for the participants there was the chance of feeling demeaned, embarrassed worried or upset, as well there was a chance to suffer from emotional stress. The participation in the study was voluntary and the students were notified that they can remove themselves from the research at any time while still completing their course without penalty.

An information session was imparted to the students in order to show them how to use the IPSims system.

Computer administered tasks and a paper questionnaire were provided to them in order to evaluate their performance within the system activities, obtain feedback about the usability experience, and recollect some extra data for the user profile.

All our recruitment and consent materials provide this contact as well as the clear explanation of the voluntary participation nature. Also we make clear that there are no consequences in case of withdrawal from the study.

Previous to the participation we provided the students with a clear explanation that all the data collected during the experiments would be treated with confidentiality and no individual's data could be identified by name.

Our experiments never identified any individual for that reason the data was coded in a way that no individual could be identified. Anonymity was granted to the students and only aggregate information in which no individual can be identified was used for the presentation of our results.

The data was not available for anyone external to the research team in this study.

We trained the participants that accepted to participate in our research in the use of the system. After the instruction session they proceed to generate their user profile in the login screen of IPSims Figure 1.2 . The students were required to complete the activities that conformed one of the scenarios presented in the IPSims and they were asked to answer the paper questionnaire in order to evaluate their performance.

69 students enrolled in two different health sciences courses were recruited. The information was treated and cleaned for research purposes and the paper questionnaire was digitalized and stored in a secure database. Figure 4.1 presents the workflow of our experimentation process.

4.2 Dimensionality Reduction

New IPSims database is depicted in Figure 3.5. Unfortunately given the low number of study subjects and the high number of variables is incorrect to make a predictive analysis for user behaviour. Also is incorrect to predict the final grade

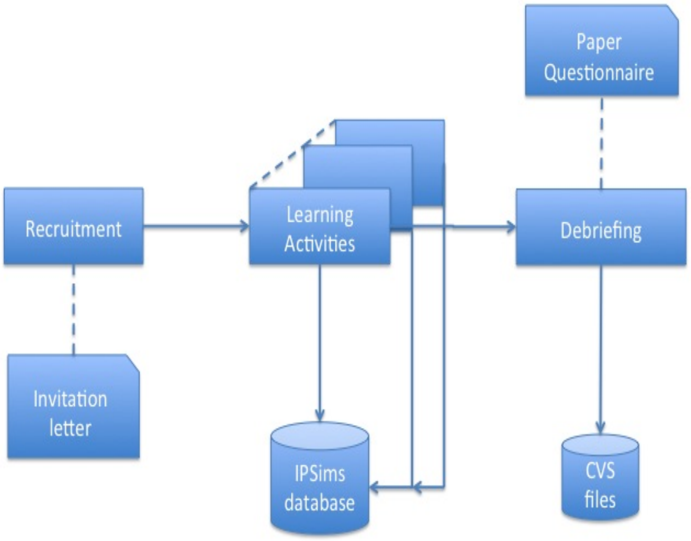


Figure 4.1: Experimentation process workflow

in the activities using all the variables present in our user profile. A dimensionality reduction process is a requirement in order to retrieve non-trivial valid results.

There is not a specific study that gives a specific or optimal value for the sample size or the ratio between study subjects and variables in order to obtain optimal results with prediction experiments. We find diverse literature related with this issue but nothing that we can generalize. The results depend of the environment where the experiments are applied, some of the authors such as: Comfrey et al. [24] related the quality of the experiments to the size of the sample while in the other hand diverse authors explain the quality of the results based on the variables-study subjects ratio Pedhazur [25]. It is hard to make a conclusion but given the reduce size of our sample and the inadequate balance in our variables-study subjects ratio we had to make use of dimensionality reduction techniques.

4.2.1 Factor Analysis

Dimensionality reduction techniques are diverse and they are divided in two big barns: Feature extraction and feature selection. We use in our study factor analysis, a feature extraction technique that allow us to maintain an acceptable study cases-variable ratio (Sheppard [26]).

After we performed factor analysis reduction a Kaiser-Meyer-Olkin measure is obtained with a value of 0.541 which means that the feature extraction results are acceptable. According to the literature values from 0.5 to 0.7 are mediocre, 0.7 to 0.8 good, 0.8 to 0.9 are excellent and greater than 0.9 are superb. Our value fit in the mediocre class. This the only apparently solution to our study case-variable ratio problem we have to based our predictive model results in the dataset generated for this technique.

Once the previous measure is obtained we proceed to analyze the correlation and component matrix, the eigenvalues and variances obtaining a final valid subset of components that is the input of our predictive model. These components are the description of 17 variables conforming the dataset of our user profile. Figure 4.2 depicts the Scree plot for our factor analysis based in principal components. The selected components are hyperplanes that capture the characteristics of our dataset although this description is a mediocre one based in our Kaiser-Meyer-Olkin value. The variables to analyze using factor analysis are shown in Table 4.1 we make use of SPSS. The tool retrieves multiple outputs and diverse criteria can be applied in order to select the final descriptive components of our dataset. For example with the use of Kaiser criterion we keep the components with eigenvalues greater than one. We can make use of the Scree test. Figure 4.2 depicts our scree plot for the analysis. In the Scree test we would select the number of components indicated by the point where the smooth in the curve decrease. We use two approaches to select two diverse datasets, one based in the output of our statistic software that retrieves eight factors as a new variables and one based in our observations of the correlation matrix, significance and variances. The tool extracts the components using principal component analysis as a extraction method, and Varimax with Kaiser normalization for the rotation method.

For the selection of variables through correlation matrix, significance and variances we discarded from our dataset the variables with significance values greater than 0.05 and correlation coefficients greater than 0.9. These values display singularity in our data then we decide to remove them from our dataset keeping just eight variables. The variables selected are:

- Surf hour per week
- High school average

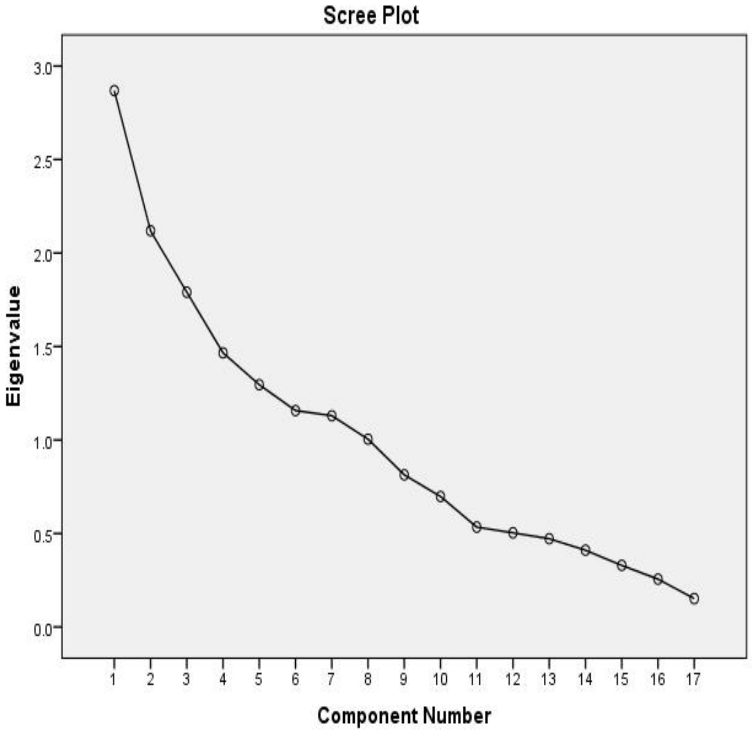


Figure 4.2: Scree plot for our factor analysis

- Pursue of post secondary education
- Age
- Computer literacy
- Likelihood of choosing a career in health sciences
- College or university before University of Ontario Institute of Technology
- Gender

Originally the variable ratio in IPSims was useless with a ratio of 1:0.62. After our dimension reduction analysis we reach a ratio of 1:8.625. The main objective of factor analysis is to discover if our original dataset can be explained appropriately in terms of a reduced dataset. Our description of the original dataset is mediocre but still valid based on the extracted components using factor analysis through principal components.

Variable
surf hour peer week
play hour peer week
marital status
number of children
citizenship
gender
year of high school graduation
high school average
post graduate education
college or university before UOIT
age
computer literacy
likelihood of choosing a career in health sciences
interest in course material
experience with computer based simulations
perceived educational value of computer based simulations
expected grade in the course

Table 4.1: Variables analyzed in factor analysis

4.2.2 Standard Deviation Analysis

We do not relay completely in the results obtained from the factor analysis for dimensionality reduction. A standard deviation analysis is generated as well in order to select an alternative subset of variables for our experiments.

This approach gives us a comparison point to compare our prediction model experiments with two different datasets.

We generate a table for this specific analysis after applying our outlier detection and substitution methods mentioned in Chapter III subsection 3.1.3.3. Table 4.2 displays the generated results.

After the calculations of our standard deviation we select the eight variables with the highest standard deviation if the $standard \quad deviation \geq 0.5$ considering this a reasonable variability of our dataset, contributing to a wider exploratory space within the solution space.

The selected variables were:

- surf hour peer week.
- play hour peer week.
- year of high school graduation.

Variable	**Standard Deviation**
surf hour peer week	15.30
play hour peer week	1.72
grade	1.64
marital status	0.45
number of children	0.42
citizenship	0.12
gender	0.21
year of high school graduation	11.12
high school average	1.23
post graduate education	0.44
college or university before UOIT	0.61
age	2.85
computer literacy	0.64
likelihood of choosing a career in health sciences	0.28
interest in course material	0.63
experience with computer based simulations	0.62
perceived educational value of computer based simulations	0.42
expected grade in the course	0.62

Table 4.2: Standard Deviation Analysis Table

- high school average.
- age.
- computer literacy.
- interest in course material.
- expected grade in the course.

These variables are part of another dataset to test our prediction model experiment.

Some of the variables histograms are depicted in Figure 4.3, Figure 4.4, Figure 4.5 and Figure 4.6. These histograms give us a general idea in the distribution of our variables of study.

Chapter V provides a rationale for the results obtained and the impact of selecting certain variables.

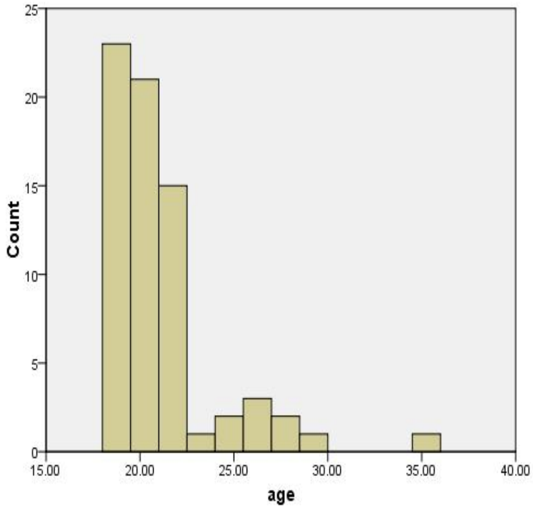


FIGURE 4.3: Distribution of the variables age

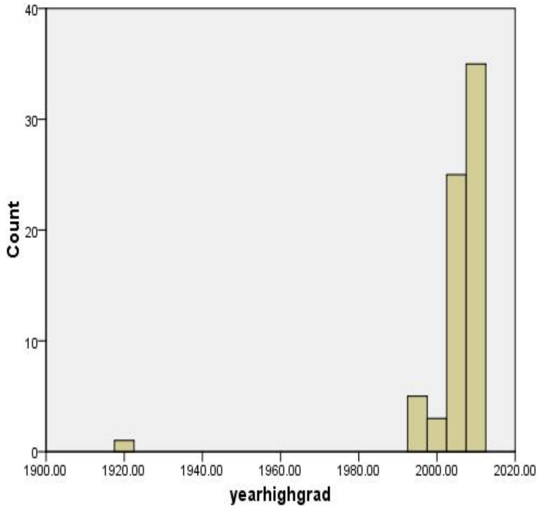


FIGURE 4.4: Distribution of the variable graduation year from high school

4.3 Cluster Analysis Experiment

For the cluster analysis experiment our intention is to observe the clusters of students trying to determine specific and clear characteristics that defines each cluster. This with the intention to identify a subset of variables that may have the highest impact in the student's grade in the IPSims activities.

Our study subjects are abstracted as high dimensional vectors conformed by

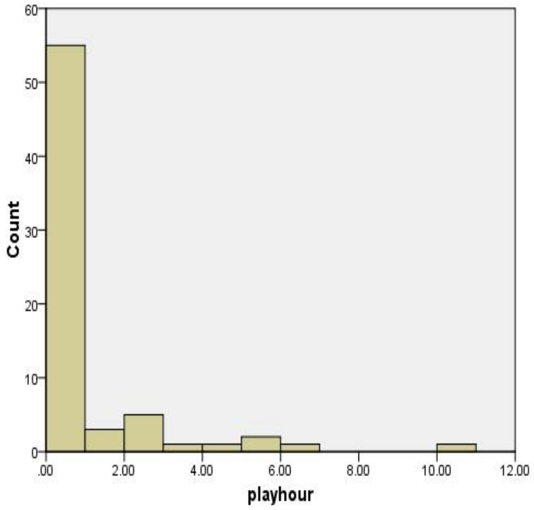


FIGURE 4.5: Distribution of the variable play hour per week

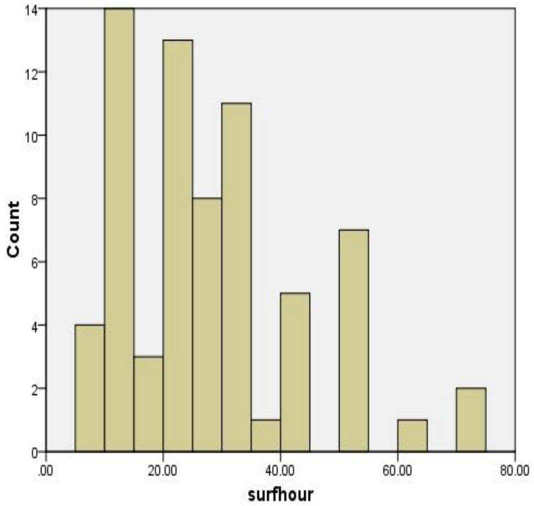


FIGURE 4.6: Distribution of the variable surf hour per week

our set of variables for the user profile. These high dimensional vectors are the input for our cluster algorithm.

The cluster algorithm applied is K-Means and it was developed in *Python*.

Our cost function based in the minimization of the cluster sum of squares between elements of a generated cluster over each iteration is:

$$\arg\min_{\mathbf{C}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in C_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

Where k represents the number of clusters defined as a parameter before the run of our K-Means algorithm. The n high dimensional vectors (69 students) are represented by $(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$ and $\mathbf{S} = \{C_1, C_2, ..., C_k\}$ represents the k partitions (clusters) of our 69 study cases.

μ_i is the calculated mean for the Euclidean distance of the elements contained in C_i

The Euclidean distance for the n high dimensional vectors (students) $\mathbf{x}_p$ and $\mathbf{x}_q$ with values for the user profile variables $\mathbf{x}_{ps}$ and $\mathbf{x}_{qs}$ with $s \leq dim$; being dim our high dimensional vectors size and $\mathbf{x}_p\mathbf{x}_q \in C_i$ the Euclidean distance is defined as:

$$\mathrm{d}(\mathbf{x}_p, \mathbf{x}_q) = \mathrm{d}(\mathbf{x}_q, \mathbf{x}_p) = \sqrt{(x_{p1} - x_{q1})^2 + (x_{p2} - x_{q2})^2 + \cdots + (x_{pm} - x_{qm})^2}$$

$$= \sqrt{\sum_{m=1}^{n} (x_{pm} - x_{qm})^2}$$

With $i \leq k$ being k the number of clusters.

Our algorithm works based on the optimization of these functions trying to generate the best quality clusters. K-Means guarantee a fast convergence at least for a local optimum the recommended number of iterations to reach the convergence is 1000. Our algorithm stop conditions is accomplished when there is no change in our cost function value which means we reach at least a local minimum or 1000 iterations condition.

We run our algorithm multiple times with distinct starting points with the intention to measure the quality of our generated partitions based on our cost function. In order to keep track of the high dimensional vectors behaviour within the K-Means algorithm a special class called Vector is generated. Our class diagram is displayed in Figure 4.7.

- The property **elements** refer to the variables which conform the high dimensional vector.
- The property **uid** (user ID) is the **uid** corresponding to an specific high dimensional vector. **uid** is not used in the calculations, it is just used to keep track of the objects and where are they located in our high dimensional space.
- The property **dimension** is the dimension of our vector.

Vector
elements
uid
dimension

Figure 4.7: Vector class definition

Cluster
vectors
dimension

Figure 4.8: Cluster class definition

After this we proceed to generate a Cluster class shown in Figure 4.8

- The property vectors refers to all the vectors that conformed the cluster.
- The property dimension refers to the dimension of the cluster and is essentially used to verify that all the vectors keep the same dimension in counter case we raise an exception.

  Empty clusters are not allowed we verify in the process that any cluster is empty.

The well known algorithm is based on the Euclidean distance and can be described by the following pseudocode:

```
$Select\ \ Centroids\ \ Randomly\ \ from\ \ the\ \ High\ \ Dimensional\ \ Vectors\ \ c_1, ..., c_k$
while $CentroidsVariation \leq minval \quad or \quad iterations = 1000$ do
    $Cluster\ \ High\ \ Dimensional\ \ Vectors.$
    for $i = 1, \rightarrow n$ do
        $Assign\ \ Vector\ \ p_i\ \ to\ \ the\ \ cluster\ \ whose\ \ centroid\ \ c_j\ \ is\ \ closest$
    end for
    $Update\ \ the\ \ cluster\ \ centroids$
    for $j = 1, ..., k$ do
        $n_j \quad \leftarrow \quad numberofpoints \quad in \quad C_j$
        $c_j \quad \leftarrow \quad \frac{1}{nj} \sum p_i \in C_j p_i$
    end for
end while
```

A trial and error approach as well as the fact to try to keep the number of clusters close to the grading scale used in the school: A, B, ... F are the criteria used to select the best suitable number of clusters. We could have saved time using the five clusters from the beginning to satisfy our needs for this particular work, however we decided to test other options in order to evaluate the quality of clusters for different values of the parameter K.

It is known that one of the weaknesses of K-Means is the correct selection of the number of clusters K. There is no a standard technique to know the optimal number of clusters. We used trial and error tests between 3 to 10 clusters obtaining five clusters as the best option within this range. This selection was appropriate given that the GPA ranking grade has five classes as well. Techniques like genetic algorithms can be implemented in large datasets in order to determine the value of K, this implies a considerable computational effort. In our case after the testing runs the measure to determine the best K value for our cluster technique are based in our cost function and the average Euclidean distance. It can be observed that our cost function is minimum when our average Euclidean distance between elements in the same clusters reaches their minimum.

```
$After \quad K - means \quad runs \quad finalization$
for all $Clusters$ do
    for $i = 1, \rightarrow ClusterSize$ do
        for $j = 1 \rightarrow VectorSize$ do
            $Distance \quad \leftarrow \quad Distance \quad (X_j \quad - \quad CentX_j)^2$
        end for
        $SumOfDistances \quad \leftarrow \quad SumOfDistances + Distance$
        end for
        $AverageDistance \leftarrow \frac{SumOfDistances}{ClusterSize}$
    end for
```

The validation of our code was based in the cost function and we used a simple approach for the testing of our K-means algorithm. Our logic follows the principle that every time we increase the number of clusters, the average distances between elements of the same cluster is decreasing just if these elements are correctly classified. When we define 69 clusters we expect that each of our observations below to one of these 69 clusters having an average distance of 0 for our cost function. If this situation happens that means that our algorithm is working properly. In the other side if we can not see this situation after define a value of N for the parameter K having N study subjects then something is not working properly in our algorithm. Fortunately our code works properly in the testing making our results trustable and valid.

After multiple runs with the algorithm the objects are identified in their clusters and located in five different tables for calculations. Each cluster is represented for a table with its own vectors (student identified by user ID) and respective variables value. The next step in the experiment is a straightforward strategy. We calculate the mean of each variable by cluster we also calculate the mean of our pivot variable. Our pivot variable is the *grade* obtained in the IPSims activities; this variable is the reference to determine if the clusters reflect any kind of pattern based on the user profile variables and the *grade*. Once the means are calculated we proceed to organize the clusters from the lowest to the highest mean *grade*. One of the tables generated for one cluster in our cluster analysis is shown in Table 4.3.

Cluster techniques are applied in order to classify a group of elements in subgroups according to similarities in the objects that below to a same class.

Our pivot variable or point of reference is the final *grade* in the IPSims activities. Based on the observed values for this variable our expectation is to generate clusters where the relationship between variables value and grades is clear. For example It is expected to see a cluster with a low grade where the values for the variables: computer literacy, likelihood choosing a career in health sciences, desire to pursue post-secondary education, the high school average and the experience with computer based simulations display low values. In the other side we expect clusters with a higher grade where the values for these variables are higher.

uid	14857	11558	7324	50969	104	1205	51454	57194	Mean
surfhour	26.11	10	20	15	15	26.11	20	10	17.77
playhour	0	0	2	5	0	0	0	0	0.87
marital	2	2	1	1	4	1	1.10	2	1.76
children	0.10	2	0	0	0	0	0	2	0.51
citizenship	1	1	1	1	1	1	1	2	1.12
yearhighgrad	1994	1994	1994	2004	2001	1995	2002	1994	1997.25
highavg	5	4	3	4	5	6	3	2	4
posteducation	2	2	1	2	2	2	1.28	2	1.78
collegeuniversity	1	1	0	0.37	1	0.37	2	1	0.84
age	20.94	20.94	19	24	29	35	26	20.94	24.48
gender	2	2	2	2	2	2	2	2	2
comskill	5	3	4	4	4	4.33	5	4	4.16
likelihood	3.92	3.92	3.92	4	3	4	4	4	3.84
interest	3.14	3.14	2	3	3	3	3	3	2.91
experience	2.74	2.74	3	3	2	2	3	3	2.68
perceived	3.08	3.08	3	3	2	3	3	3	2.89
grade	10	12	9	10.33	11	11	8	8	9.91

TABLE 4.3: Cluster Analysis Table for a Generated Cluster

Some of this expected assumptions are display by the experiments, some others are not and this is our criteria to select the subset of variables that may reflect the biggest impact in the students final grade in the IPSims activities.

