Assessing Students’ Performance Prediction Models Utilizing Machine Learning

Student's Name: Karl Shen

Professor’s Name: Zhichun Li

Course: HUDK 4050

Date:12/15/2021

Introduction

Researchers typically utilize data mining techniques as the best approach for revealing some hidden patterns in huge, unintelligible datasets. The approaches are utilized in diverse fields such as healthcare, communications, education, accounting, pharmacy, and human disease among others to reveal hidden trends in large datasets. Data mining techniques are largely used in the field of educational data mining (EDM) to extract large educational datasets to obtain meaningful data and information, which is often analyzed, evaluated, and interpreted to make an informed decision during decision making-process (Alturki, Hulpuș, & Stuckenschmidt, 2020). Innumerable variables often impact the choice to predict and track learners’ success in different colleges or schools. Fortunately, most decision-makers in schools utilize these variables to make informed decisions or strategies that improve learner educational performance. The rapid use of data mining techniques in the education environment has led to the establishment of Educational Data Mining (EDM) as a core approach for making an informed decision (Alturki, Hulpuș, & Stuckenschmidt, 2020). EDM is based on the study of scholarly data generated through an educational setting. The EDM typically uses regression, classification, clustering, time series analysis, neural networks, and connection rule mining to obtain meaningful information for a hidden dataset. The EDM help in the early detection of hidden trends in the educational sector (Khasanah, 2017). For many years, innumerable studies have been conducted by scientists and instructors to examine the factors that impact learners' performance either positively or negatively. In most cases, it would be challenging to measure learners' academic performance since learners' performance is impacted by innumerable factors related to psychological, socioeconomic, and environmental factors (Khasanah, 2017). Most research holds a special role in evaluating learners’ performance, but learners’ exam results are massively impacted by three major factors, which include academic environment, socioeconomic and demographics. Demographic variables students' personal attributes such as gender, age, weight, disabilities, eating habits, sibling structure, race, family system, and residential area. The academic climates are those variables that have a direct impact on a learner's academic performance, which include the type of school, location of that school, method of training, continuous evaluation assessment, learning environment, final results at the end of the semester, and the type of community, including extra-curricular activities in that institution (Khasanah, 2017). The socio-economic factors are those variables related to family income, parental occupation, social status, and the education level of the parents.

Importantly, early prediction of learners' success improves learners' retention and the assessment techniques utilized by the learners. It also supports education officials and instructors by offering more knowledge regarding learners' abilities to master new ideas and the strategy to assist learners who are performing better in class (Shakeel & Butt, 2015). Evaluating students' learning performance is substantial for an educational system, as it helps provide directives on how to customize or design a learning environment tailored to every learner's specific needs or even to guide students to pursue technical courses or subjects in school. This makes it critical to monitor learners' performance to acknowledge some potential retardation and to proactively intervene to improve their performance (Shakeel & Butt, 2015). Therefore, this report aims at evaluating educational data generated from learners' transcripts or class performance and predicting their success for appropriate measures to be taken to meet students' expectations, including creating an integrated framework that permits machine learning to be utilized to predict learners’ achievement and performance. To evaluate educational data to predict learners’ performance and create a model, we aim to explore peer reviews on crafting models to predict learners’ performance in schools. This approach is aimed to embark the study's goals, which include;

1. To define present prediction approaches, including tools utilized to predict learners’ achievements.
2. To assess and classify the researchers who used these learning models to assess student performance.
3. To assess and classify the kind of variable utilized for the predictive procedures.

For that matter, predicting learners' performance is essential for instructors, students, and school administration officials. In nutshell, ‘‘knowledge exploration through data mining " will assist instructors to better design their classes, develop their training methods, recognize learning disabilities whereas learners will be offered a preliminary assessment that will enhance their accomplishment in school.

