Project Proposal

Title: Mining Student Data Using Machine Learning to Predict Academic Performance

Submitted by: Irving Manaog (imanaog3)

A description of the phenomenon to be investigated, including the research questions to be answered.

Predicting students' academic performance is an essential tasks for education authorities. This task is essential because it could help plan and focus on actions necessary to help struggling students improve their grades and increase their graduation rate. The main objective of my research project is to predict student's academic performance using data mining or machine learning algorithms. In particular, my research intends to answer the following questions:

- What are the impactful attributes that are necessary to predict students' academic performance?
- Which of the machine learning and data mining algorithms are suitable for modelling to predict students' performance?
- Based on the applicable machine learning and data mining algorithms, which one provides highest or better prediction accuracy?

A description of background literature in the area that leads to your research question.

There are several research papers and literatures that gave me ideas on what to pursue as my research topic. Listed below are some of these literatures:

Title: Mapping Student's Performance Based on Data Mining Approach (A Case Study)
Authors: Harwati, A.P. Alfiani, F. A Wulandari
Article Summary:
This research paper is about mapping students' demographic data (gender, origin, GPA, grade from certain courses) using K-mean clustering algorithm to reveal hidden patterns and classification. The datasets is composed of about 300 students. The authors emphasized that, to improve the students' academic performance, it is important that data is available for analysis because students have different levels of motivation, different understandings of teaching and learning, and environment conditions are different from one another. Several research to map students' performance using variety of variables and methods like Decision Trees and Naïve Bayes are not the only method to determine performance. Some studies showed that data mining techniques are not only useful to map students based on demographic variables, the understanding of the learning process, activity levels and other variables, but are also useful to predict the performance level of student based on the combination of data such as CGPA, test scores, time spent studying, and school attendance. **What's interesting with this research is that the researchers used two main variables, students' demographic and cognitive variables, to cluster students.**

Title: Preventing Student Dropout in Distance Learning Using Machine Learning Techniques
Authors: S.D. Kotsiantis, C.J. Pierrakeas, and P.E. Pintelas
Article Summary:
The paper is about using machine learning algorithms to predict and prevent student dropouts from distance learning. It attempted to identify the best algorithm to predict students' dropout. A prototype web based support tool, which can automatically recognize those students with high probability of

dropout, was developed by the authors. The paper's assumption was that dropout rates in university level distance education are higher than those students ground or conventional universities. The authors implied that dropout is usually caused by academic, professional, family, health, and personal reasons and varies depending on the education system adopted by the institution providing distance education. The authors used the Hellenic Open University's (HOU) 'Informatics' course dataset. HOU offers distance education at a university level. The 'Informatics' (INF) course was composed of 12 modules and lead to a Bachelor's degree. The INF dataset has a total of 354 students records for 'Introduction to Informatics' (INF10) module. Students of the INF10 module have to submit 4 written assignments, participate in 4 optional face-to-face consulting meetings with their tutors and complete a final exam. To predict students' dropout from distance learning, the authors used 6 types of machine learning: Decision Trees, Neural Networks, Naïve Bayes algorithm, Instance-based learning, Logistic Regression, and Support Vector Machines (SVM) algorithm. The experiments happened in 2 phases. The first phase or training phase, the algorithms were trained using the data collected from the academic year 2000-2001. This training set was divided into 5 consecutive steps. The first step included the demographic data and the resulting class (dropout or not). The second step included both the demographic data and the data from the first face-to-face meeting and the resulting class. The third step included data used for the second and the data from the first written assignment. The fourth step included data used from the third step and data from the second face-to-face meeting. The fifth step included the basic attributes from the dataset. **One of the interesting parts of this research is that the authors not only used the academic performance records but also non-academic data as part of determining the overall performance of the student.**

Title: Mining Student Data by Ensemble Classification and Clustering for Profiling and Prediction of Student Academic Performance

Authors: Ashwin Satyanarayana, Gayathri Ravichandran

Article Summary:

This research paper is about applying ensemble methods to students' dataset in the context of supervised learning to increase the accuracy and stability of student's performance prediction. The authors presented a hybrid methods based on ensemble classification and clustering that enables educators to predict students' academic performance and then group each student in a well-defined cluster for further advising. The authors used variants of ensemble classification such as Decision Trees, Naïve Bayes, and Random Forest to improve the quality of student data by removing noisy instances to improve predictive accuracy. The authors also used the bootstrap approach for averaging, which consist of k-means clustering algorithm to converge the training data and by averaging similar clusters to attain a single model. **To me, one of the most interesting parts of this research is the empirical comparison of the authors' methods with other ensemble methods on real-world education datasets**.