Figure 4.9 and Figure 4.10 are plots of some of our results where the expected patterns can be observed.

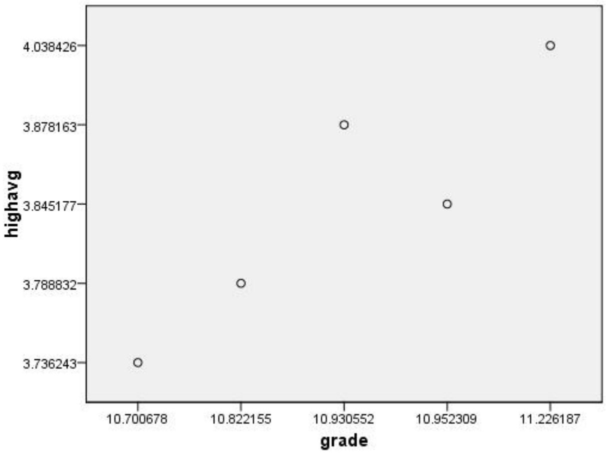


FIGURE 4.9: High school average vs grade by cluster

After analyzing and observing the generated plots we discovered some interesting patterns, which could support our assumptions of a descriptive subset of

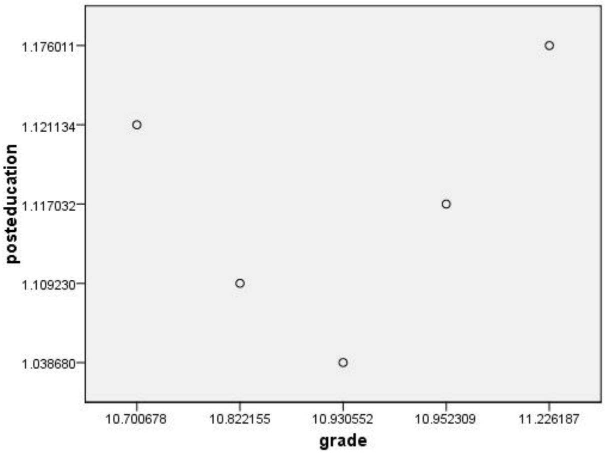


FIGURE 4.10: Intention to continue post graduate studies after UOIT vs grade

elements that may have the biggest impact in the final grade activity.

As an example of observed patterns Figure 4.9 depicts that the cluster with the lowest high school average is the one displaying the lowest grade in the activities as well as the cluster with the highest high school average is the cluster with the highest average grades in the activities in IPSims. We can observe almost a linear pattern that indicates that the high school average is a strong candidate to be one of the variables that may have the biggest impact in the grade of the students for the IPSims activities. These kinds of plots give us a positive indicator that it might be possible to find a subset of variables that may have the biggest impact in the student activities.

The subset of variables that may have the biggest impact in the student's final grade is a reference point to generate mining rule associations. The same logic is followed to analyze all the variables in the cluster in order to find which variables could be the ones that may have a real impact in the final grade of the activities within IPSims.

Another interesting visualization tool is the generation of 3D plots. These plots allow us to locate some interesting regions where a correlation between tuples of variables and the grade can be observed.

Figure 4.11 and Figure 4.12 are examples of this 3D plots depicting the interesting regions between tuples.

The use of surfaces is a supportive tool to analyze tendencies between pair of variables (Figure 4.13, Figure 4.14 and Figure 4.15).

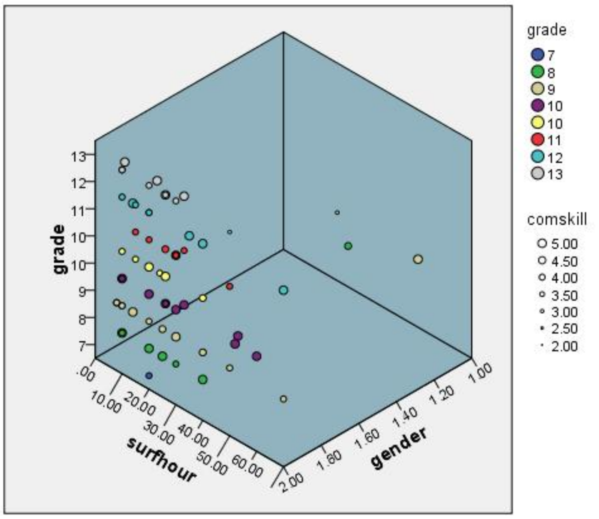


FIGURE 4.11: Grade vs surf hour vs gender 3D Plot

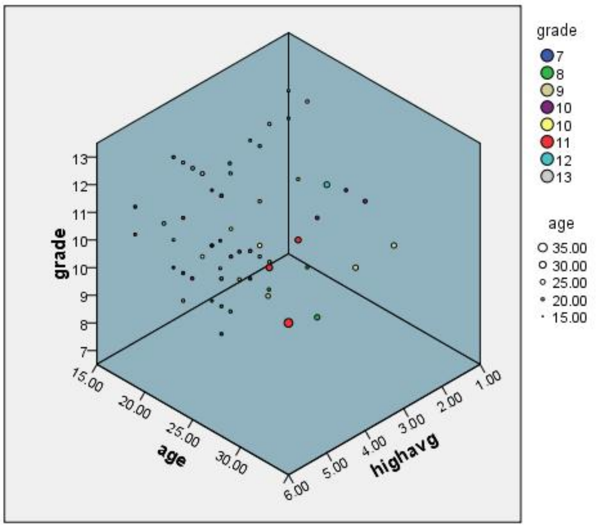


FIGURE 4.12: Grade vs high school average vs age 3D plot

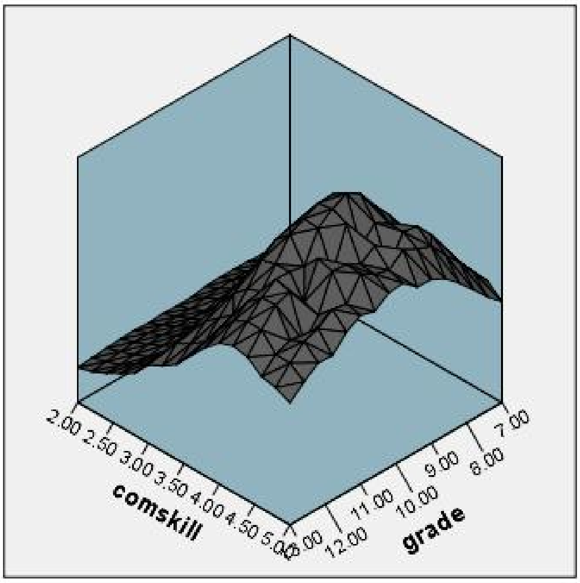


FIGURE 4.13: Surface grade vs computer literacy

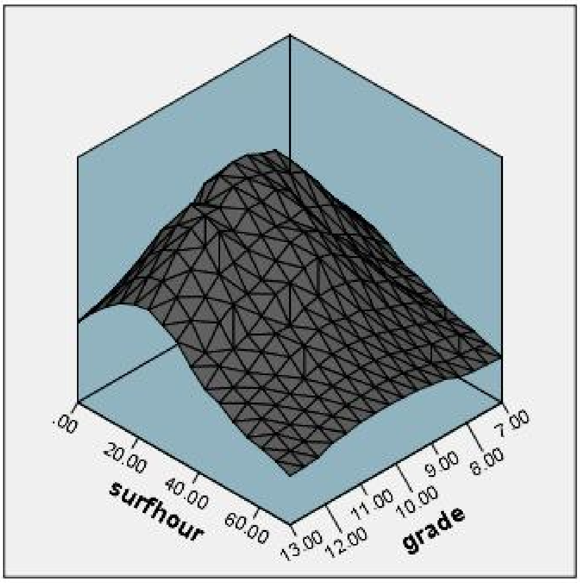


FIGURE 4.14: Surface grade vs surf hour peer week

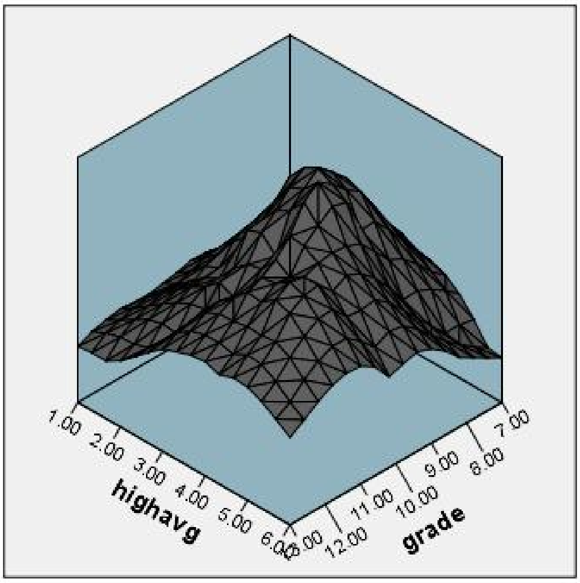


FIGURE 4.15: Surface high school average Vs grade

The analysis of surfaces with three variables depicts unclear and undetectable patterns.

Cluster analysis experiments are usually really intuitive and the success of cluster analysis research relay on the adequate interpretation and representation of the results. In our case we are showing more than acceptable plots and measures for cluster quality evaluation. We are displaying some interesting patterns that can lead to association rules and identification of specific groups of students.

The described process for our cluster analysis experiment is applied to the complete user profile dataset and the dataset generated through standard deviation analysis for dimensionality reduction. Our intention with this experiment is to show that there is a subset of variables that may have the biggest impact in the final grade of the students within the IPSims activities. We apply the algorithm in two different dataset with the intention to have two different approaches to compare our results. It is not a surprise that the cluster analysis for the entire user profile dataset provides more descriptive results. It may be unnecessary to apply the experiment over a reduced dataset like in the case of our standard deviation analysis dataset; but it is interesting to observe the behaviour of our study subjects under different conditions.

4.4 Prediction Model Experiment

The datasets explained in the dimensionality reduction section are the ones to be used in this experiment. The intention with this experiment is to generate a prediction model based on neural networks that may be capable to predict the final grade of a student within the IPSims activities based on the user profile.

For this approach it is important to select a valid training and testing set. Being the training set the most important one; the success of our result depends on the selection of our training set.

Our intention is to classify our observations into classes representing our grading scale A, B, C, D, F or our obtained grade from 0 to 19. This means that we have to be able to identify five different sets of observations for the scale A,B,C,D,F or 19 sets for IPSims grading scale both of them based on the user profile variables after dimensionality reduction is applied.

The classification approach depends on the coding that we use for our target variable (grade within the IPSims activities). We can decide to represent our target variable as a number from 0 to 19 or as a class in the range A, B, C, D, F. It is clear that using an A, B, C, D, F approach we are making wider the tolerance of error for the model.

Our prediction classifier is a neural network model which defines a function $f : X \to Y$ where X represents our student user profile and Y is the predicted class for the object.

The neural network model can be visualized as a linear weighted sum with the form: $f(x) = C(\sum_i w_i g_i(x))$; where C represents an activation function and $i \in \zeta^+$ being w_i the weights of the edges connecting the network layers and g_i is a collection of functions.

We need to define a general cost function for our specific case.

Let's define X_0, X_1, X_2, X_3 , X_4being these sets of points (sets of students) in a high dimensional space of size n.

$X_0, X_1, X_2, X_3, X_4 \in \Re^n$;

These set of points represent the grades A, B, C, D, and F respectively (if we define 20 X_i then the X's represents the classes from 0 to 19 for the IPSims grading scale).

It is our intention to differentiate X_i with $0 \leq i \leq 4$ from the rest of the non-selected classes.

The non selected classes are defined by X_j with:

$X_j = X_0 \bigcup X_1 \bigcup X_2 \bigcup X_3 \bigcup X_4 - X_i;$

then we can differentiate X_i from:

X_j if there is $m_1, m_2, ..., m_{n+1} \in \Re$ that satisfies the equations:

$\forall x \in X_i; \quad \sum_{k=1}^{n} m_k x_k \geq m_{n+1}$

$\forall x \in X_j; \quad \sum_{k=1}^{n} m_k x_k < m_{n+1};$

where x_k is one of our 69 students.

It is our intention to identify the hyperplanes that differentiates the classes from each other in order to generate a model that may predict the grades accurately.

The previous described function requires just one output given that our abstraction for the classification process is: $x \in X_i$ or $x \notin X_i$ (one or another not both). Then we can classify our element using just one output.

Also we can use five outputs if we visualize our classification process as: $x \in X_0 \Rightarrow x \notin X_1 X_2 ... X_n$ This case can be generalized for our two different grading scales, based on the A,B,C,D,F notation or 0 to 19 IPSims grading rank.

The format of the target variable defines the approach to be used. After some experimentation we decided to use one output given the fact that it shows best results compared to the other architectures.

The selection of the training set is based on the elements showing more diversity in their variable values this provide us with a wider exploration of our solution space for our neural network system training stage.

We select the training set with the use of an iterative greedy algorithm. We generate random datasets keeping the datasets that presents best results in our predictions.

Our dataset will be conformed by 75% percent of our study cases.

The greedy algorithm works as follows:

```
**for** $i = 0, \rightarrow i = (n * 0.75)$ **do**
```

```
        $Element \leftarrow Random(1, dataSet[i])$
        $TrainingSet.append(Element)$
        $DataSet.pop(i)$
    end for
    for $j = 0, \rightarrow j = m$ do
        $TestingSet.append(dataSet[j])$
    end for
```

Different neural network architectures are tested in order to select the best network to predict the final grade in the activities. Cross validation approach is used over the generated training sets by the greedy algorithm obtaining the best results with the distributed training algorithm. At the end we come with an architecture based on eight input nodes with bias, a hidden layer with six neurons and one output being this architecture with the best prediction results. We based the selection of our final architecture in a trial and error approach keeping the one with the best prediction measures.

The other tested architectures are:

- Eight inputs with no bias, six neurons in the hidden layer and five outputs.
- Eight inputs with bias, six neurons in the hidden layer and five outputs.
- Eight inputs with no bias, six neurons in the hidden layer and one output.
- **Eight inputs with bias, six neurons in the hidden layer and one output.**

The neural networks are constructed using the rule of thumb of neural networks: The number of inputs has to be equal to the number of variables, the number of neurons in the hidden layer has to be more less 75% of the number of inputs and the number of outputs is preferably to be equal to the number of classes. As mentioned before in our work one output neuron presents the best prediction results.

Our neural network throws a hyperplane in our high dimensional space trying to split the study cases in their corresponding group (grade) based on the selected subset of variables obtained from the user profile. All the neurons in the neural network are arranged as a feed forward neural network (perceptron) fully connected. Figure 4.16 depicts a neural network example and Figure 4.17 shows a representation of a feed forward network.

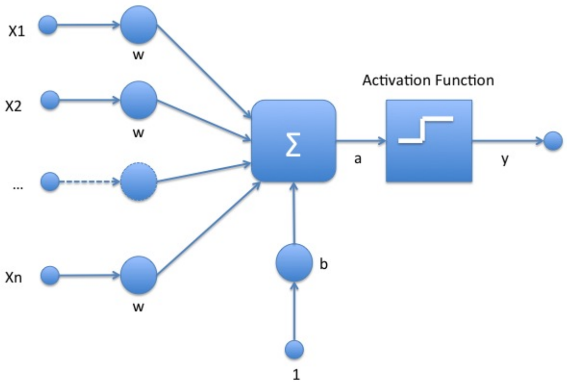


FIGURE 4.16: A diagram of a Perceptron

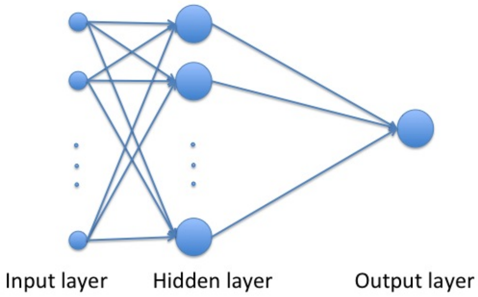


FIGURE 4.17: Feed forward network representation

Different training algorithms are used some of them displaying good results but others in counterpart throws really bad results. It is still a concern the low numbers of study cases having just 69 subjects 75% are used for the training and 25% for the testing. Some of the prediction model training algorithms display really good results and some others show a poor performance.

When more students use the system more predictions can be made with these models and with this we might be able to confirm the accuracy and validity of our results.

Neural networks are selected over other tools by their simplicity. Normally we understand the computational process as a finite number of structured steps that accomplish with a goal. If the computer does not know the steps the problem can not be solved.

In order to be able to generate what we call an algorithm we need to understand the nature of our problem and how it could be solved. When we do not understand the solution to a problem we are not able to generate the steps that will guide the computer to solve the problem. When our computational capabilities are limited in the understanding of a given problem then we can address techniques like neural networks.

Neural networks are an alternative that allows the computer to solve problems that programmer does not understand or does not know how to solve. A neural network is trained with an input dataset therefore this training provides to the neural network with the ability to recognize patterns and generate general activation rules to respond in an accurate way to unexpected inputs (new cases).

According to the training and the recognized patterns by the neural network in this stage, the output neurons are activated displaying the best learned value for the given input. We do not understand how the selected input variables affect or determine the final grade of the user. We do not know either how to generate a sequence of finite instructions that indicates the computer how to predict a final grade based on the input variables. Then our natural choice for simplicity and definition is the use of neural networks.

There is one big concern with the selection of this tool and is that the number of training elements is small. Neural networks require a large diversity in the training set in order to receive an appropriate training. In our study we lack of a big number of study cases we tried to overcome this issue with the dimensionality reduction stage and the selection of a random training data set that displays the biggest variance within variable values. This allows us to expand our exploring space within the solutions universe. Without the help of this tools most likely our exploring space would be reduce and the predictions of our neural network might display trivial or inaccurate results.

The tools used for the construction and training as well testing of the neural networks were *Python* libraries such as *numpy* [27], *ffnet*[22], *networkx*[28],*matplotlib*[29].

The tested training algorithms are:

- Distributed training algorithm.

- Conjugate gradient training algorithm.
- Genetic optimization training algorithm.

4.4.1 Training Algorithms

Genetic Optimization Training Algorithm

This algorithm works based on a Fortran subroutine called PIKAIA generated in HAO (High Altitude Observatory) it is a general purpose function optimization based on a genetic algorithm. Paul Charbonneau and Barry Knapp wrote PIKAIA in 1995 both then at HAO/NCAR. Version 1.2 was released in April 2012. We are using the version 1.0 wrapped in Python.

More documentation and the implementation can be found at HAO website [30].

Distributed Training Algorithm

This algorithm has two modalities a multiprocessing version used when we have more than one core or processors in our system and the single process version that works when we just have available a single core or processor.

The next pseudocode shows how the training algorithm works. The difference between the single processing or multiprocessing is the initialization of our pool of processors with a single thread or multiple threads:

B_k represents $\nabla^2 f_k$ which is the Hessian, d_j denote the search directions and the sequence of iterations that it generates by z_j

```
$Given \quad initial \quad point \quad X_0$
for $k = 0, \rightarrow n$ do
    $Define \quad tolerance \quad \epsilon_k \leftarrow min(0.5, \sqrt{\|\nabla f_k\|})\|\nabla f_k\|$;
    $Z_0 \leftarrow 0, \quad r_0 \leftarrow \nabla, d_0 \leftarrow -r_0 \leftarrow -\nabla f_k$;
    for $j = 0, \rightarrow m$ do
        if $d_j^T \quad B_k \quad d_j \leq 0$ then
            if $j = 0$ then return $p_k \leftarrow -\nabla f_k$
            elsereturn $p_k \leftarrow z_j$
            end if
            $\alpha_j \leftarrow \frac{r_j^T r_j}{d_j^T B_k dj}$
            $z_{j+1} \leftarrow z_j + \alpha_j d_j$
            $r_{j+1} \leftarrow r_j + \alpha_j B_k d_j$
        **end if**
        **if** $\|r_{j+1}\| l\epsilon_k$ **then return** $p_k \leftarrow z_{j+1}$
        **end if**
        $\beta_{j+1} \leftarrow \frac{r_{j+1}^T r_{j+1}}{r_j^T r_j}$
        $d_{j+1} \leftarrow -r_{j+1} + \beta_{j+1} d_j$
    **end for**
**end for**
$x_{k+1} \leftarrow x_k + \alpha_k p_k, where \quad \alpha_k \quad satisfies \quad the \quad Wolfe, \quad Goldstein$
$or \quad Armijo \quad backtracking \quad conditions(using \quad \alpha_k \leftarrow 1 \quad if \quad possible)$
```

Pseudocode extracted from Nocedal et al. [31].