**Related Work**

It is substantial to assess educational settings as the prediction of learners' academic performance is a very critical factor in improving the academic learning standard, helping learners as they study different subjects, and offering instructors a diverse alternative while training students. Several studies have been conducted on this topic to predict learners’ performance. According to Ghorbani & Ghousi (2020), effective models are needed to perform better with fewer classes and nominal features. This was after Ghorbani & Ghousi (2020) consolidated and compared diverse resampling approaches, including SMOTE, SMOTE-Tomek, SVM-SMOTE, Random Over Sampler, Borderline SMOTE, and SMOTE-ENN to handle the disparity in data while predicting learners’ success using two different datasets. The author also investigated the differences between binary and multiclass classification, including feature structure. In addition, Ghorbani & Ghousi (2020) in the machine learning classifiers utilized Decision Tree, K-Nearest-Neighbor, Random Forest, XG-boost, Support Vector Machine (Radial Basis Function), Artificial Neural Network, Nave Bayes, and Logistic Regression. The Shuffle 5-fold cross-validation and Random hold-out approaches were utilized as model validation techniques. Ghorbani & Ghousi’s (2020) evaluation criteria revealed that a good model would have fewer nominal and classes features, yet this work did not provide an effective alternative to the use of EDM technique to solve data issues in the educational sector.

Almasri, Alkhawaldeh, & Çelebi (2020) insinuated a coherent structure aimed to create a new supervised cluster-based (CB) classifier model, which was to organize learners’ archival records into a sequence of homogeneous clusters for predicting learners’ performance. Based on Almasri, Alkhawaldeh, & Çelebi's (2020) experimental findings, the cluster-based model obtained a more accurate performance outcome that deploys an effective feature selection strategy to choose appropriate features for decision making. Through the use of a diverse relevant feature, the authors' model obtained ninety-six percent accuracy output. On the same line, Popescu, and Leon (2018) suggested another technique that encompasses using an advanced regression algorithm that was used to predict learners' grades based on learners' behaviors on microblogging tools, forums, and wiki. This method was called "Large Margin Nearest Neighbor Regression" (LMNNR). The finding by Popescu and Leon (2018) was outperforming and excellent compared to other regression algorithms approaches used by other researchers. The authors obtained excellent correlation coefficients as eighty-five percent of predictions were within one or more points of the actual students’ grade, which outperformed conventional regression algorithms. Another study conducted by Chen et al. (2018), which was centered on short courses and giving one outcome assigned by educators, revealed that there is a lack of performance data, including limited learners’ enrollments. The author exhibits that student activity recorded while engaging with learning material and collaboration with teachers and peers in Social Learning Networks (SLN) is an important factor that can be used to predict student learning performance. They posit that the SLN features emerge effective predictive set that can be utilized for the long-term as the network matures. In addition, Lau et al. (2019) also suggested a method that blends conventional statistical analysis with neural networks to predict learners' performance. Conventional statistical evaluations are utilized to assess and identify variables that can easily impact learners' outcomes whereas the neural network is exhibited by two hidden neuron layers, eleven input variables, and one output layer during the study. The neural network structure's performance is tested using regression, uncertainty matrix, error histogram, areas under the receiver operating attributes curve, and error performance. The finding by Lau et al. (2019) revealed that the neural network model is more effective in predicting the accuracy of analysis, yet has some limitations.

Moreover, Xu, Moon, and Van-Der- Schaar (2017) created a novel machine learning technique for predicting learners' academic performance in universities. The approach had two main structures for making predictions. One is called a bilayered structure, which included a cascade of ensemble predictors and multiple base predictors while the other one was a data-driven approach based on probabilistic matrix factorization and latent factor models to discover the relevancy of the course, yet this method had some limitations. According to Alturki, Hulpuș, & Stuckenschmidt (2020), the features often utilized to predict learners' academic accomplishment can be grouped into three classes. This includes the use of demographics, post-enrollment, and pre-enrollment attributes. Although most researches rely on utilizing demographical features to predict student academic accomplishment, yet the level to which such variables are useful is not clear. According to Alturki, Hulpuș, & Stuckenschmidt (2020), educators and researchers heavily use gender and age to predict learners' performance. However, some scientists posit that gender and age do not considerably influence the overall prediction of learners' accomplishments.

