Proposal Title

Ramakrishna Danda
Summer 2021

Which Domain?

Student education data. Educational data analysis concerns of methods to discover hidden patterns from student education data. This analysis would help understand the influence of parameters on students' performance in exams.

What domain is this data going to come from? Please list 10 references (with a brief annotation) to use to make sense of what you're doing with these data.

1) Why is Educational Data Mining important in the research?

https://towardsdatascience.com/why-is-educational-data-mining-important-in-the-research-e78ed1a17908

2) Educational Data Mining (EDM)

https://www.cmu.edu/datalab/getting-started/what-is-edm.html

3) Mining Educational Data to Analyze Students Performance

https://thesai.org/Downloads/Volume2No6/Paper%209-Mining%20Educational%20Data%20to%20Analyze%20Students%20Performance.pdf

4) International Journal of Database Theory and Application. Mining education data.

http://article.nadiapub.com/IJDTA/vol9_no8/13.pdf

5) Educational Data Mining & Students' Performance Prediction

https://pdfs.semanticscholar.org/b280/216a1d63015afc6a3d1aac9595aeb2b7dd5a.pdf

6) Educational Data Mining: Student Performance Prediction in Academic

https://www.researchgate.net/profile/Mukesh-Kumar-234/publication/332369964_Educational_Data_Mining_Student_Performance_Prediction_in_Academic/links/5cb03346299bf120975f8dc1/Educational-Data-Mining-Student-Performance-Prediction-in-Academic.pdf

7) Educational Data mining for Prediction of Student Performance Using Clustering Algorithms

https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.567.8824&rep=rep1&type=pdf

8) Review on Prediction Algorithms in Educational Data Mining

https://acadpubl.eu/jsi/2018-118-7-9/articles/8/77.pdf

9) An Educational Data Mining Model for Predicting Student Performance in Programming Course

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.403.279&rep=rep1&type=pdf

10) Implementation Educational Data Mining For Analysis of Student Performance Prediction with Comparison of K-Nearest Neighbor Data Mining Method and Decision Tree C4.5

http://iocscience.org/ejournal/index.php/mantik/article/view/770/517

11) A Comparative Study to Predict Student's Performance Using Educational Data Mining Techniques

https://iopscience.iop.org/article/10.1088/1757-899X/215/1/012036/pdf

Which Data?

Data for the project will come from **uci.edu** & **ju.edu.jo** data that is accessible from reference 1 and 2 below; while supported by the rest of below data sets to be used as a supplement as project progress into Predictive methods.

1) Student Performance Data Set

https://archive.ics.uci.edu/ml/datasets/student+performance

2) Students' Academic Performance Dataset xAPI-Educational Mining Dataset

https://www.kaggle.com/aljarah/xAPI-Edu-Data

3) Exam scores for students at a public school

http://roycekimmons.com/tools/generated_data/exams

4) U.S. U.S. Education Datasets: Unification Project Education Datasets: Unification Project

https://www.kaggle.com/noriuk/us-education-datasets-unification-project

6) Student attendance in class & their performance in the examinations

https://www.kaggle.com/tanmoyie/grading-of-the-students-in-the-exam-ipe101-rawDatasets: Unification Project

7) Brazilian Law School Students Study, 1960 (ICPSR 7045)

https://www.icpsr.umich.edu/web/ICPSR/studies/7045/versions/V2/variables

8) Factors about students performance

https://www.kaggle.com/dariushatesteemo/factor-student-performance-affecting?select=Data.xlsx

9) Student Time Management Performance Dataset

https://www.kaggle.com/xiaowenlimarketing/international-student-time-management

11) Grading of the students in the exam OR

https://www.kaggle.com/tanmoyie/grading-of-the-students-in-the-exam-or

13) Student Growth

https://data.delaware.gov/Education/Student-Growth/kqmb-6xbs

Research Questions? Benefits? Why analyze these data?

How are you proposing to analyze this dataset? This is about your approach. Here, you'll be proposing your research questions as well as justifications for why you'd offer these data in this way.

Research question is to uncover week and strong relationship between datapoints of learner, discover, predict behaviors of students, contributing factors for the said outcome, and their academic achievements, related to student success with school environment and school management. The primary purpose of this analysis is also to understand how such analysis and techniques can help improving students' performance in higher education domain structure, understand effective support that works for students and essentially provide tools for institutions/researchers by providing empirical and reproducible evidence for finetuning the learning frameworks that eventually enables stronger learning systems.

What Method?

What methods will you be using? What will those methods provide in terms of analysis? How is this useful?

This analysis of educational Data to Predict Student's academic Performance is by the entire gamut of preparing the data, EDA with visualizations, feature selection/engineering with dimensionality reduction, model selection & evaluation of each model performance within the Ensemble Methods.

Classification: Classifier-training algorithm uses pre-classified examples to determine the set of parameters required for proper discrimination. In this use case of student data, plan is to classify the students into different categories of performers(high), average performers(middle) & underperformers(low).

Clustering: Using clustering techniques we can further identify data space and could discover overall distribution pattern and correlations among data attributes. This enables the analysis of student data into each bucketing block for similar patterns found with in the space.

Regression: Relationship between one or more independent variables and dependent variables and if feature engineering can be employed from maze of features within student data.

Association Rules: Methods to find frequent item set among large amounts of student information systems.

Decision Trees: Generally, the most important method based on the tree-shaped structures that represent sets of decisions for each student behaviors and their patterns.

Nearest Neighbor method etc., classifies each record in a dataset based on a combination of the classes of the k record(s) most like it in a historical dataset.

Ensemble Methods: Bagging, Boosting and Random Forest: Ensemble evaluation of features that will have impact on performance of the students, and to improve the performance of student's prediction model. For which Boosting dependent method, the output of a learner is used in the creation of the next learner. While in independent method of Bagging methods resample the original data into samples of data, then each sample will be trained by a different classifier of Decision Trees as explained above as an example and Naïve Bayesian methods.

Will be employing classification quality measures of accuracy, precision, recall and f-measure with confusion matrix helping the above measure calculations.

Potential Issues?

What challenges do you anticipate having? What could cause this project to go off schedule?

Considerable data volume collections are a challenge to get within each set of time frame that would be statistically significant; that could also be a factor such models are not portable from one set of student data to others. Retraining & retrofitting models are essentially needed to refactor new data set because of overfitting to the training data. Because every student situation is different, and no students has the exact same issues of why they are falling behind to make deterministic conclusions on grades.

Concluding Remarks

Tie it all together. Think of this section as your final report's abstract.

Educational institutions in collaboration with institutional research plans to create profile of each student to understand if student would complete the program on time & with acceptable grades, and if they would drop off program midterms/year. And or students looking for other opportunities in a different institution by transferring. Creating such student profile ahead of time using predictive analytics by Educational Orgs help both students in custom approach in resolving issues & difficulties in completing the program and intervein if such students would be counseled for a better outcome(grades) of their educational goals. At the same time help their own institutions on loosing monitory value in foregone revenue otherwise would have gained if student does not drop off midway because of non-satisfactory grades obtained or planning to transfer to other institutions, because such Students did not learn to fit into the society of the institution removed themselves from the institution or many other factors that could be mitigates if intervened early. This project is to understand and analyze data sets of such use cases and provide usable models in identifying students to be intervened and help both students and Institutions.