Conjugate Gradient Algorithm

This algorithm is based on the Polak-Ribiere method extracted from Nocedal et al.[31].

```
$Given \quad x_0$
$Evaluate \quad f_0 \leftarrow f(x_0), \nabla f_0 \leftarrow \nabla f_(x_0)$
$p_0 \leftarrow -\nabla f_0, k \leftarrow 0$
**while** $\nabla f_k \neq 0$ **do**
    $Compute\alpha_k \quad and \quad set \quad x_{k+1} \leftarrow x_k \quad + \quad \alpha_k p_k$
    $Evaluate\nabla f_{k+1}$
    $\beta_{k+1}^{FR} \leftarrow frac\nabla f_{k+1}^T(\nabla f_{k+1} - \nabla f_k)\|\nabla f_k\|^2$
    $p_{k+1} \leftarrow -\nabla f_{k+1} + \beta^F R_{k+1} p_k$
    $k \leftarrow k + 1$
**end while**
```

Where according to Nocedal et al. [31] α_k represents the step length (which minimizes ϕ along the search direction p_k), we need to perform a line search that identifies an approximate minimum of the nonlinear function f along p_k.

Nocedal et al.[31] mentioned: "Unlike the linear conjugate gradient method, whose convergence properties are well understood and which is known to be optimal as described above, nonlinear conjugate gradient methods possess surprising, sometimes bizarre, convergence properties".

It is out of the scope of this work to analyze in a deeply form the properties of the training algorithms. The intention for the selection of different algorithms is to have variability in our prediction results in order to select the ones that behave more accurate for our input dataset.

The measures to rate the success of every algorithm are the R-squared measure, the maximum squared error and accuracy rate. Multiple runs are performed and average values are calculated also the runs of every algorithm are being displayed in a plot. The distributed training algorithm is the one that displayed best results under this criterion.

The neural networks are executed with the selected inputs for the greedy algorithm, then with that input each of our three neural network architectures are trained and tested with each of our three different training algorithms. The R-square measure and maximum squared error are recorded. After save these values we proceed to select a different input set using our greedy algorithm we repeat the process.

The following formula is the one to measure the accuracy rate based in the precision/recall analysis for a not binary classification:

$$Accuracy = \frac{\#registersindatasetwhereM(t)=C(t)}{\#registersindataset}$$

Where *M(t)* is the predicted value and *C(t)* is the original value for the record *t* in our dataset.

The classification error for our model can be defined as the proportion of mistaken classifications for our dataset: *err* = 1 - *Accuracy.*

A binary version of this formula is applied in order to measure precision and recall over a binary form of our experiments.

We define a threshold to divide the students that pass or fail the course. IPSims grading scale goes from 0 to 19 therefore we define five division classes for the ranking A,B,C,D and F as depicted in Table 4.4:

A	18-19
B	13-17
C	9-13
D	5-9
F	0-4

Table 4.4: Grading rank Table

The students getting an F fail the activity. Based on this class division and the binary version of the precision/recall analysis:

$$Precision = \frac{TruePositives}{\#PredictedPositive} = \frac{TruePositive}{TruePositives+FalsePositives}$$

$$Recall = \frac{TruePositives}{\#ActualPositives} = \frac{TruePositives}{TruePositives+FalseNegatives}$$

$$Accuracy = \frac{TruePositive+TrueNegatives}{TruePositives+TrueNegatives+FalsePositives+FalseNegatives}$$

Where our true positives are the values where we predict that the student will pass the course based on the user profile and our assumption is true.

Our false positive means that we predict that the student will pass the activities within the IPSims based on the user profile but our assumption is false the user actually fails.

The false negative is the prediction that our student will fail the activities within the IPSims when the student actually passes the activities.

True negative indicates that we predict the student will fail and actually the student fails the activities within the IPSims.

We evaluate our model results with these formulas in order to obtain measures for the quality of our predictions. These measures help us to evaluate the success of our training stage.

Hamalainen et al. [8] mention that in educational data mining: "The training sets are so small that they cannot capture the real distribution and the resulting classifier is seriously biased. Therefore, we should somehow estimate the generalization error on unseen data. A common solution is to reserve a part of the data as a test set." In different of the analyzed works in the related literature there is not a complete description in how the system is trained, or which percentage is using for training and testing. Our work clearly defines that 75% of our records are used as a training set while the other 25% of the cases will be the testing set. We present a clear evaluation of the accuracy and effectiveness of our model for the presented cases.

The Table 4.5 show some of the averages obtained for the measures R-squared and maximum squared error when using factor analysis reduction dataset.

Algorithm	R-squared	Maximum Squared Error
Distributed Training Algorithm	0.70-0.80	0.15-0.24
Conjugate Gradient Training Algorithm	0.28-0.55	0.32-0.78
Genetic Optimization Training Algorithm	0.27-0.31	0.89-0.96

Table 4.5: R-squared and MSE for each training algorithm using our correlation, significance and variance analysis dataset

The neural networks are trained multiple times with different algorithms and with the same training set being tested 12 different times in order to observe how the neural network behave with the dataset, the results from each run had a big difference in R-squared and maximum squared error measures between each training. Figure 4.18 depicts the process to evaluate the quality of our neural networks. The evaluation process has a stage to determine the quality of our predictions. Table 4.5 displays some minimum and maximum values for the evaluation process using one of our datasets. We are looking forward for next courses where the IPSims will be used in order to keep testing our prediction model and the validity of our results.

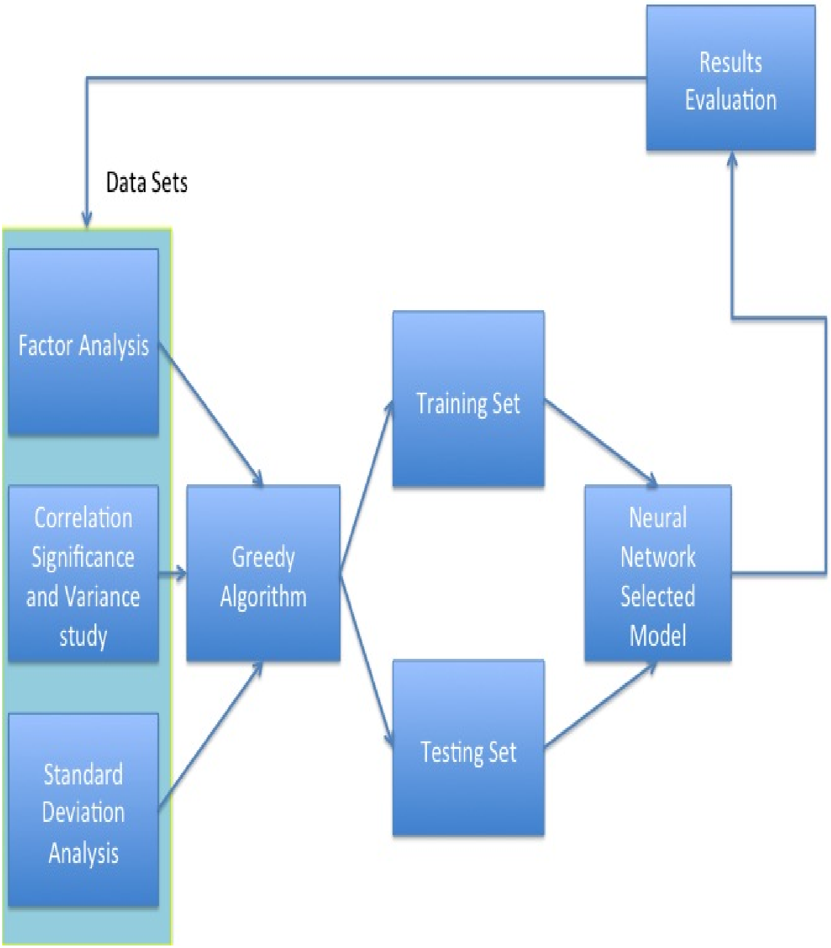

Results Evaluation
Data Sets
Factor Analysis
Correlation Significance and Variance study
Standard Deviation Analysis
Greedy Algorithm
Training Set
Testing Set
Neural Network Selected Model

Figure 4.18: Plot generated for the conjugate gradient algorithm

The Figures 4.19, 4.20 and 4.21 are plots randomly selected from our 12 different runs for each algorithm. The distributed training algorithm always shows better results.

As observed we provide three valid measures for predictive models that ranks our predictions. We observed good results for the distributed training algorithm but we still required more test cases to analyze the model behaviour with more unexpected inputs not presented in our actual dataset.

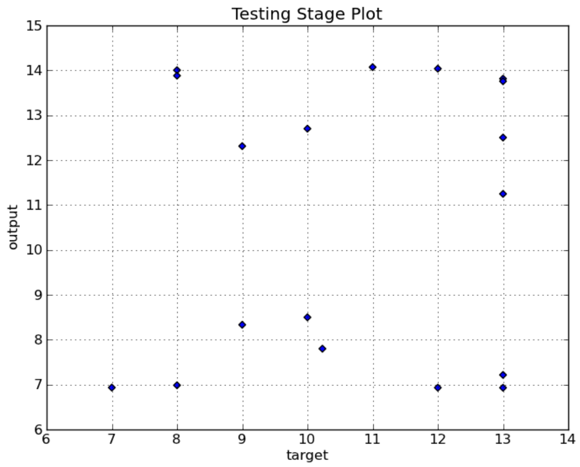


Figure 4.19: Plot generated for the conjugate gradient algorithm

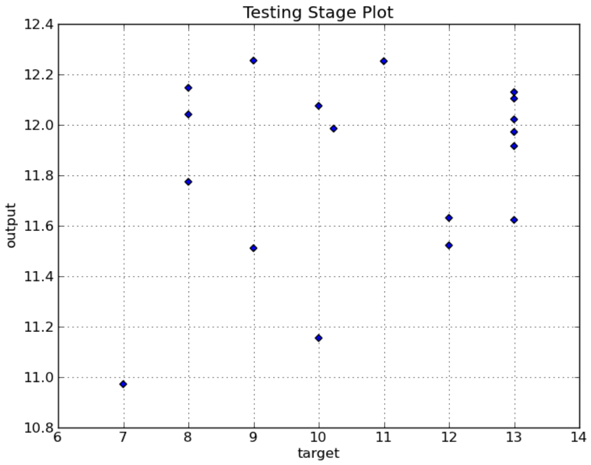


Figure 4.20: Plot generated for the genetic training algorithm

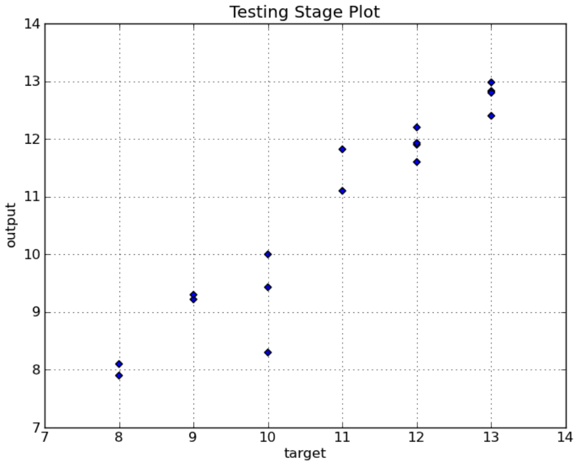


FIGURE 4.21: Plot generated for the distributed training algorithm

Chapter 5

Conclusions

5.1 Summary

The present thesis work showed some interesting results regarding the IPSims. We presented complete data mining process in a digital media learning environment.

Our first proposal was the design of a data mining framework for the IPSims that allow us to obtain an input dataset for our machine learning algorithms. This proposed framework has become an efficient data mining system for the IPSims in order to realize diverse studies and analysis related to the educational data stored in IPSims.

In our study process diverse computer science and math areas were involved such as: databases, artificial intelligence, statistics, information visualization. Our thesis work presents a data mining system capable to unify, pre-process, extract and analyze data from the educational field in order to retrieve valuable knowledge that could be used for educational purposes in further studies.

Being our two main premises the detection of an identifiable subset of variables which may have the biggest impact in the final grade in the IPSims activity and the generation of an accurate prediction model of the final grade based on the user profile variables; the present work presented a justification and validation of the obtained results involving the two main objectives.

We applied diverse measures to evaluate the quality of our clusters generated by the cluster technique and also we used statistic measures in the evaluation of the predictions for the neural network predictive model being these measures satisfactory enough for this specific and reduced dataset for experimentation.

We generated reduced datasets for the user profile variables based in factor analysis for dimensionality reduction and standard deviation analysis. Our resulting dataset for the factor analysis technique was evaluated by diverse measures showing that the new reduced subset of elements is a mediocre but still an acceptable subset that represents the characteristics of our original user profile.

We provide a data mining system for a real scenario where we are giving valuable tools for the instructors an students in order to generate feedback for the users that tend to fail in the IPSims activities identifying the group of students that have a tendency to do bad in digital media learning environments activities. Based on database normalization rules we generate a new data model where the acquired data was stored. As a result the extraction and organization of the information for our system is more efficient.

The presentation of a validated user interface for user error prevention was generated. With this we reduced the inconsistencies in the acquired data from the students we also applied Testing techniques and tools in order to tune up the system with the intention to give a satisfactory usability experience to the users the entire process involved the generation of *Python* classes that can be re-used for further experiments or studies related with IPSims.

We have described, explained and probed how the data mining process was applied on this study. This process involved multiple challenges to the author. This document is the materialization of a complete work on data mining in a digital media learning environment setting a starting point for further research studies in the IPSims system.

5.2 Discussions

During our research process there were a lot of aspects that influenced our results, aspects such as: The collection of the data from IPSims and the paper questionnaire, inconsistencies in the data model, noise in the data collected, the distribution of our variables, an approach to select adequate training sets, dimensionality issues and neural networks overfitting aspects and some other that arose while this work was being developed. In order to address these issues we required diverse statistical and computational techniques, which are explained in the following subsections.

5.2.1 Overfitting

Figures 5.1, 5.2 and 5.3 depicted the results of our training stage for each training algorithm using the dataset generated by the standard deviation analysis.

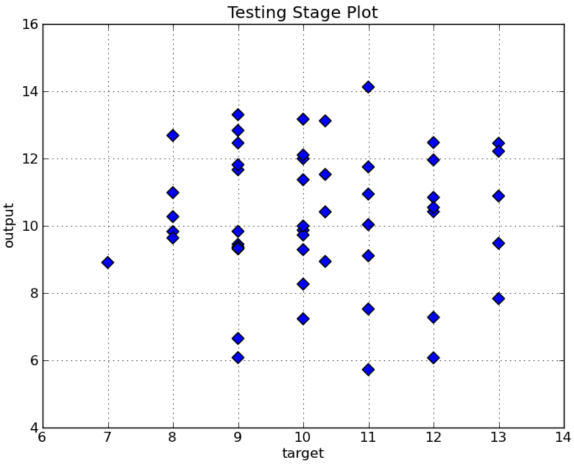


Figure 5.1: Training plot for the genetic training algorithm using standard deviation analysis dataset

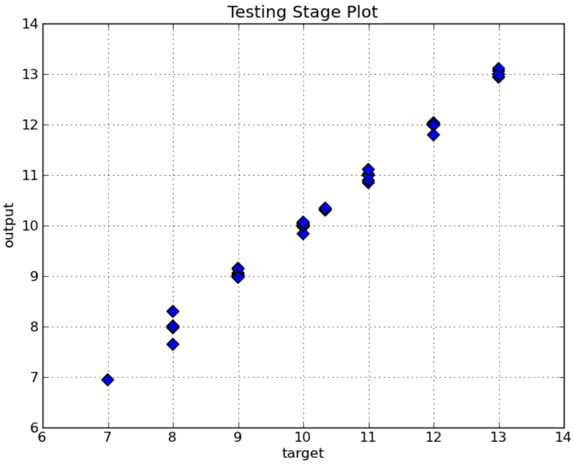


Figure 5.2: Training plot for the conjugate gradient training algorithm using standard deviation analysis dataset

The figures displayed a successful training stage for the conjugate gradient and distributed algorithm with values $R - squared \approx 0.99999$ for the conjugate

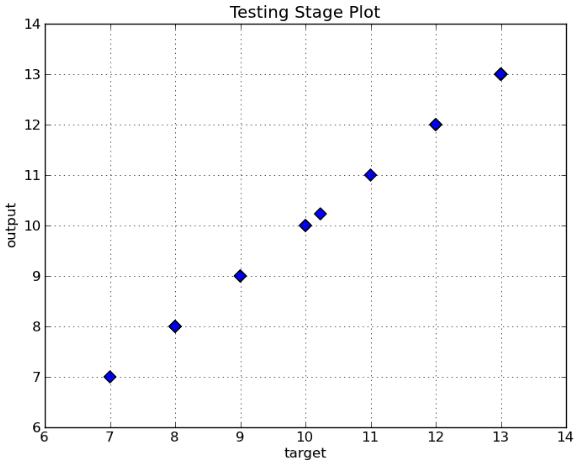


FIGURE 5.3: Training plot for the distributed training algorithm using standard deviation analysis dataset

gradient algorithm and $R - squared \approx 0.99564$ for the distributed algorithm in the other side the genetic algorithm displayed a poor performance in the learning process.

The results for the learning process especially for the conjugate gradient algorithm suggest that we could obtained desire predictions with a high accuracy. Unfortunately for this specific dataset generated with our standard deviation analysis the models present overfitting in all our experiments. This phenomenon might be due to diverse reasons that we can analyze in further studies some possible explanations could be as follows:

- Standard deviation analysis did not take into account relationships between variables.
- The reduction might have be confined to a local minimum not representing the actual minimum of our solution space.
- The neural network was over trained.
- Low number of study subjects.

Our intention when used standard deviation analysis was to select the variables with the highest standard deviation, this approach might guarantee a wider exploratory space over our solution space; which in our assumptions would gave us a good subset of variables to train our neural network. In the practice this assumption resulted false.

Different training sets were tried, and diverse iterations were addressed and the results were always the same for this dataset. Overfitting was always present. Our overfitting assumption applies just for the conjugate gradient algorithm and the distributed algorithm. Our genetic training algorithm does not even pass our test for the training stage based on *R-squared* and *mean square error* values for a reliable learning process.

Figures 5.4, 5.5 and 5.6 present the testing stage for the models where the overfitting is evident for the plots of conjugate gradient and distributed algorithm. Genetic algorithm plot depicts extremely poor results with $R - squared \leq 0.5$.

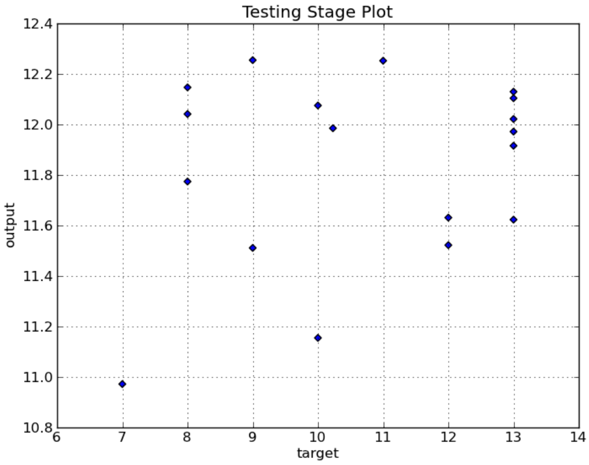


Figure 5.4: Testing plot for the genetic training algorithm using standard deviation analysis dataset

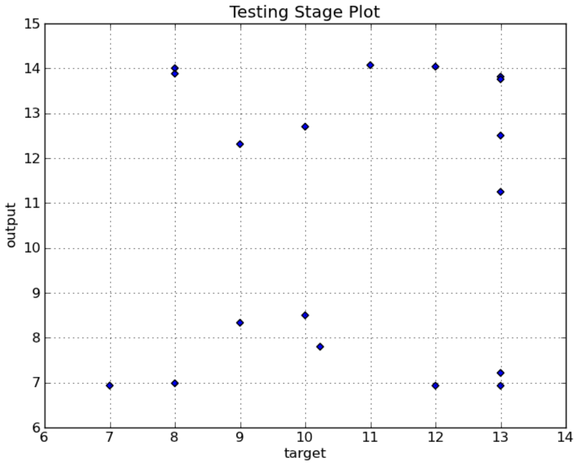


Figure 5.5: Testing plot for the conjugate gradient training algorithm using standard deviation analysis dataset

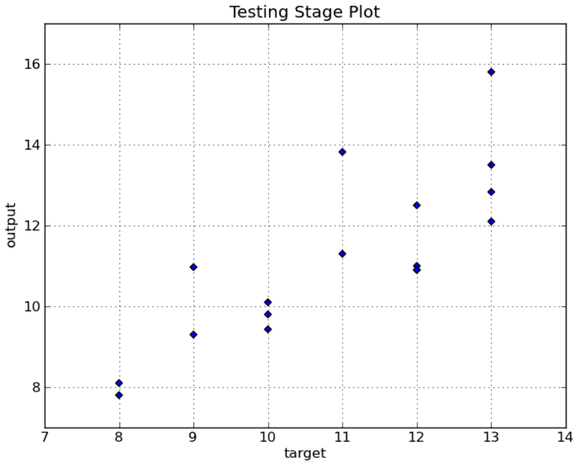


FIGURE 5.6: Testing plot for the distributed training algorithm using standard deviation analysis dataset

When we talk about overfitting we refer to the phenomena where the neural network presents excellent results for the training stage with minimal error measures and highest accuracy values, but the testing stage fails completely; obtaining predictions with high error measure values.

The factor analysis dataset presents slightly improved results than the standard deviation analysis; with the exception of the genetic training algorithm, this specific training algorithm did not performed appropriately in the training stage Figure 5.7 depicts one of the training iterations for the genetic training algorithm. We can observe an unsuccessful training stage.