Based on the performance of different data mining techniques, there is no single approach that works effectively in all academic settings. For that matter, scientists and educators frequently explore two or more data mining approaches to reveal concealed trends in enormous indiscernible datasets, which derives high-performance accuracy (Alturki, Hulpuș, & Stuckenschmidt, 2020). Obsie and Adem (2018) created a system to help universities and colleges predict their learners’ academic performance after graduation from the institution. The authors utilized learners’ scores for non-core and core courses throughout the semesters. They revealed that Linear Regression and Support Vector Regression were more effective than the Neural Networks approach. On the same line, Khasanah (2017) carried out Feature Selection to identify and determine the attributes that had a higher impact on learners’ performance using learners' demographics, pre-and post-enrollment information. The authors revealed that Decision Trees was outperformed by Bayesian Network and that learners’ learners’ GPA and attendance in semester one are the two primary features that can be used to predict students’ academic accomplishments. In addition, Shakeel and Butt (2015) utilized pre-and post-enrollment information, including learners’ demographics to predict learners who may drop out of school. The authors posit that Naïve Bayes is the best and effective approach, and Random Forest being the second followed by J48, and Logistic Regression being the fourth. From these peer-reviewed, one of the research gaps identified is that none of these researchers examined the impact of the orientation year on forecasting learners’ scholarly performance. McMullen (2014) addresses the significance of the orientation year on learners’ success, as the authors offered statistical evidence that learners in an educational setting in Saudi Arabia hold that they complete their high school programs without essential knowledge for joining their academic majors in the future.

**Proposed methodology**

Prediction of learners’ achievement and performance is a major issue that can be assessed using EDM. This work addresses the significance of an unspecified variable defining learners in terms of their performance as either pass or fail in their exams. This report’s peer review covers predicting learners’ attrition, progress, and errors. Everyone in the education setting needs an early warning framework to predict learning outcomes and to adjust approaches to improve learners' performance. On classification and regression issues, various learning algorithms give different outcomes because the learning outcome is derived from varying algorithms. Multiple learning algorithms can enhance final prediction output, as opposed to using one algorithm. In the same line, Ensemble learning is aimed to enhance predictive accuracy through blending projections from innumerable learning algorithms. Factually, researchers have widely utilized ensemble learning in machine learning to enhance efficiency through gathering each algorithm on several regression and classification tasks. Ensemble strategies such as averaging, voting, bagging, piling, and boosting have been used in machine learning. Many models are utilized by researchers to resolve binary classification issues, such as clustering algorithm, Bayesian process algorithm, the regression algorithm, kernel-based algorithm, decision tree algorithm, and many more. Stacking can be utilized in a combination of other classifiers to predict learners’ performance.

#### Data collection

The data used were from empirical reviews. Actually, for the effective creation of the proposed work prediction model of learners' performance, it was critical to examine empirical reviews or data collected and used by other scholars about learners' archival academic details and label them before these large data could be utilized to establish prediction model by the supervised classification algorithm.

**Source of data**

The data sources can be grouped into three main classes, which include the digital systems utilized by schools such as course learning management system (LMS) and Student information management system (SIS), manual collection through surveys and questionnaires, public dataset repository of Machine Learning like Kaggle.

**Step 1: Pre-processing and Balancing**

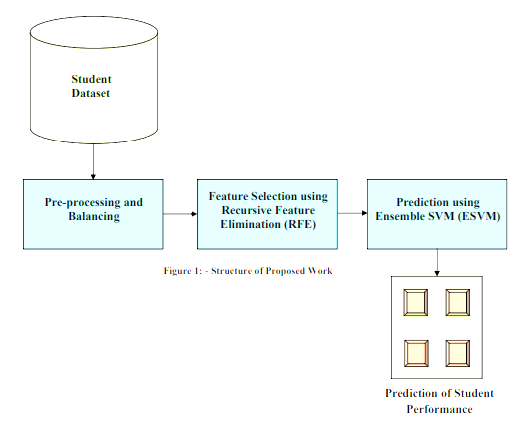
This is the most important phase in Machine Learning since the data quality and the sensitive information derived from it may directly influence the ability of the model to learn; therefore, data pre-processing before feeding it into the model is substantial. The process aims to change the raw data into an appropriate format that can be easily and effectively used by the mining algorithms. The following tasks are performed; data integration, data cleaning, and discretization. Data balancing is then performed after data pre-processing to resolve any class imbalance issue in the data. The insinuated work presumes that the Cross-Validation technique is utilized to measure the test error correlated with the framework so as to appraise its accuracy.

**Step 2: Feature Selection using Recursive Feature Elimination (RFE)**

For classification purposes, innumerable attributes in the learners’ output dataset might be inappropriate. Therefore, the main purpose of feature selection is to pick an appropriate subset of features, which can be presented in the student’s input data to eliminate irrelevant data and to reduce the disparity in the feature space. RFE ranks various features using variable grading approaches, with the highest graded features being chosen and included in the learning algorithm. The insinuated work is a starting point.