Title: An Approach of Improving Student's Academic Performance by Using K-means Clustering Algorithm and Decision Trees

Authors: M.H.I Shovon and M. Haque

Article Summary:

The authors of this paper used a hybrid methods based on Decision Trees and Data Clustering to predict students' GPA to help teachers take necessary actions to improve student academic performance. The authors emphasized that GPA is the most common factor used by the academic planners to evaluate progression in an academic environment. Several factors could impact student's ability to attain or maintain high GPA. These factors could be targeted by the teachers to develop strategies to improve student learning and improve their academic performance by way of monitoring the progress of their

performance. Utilizing decision trees and clustering algorithms in mining educational data could help discover key attributes for students' future performance. The authors used clustering algorithm to partition students into homogeneous group based on their common characteristics and abilities. Decision tree algorithm was used to explain variables like attendance ratio and grade ratio. The authors combined both decision trees and clustering algorithms to discover hidden information from the dataset which was used to determine factors affecting student's performance. **What I like about this research paper is that the authors combined decision tree and clustering algorithms to predict the student's GPA based on previous performance.**

A description of the research methodology that will be used, including the independent and dependent variables, internal and external validity, and the connection between these details and the research question.

The primary research method for my research project is literature review and correlational modelling.
This research project will first review existing literatures to determine attributes and variables that impact student's academic performance. Based on this understanding, a classification of method will be created to categorize these attributes and variables for the purpose of determining the impact of each to student's academic performance. Next, is to determine available machine learning and data mining algorithms that are applicable to these variables. Finally, once the impactful attributes and variables are identified together with the machine learning and data mining algorithms applicable to the dataset at hand, a model implementation of each algorithm will be applied to the dataset which includes data exploration analysis, separation of test and training datasets and the identification of cross validation techniques. The intention is to determine which attributes at each selected machine learning algorithm is more effective in determining in predicting student's performance.

A description of the data that will be needed or obtained, including spring-back plans if the data cannot be obtained.

For my research project, I intend to use publicly available students' performance datasets. There are not too many datasets of this kind, some datasets are either too "old" or lack attributes that would be enough to determine factors that are affecting students' academic performance. I found 2 datasets that might be able to use for my modelling tasks, I'm in the process of evaluating these 2 datasets. If these datasets are not able to serve the purpose of my research project, I will have to create my own datasets through surveys.

All project proposals you should cover the following information:

- ***A task list of the tasks that must be completed to execute and deliver your project. Make sure to include the required tasks as well, such as the intermediate milestones and final paper.***
- ***A calendar describing weekly milestones from the start of Week 5 through the end of Week 11.***
- ***Descriptions for your two intermediate milestones. You should consult the assignment page for those for a better idea of what to include.***

Project (tentative) Tasks/Outline and Milestones:

Deliverables	Schedule	Submission For
Part 1: • Identify and Select Datasets • Review Related Literatures • Refine Research Problems and Questions • Refine Research Methodology	Week 5 to Week 7	Weekly Status Check 1 and Weekly Status Check 2
Milestone 1: • Preview Research Methodology	Week 7	Intermediate Milestone 1
Part 2: • Data Exploration, Cleaning, and Preparation • Data Mining Model Selection • Data Model Implementation • Preliminary Results Discussions • Preliminary Conclusion and Recommendations	Week 8 and Week 9	Weekly Status Check 3 and Weekly Status Check 4
Milestone 2: • Data Presentation • Data Model Presentation • Preliminary Results Presentation • Preliminary Conclusion and Recommendation Presentation	Week 9	Intermediate Milestone 2
Part 3: • Finalize Result Summary and Discussions • Finalize Conclusion and Recommendations	Week 10	Weekly Status Check 5
Part 4: • Finalize Project Report and Presentation	Week 11	Final Project Project Paper Project Presentation