Figure 5.8 and Figure 5.9 depicts the training and testing stage for the distributed algorithm using the variables generated by the factor analysis, this subset of variables display clearly better results.

Figure 5.10 and Figure 5.11 displays the same stages for the conjugate gradient algorithm where the predictions obtained has considerably lower quality compared against our distributed training algorithm results.

Based on this plots we can observe slightly better results in the predictions made using the factor analysis data set. The fact that factor analysis take in account the relationship between variables might be the fact that makes factor analysis a better dataset for our research purposes but in counter case the fact that the generate factors are new generated variables affect the exploration of our original solution space.

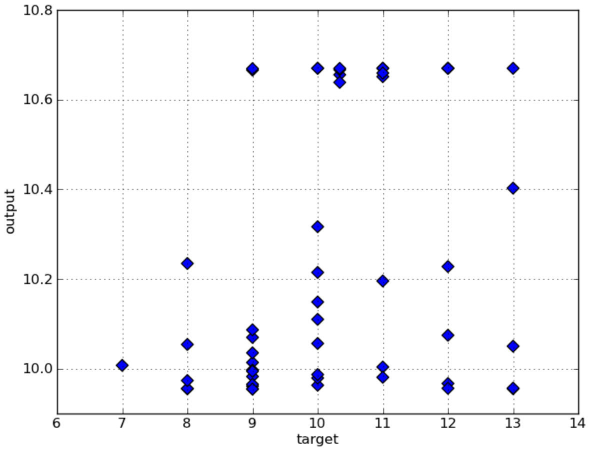


Figure 5.7: Training plot for the genetic training algorithm using the components from factor analysis

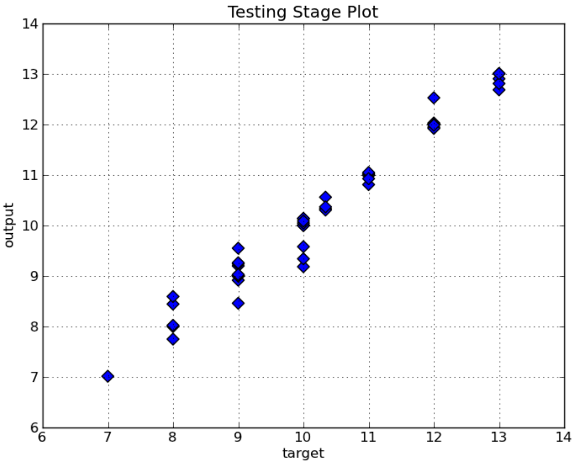


Figure 5.8: Training plot for the distributed training algorithm using the components from factor analysis

Fortunately even when the distributed training algorithm and the conjugate gradient algorithm present overfitting when tested with our standard deviation analysis dataset and factor analysis dataset. In our experiments where the correlation, significance and variance dataset was used the results were more satisfactory when the distributed training algorithm was used.

In most of the cases when the conjugate gradient algorithm was used overfitting was present. But after a number of iterations over our experiments and training algorithms we generated some models that accomplished with acceptable results specifically using the distributed training algorithm.

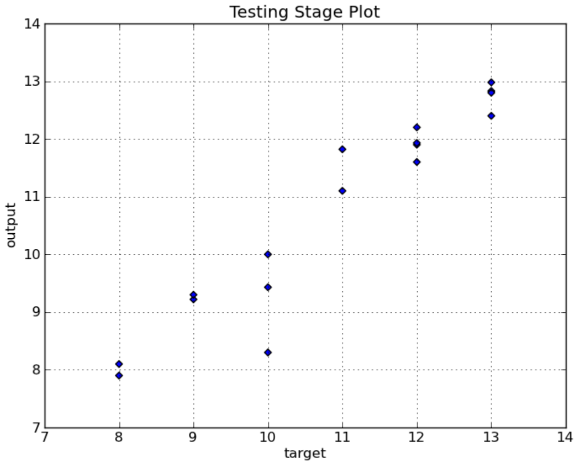


Figure 5.9: Testing plot for the distributed training algorithm using the components from factor analysis

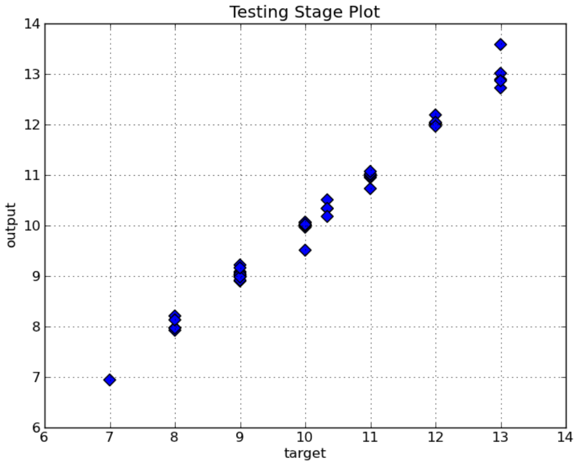


Figure 5.10: Training plot for the conjugate gradient training algorithm using the components from factor analysis

Figures 5.12, 5.13 and 5.14 presents the plots for one of our training results using the factor analysis reduction dataset.

Even when we can observe overfitting again in our conjugate gradient algorithm. We can see for this iteration that our distributed training algorithm displays satisfactory results in the predictions.

We assumed that the generated dataset based on the correlation, significance and variance analysis accomplish with better results given the fact that we take in account variability and relationship between variables. Respecting the original

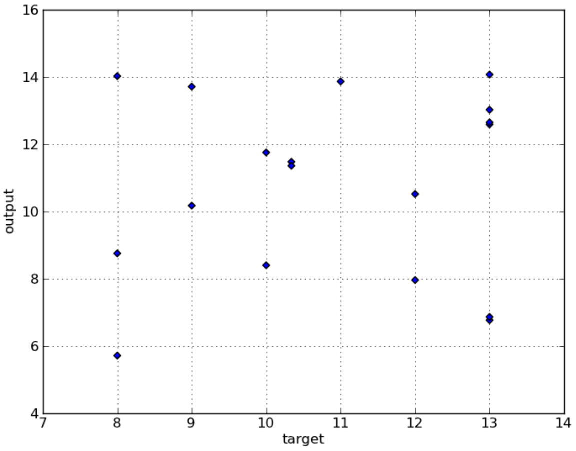


Figure 5.11: Testing plot for the conjugate gradient training algorithm using the components from factor analysis

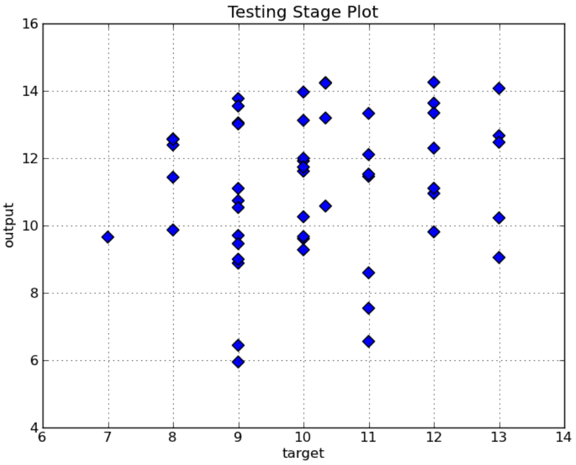


Figure 5.12: Training plot for the genetic training algorithm using correlation, significance and variance analysis dataset

values of our variables and discarding the ones that does not show enough variability and does not reflect a strong correlation with other present variables. We could say that this analysis is an intermediate point between our two previous techniques for dimensionality reduction.

Figures 5.15, 5.16, 5.17 depicts the results of the testings in this experiment run.

Given the low number of study subjects overfitting is a common factor present in our predictions. This thesis work provides the basis for knowledge discovery within IPSims and a starting point for the analysis of the sources that might

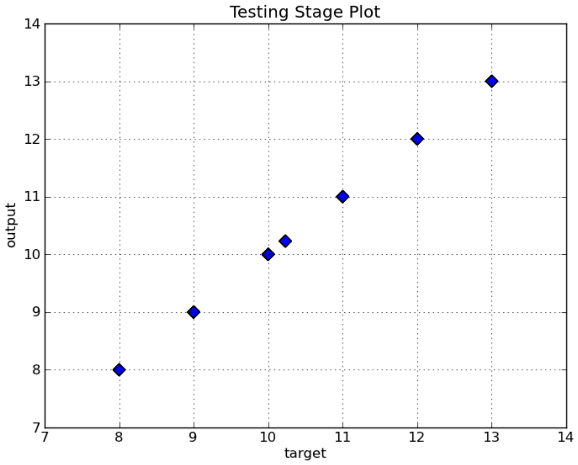


FIGURE 5.13: Training plot for the conjugate gradient training algorithm using correlation, significance and variance analysis dataset

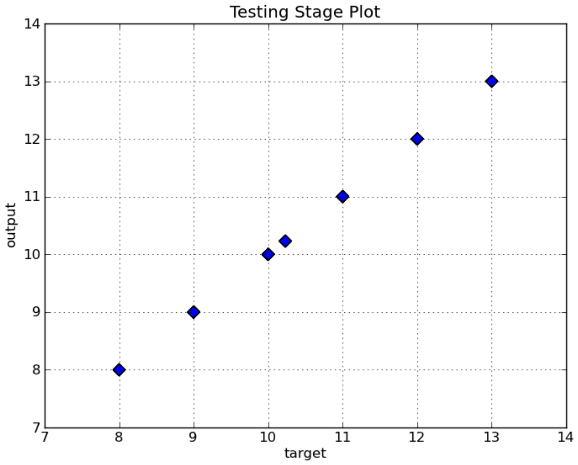


FIGURE 5.14: Training plot for the distributed training algorithm using correlation, significance and variance analysis dataset

generate diverse issues that could obstruct the knowledge discovery process in IPSims. We also identified some of these problems and knowing the nature of the issues we could propose future solutions to them. One of these problems is the overfitting we thought that an increment in the number of study subjects for IPSims is required in order to minimize the overfitting when prediction models are used as well as more complemented testing and training set to validate our results.

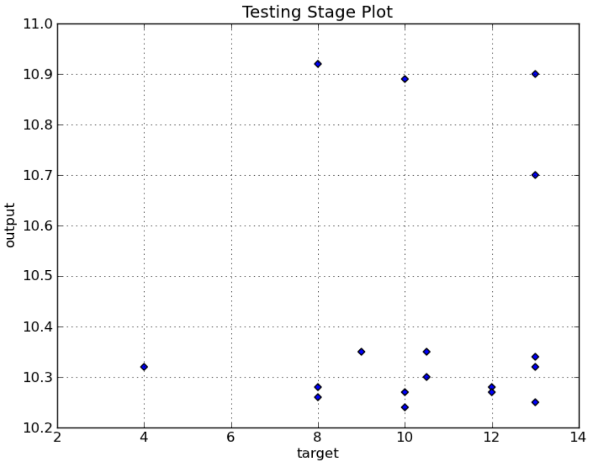


FIGURE 5.15: Testing plot for the genetic training algorithm using correlation, significance and variance analysis dataset

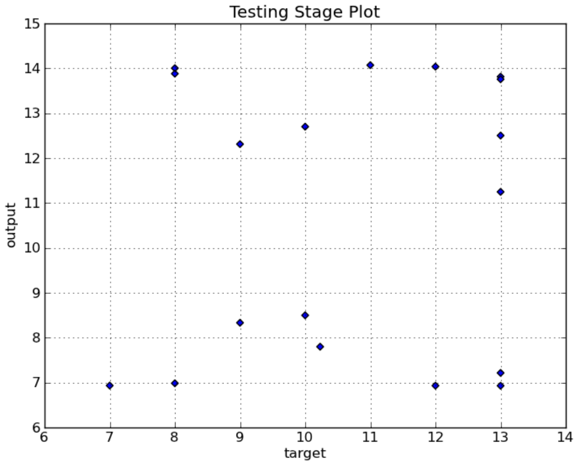


FIGURE 5.16: Testing plot for the conjugate gradient training algorithm using correlation, significance and variance analysis dataset

5.2.2 Clustering

We were able to observe diverse patterns and we also identified a group of variables that may have the biggest impact in our final grade for the IPSims activities. The cluster analysis was realized with the use of our both datasets. The one original user profile and the dataset obtained from standard deviation analysis.

The used of the original user profile retrieved more descriptive elements given the higher number of variables. We were able to observe interesting patterns with

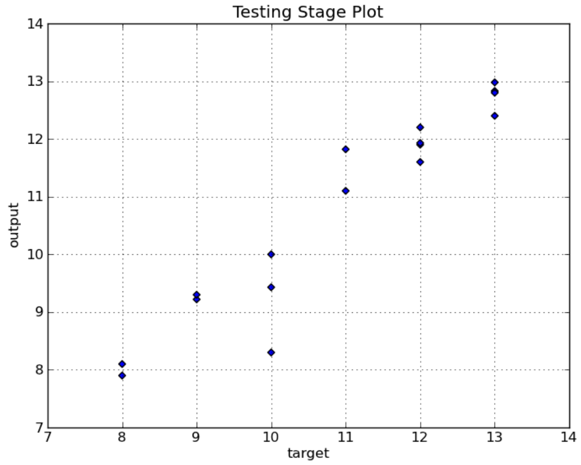


FIGURE 5.17: Testing plot for the distributed training algorithm using correlation, significance and variance analysis dataset

the use of both datasets and we defined some rules that indicate which ones are the variables that may have the biggest impact in the grade for the activities.

The interesting patterns observed with the use of our entire user profiles are:

- The cluster of students that showed the highest value for the variable that indicates if the students course college or university education before University of Ontario Institute of Technology is the cluster that displayed the highest grade in the activities. Figure 5.18 displays the result.
- The cluster of students that displayed the highest average value for computer literacy surprisingly showed the lowest average grade in the other side the cluster that displayed the lowest average value for the computer literacy obtained the highest grade. Figure 5.19 displays the result.
- The cluster of students that displayed the lowest expectancy for the course in the grade obtained the lowest average grade for the activities. Figure 5.20 displays the result.
- The cluster displaying the highest value in their experience with computer based simulations surprisingly displayed the lowest grade in the activities while the cluster with the lowest experience in computer based simulations reflected the highest grade. Figure 5.21 displays the result.
- The cluster displaying the lowest high school average grade is the cluster with lowest average grade in the activities while the cluster with the highest high

school average reflected the highest grade in the activities, this plot showed almost a linear pattern. Figure 5.22 displays the result.

- The cluster with the highest interest in the course material is the cluster that displayed the lowest grade and the cluster with the lowest interest in the course material obtained the highest grade in the activities. Figure 5.23 displays the result.

- The cluster that displayed the highest likelihood in choosing a career in health sciences is the cluster that displayed the lowest results in the activities. Figure 5.24 displays the result.

- The cluster that showed the highest average for the perceived value of computer based simulations displayed the lowest grade in the activities. Figure 5.25 displays the result.

- The cluster that showed the lower rate for play hour peer week is the one that displayed the lowest average grade in the activities. Figure 5.26 displays the result.

- The cluster that showed the highest intention to pursue graduate education after University of Ontario Institute of Technology is the one that displayed the highest grade in the activities. Figure 5.27 displays the result.

We observed certain tendencies in the presented plots that lead us to the generation of a subset that may have the biggest impact in the students performance in IPSims activities. Based in our observations we can say that there are 10 variables that reflected the biggest impact in the activities. These variables are: college or university before University of Ontario Institute of Technology, computer literacy, expected grade in the course, experience with computer based simulations, high school average, interest in the course material, likelihood in choosing a career in health sciences, perceived value for computer based simulation, play hour peer week and intention to pursue a graduate education after University of Ontario Institute of Technology. These are the variables selected after our cluster analysis using the entire user profile. These variables may have the biggest impact in the student performance in the IPSims activities.

In order to determine the real impact of these 10 variables in the student's performance, a further analysis is required where we have to generate user profiles based on these observations. It seems that these 10 variables form a set of implicit user profiles which we verified manually by visually observing wether these variables matched with the study subjects to verify that effectively these variables are having the impact displayed in the plots for our cluster analysis.Unfortunately not all of these implicit user profiles matched our manual observations. Nevertheless those implicit user profiles may provide instructors and students with valuable insight as to the predicted performance in the course. For example.

An implicit user profile with good chances of succeeding in the IPSims activities is composed of the highest possible value for the variable that indicates enrolment in college or university before University of Ontario Institute of Technology, lowest value for the variable containing the computer literacy, lowest value in the experience with computer based simulations, highest values in the high school average variable, lowest interest in the course material, highest intention to pursue graduate education after University of Ontario Institute of Technology, low value in perceived value for computer based simulations, high expectancy in the course grade.

Let's not forget that we still have an original database with more variables describing our study subjects. Variables that were discarded might be added.

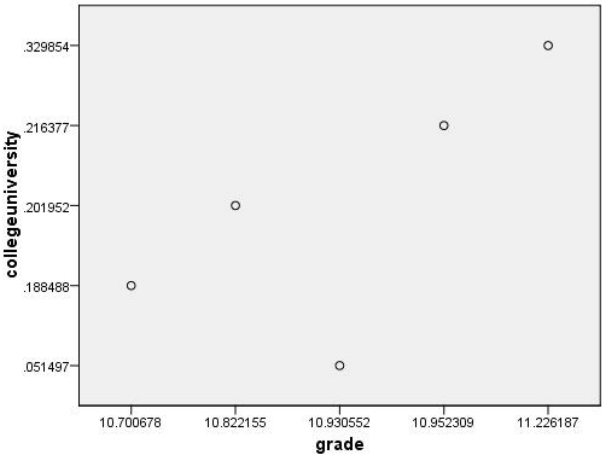


FIGURE 5.18: College or university education before UOIT vs. grade by Cluster

The correlations observed in our dataset generated by standard deviation analysis were the next ones:

- The cluster of students expecting a low grade in in the course are the students that obtained the highest average grade in the activities. Figure 5.28 displays the result.

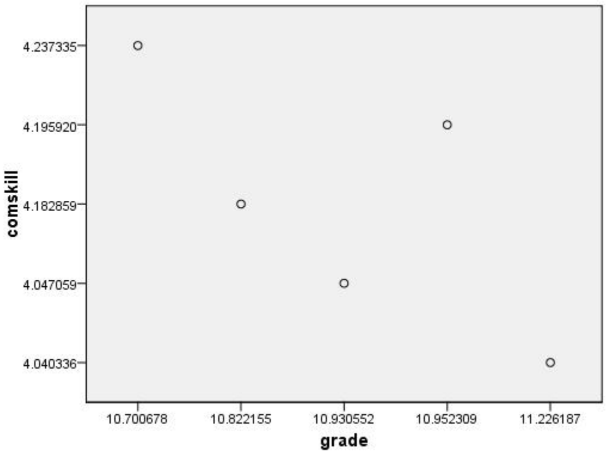


FIGURE 5.19: Computer literacy vs. grade by Cluster

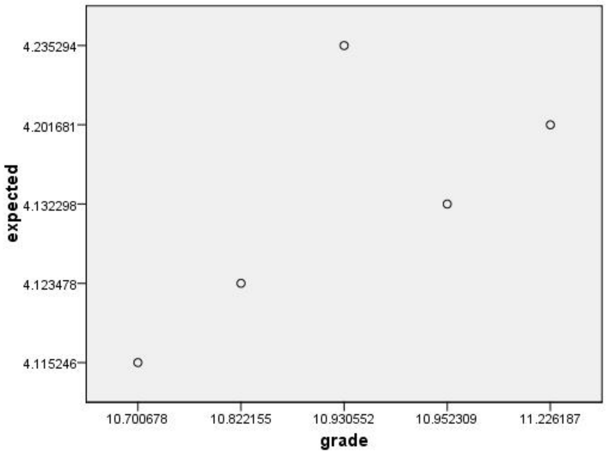


FIGURE 5.20: Expected grade in the course vs. grade by Cluster

- The cluster of students that showed the most interested in the course material are the ones that obtained the lowest average grade in the activities. Figure 5.29 displays the result.

- The group of students with the lowest average play hour week were the group of students with the lowest average grade while the group that reflected the highest average time of playing hour week is the group with the highest average grade. Figure 5.30 displays the results.

- The group with the lowest time of surfing the web hour week is the group with the highest average grade. Figure 5.31 displays the result.

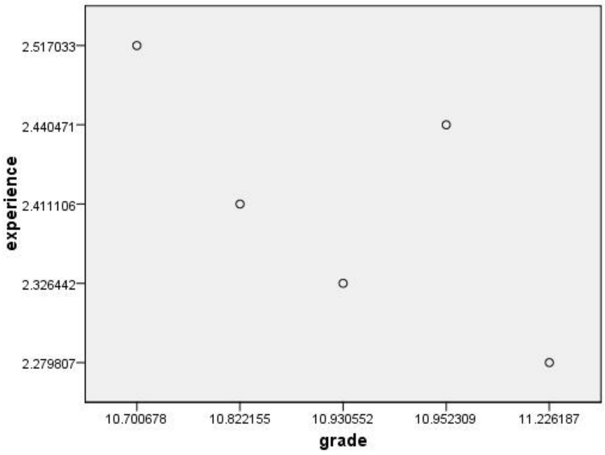


FIGURE 5.21: Experience vs. grade by Cluster

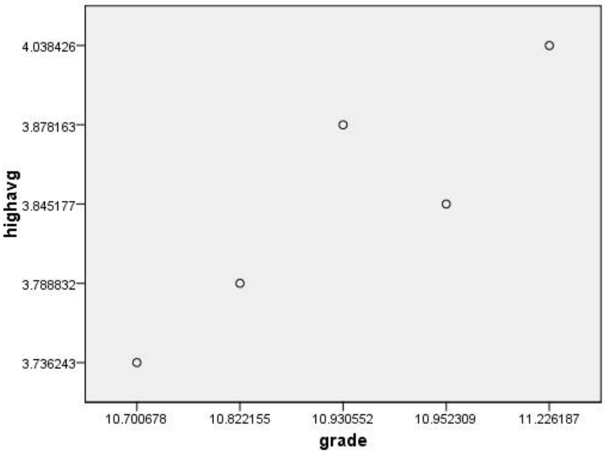


FIGURE 5.22: High school average vs. grade by Cluster

- The students that graduated more recently from high school are the ones that reflected the lowest average grade. Figure 5.32 displays the result.