**Step 3: Ensemble Learning Paradigm using Ensemble SVM (ESVM)**

Ensemble techniques are approaches for building multiple models and integrating those models to obtain an effective performance or more accurate results. Ensemble models typically produce a more effective and productive framework as compared to a single model. The final stage encompasses deploying the ensemble support vector machine (ESVM) classification approach that was established by blending innumerable structures of SVM classifiers to obtain a high classification accuracy and generalization efficiency. The insinuated SVM ensemble learning model consists of five distinctive kernel functions and two distinctive SVM classifier frameworks. This is because most schools currently need an effective learners’ performance prediction model. This report also proposed a model for forecasting learners’ accomplishments based on the supervised learning technique ensemble SVM.



**Performance evaluation**

It is very substantial to evaluate classifier efficiency to compare and select the most appropriate approach. Factually, innumerable approaches can be used to measure and evaluate machine learning algorithms' outputs. We, therefore, utilized prediction sensitivity, accuracy, F1-score, and precision, including a statistical evaluation approach for a more reliable and efficient analysis and comparison. Evaluating and analyzing the classifiers’ output is a critical approach in creating an effective model. Although the assessment techniques are easy and simple to utilize, the outcome may be misleading in one way or the other. The desire to obtain the right model or framework based on their effectiveness is a major problem as there are several evaluation methods. The five frequently utilized approaches for evaluating classification consistency include correctly classified instances (CCI), Incorrectly Classified Instances (ICI), Precision, Recall, and F-Measure.

**Conclusion**

Studies revealed that students' performance is directly impacted by the learning environment, including demographic and socioeconomic status. This helps predict student performance to improve learning outcomes. The prediction of learners’ achievement and performance can be conducted using innumerable educational data mining approaches to offer accurate and precise educational information suitable for making an informed decision. At the same time, there are multiple prediction models that can be utilized, yet they have some limitations. Therefore, this study has proposed an effective model to predict learners’ success and performance. The study has aimed to provide a model for enhancing education quality. The data mining techniques often offer useful insights that can be utilized to predict the final students’ academic performance. The proposed prediction model can assist educators, administrative officials, and students in many ways, including early detection of performance issues and enhancing the learning environment for disabled students.

References

Almasri, A., Alkhawaldeh, R. S., & Çelebi, E. (2020). Clustering-Based EMT Model for Predicting Student Performance. *Arabian Journal for Science and Engineering*, *45*, 10067-10078.

Alturki, S., Hulpuș, I., & Stuckenschmidt, H. (2020). Predicting academic outcomes: A survey from 2007 till 2018. *Technology, Knowledge and Learning*, 1-33.

Chen, W., Brinton, C. G., Cao, D., Mason-Singh, A., Lu, C., & Chiang, M. (2018). Early detection prediction of learning outcomes in online short-courses via learning behaviors. *IEEE Transactions on Learning Technologies*, *12*(1), 44-58.

Ghorbani, R., & Ghousi, R. (2020). Comparing different resampling methods in predicting Students’ performance using machine learning techniques. *IEEE Access*, *8*, 67899-67911.

Khasanah, A. U. (2017, June). A comparative study to predict student’s performance using educational data mining techniques. In *IOP Conference Series: Materials Science and Engineering* (Vol. 215, No. 1, p. 012036). IOP Publishing.

Lau, E. T., Sun, L., & Yang, Q. (2019). Modeling, prediction, and classification of student academic performance using artificial neural networks. *SN Applied Sciences*, *1*(9), 1-10.

Obsie, E. Y., & Adem, S. A. (2018). Prediction of Student Academic Performance using Neural Network, Linear Regression and Support Vector Regression: A Case Study. *International Journal of Computer Applications*, *180*(40), 39-47.

Popescu, E., & Leon, F. (2018). Predicting academic performance based on learner traces in a social learning environment. *IEEE Access*, *6*, 72774-72785.

Shakeel, K., & Butt, N. (2015, May). Educational data mining to reduce the student dropout rate by using classification. *Proceedings of the 253rd OMICS International Conference on Big Data Analysis & Data Mining,* Lexington, Kentucky.

Xu, J., Moon, K. H., & Van Der Schaar, M. (2017). A machine learning approach for tracking and predicting student performance in degree programs. *IEEE Journal of Selected Topics in Signal Processing*, *11*(5), 742-753.