The presented plots were useful for the analysis and identification of the explained patterns.

We observed almost a linear relationship with the exception of some clusters that did not follow the tendencies. In some of the cases an increase or decrease in the slope of our line in an almost linear pattern were observed.

Based on the plots for this reduced dataset our observations indicate that in this reduced dataset the variables that may have the biggest impact in the students

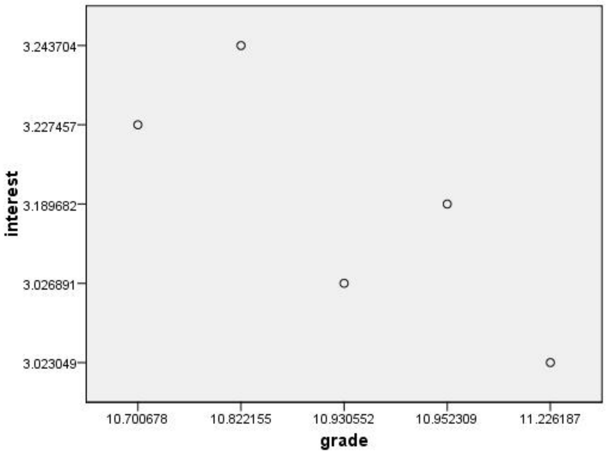


FIGURE 5.23: Interest in course material vs. grade by Cluster

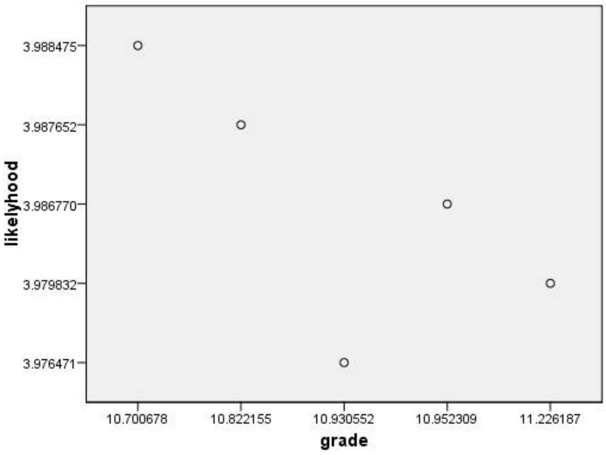


FIGURE 5.24: Likelihood for choosing a career in health sciences vs. grade by Cluster

performance are: expected grade in the course, interest in the course material, play hour peer week and surf hour peer week. In this reduce dataset we observe that the variable play hour peer week are part of the possible biggest impact variables for the student's performance this variable did not appear in our analysis for the entire user profile dataset.

This experiment was redundant but our intention was to observe the behaviour of the cluster algorithm with a reduced dataset. It is interesting to observe that in the reduced dataset the variable play hour peer week emerged as a high impact variable for the student performance while in our entire user profile the variable play hour did not present any interesting pattern.

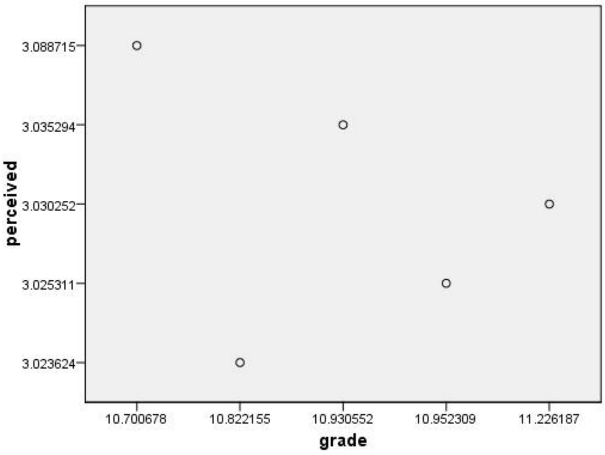


Figure 5.25: Perceived value of computer based simulations vs. grade by Cluster

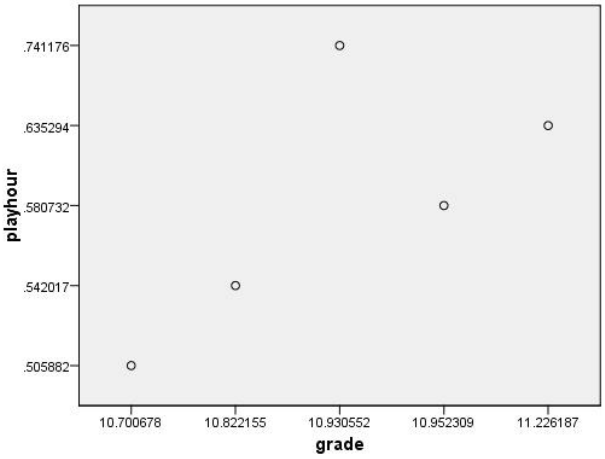


Figure 5.26: Play hour peer week vs. grade by Cluster

The cluster analysis displayed some counter intuitive results that may challenge our logic, but as complex as the human mind is as diverse factors can be affecting the results in our observations. For example seems contradictory that the students with the highest interest in the course material are obtaining the lowest grades in the activity. We have to sharpened our analysis and think that these group of students might does not like computer based simulations or as they are interested in the course material they are not interested in these kind of activities that are not commonly used in conventional health science courses. Another example is the variable computer literacy where the students with highest values in this variable presented the lowest grades in the activities. Once again we have

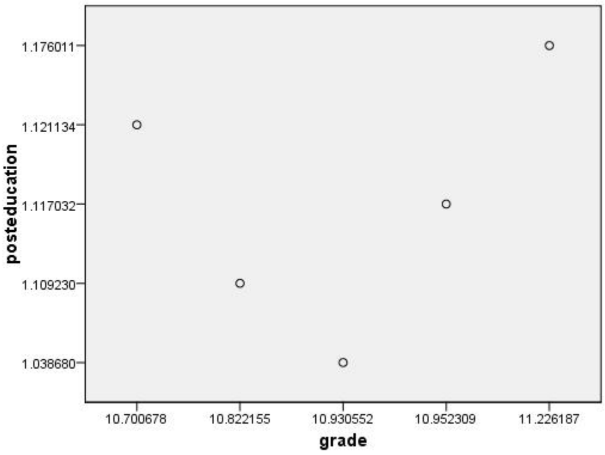


FIGURE 5.27: Intention to pursue graduate education after UOIT vs. grade by Cluster

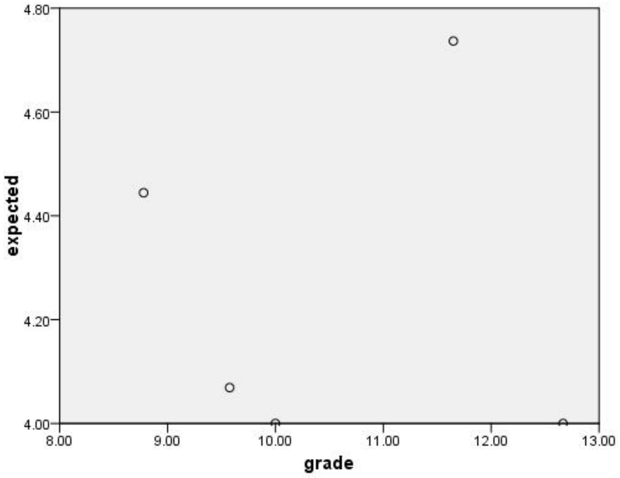


FIGURE 5.28: Expected grade for the course vs. grade by Cluster

to narrow further our analysis and think that these group of students might feel these kind of activities boring or not worth it of a big effort of their part given their experience with the use of computer. Then we have to add more and more factors that may affect the observed results.

Further analysis with a higher number of students can provide a more solid justification for our assumptions of a reduced dataset that has the biggest impact in the students performance in the activities with IPSims. At the moment we would say that the set of variables that may have the biggest impact in the student performance are the ones presented using the entire user profile, given the fact

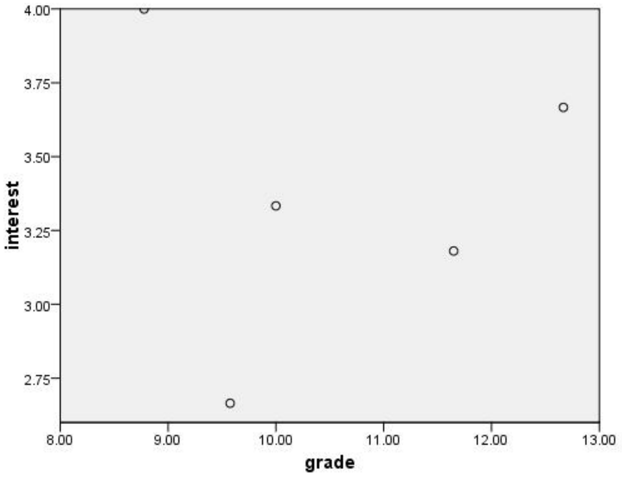


FIGURE 5.29: Interest in course material vs. grade by Cluster

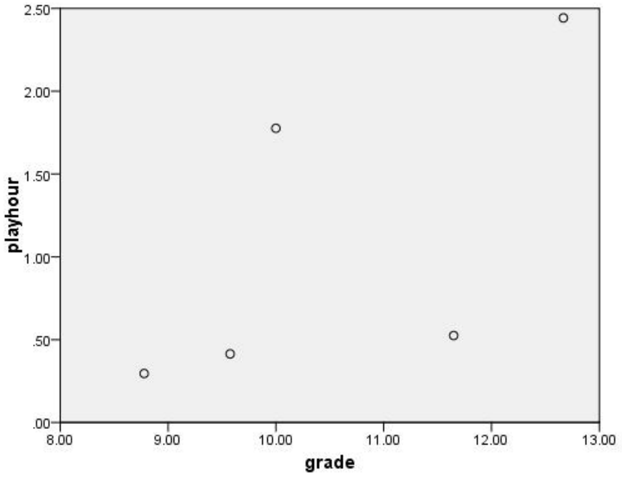


FIGURE 5.30: Play hour peer week vs. grade by Cluster

that this dataset provides more information about our users against the dataset generated for the standard deviation analysis.

Figure 5.33 is a 3-D plot showing the organization of our elements by cluster, age and surf hour. We observed in some of the plots small conglomerations that indicate possible correlations between the variables values and the grade.

In Figure 5.34 we can observe the 3-D plot for the high school average, age the cluster where the elements were classified. The grade is represented by the colour of the elements.

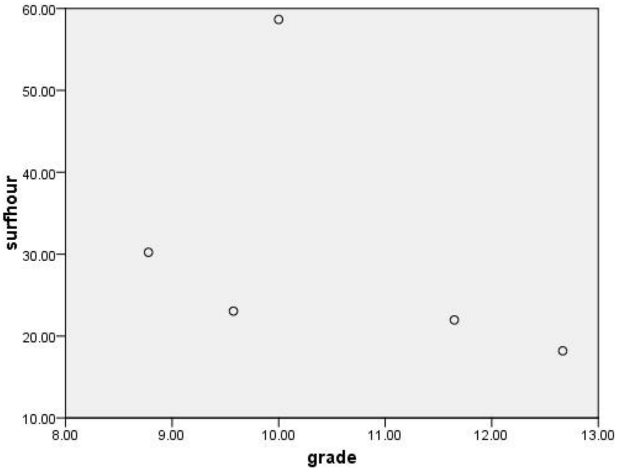


FIGURE 5.31: Surf hour peer week vs. grade by Cluster

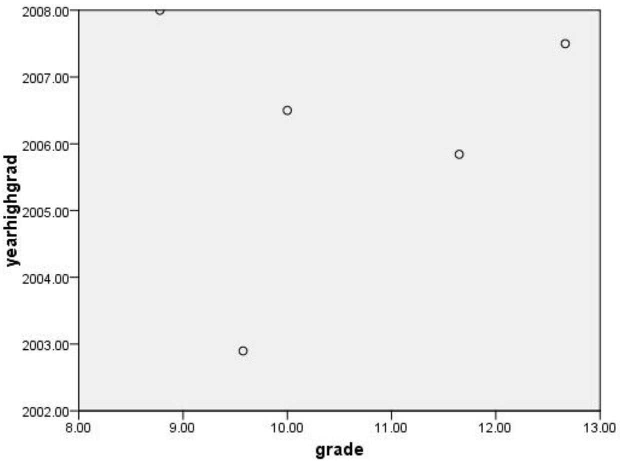


FIGURE 5.32: College or university education before UOIT vs. grade by Cluster

We generated more plots for diverse groups of variables in order to analyze the nature of our generated groups and possible generated rules or defined patterns. Figure 5.35, figure 5.36 and figure 5.37 are some other examples of our generated plots.

In order to amplify the analysis of our results we generated two dimensional plots. Figure 5.38 depicts the distribution for the variable high school average by cluster.

These types of plots supported our analysis in the recognition in how the variables were grouped in a given cluster by the analyzed variable value. Some of

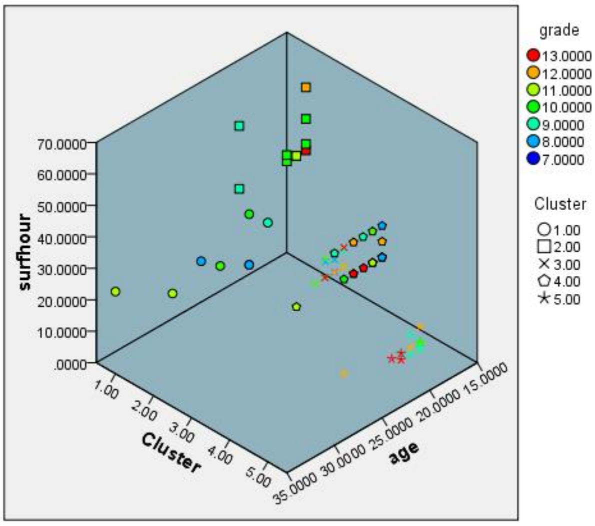


Figure 5.33: 3-D plot for the age, cluster and surfhour

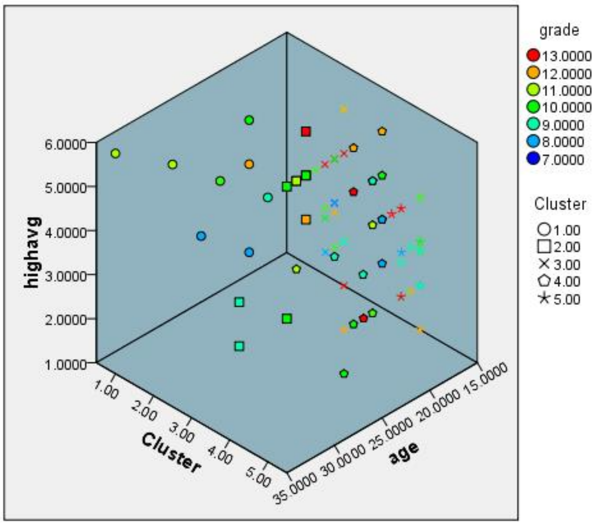


Figure 5.34: 3-D plot for the age, cluster and high school average

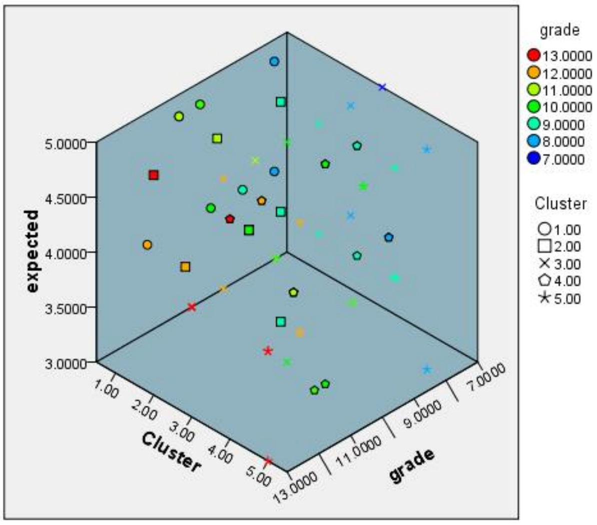


Figure 5.35: 3-D plot for cluster, expected grade in the course and grade

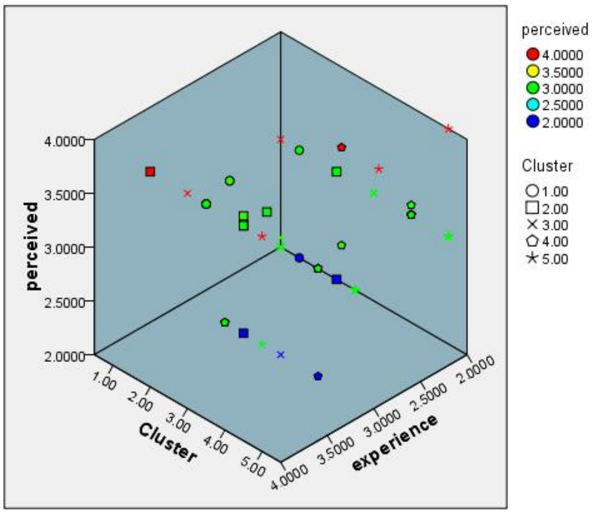


Figure 5.36: 3-D plot for cluster, perceived value for computer based simulations and experience with computer based simulations

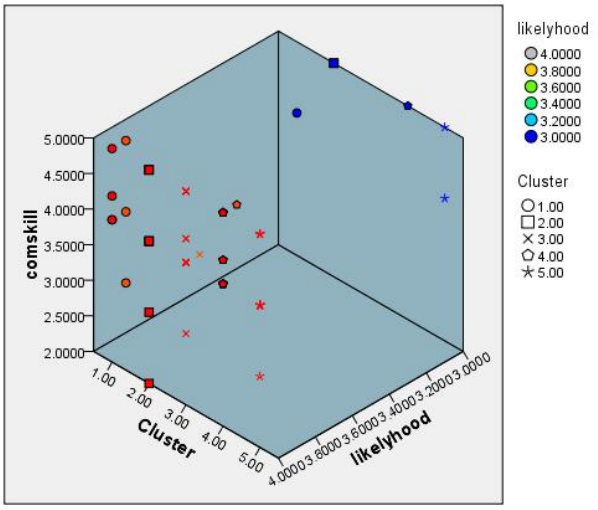


Figure 5.37: 3-D plot for cluster, computer literacy and likelihood in choosing a career in health sciences

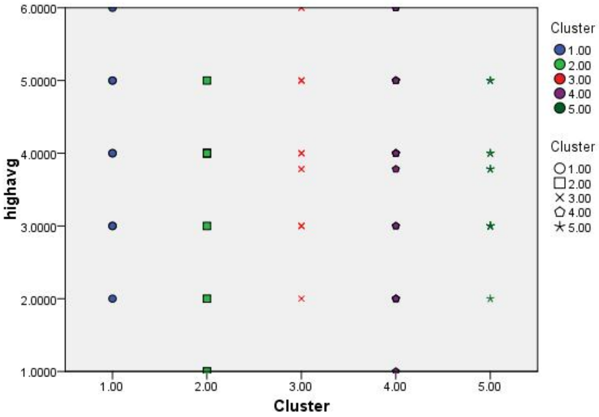


Figure 5.38: Distribution by cluster and high school average

the plots presented interesting conglomeration of values by variable, some other did not reflect any correlation or interesting pattern. For example Figure 5.39 depicts an interesting conglomerate of values by cluster related to the surf hour peer week of the users.

Figure 5.40 and Figure 5.41 did not reflect any detectable or easy to understand pattern the distributions looked homogeneous and there were no clear conglomeration of elements.

These plots were really useful to reinforce our previous observations based on the 2-D plots. Unfortunately sometimes the conglomerates appeared as a single

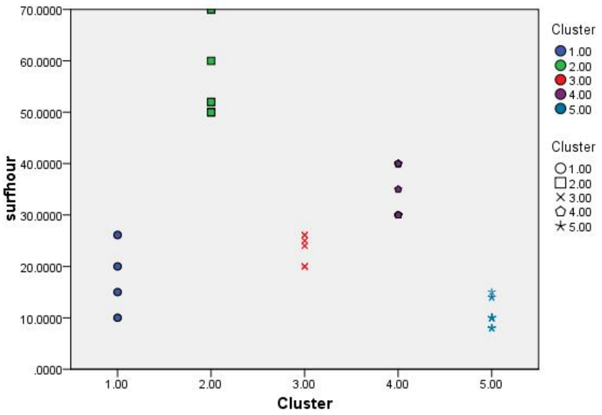


FIGURE 5.39: Surf hour peer week distribution by cluster

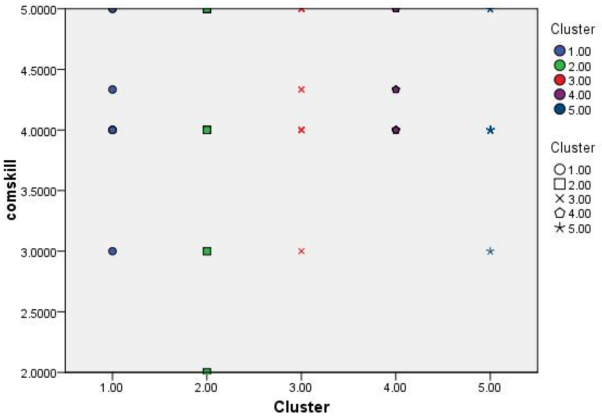


FIGURE 5.40: Computer literacy distribution by cluster

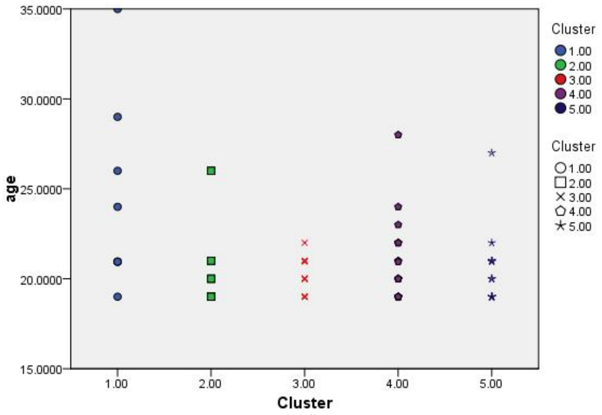


FIGURE 5.41: Age distribution by cluster

point in the distribution for this reason we were unable to detect how many elements were contained in a single point. For example in Figure 5.39 we knew that in cluster four with value 40 for the surf hour there were a conglomerate of elements. But we were unable to determine how many elements were contained in that conglomerate. These plots were more a supportive technique to visualize the behaviour of our clustering algorithm.

5.3 Contributions

We developed a data mining system for a digital media learning environment called IPSims. This developed data mining system supported tasks like the extraction of diverse information in order to generate research for University of Ontario Institute of Technology based on a system that was being used in diverse health science courses this with the intention to understand diverse aspects of the participant students such as: misconceptions, profile of students that are likely to fail, prediction of a failing profile based on diverse aspects like navigation preferences, time spent on internet, activity within the IPSims. We applied database management tasks for the IPSims. Diverse issues were generated for the IPSims database. In some cases the database was not recording values or the information was missing, with the use of diverse tools for Postgres database administration, the database was tuned, cleaned and backed up in order to preserve the integrity and security of the students recorded information. Then we can say that a data warehouse that accomplished with database normalization was provided in order to store the information originated from IPSims and the paper questionnaire.

We provided the IPSims with user interface maintenance, some aspects on the user interface were incomplete or out of service. It was a required task for this thesis to fix and manage these issues with the intention to have a complete running system to start the required experiments. The generation of an accurate prediction model constructed on Python measured using cross validation and statistic measures. This model predicts accurately the final grade that will be obtained for the user based on some of the user profile variables. The results were good for the few test cases it is required to recruit more students within the next academic terms to continue testing the acquired predictive model. We generated based in clustering techniques clusters for identification of groups of students. Within this generated clusters information visualization techniques were applied in order to identify in an easy way the auto generated groups of students based on their user profiles.

A descriptive subset of variables based on a factor analysis reduction and standard deviation analysis. The result of this contribution is a reduce set of variables that keep in an mediocre but still acceptable form the characteristics of the complete set of variables for the user profile. The reduced datasets were input for the prediction model proposed in this work. A number of variables from the user profile were proposed as the ones that may have the biggest impact in the student performance in the activities with IPSims. The proposed variables were selected based on the observations for the cluster analysis experiment. Based on the previous accomplishments and contributions we published four scientific papers in international conferences.

It was our intention to generate a complete data mining framework that followed an established process for knowledge discovery. Since the data selection passing through data pre-processing, data transformation, data mining algorithms and finally evaluation and interpretation with the intention to provide knowledge from an educational environment called IPSims.

We accomplished with this original goal, but in order to prove the effectiveness of our main contribution we decided to apply our data mining framework to a two specific cases where we discovered the variables that might have the biggest impact in the user performance in the system and a predictive model that based in the user profile predicts the final grade within the IPSims activities.

The entire process of proposing a data mining framework lead to a diverse contributions, like reduced datasets of variables derived from formal processes based in statistical techniques, a completely new database architecture that consolidate our data warehousing for the extraction of the desire information in the system. The treatment of the variables related with user navigation preferences that even if they were not use in the analysis they are valuable and potential source of knowledge for further studies.

We provided with valuable tools and knowledge that can be exploit in diverse forms for the final users, the generation of association rules, the generation of specific user profiles that reflect the probability to become successful or unsuccessful in the activities, a prediction model for the final grade, and so much more possibilities according to the needs and requirements of the instructors, the faculty and the students.

5.4 Closing Remarks

Even when all the experiments and results in our study are justified and validated using our actual dataset. Our results are still inconclusive. We cannot extent our assumptions and results to an extended dataset until we obtain more study subjects to test all our experiments. In section II diverse of the reviewed papers mentioned that one of the principal issues when working with educational data mining is the small amount of available data. In our case this was the principal obstacle in our research.

Through the use of cluster analysis techniques we generated a subset of elements from the user profile that may have the biggest impact in the student performance in IPSims activities. These results are validated for the actual dataset but our results are not conclusive given the low number of study subjects. In order to reinforce our assumptions we require the enrollment of more individuals in our experiments.

The provided prediction model that works more accurately was the one using the dataset generated by the use of our correlation, significance and variance analysis. We are optimistic with this result but at the same time we are skeptics of it we have in mind that our testing cases were less than 20. In the specific test of this case and nature of our greedy algorithm for training and testing inputs we can not guarantee that this predictive model can predict unexpected cases; or even in our experiments showing acceptable results there is the possibility that the testing cases were really similar to the training cases. The only way to prove the validity of this prediction model out of this reduced dataset is recruiting more students in order to increase our training and testing cases.

The increase in the number of study subjects is required in order to be able to include all our user profile variables in our experiments. Once we can include all the user profile variables we can provide a wider set of knowledge and discoveries within the educational data in IPSims.

Since the beginning we were aware of the problems that appeared given the small number study subjects. But as mentioned before the original intention of this work is to provide a neat process in how data mining could be accomplished in our digital learning media environment.

We are happy with the generated framework and we expect this document works in the future as a starting point for further researches that involved IPSims and its information.

5.5 Future Work

Future analysis will be performed with new recruited students in order to validate our results with a bigger number of study cases.

The data related to decision sequela and times was treated but not used for the low number of study cases. Another factor that prevented the use of these information were the inconsistencies in the user activities in the system. Future research will focus in the use of these information to generate studies related with student navigation preferences, usability analysis and navigation predictions.

With the increase of the number of the study subjects the original dataset for the IPSims can get involved in our prediction model experiments, a bigger set of variables may reflect in a more accurate prediction model.

The application of different prediction techniques as well as cluster techniques in order to compare them and select the ones that fit best with this study is another opportunity for research in the IPSims.

Appendix A

University of Ontario Institute of Technology, Research Ethics Board file #: REB09-27

Principal Investigators: Drs. Miguel Vargas Martin and Jayshiro Tashiro

Student Investigators: Meaghen Regts and Arturo Fernandez Espinosa

Instrument

Objective 1

- Rating of Web-based Simulation Designs
- Preferences for Learning Resources, and Educational Scaffolding
- Rating of Simulation Templates Usability
- Satisfaction with Simulation Template Designs

Objective 2

- Disposition to Engage in Effortful Cognitive Endeavor (Need for Cognition and Ambiguity Tolerance)
- Expectancy-Value Questionnaire
- Performance Evaluation in Interprofessional Learning Activities
- Satisfaction with Realism of Simulations, Delivery Modality, and Content A2-26

All of the research instruments that are herein proposed are public domain research tools. They are not copyrighted nor do they require licensing or have a royalty structure. None require permission of the author for usage in research.

State 1: Rating of Web-based Simulation Designs

Background Information

For the (APPROPRIATE DATES) Semester, how many (courses OR professional development programs) are you taking?

Circle Number of Courses

1 2 3 4 5 6

Of the courses OR professional development programs you are taking this semester, how many require that you work within a Web environment each week?

Circle Number of Courses

1 2 3 4 5 6

Rate Your Computer Skills.

Poor 1 2 3 4 5 6 Excellent

How many hours a day do you spend on the computer?

1 2 3 4 5 6 > 6

Of these hours spent on the computer, how many are for course OR professional development work?

1 2 3 4 5 6 > 6

Stage 1: Preferences for Learning Resources, and Educational Scaffolding

Please rate how much you like having both the Course Tools and My Tools menus as well as icons for representing the tools in your WebCT Course Home Page.

Dislike Very Much 1 2 3 4 5 6 Like Very Much

Please rate how much you would like your course assignments provided within the WebCT Course so you could complete all assigned exercises online.

Dislike Very Much 1 2 3 4 5 6 Like Very Much

Please rate how much you would like your assigned readings all provided within the WebCT Course so you didnt need a textbook or readings provided in hardcopy.

Dislike Very Much 1 2 3 4 5 6 Like Very Much

Please rate how much you would like all of your courses to be totally online without face-to-face instruction by a faculty member, but with a faculty member available online.

Dislike Very Much 1 2 3 4 5 6 Like Very Much

Please rate how much you would like some of your courses to be totally online without face-to-face instruction by a faculty member, but with a faculty member available online.

Dislike Very Much 1 2 3 4 5 6 Like Very Much

Please choose the percentage of face-to-face instruction by a faculty member you would like for a course.

- 100% = Faculty present in face-to-face instruction each week of the semester.
- 75% = Faculty present in face-to-face instruction about 9 weeks of the semester, with online work for remaining weeks.
- 50% = Faculty present in face-to-face instruction about 6 weeks of the semester, with online work for remaining weeks.
- 25% = Faculty present in face-to-face instruction about 3 weeks of the semester, with online work for remaining weeks.
- 0% = Course is totally online, with faculty providing support online.

Select Preference

- 100%
- 75%
- 50%

- 35%
- 0%
- Depends on Course

If face-to-face instruction depends on course, list courses where more face-to-face instruction would be desirable: ______________________________

Stage 1: Rating of Simulation Templates Usability

Please rate the navigation and general usability of Simulation Template 1.

Not at all User-Friendly 1 2 3 4 5 6 Very Friendly

Please rate the navigation and general usability of Simulation Template 2.

Not at all User-Friendly 1 2 3 4 5 6 Very Friendly

Please rate the navigation and general usability of Simulation Template 3.

Not at all User-Friendly 1 2 3 4 5 6 Very Friendly

Please rate the navigation and general usability of Simulation Template 4.

Not at all User-Friendly 1 2 3 4 5 6 Very Friendly

Please rate the navigation and general usability of Simulation Template 5.

Not at all User-Friendly 1 2 3 4 5 6 Very Friendly

Please rate the navigation and general usability of Simulation Template 6.

Not at all User-Friendly 1 2 3 4 5 6 Very Friendly

List and describe the types of navigation and usability features that would be present in your ideal simulation:

Stage 1: Satisfaction with Course Template Designs

Relative to simulations you have used, please describe your satisfaction with Simulation Template 1.

Not At All Satisfied 1 2 3 4 5 6 Very Satisfied

Relative to simulations you have used, please describe your satisfaction with Simulation Template 2.

Not At All Satisfied 1 2 3 4 5 6 Very Satisfied

Relative to simulations you have used, please describe your satisfaction with Simulation Template 3.

Not At All Satisfied 1 2 3 4 5 6 Very Satisfied

Relative to simulations you have used, please describe your satisfaction with Simulation Template 4.

Not At All Satisfied 1 2 3 4 5 6 Very Satisfied

Relative to simulations you have used, please describe your satisfaction with Simulation Template 5.

Not At All Satisfied 1 2 3 4 5 6 Very Satisfied

Relative to simulations you have used, please describe your satisfaction with Simulation Template 6.

Not At All Satisfied 1 2 3 4 5 6 Very Satisfied

Briefly describe the most satisfying elements of a simulation you really enjoyed:

In addition an updated version of the Purdue Scale of usability will be administered.

Reference: BEHAVIOUR & INFORMATION TECHNOLOGY, 1997, VOL. 16, NO. 4/5, 267 - 278

Objective 2: Disposition to Engage in Effortful Cognitive endeavor (Need for Cognition and Ambiguity Tolerance)

We will use two dispositional measures. These are provided immediately following this page. The measure is the Need for Cognition. The second is Ambiguity Tolerance. These instruments have a large literature base, with good evidence of both reliability and construct validity.

The Need for Cognition Scale

Please indicate the extent to which you agree or disagree with each of the following items, using the scale below. There are no correct answers, we are only interested in how you feel about the statements. Write a number between +4 and -4 in the blank by each item to indicate your agreement/disagreement with it.

+4 = very strong agreement

+3 = strong agreement

+2 = moderate agreement

+1 = slight agreement

0 = neither agreement not disagreement

-1 = slight disagreement

-2 = moderate disagreement

-3 = strong disagreement

-4 = very strong disagreement

_____1. I would prefer complex to simple problems.

_____2. I like to have the responsibility of handling a situation that requires a lot of thinking.

_____3. Thinking is not my idea of fun.

_____4. I would rather do something that requires little thought than something that is sure to challenge my thinking abilities.

_____5. I try to anticipate and avoid situations where there is a likely chance I will have to think in depth about something.

_____6. I find satisfaction in deliberating hard and for long hours.

_____7. I only think as hard as I have to.

_____8. I prefer to think about small, daily projects to long-term ones.

_____9. I like tasks that require little thought once Ive learned them.

_____10. The idea of relying on thought to make my way to the top appeals to me.

_____11. I really enjoy a task that involves coming up with new solutions to problems.

_____12. Learning new ways to think doesnt excite me very much.

_____13. I prefer my life to be filled with puzzles that I must solve.

_____14. The notion of thinking abstractly is appealing to me.

_____15. I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought.

_____16. I feel relief rather than satisfaction after completing a task that required a lot of mental effort.

_____17. Its enough for me that something gets the job done; I dont care how or why it works.

_____18. I usually end up deliberating about issues even when they do not affect me personality.

The AT-20 Scale

Please do not spend too much time on the following items. There are no right or wrong answers and therefore you first response is important. Mark T for true and F for false. Be sure to answer every question.

_____1. A problem has little attraction for me if I dont think it has a solution.

_____2. I am just a little uncomfortable with people unless I feel that I can understand their behaviour.

_____3. Theres a right way and a wrong way to do almost everything.

_____4. I would rather bet 1 to 6 on a long shot than 3 to 1 on a probable answer.

_____5. The way to understand complex problems is to be concerned with their larger aspects instead of breaking them into smaller pieces.

_____6. I get pretty anxious when Im in a social situation over which I have no control.

_____7. Practically every problem has a solution.

_____8. It bothers me when I am unable to follow another persons train of thought.

_____9. I have always felt there was a clear difference between right and wrong.

_____10. It bothers me when I dont know how other people react to me.

_____11. Nothing gets accomplished in this world unless you stick to some basic rules.

_____12. If I were a doctor, I would prefer the uncertainties of a psychiatrist to the clear and definite work of someone like a surgeon or X-ray specialist.

_____13. Vague and impressionistic pictures really have little appeal for me.

_____14. If I were a scientist, it would bother me that my work would never be completed (because science will always make new discoveries).

_____15. Before an examination, I feel much less anxious if I know how many questions there will be.

_____16. The best part of working a jigsaw puzzle is putting in the last piece.

_____17. Sometimes I rather enjoy going against the rules and doing things Im not supposed to do.

_____18. I dont like to work on a problem unless there is a possibility of coming out with clear-cut and unambiguous answer.

_____19. I like to fool around with new ideas, even if they turn out later to be a total waste of time.

_____20. Perfect balance is the essence of all good composition.

Objective 2: Expectancy-Value Questionnaire

The proposed measure of expectations for success within a suite of learning activities and the value placed in such success will be measured by the Expectancy-Value questionnaire. This questionnaire has been studied extensively, and it has good reliability and construct validity. Dr. Tashiro used the Expectancy-Value survey in a major study factors shaping retention in science majors at 12 American Universities. This study examined factors shaping retention of women and minority students in science majors. Over 5,000 students were surveyed. A complementary survey was administered to several hundred faculty members. These quantitative surveys were complemented by detailed case studies of each university.

Please note, that the survey included was originally designed for undergraduate chemistry majors. We have revised this questionnaire by editing as follows:

- We will not ask for Social Security Number or Social Insurance Number
- Item 2, School code has been dropped.
- Item 6 has been modified to be suitable for Canadian students.
- Other items have been modified to better fit Canadian populations and educational systems.

Combined Expectancy Value Instrument and Demographic Survey

Originally developed with National Science Foundation funding and used in a study of 12 American universities. This version was designed for students taking courses Introduction to Health Informatics and Theory and Practice of Patient-Centred Care.

1.What are the last three digits of your student ID? ____

2.Todays date (month/day/year): ____

3.Sex:? Male Female

4.Birthday (month/day/year): ____

5.Marital status:

?Single ? Married, living with spouse

Married, not living with spouse

6.# of children:

None

1

2

3

4 or more

7. Canadian citizen: Yes No

8. In what year did you graduate from high school?

2006 2005 2004 2003 2002 2001 2000 1999 1998
1997 1996 1995 1994 or sooner

9. Mark the one that best describes your average high school marks?

70% or less

70%-75%

75%-80%

80%-85%

85%-90%

90%-95%

95%-100%

10. Which courses did you take in your last year of high school (or university preparation program at college)?

English

Calculus

Algebra and Geometry

Physics

Chemistry

Biology

Other? Please specify: ____

11. Did you have college or university education before admission to your Nursing, Medical Laboratory Science, or Health Sciences program?

Yes. Please specify: ____

No

12. Which cohort are you in?

2003-2004

2004-2005

2005-2006

2006-2007

13. Are you enrolled as a:

Full-time student

Part-time student

INSTRUCTIONS: For all items that have a rating scale, mark one number only. On all other types of items, follow the directions given. Remember, it is very important to complete all the items on the questionnaire! Please note that when an item refers to COURSE, it refers to the course identified by Dr. Jay Shiro Tashiro.

14. How successful do you think you would be in a career which required knowledge of {COURSE}?

not at all successful 1 2 3 4 5 6 7 very successful

15. If you were to take a similar course as {COURSE} next semester, how well do you think you would do?

not at all well 1 2 3 4 5 6 7 very well

16. How well would you expect to do in advanced course in your program?

not at all well 1 2 3 4 5 6 7 very well

17. How well would you expect to do in HLSC 3800U Critical Appraisal for Health Sciences?

OR if you have already taken it, how well did you do?

not at all well 1 2 3 4 5 6 7 very well

18. How well would you expect to do in another advanced course in your program?

not at all well 1 2 3 4 5 6 7 very well

19. How well would you expect to do in COURSE?

not at all well 1 2 3 4 5 6 7 very well

20. Compared to other students in your class, how well do you expect to do in {COURSE} this semester?

not at all well 1 2 3 4 5 6 7 very well

21. How well do you expect to do on your next {COURSE} test?

not at all well 1 2 3 4 5 6 7 very well

22. If you are taking other COURSE courses this semester, how well do you think you will do in these courses?

not at all well 1 2 3 4 5 6 7 very well

23. How good at {COURSE} are you?

not at all good 1 2 3 4 5 6 7 very good

24. If you were to rank all the students in this class from the worst to the best in {COURSE}, where would you put yourself?

the worst 1 2 3 4 5 6 7 the best

25. In comparison to most of your other academic subjects, how are you at {COURSE}?

not at all good 1 2 3 4 5 6 7 very good

26. How good at {COURSE} does your mother/female guardian think you are?

not at all good 1 2 3 4 5 6 7 very good

27. How good at {COURSE} does your father/male guardian think you are?

not at all good 1 2 3 4 5 6 7 very good

28. How good at {COURSE} does your professor in this course think you are?

not at all good 1 2 3 4 5 6 7 very good

29. In general, how difficult is {COURSE} for you?

not at all difficult 1 2 3 4 5 6 7 very difficult

30. Compared to most other students in your class, how difficult is {COURSE} for you?

not at all difficult 1 2 3 4 5 6 7 very difficult

31. Compared to most other school subjects that you have taken or are taking, how difficult is {COURSE} for you?

not at all difficult 1 2 3 4 5 6 7 very difficult

32. How difficult does your mother/female guardian think {COURSE} is for you?

not at all difficult 1 2 3 4 5 6 7 very difficult

33. How difficult does your father/male guardian think {COURSE} is for you?

not at all difficult 1 2 3 4 5 6 7 very difficult

34. How difficult does your professor in this course think {COURSE} is for you?

not at all difficult 1 2 3 4 5 6 7 very difficult

35. How hard do you have to try to get good grades in {COURSE}?

not at all hard 1 2 3 4 5 6 7 very hard

36. How hard do you have to study for {COURSE} tests to get a good grade?

not at all hard 1 2 3 4 5 6 7 very hard

37. To do well in {COURSE} I have to work(Mark one).

Much harder in {COURSE} than in other subjects.

Somewhat harder in {COURSE} than in other subjects.

A little harder in {COURSE} than in other subjects.

The same as in other subjects.

A little harder in other subjects than in {COURSE}.

Somewhat harder in other subjects than in {COURSE}.

Much harder in other subjects than in {COURSE}.

38. How much time do you spend on {COURSE} homework? (Mark one).

An hour or more a day

30 minutes a day

15-30 minutes a day

About 1 hour a week

About 30 minutes a week

About 30 minutes every two weeks

I rarely do any {COURSE} homework

39. How hard do you try in {COURSE}?

not at all hard 1 2 3 4 5 6 7 very hard

40. Compared to most other students you know, how much time do you have to spend working on your {COURSE} assignments?

not some much time 1 2 3 4 5 6 7 a lot of time

41. How useful is learning the basics in {COURSE} for what you want to do after you graduate and go to work?

not at all useful 1 2 3 4 5 6 7 very useful

42. How useful do you think the things you have learned from the basics in {COURSE} for your other school courses?

not at all useful 1 2 3 4 5 6 7 very useful

43. How useful is what you would learn in university {COURSE} for what you will do when you finish school and go to work?

not at all useful 1 2 3 4 5 6 7 very useful

44. How useful is what you would learn in advanced {COURSE} for your daily life outside of school?

not at all useful 1 2 3 4 5 6 7 very useful

45. I feel that being good at solving problems which involve knowledge of {COURSE} is:

not at all important 1 2 3 4 5 6 7 very important

46. How important is it to you to get good grades in {COURSE}?

not at all important 1 2 3 4 5 6 7 very important

47. How upset would you be if you got a low grade in {COURSE}?

not at all upset 1 2 3 4 5 6 7 very upset

48. In general, I find working on {COURSE} assignments:

very boring 1 2 3 4 5 6 7 very interesting

49. How much do you like working with {COURSE}?

not at all 1 2 3 4 5 6 7 very much

50. How much do you like your professor in this course?

not at all 1 2 3 4 5 6 7 very much

51. How upset do you think your mother/female guardian would be if you got a low grade in {COURSE}?

not at all upset 1 2 3 4 5 6 7 very upset

52. How upset do you think your father/male guardian would be if you got a low grade in {COURSE}?

not at all upset 1 2 3 4 5 6 7 very upset

53. In {COURSE}, most of the time, how well do you do in each of the following things?

a. when the teacher calls on you for an answer in class:

not at all well 1 2 3 4 5 6 7 very well

b. when taking a test you have studied for:

not at all well 1 2 3 4 5 6 7 very well

c. when doing {COURSE} homework problems:

not at all well 1 2 3 4 5 6 7 very well

54. How have you been doing in this course, so far this semester?

not at all well 1 2 3 4 5 6 7 very well

55. What is the lowest grade you would be satisfied with in this course? (Mark one).

A A- B+ B B- C+ C C- D+ or lower

56. How hard do you think next semester {COURSE} would be for you?

not at all hard 1 2 3 4 5 6 7 very hard

57. How hard do you think advanced {COURSE} would be for you?

not at all hard 1 2 3 4 5 6 7 very hard

58. Compared to most other subjects you may take in university, how hard do you think advanced {COURSE} would be for you?

not at all hard 1 2 3 4 5 6 7 very hard

59. If you decide to take {an advanced course in your program}, how hard do you think it would be for you?

not at all hard 1 2 3 4 5 6 7 very hard

60. When you take HLSC 3800U Statistics and Critical Appraisal for Health Science how hard do you think it would be for you?

OR if you are taking it right now, how hard is it for you?

OR if you have already taken it, how hard was it for you?

not at all hard 1 2 3 4 5 6 7 very hard

61. If you decide to take {another advanced course in your program}, how hard do you think it would be for you?

not at all hard 1 2 3 4 5 6 7 very hard

62. How well do you think your father/male guardian expects you to do in {COURSE} this semester?

not at all well 1 2 3 4 5 6 7 very well

63. How well do you think your mother/female guardian expects you to do in {COURSE} this semester?

not at all well 1 2 3 4 5 6 7 very well

64. How well do you think your professor in this course expects you to do in {COURSE} this semester?

not at all well 1 2 3 4 5 6 7 very well

65. Is the amount of effort it will take to do well in your {COURSE} this semester worthwhile to you?

not at all worthwhile 1 2 3 4 5 6 7 very worthwhile

66. Is the amount of effort it will take to do well in advanced {COURSE} worthwhile to you?

not at all worthwhile 1 2 3 4 5 6 7 very worthwhile

67. How much does the amount of time you spend on {COURSE} keep you from doing other things you would like to do?

takes away no time 1 2 3 4 5 6 7 takes away a lot of time

Please rate on a scale of 1 to 7 how much each of the following people have encouraged or discouraged you to continue in {COURSE}:

68. My mother/female guardian:

strongly discouraged me 1 2 3 4 5 6 7 strongly encouraged me

69. My father/male guardian:

strongly discouraged me 1 2 3 4 5 6 7 strongly encouraged me

70. My brothers (skip if you have no brothers):

strongly discouraged me 1 2 3 4 5 6 7 strongly encouraged me

71. My sisters (skip if you have no sisters):

strongly discouraged me 1 2 3 4 5 6 7 strongly encouraged me

72. Other family members:

strongly discouraged me 1 2 3 4 5 6 7 strongly encouraged me

73. My friends:

strongly discouraged me 1 2 3 4 5 6 7 strongly encouraged me

74. My high school teachers:

strongly discouraged me 1 2 3 4 5 6 7 strongly encouraged me

75. My high school counselor:

strongly discouraged me 1 2 3 4 5 6 7 strongly encouraged me

76. My university professors:

strongly discouraged me 1 2 3 4 5 6 7 strongly encouraged me

77. My student advisor in university:

strongly discouraged me 1 2 3 4 5 6 7 strongly encouraged me

78. Would you take more {COURSE} if you did not have to?

definitely would not take more 1 2 3 4 5 6 7 definitely would take more

79. If it were your decision alone, how much more {COURSE} would you take?

I would not take any more {COURSE}

I would take one or two more {COURSE}

I would take {COURSE} in my 4th-Year

I would take {COURSE} through undergraduate, plus some graduate work

I would take {COURSE} through a masters degree

I would take {COURSE} all the way through a doctoral degree

80. a) How many courses like {COURSE} did you take or plan to take during your first year (not including this course):

None 1 2 3 4 or more

b) How many courses like {COURSE} did you take or plan to take during your second year (not including this course):

None 1 2 3 4 or more

c) How many courses like {COURSE} did you take or plan to take during your third year (not including this course):

None 1 2 3 4 or more

d) How many courses like {COURSE} are you taking or plan to take during your fourth year (not including this course):

None 1 2 3 4 or more

89. In the past, how often have you performed very well on {COURSE} tests?

not at all often 1 2 3 4 5 6 7 very often

90. In the past, how often have you performed very poorly on {COURSE} tests?

not at all often 1 2 3 4 5 6 7 very often

91. How much does your mother/female guardian use knowledge of {COURSE}?

a little 1 2 3 4 5 6 7 a lot

92. How useful do you think women find knowledge of basic {COURSE} in their jobs?

not at all useful 1 2 3 4 5 6 7 very useful

93. How useful do you think women find knowledge of advanced {COURSE} in their jobs?

not at all useful 1 2 3 4 5 6 7 very useful

94. How useful do you think women find basic {COURSE} in their everyday activities?

not at all useful 1 2 3 4 5 6 7 very useful

95. How much does your father/male guardian use knowledge of {COURSE}?

a little 1 2 3 4 5 6 7 a lot

96. How useful do you think men find knowledge of basic {COURSE} in their jobs?

not at all useful 1 2 3 4 5 6 7 very useful

97. How useful do you think men find knowledge of advanced {COURSE} in their jobs?

not at all useful 1 2 3 4 5 6 7 very useful

98. How useful do you think men find basic {COURSE} in their everyday activities?

not at all useful 1 2 3 4 5 6 7 very useful

99. In general, I think women are

much worse than men at {COURSE} 1 2 3 4 5 6 7 much better than men at {COURSE}

100. In general, I think men are

much worse than women at {COURSE} 1 2 3 4 5 6 7 much better than women at {COURSE}

101. In general, I think people of my ethnicity are

much worse than others at {COURSE} 1 2 3 4 5 6 7 much better than others at {COURSE}

102. Please indicate which of the following you plan to do after you graduate from college.

a) Continue your education (please mark all that apply)

Masters degree

Doctoral degree (PhD. Or EdD.)

Doctoral degree (M.D. or other medical degree)

Law or other professional degree

Other

b)Look for a job

c)Go into business

d)Military service

e)Public service (Peace Corps, etc.)

f)Other plans

IN THIS SECTION, WE WOULD LIKE TO ASK YOU ABOUT THE REASONS FOR HOW YOU HAVE PERFORMED ON SOME OF YOUR {COURSE} TESTS. First, think about a time when you did very well on a {COURSE} test. The items below concern your impressions or opinions of the cause or causes of your performance. Mark one number for each of the questions. For example, the first item below asks to what extent the cause of your performance was something that reflects an aspect of the situation, or something that reflects an aspect of yourself. If you believed your performance was totally due to something about yourself, you would mark 9. If you believed your performance was mostly due to yourself but partly to the situation, you might mark 7 or 6. If you believed your performance was almost totally due to the situation, you might mark 2, and so on.

103. Was the cause(s) of your good performance something:

That reflects an aspect of the situation 1 2 3 4 5 6 7 8 9 Reflects an aspect of yourself

Not manageable by you 1 2 3 4 5 6 7 8 9 Manageable by you

Temporary 1 2 3 4 5 6 7 8 9 Permanent

You cannot regulate 1 2 3 4 5 6 7 8 9 You can regulate

Over which others have no control 1 2 3 4 5 6 7 8 9 Over which others have control

Outside of you 1 2 3 4 5 6 7 8 9 Inside of you

Variable over time 1 2 3 4 5 6 7 8 9 Stable over time

Not under the power of other people 1 2 3 4 5 6 7 8 9 Under the power of other people

Something about others 1 2 3 4 5 6 7 8 9 Something about you

Over which you have no power 1 2 3 4 5 6 7 8 9 Over which you have power

Changeable 1 2 3 4 5 6 7 8 9 Unchangeable

Other people cannot regulate 1 2 3 4 5 6 7 8 9 Other people can regulate

NOW, think about a time when you did very poorly on a {COURSE} test. The items below concern your impressions or opinions of the cause or causes of your performance. Mark one number for each of the following questions.

104. Was the cause(s) of your poor performance something:

That reflects an aspect of the situation1 2 3 4 5 6 7 8 9 Reflects an aspect of yourself

Not manageable by you 1 2 3 4 5 6 7 8 9 Manageable by you

Temporary 1 2 3 4 5 6 7 8 9 Permanent

You cannot regulate 1 2 3 4 5 6 7 8 9 You can regulate

Over which others have no control 1 2 3 4 5 6 7 8 9 Over which others have control

Outside of you 1 2 3 4 5 6 7 8 9 Inside of you

Variable over time 1 2 3 4 5 6 7 8 9 Stable over time

Not under the power of other people 1 2 3 4 5 6 7 8 9 Under the power of other people

Something about others 1 2 3 4 5 6 7 8 9 Something about you

Over which you have no power 1 2 3 4 5 6 7 8 9 Over which you have power

Changeable 1 2 3 4 5 6 7 8 9 Unchangeable

Other people cannot regulate 1 2 3 4 5 6 7 8 9 Other people can regulate

In the following sections we are interested in learning some of your impressions of the course in which you received this questionnaire. Please refer only to this course in filling out the sections below. In the following section we are interested in the difficulty of the course:

105. How difficult is it to understand the assigned reading materials?

very easy 1 2 3 4 5 6 7 very difficult

106. How difficult are the problem sets?

very easy 1 2 3 4 5 6 7 very difficult does not apply

107. How difficult are the writing assignments?

very easy 1 2 3 4 5 6 7 very difficult does not apply

108. How difficult are the exams in {COURSE}?

very easy 1 2 3 4 5 6 7 very difficult

109. How difficult is it to understand the terminology used in {COURSE}?

very easy 1 2 3 4 5 6 7 very difficult

110. How would you describe the professors spoken accent?

Definitely standard English 1 2 3 4 5 6 7 Definitely non-standard English

111. How difficult is it to understand the professors spoken language?

very easy 1 2 3 4 5 6 7 very difficult

112. How difficult is the course overall?

very easy 1 2 3 4 5 6 7 very difficult

In the following section we are interested in how well the course meets your expectations. For each of the following course characteristics, please indicate the extent to which it matches the expectations you had when you first entered the course:

113. Readability of assigned readings:

not at all close to my expectations 1 2 3 4 5 6 7 very close to my expectations

114. Work load:

not at all close to my expectations 1 2 3 4 5 6 7 very close to my expectations

115. Overall level of difficulty:

not at all close to my expectations 1 2 3 4 5 6 7 very close to my expectations

We would like to know if you are aware of the reasons for your instructors choices to teach you in a particular way. Your responses should reflect your general level of awareness and not specific feelings about this specific course.

116. The logic of the course:

not at all clear 1 2 3 4 5 6 7 very clear

117. The reasons for the choice of the text or other readings:

not at all clear 1 2 3 4 5 6 7 very clear

118. The reasons for the course format (lecture, laboratory, discussion, etc.):

not at all clear 1 2 3 4 5 6 7 very clear

119. The reasons for the choice of assignments:

not at all clear 1 2 3 4 5 6 7 very clear

120. How the level of difficulty was chosen:

not at all clear 1 2 3 4 5 6 7 very clear

121. Why group activities are used:

not at all clear 1 2 3 4 5 6 7 very clear

122. Why writing assignments are used:

not at all clear 1 2 3 4 5 6 7 very clear

When student discussions occur in this course, what do they typically focus on? Please indicate the approximate percentage of time devoted to each of the items below:

123. Textbook material:

$< 10\%$ 20%-30% 40%-50% 60%-70% 80%-90% 10%-20% 30%-40% 50%-60% 70%-80% $> 90\%$

124. Non-textbook material:

<10% 20%-30% 40%-50% 60%-70% 80%-90% 10%-20% 30%-40% 50%-60% 70%-80% $> 90\%$

125: Ideas raised by the instructor:

<10% 20%-30% 40%-50% 60%-70% 80%-90% 10%-20% 30%-40% 50%-60% 70%-80% $> 90\%$

126: Ideas raised by the students:

<10% 20%-30% 40%-50% 60%-70% 80%-90% 10%-20% 30%-40% 50%-60% 70%-80% $> 90\%$

In this section we are interested in the nature and quality of language used in introductory university textbooks. Please rate the textbook used in this course on the scales below:

127. Organization of topics:

easy for most students 1 2 3 4 5 6 7 difficult for most students

128. Writing style used in the text:

easy for most students 1 2 3 4 5 6 7 difficult for most students

129. Vocabulary used in the text:

easy for most students 1 2 3 4 5 6 7 difficult for most students

130. Mathematical symbols used in the text:

easy for most students 1 2 3 4 5 6 7 difficult for most students

131. Examples used in the text:

easy for most students 1 2 3 4 5 6 7 difficult for most students

132. Problem sets used in the text:

easy for most students 1 2 3 4 5 6 7 difficult for most students

133. If course grades were assigned today, what grade do you think you would get?

A B+ C+ D+ or lower ? A- B C

B- C-

Please mark a number to indicate how much you agree or disagree with the following statements:

134. I am active in organizations or social groups that include mostly members of my own ethnic group.

strongly agree somewhat agree somewhat disagree strongly disagree

135. I participate in cultural practices of my own group, such as special food, music, or customs.

strongly agree somewhat agree somewhat disagree strongly disagree

136. Do you speak any languages other than English?

Yes. Please specify:____

If yes, which language do you prefer?____

No

137. Which option best describes how you learned your language(s). Mark one only.

Learned only English

Learned English first, then a second language

Learned another language first, then English

Learned English and another language at the same time

138. Since you have been in university, about how much time do you typically spend per week in each of the following activities:

Activity	**Hours Per Week**							
	None	Less than 1	1-2	3-5	6-10	11-15	16-20	Over 20
Classes/labs	_	_	_	_	_	_	_	_
Studying/homework	_	_	_	_	_	_	_	_
Socializing with friends	_	_	_	_	_	_	_	_
Talking with faculty outside of class	_	_	_	_	_	_	_	_
Exercising/sports	_	_	_	_	_	_	_	_
Reading for pleasure	_	_	_	_	_	_	_	_
Partying	_	_	_	_	_	_	_	_
Working (for pay)	_	_	_	_	_	_	_	_
Volunteer work	_	_	_	_	_	_	_	_
Student clubs or groups	_	_	_	_	_	_	_	_
Watching TV	_	_	_	_	_	_	_	_
Commuting to campus	_	_	_	_	_	_	_	_
Religious services/meetings	_	_	_	_	_	_	_	_
Hobbies	_	_	_	_	_	_	_	_
Child or family obligations	_	_	_	_	_	_	_	_

139. Which option below best describes where you are living this semester?

With parents or relatives

Your own home or apartment

UOIT/DC residence

Off-campus student housing

Other

140. How many kilometres is this university from your permanent home? Mark one only.

5 or less

51-100

6-10

101-500

11-50

More than 500

141. What is your best estimate of your (or your parents, if you are being supported by them) total income last year? (Consider all sources before taxes):

Less than $10,000

$10,000-19,999

$20,000-29,999

$30,000-39,999

$40,000-49,999

$50,000-59,999

$60,000-74,999

$75,000-99,999

$100,000-149,999

$150,000 or more

142. What is the highest level of education obtained by your parents/guardian? Mark one in each column.

Father or Male Guardian Mother or Female Guardian

8th grade or less

Some high school

High school graduate

Some college or university

College or university degree

Some graduate school

Graduate degree

This is the end of the Expectancy-Value questionnaire

THANK YOU VERY MUCH FOR YOUR COOPERATION

Objective 2: Competency Performance Evaluation in Simulation Learning Activities

Below, we show the five types of performance evaluations for learning activities related to understanding Interprofessional care planning and delivery. Each type of evaluation provides an estimate of performance for a particular type of learning outcome. These estimates will be used to test the hypotheses related to differences among the control and treatment groups.

Assessment	Allocation of 200 Total Course Points
Discussion Forum	Four (4) Discussion Forum assignments @ 5 points each for a tota
Simulations: Case Studies	Six (6) Cases Studies @ 10 points each for a total of 50 points. Tl

Treatment groups will complete the Discussion Forum assignments and also work within the eight (8) simulations.

Control groups will complete the four Discussion Forum assignments and eight (8) sets of paper and pencil learning activities related to the same content as covered within the simulations.

Objective 2: Satisfaction with {Simulations OR Paper-pencil Learning Activities}, Realism of {Simulations OR Paper-pencil Learning Activities}, Delivery Modality for {Simulations OR Paper-pencil Learning Activities}, and Content or {Simulations OR Paper-pencil Learning Activities}

On average, how much time per week did you work within the simulations OR paper-pencil learning activities?

- 0 hours
- 1 hour
- 2 hours
- 3 hours
- More than 3 hours

How satisfied are you with the {simulations OR paper-pencil learning activities}?

Not At All Satisfied 1 2 3 4 5 6 Very Satisfied

Rate your satisfaction with the Study Plan provided with the {simulations OR paper-pencil learning activities}.

Not At All Satisfied 1 2 3 4 5 6 Very Satisfied

Rate your satisfaction with the realism of the {simulations OR paper-pencil learning activities}.

Not At All Satisfied 1 2 3 4 5 6 Very Satisfied

Rate your satisfaction with the relevance of each {simulations OR paper-pencil learning activities}.

Not At All Satisfied 1 2 3 4 5 6 Very Satisfied

Rate how well the {simulations OR paper-pencil learning activities} improved your learning.

Not At All Satisfied 1 2 3 4 5 6 Very Satisfied

What types of {simulations OR paper-pencil learning activities} would you like to see for this {course OR professional development program}:

How satisfied are you with the ratio of online work to face-to-face instruction for this {course or professional development program}?

Not At All Satisfied 1 2 3 4 5 6 Very Satisfied

Rate your satisfaction with the instructional support provided by the faculty member?

Not At All Satisfied 1 2 3 4 5 6 Very Satisfied

Rate your satisfaction with the instructional support provided within the online components of the course?

Not At All Satisfied 1 2 3 4 5 6 Very Satisfied

Rate your satisfaction with the suitability of the delivery modality for the simulations OR paper-pencil learning activities.

Not At All Satisfied 1 2 3 4 5 6 Very Satisfied

Rate this {course OR professional development program} in terms of difficulty relative to other {courses or programs} you have completed.

- Much easier than most courses
- A little easier than most courses
- About the same as most courses
- More difficult than most courses
- Much more difficult than most courses

How satisfied are you with the content of the simulations OR paper-pencil learning activities?

Not At All Satisfied 1 2 3 4 5 6 Very Satisfied

Did the instructions provide a kind of map for you to follow as you worked through {simulations OR paper-pencil learning activities}?

No, Not At all 1 2 3 4 5 6 Yes, Very Much So

Did the simulations OR paper-pencil learning activities provide you with a kind of map for you to follow as you worked through the assignments in the module?

No, Not At all 1 2 3 4 5 6 Yes, Very Much So

Did the Discussion Forums help you learn the course material?

No, Not At all 1 2 3 4 5 6 Yes, Very Much So

How satisfied are you with the Discussion Forums of the course?

Not At All Satisfied 1 2 3 4 5 6 Very Satisfied

What overall performance evaluation do you expect for your work in the {simulations OR paper-pencil learning activities}?

Excellent

Good

Average

Below Average

Poor

Appendix B

University of Ontario Institute of Technology, Research Ethics Board file #: REB09-27

Principal Investigators: Drs. Miguel Vargas Martin and Jayshiro Tashiro

Student Investigators: Meaghen Regts and Arturo Fernandez Espinosa

Research Phase: Objective 1 Usability Study

Document Type: Consent Form for Usability Study

Format: On UOIT Letterhead

[REB FILE NUMBER]

Dear Health Sciences Student OR Nursing Staff member:

Please read this consent form, checking off each item of the form. If you decide that you want to participate in the research, then sign the form and follow the instructions to return the form to the Research Associate.

- We are inviting you to participate in a research project that will examine preferences for the design of simulations to help improve competencies in interprofessional care.
- This research has been approved by the UOIT Research Ethics Board. The project is being implemented by Dr. Jay Shiro Tashiro, Associate Professor of Health Sciences. His contact information is provided below:

  Jay Shiro Tashiro, PhD, BSN, RN

  Building UA, Office 3015

  Telephone Number (905) 721.8668, Extension 3616

  Jay.Tashiro@uoit.ca

- The research will be conducted during the period APPROPRIATE DATES.
- If you wish to participate, you would be involved in two sessions.
- The first session will last about 90 minutes. During this session, you and about 20 other students will be asked to complete two tasks. First, you will complete a questionnaire about your preferences for computer-based simulation interfaces. Second, the session group will be shown a series of simulation designs and as a group we will discuss which designs are more user-friendly for you. This discussion will be recorded on audio tapes. The questionnaire is anonymous and the discussion will be recorded but without identification of you or any others in the session.
- The second session will last about 60 minutes. During this session, you will be able to view and comment on models of simulation interfaces that were developed after analysis of the input gained from research subjects responses in the first session. This second session also will be recorded on audio tapes for analysis of student commentary. No one will be identified during these commentaries.
- Your participation is completely voluntary and refusal to participate will involve no penalty whatsoever. If you participate, you also have the option of discontinuing participation at any time, again without penalty of any kind.
- All data collected during the session will be treated with confidentiality and no individuals data will be identified by name.
- The questionnaires and taped discussions will be reviewed by a Research Associate, who is not a faculty member and who will code the data so that no individual could be identified.
- Dr. Tashiro will then conduct analyses and will be blind to any individuals identity. Furthermore, data analyses and summaries will never identify any individual.
- Collected data are stored in a repository managed by the Faculty of Health Sciences. You have the right to examine the data analyses and summaries.
- Based on similar research conducted by Dr. Tashiro, we do not anticipate any risks to you. The questionnaire will not contain questions of a personally intrusive nature. The Web-based course interfaces will not contain disturbing images.

- Because your participation is voluntary and you can leave at any time during the research sessions, you can monitor your discomfort level and remove yourself from the research at any time. You may also speak with Dr, Tashiro for a debriefing of the research participation experience at any time.
- The benefits of the research evolve mostly from participation in identifying preferences for elements of simulations nested within Web-based interfaces.
- These preferences will then be incorporated into simulation designs that will be studied to see if they improve dispositions to learn and overall learning outcomes. Consequently, your input may shape processes that improve education and training of healthcare students and practitioners.
- You may also contact a Grants Officer who can provide answers to pertinent questions about research subjects rights (or 905.721.8668 Extension 2156).
- Results of the research will be published in professional journals as well as presented at national and international conferences that focus on educational research. Again, we emphasize that no individual student can be identified from the types of data analyses and summaries that are sued for journal articles and conference presentations. If you want to be informed of articles and presentation containing the research results, please check the following statement.
- I want to be so informed (check box at beginning of this statement).

If you decide to participate, please sign this consent form in the space provided below.

I have read this consent form and understand the intent of the research and my role as a participant in the research. I know that I can ask questions about the research in the future and that I can withdraw from the research at any time without consequences or penalties of any kind. I act with free and informed consent to participate in the research by signing this consent form.

Signature:

Please Print Your Name:

Researcher:

Miguel Vargas Martin, PhD, PEng

Jay Shiro Tashiro, PhD, BSN, RN

Please return this form directly to the Research Associate. The Research Associate will record your willingness to participate, and contact you to let you know more about your sessions. A copy of the form, signed by the researchers, will be returned to you.

Appendix C

University of Ontario Institute of Technology, Research Ethics Board file #: REB09-27

Principal Investigators: Drs. Miguel Vargas Martin and Jayshiro Tashiro

Student Investigators: Meaghen Regts and Arturo Fernandez Espinosa

Research Phase: Objective 2 Dispositions and Learning Outcomes Study

Document Type: Consent Form for the Learning Outcomes and Dispositions Study

Format: On UOIT Letterhead

[REB FILE NUMBER]

Dear Health Sciences Student:

Please read this consent form, checking off each item of the form. If you decide that you want to participate in the research, then sign the form and follow the instructions to return the form to the Research Associate.

- We are inviting you to participate in a research project that will examine learning outcomes and your dispositions to learn competencies in interprofessional care.
- This research has been approved by the UOIT Research Ethics Board.
- The project is being implemented by Dr. Jay Shiro Tashiro, Associate Professor of Health Sciences. His contact information is provided below:

  Jay Shiro Tashiro, PhD, BSN, RN

  Building UA, Office 3015

  Telephone Number (905) 721.3111, Extension 3616

Jay.Tashiro@uoit.ca

- The research will be conducted during the period APPROPRIATE DATES.

- If you wish to participate, you would complete a questionnaire at the beginning and end of the research period. This questionnaire measures: (1) your disposition to engage in critical thinking; (2) your expectations for success and value placed on success in the use of PHIPA simulations, (3) your satisfaction with simulations, the realism of simulations, the simulation delivery on the Web, and the simulation content).

- The questionnaire also contains a short demographic survey that provides researchers with information related to your work and other activities (study time, working, socializing with friends, and so on), age, and general academic performance in prior courses.

- In addition, the research would examine your learning outcomes as measured by your performance working on learning activities in the simulations.

- Your participation in the research is completely voluntary and refusal to participate will involve no penalty whatsoever. Specifically, your grade in the course OR your work performance evaluation is in no way influenced by your decision to participate or by your responses on the questionnaire and demographic survey.

- If you participate, you also have the option of discontinuing participation at any time, again without penalty of any kind.

- All data collected during your work in simulations will be treated with confidentiality and no individuals data will be identified by name. The questionnaire and performance in learning activities will be placed into a database for each participant and coded so that no individual could be identified. This work will be completed by a Research Associate, who is not a faculty member.

- The coded data will then be analyzed by the Research Associate and presented as data summaries to Dr. Tashiro. In this manner, the faculty member for the course remains blind to any individuals identity because data analyses and summaries will never identify any individual.

- Collected data will be stored in a repository managed by the Faculty of Health Sciences. You have the right to examine the data analyses and summaries.

- Based on similar research conducted by Dr. Tashiro, we do not anticipate any risks to you. The questionnaire will not contain questions of a personally intrusive nature. The graded assignments are part of your routine participation in the course.
- Because your participation is voluntary and you can leave at any time during the research sessions, you can monitor your discomfort level and remove yourself from the research at any time while still completing the course without penalty. You may also speak with Dr, Tashiro for a debriefing of the research participation experience at any time.
- The benefits of the research evolve mostly from your participation in research that helps create evidence-based frameworks for educational methods and materials in the Health Sciences. Such frameworks can then be incorporated into course designs and professional development activities in order to create education and training programs that really work to improve healthcare students and practitioners learning as well as identifies elements of educational materials that are likely to improve dispositions to learn. Consequently, your input may shape processes that improve courses for UOIT a well as for other universities and colleges OT professional development training for healthcare providers.
- You may also contact a Grants Officer who can provide answers to pertinent questions about research subjects rights (905.721.8668 Extension 2156).
- We want to thank you for considering participation in the research. If you decide to participate, please sign the consent form and give it to the Research Associate. The Research Associate will receive your form, record your willingness to participate, and contact you to let you know more about your sessions.
- Results of the research will be published in professional journals as well as presented at national and international conferences that focus on educational research. Again, we emphasize that no individuals can be identified from the types of data analyses and summaries that are used for journal articles and conference presentations. If you want to be informed of articles and presentation containing the research results, please check the following statement.
- I want to be so informed. (check box at beginning of this statement).

If you decide to participate, please sign this consent form in the space provided below.

I have read this consent form and understand the intent of the research and my role as a participant in the research. I know that I can ask questions about the research in the future and that I can withdraw from the research at any time without consequences or penalties of any kind. I act with free and informed consent to participate in the research by signing this consent form.

Signature:

Please Print Your Name:

Researcher:

Miguel Vargas Martin, PhD, PEng

Jay Shiro Tashiro, PhD, BSN, RN

Please return this form to the Research Associate, who will record your willingness to participate. A copy of the form, signed by the researchers, will be returned to you.

Appendix D

University of Ontario Institute of Technology, Research Ethics Board file #: REB09-27

Principal Investigators: Drs. Miguel Vargas Martin and Jayshiro Tashiro

Student Investigators: Meaghen Regts and Arturo Fernandez Espinosa

Objective 1: Usability Studies Our approach to recruitment will be to send letters of invitation to all Health Sciences students. The first strategy will be to seek permission from faculty members to visit their courses and pass out a recruitment letter to students. The second strategy, and really a contingency if we still do not have a sufficient sample and program distribution, will be to email the recruitment letter students in Health Sciences programs.

Document Type: Letter of Invitation for Usability Study

Format: On UOIT Letterhead

Dear Health Sciences Student:

We are inviting you to participate in a research project that will examine students OR nurses preferences for the design of simulations to help improve understanding of interprofessional education leading to improved patient care. This research has been approved by the UOIT Research Ethics Board ADD FILE NUMBER. The project is being implemented by Dr. Jay Shiro Tashiro, Associate Professor of Health Sciences. His contact information is provided below:

Jay Shiro Tashiro, PhD, BSN, RN

Building UA, Office 346

Telephone Number (905) 721.8668, Extension 3616

Jay.Tashiro@uoit.ca

The research will be conducted during the period ADD DATES AS APPROPRIATE. If you wish to participate, you would be involved in two sessions. The first session will last about 90 minutes. During this session, you and about 20 other students will be asked to complete two tasks. First, you will be complete a questionnaire about your preferences for Web-based and simulation interfaces. Second, the session group will be shown a series of simulation designs and as a group we will discuss which designs are more user-friendly for you. This discussion will be recorded on audio tapes. The questionnaire is anonymous and the discussion will be recorded but without identification of you or any others in the session.

The second session will last about 60 minutes. During this session, you will be able to view and comment on models of simulations that were developed after analysis of the input gained from responses in the first session. This second session also will be recorded on audio tapes for analysis of research subjects commentary. Research subjects will not be identified during these commentaries.

Your participation is completely voluntary and refusal to participate will involve no penalty whatsoever. If you participate, you also have the option of discontinuing participation at any time, again without penalty of any kind. All data collected during the session will be treated with confidentiality and no individuals data will be identified by name. The questionnaires and taped discussions will be reviewed by a Research Associate who does not know any of the research participants and who will code the data so that no individual could be identified. Dr. Tashiro will then conduct analyses and will be blind to any individuals identity. Furthermore, data analyses and summaries will never identify any individual. Collected data are stored in a repository managed by the Faculty of Health Sciences. You have the right to examine the data analyses and summaries.

Based on similar research conducted by Dr. Tashiro, we do not anticipate any risks to you. The questionnaire will not contain questions of a personally intrusive nature. The Web-based interfaces will not contain disturbing images. Because your participation is voluntary and you can leave at any time during the research sessions, you can monitor your discomfort level and remove yourself from the research at any time. You may also speak with Dr. Tashiro for a debriefing of the research participation experience at any time.

The benefits of the research evolve mostly from your participation in identifying preferences for elements of simulation interfaces. These preferences will then be incorporated into simulations that will be studied to see if they improve dispositions to learn and overall learning outcomes. Consequently, your input may shape processes that improve courses for UOIT a well as for other universities and colleges OR professional development raining modules.

You may also contact a Compliance Officer who can provide answers to pertinent questions about research subjects rights (905.721.8668 Extension 3693).

Thank you for considering participation in the research. If you decide to participate, please complete the section below and Dr. Tashiro will contact you to confirm your session. When you arrive at the session, you will be given a consent form and Dr. Tashiro or a Research Associate will review this with you. If you want to participate, she will show you what to check off and sign. Again, all information is strictly confidential.

NOTE: TELEPHONE RECRUITMENT IS IMPLEMENTED BY USING THE TEXT OF THIS LETTER AS A SCRIPT.]

Appendix E

University of Ontario Institute of Technology, Research Ethics Board file #: REB09-27

Principal Investigators: Drs. Miguel Vargas Martin and Jayshiro Tashiro

Student Investigators: Meaghen Regts and Arturo Fernandez Espinosa

Objective 2: Learning Outcomes and Dispositions Study For this research effort, we will pass out letters of invitation on the first day of class in each of two sections in four courses that are currently under development and that will be selected based on their content of interprofessional collaborative patient-centred care.

Research Phase: Learning Outcomes and Dispositions Study

Document Type: Letter of Invitation for Learning Outcomes and Dispositions Study

Format: On UOIT Letterhead

Dear Health Sciences Student:

We are inviting you to participate in a research project that will examine students learning outcomes and their dispositions to learn in COURSE NAME OR PROFESSIONAL DEVELOPMENT ACTIVITY. This research has been approved by the UOIT Research Ethics Board ADD FILE NUMBER. The project is being implemented by Dr. Jayshiro Tashiro, Associate Professor of Health Sciences. His contact information is provided below:

Jayshiro Tashiro, PhD, BSN, RN

Building UA, Office 3015

Telephone Number (905) 721.3111, Extension 3616

Jay.Tashiro@uoit.ca

The research will be conducted during the PERIOD ADD DATES AS APPROPRIATE. If you wish to participate, you would complete a questionnaire at the beginning and end of the semester. This questionnaire measures: (1) your disposition to engage in critical thinking; (2) your expectations for success and value placed on success in the learning activity, (3) your satisfaction with the realism of simulations, the delivery of simulations on the Web, and the simulation content). The questionnaire also contains a short demographic survey that provides researchers with information related to your school OR work and extracurricular activities (study time, working, socializing with friends, and so on), age, and general academic performance in prior courses. In addition, the research would examine your learning outcomes as measured by your performance on learning activities in the simulations.

In addition, all research subjects will be completing learning activities related to PHIPA regulations. If you volunteer to participate in the research, your work on a series of PHIPA modules will used as research data.

Your participation in the research is completely voluntary and refusal to participate will involve no penalty whatsoever. Specifically, your grade in the course OR your work performance evaluation is in no way influenced by your decision to participate or by your responses on the questionnaire and demographic survey or by your work on the PHIPA modules. If you participate, you also have the option of discontinuing participation at any time, again without penalty of any kind. All data collected during the session will be treated with confidentiality and no individuals data will be identified by name. The questionnaire and performance on PHIPA modules will be placed into a database for each participant and coded so that no individual could be identified. This work will be completed by a Research Associate who does not know you. The coded data will then be analyzed by the Research Associate and presented as data summaries to Dr. Tashiro. In this manner, the faculty member for the course OR your staff educator remains blind to any individuals identity because data analyses and summaries will never identify any individual. Collected data will be stored in a repository managed by the Faculty of Health Sciences. You have the right to examine the data analyses and summaries.

Based on similar research conducted by Dr. Tashiro, we do not anticipate any risks to you. The questionnaire will not contain questions of a personally intrusive nature. The graded assignments are part of your routine participation in the course. Because your participation is voluntary and you can leave at any time during the research sessions, you can monitor your discomfort level and remove yourself from the research at any time while still completing the course OR professional development activities without penalty. You may also speak with Dr, Tashiro for a debriefing of the research participation experience at any time.

The benefits of the research evolve mostly from participation in research that helps create evidence-based frameworks for PHIPA simulations for healthcare students and providers. Such frameworks can then be incorporated into educational and training materials in order to create educational options that really work to improve learning and identifies elements of educational materials that are likely to improve dispositions to learn. Consequently, your input may shape processes that improve courses for UOIT a well as for other universities and colleges OR professional development activities for healthcare providers.

You may also contact a Compliance Officer who can provide answers to pertinent questions about research subjects rights (905.721.8668 Extension 3693).

Thank you for considering participation in the research. If you decide to participate, please sign the consent form and place the form in the envelope we have provided. Then, simply give the envelope to the Research Associate who passed out this letter of invitation to participate in the research. The Research Associate will record your willingness to participate, and contact you to let you know more about the research.

NOTE: TELEPHONE RECRUITMENT IS IMPLEMENTED BY USING THE TEXT OF THIS LETTER AS A SCRIPT.]

Bibliography

[1] J. Han and M. Kamber. *Data Mining Concepts and Techniques second edition*. Morgan Kaufmann Publishers, 2006.

[2] M. Regts, A. Fernandez, M. Vargas Martin, and J. Tashiro. A knowledge management methodology for studying health science students development of misconceptions. *DBKDA '12 The Fourth International Conference on Advances in Databases, Knowledge and Data Applications*, I:112–119, 2012.

[3] J. Tashiro, C.K.P. Hung, and M. Vargas Martin. Evidence-based educational practices and a theoretical framework for hybrid learning. *ICHL '11 Proceedings of the 4th International Conference on Hybrid Learning*, pages 51–72, 2011.

[4] A. Fernandez, M. Regts, J. Tashiro, and M. Vargas Martin. Neural network prediction model for a digital media learning environment. *Fourth International Conference on Advances in Databases, Knowledge, and Data Applications*, pages 120–123, 2012.

[5] A. Fernandez, M. Regts, J. Tashiro, and M. Vargas Martin. Cluster analysis on user profile variables for a digital media learning environment. *The 11th International Conference on Information and Knowledge Engineering IKE '12*, page To be published, 2012.

[6] A. Fernandez, M. Regts, J. Tashiro, and M. Vargas Martin. Neural network prediction model for a digital media learning environment. *The 11th International Conference on Information and Knowledge Engineering IKE '12*, page To be published, 2012.

[7] J. Vreeken. Making pattern mining useful. *SIGKDD Explorations*, 12:75–76, 2010.

[8] W. Hamalainen and M. Vinni. Classifiers for educational data mining. *Handbook in Educational Data Mining*, 2010.

[9] R.O. Duda, P.E. Hart, and Stork D.G. *Pattern classification.* Wiley-Interscience Publication, 2000.

[10] R. Kabra. Performance prediction of engineering students using decision trees. *International Journal of Computer*, 36:8–12, 2011.

[11] S. Sembiring, M. Zarlis, D. Hartama, and E. Wani. Prediction of student academic performance by an application of data mining techniques. *2011 International Conference on Management and Artificial Intelligence*, 6:110–114, 2011.

[12] M. Ramaswami. A chaid based performance prediction model in educational data mining. *International Journal of Computer Science Issues*, 7:10–18, 2010.

[13] G.P. Suthan and S. Baboo. Hybrid chaid a key for mustas framework in educational data mining. *IJCSI International Journal of Computer Science Issues*, 8:356 – 360, 2011.

[14] A. Raychaudhuri, M. Debnath, S. Sen, and B.G. Majumder. Factors affecting student's academic performance: A case study in agartala municipal council area. *Bangladesh e-journal of sociology*, 7:34–41, 2010.

[15] M.K. Barbour and T.C. Reeves. The reality of virtual schools: A review of the literature. *Computers & Education*, 52:402–416, 2009.

[16] C. El Moucary, M. Khair, and W. Zakhem. Improving student's performance using data clustering and neural networks in foreign-language based higher education. *The Research Bulletin of Jordan ACM*, II:127–134, 2011.

[17] Z.S. Erdogan and M. Timor. A data mining application in a student database. *Journal of Aeronautics and Space Technologies*, 2:53–57, 2005.

[18] A. Shaeela, M. Tasleem, S. Raza Ahsan, and K.M. Inayat. Data mining model for higher education system. *European Journal of Scientific Research*, 43:24–29, 2010.

[19] A.M. M. Tair and A.M. El-Halees. Mining educational data to improve student's performance: A case study. *International Journal of Information and Communication Technology Research*, 2:140–146, 2012.

[20] A. El-Halees. Mining students data to analyze learning behavior: A case study. *The International Arab Conference on Information Technology*, pages 158–162, 2008.

[21] D.L. Driscoll, A. Appiah-Yeboah, S. Philip, and D.J. Rupert. Merging qualitative and quantitative data in mixed methods research: How to and why not. *Ecological and Environmental Anthropology (University of Georgia)*, 3: 18, 2007.

[22] M. Wojciechowski. ffnet 0.7.1. http://ffnet.sourceforge.net/, 2011.

[23] D. Varazzo. psycopg2. http://psycopg2.sourceforge.net, December 2011.

[24] A.L. Comfrey and H.B. Lee. *A First Course in Factor Analysis*. Lawrence Erlbaum Associates, 1992.

[25] E.J. Pedhazur. *Multiple Regression in Behavioral Research: Explanation and Prediction*. Harcourt Brace College Publishers, 1997.

[26] A.G. Sheppard. The sequence of factor analysis and cluster analysis: Differences in segmentation and dimensionality through the use of raw and factor scores. *Tourism Analysis 96*, pages 49–57, 1996.

[27] scipy.org. numpy 1.6.1. http://numpy.scipy.org, July 2011.

[28] A. Hagberg, D. Schult, and P. Swart. networkx 1.6. http://networkx.lanl.gov/, November 2011.

[29] J. Hunter. matplotlib 1.1.0. http://matplotlib.sourceforge.net, October 2011.

[30] P. Charbonneau and B. Knapp. Pikaia. http://hao.ucar.edu/modeling/pikaia/, 2012.

[31] J. Nocedal and J.S. Wright. *Numerical Optimization*. Springer, 2 edition, 2006.