Student GPA Prediction and Course Recommendation Using Ml Algorithms

[1] Thaduturi Karthik, [2] A. Naga Durga Bhavani, [3]Dr. Bommana Indu
Student in Dept. Of Master of Computer Applications, Miracle Educational Society Group of Institutions
, Assistant Professor, Miracle Educational Society Group of Institutions
, Associate Professor, Miracle Educational Society Group of Institutions

ABSTRACT
Predicting student academic performance is vital for early intervention and ensuring timely graduation, thus predictive student academic performance evaluation is essential. This project presents a machine learning based framework that predicts degree completion is achieved by evaluation of historical academic data over a period of time. In contrast to traditional models which concentrate on individual course outcomes, this one features bilayered prediction structure with SVM, Random Forest, Logistic Regression, and Ensemble based Progressive Prediction (EPP) model. Using UCLA undergraduate data, the system was found to predict outcomes with greater accuracy, especially when EPP was utilized. The model also provides personalized academic advising by selecting prospective courses depending on the expected GPA. This advancement assists institutions in detecting at-risk learners prompting them to develop plans for academic enhancement thereby increasing the level of on-time graduation and subsequently reducing the financial strain of student debt.

Keywords: GPA, Random Forest, Predictive Modeling, Machine Learning, Ensemble Progressive Prediction

INTRODUCTION:

The expense of attaining higher qualifications has constantly escalated thus necessitating timely graduation. Further, delayed attainment encumbers the student with increasing exposure to debt. In any scenario, a predictive system that is not static and evolves with the progress of students' academics and proactively identifies potential delays serves as a solution. Such tailored monitoring aids in setting milestones that are enrolled with greater predictive accuracy, enabling one to set targets and subsequently evaluate performance across different courses and environments irrespective of student profiles or academic tracks. This project intends to develop a machine learning model that predicts academic results over the duration of a degree based on the student's diversity, course relevance, and progression patterns. This project aims to use historical data and ensemble modeling to make real-time, self-correcting updates to the model in order to improve accuracy. Through the prediction of relevant at-risk students, the model assists educational institutions in defining future course selection policies, contributing to improved educational outcomes and effective resource management.

GAP IDENTIFIED BASED ON LITERATURE SURVEY:

Although educational data mining (EDM) has significantly advanced, most attempts analyze data from a single course, or rely on very specific learning environments, like MOOCs or ITSs. The gap in research is related to long-term academic outcome predictions, like graduation success, across a degree program. Existing models lack flexibility in accommodating students' diverse backgrounds, such as major, career goals, combinations of electives, and even chosen courses. A lot of these models also disregard

the temporal dimension of student performance, where prior courses tend to be treated equally.

Moreover, a significant gap is the absence of dynamic personalized recommendation engines that evolve with student progress within the degree program. Literature emphasizes the need for models that can filter relevant academic indicators, continuously monitor performance, and offer real-time insights. Most frameworks do not integrate ensemble learning or layered structures capable of handling heterogeneous educational data.

This project addresses these key gaps by:

- Incorporating a bilayered structure for progressive performance prediction.
- Using ensemble learning for increased robustness and accuracy.
- Identifying key academic indicators dynamically, rather than uniformly weighing all prior courses.
- Providing course recommendations tailored to predicted performance trends.

PROBLEM STATEMENT:

Current student performance prediction models are ineffective in tracking long-term academic success due to static approaches, uniform course weighting, and lack of adaptability to individual trajectories.

Key Challenges:

1. Variability in student backgrounds and academic paths.
2. Limited course-level predictive correlation insights.
3. Difficulty in dynamically updating predictions based on ongoing progress.
4. Overfitting due to excessive feature noise from irrelevant courses.
5. Lack of real-time, actionable interventions and course recommendations.

PROPOSED METHOD:

This project implements a machine learning framework that tracks student data over time to analyze and recommend appropriate upcoming courses based on predicted future performance. It entails data preprocessing on the UCLA dataset, training various ML models: SVM, Random Forest, and Logistic Regression, and applying an Ensemble-based Progressive Prediction (EPP) model on the integrated predictions. The system supports real-time updates, permitting constant recalibration of forecasts as grade-level milestones are achieved. Through consideration of temporal context and student context evolution, the bilayered structure refines accuracy by accounting for course relevance. Predictions offer guidance to faculty and administrators on timely interventions for students at risk of dropping out, while the model withholds highly relevant suggestions on courses, thus elevating student success and lowering dropout rates.

ARCHITECTURE:

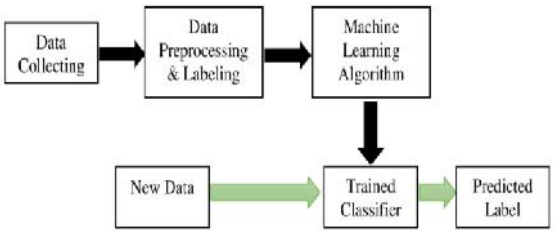

Fig-1 Architecture

DATASET:

The dataset used is sourced from UCLA and provides access to the academic records of postgraduate students. It has numerical and categorical data points including GPA, self-learning capabilities, awarded certifications, talent tests, preferred courses, career interests, and other extracurricular activities. The dataset undergoes preprocessing with the encoding of categorical features alongside normalization of numerical ones. It is then split into training and testing sets using an 80:20 ratio. This structured dataset enables

the training of multiple ML algorithms and supports performance prediction with real-time updates. The diverse attributes in the dataset provide a holistic view of each student's academic journey, enabling more accurate forecasting.

METHODOLOGY:

- **Data Collection & Loading:**
  - The UCLA undergraduate dataset is collected and uploaded through the system interface.
  - The dataset includes 21 features related to student skills, preferences, and academic performance.
- **Data Preprocessing:**
  - Categorical data is encoded using LabelEncoder for consistent machine processing.
  - Normalization is applied to numerical values using the Normalizer function to bring all features to a similar scale.
  - The dataset is split into feature variables (X) and the target variable (Y).
- **Dataset Splitting:**
  - The dataset is divided into training (80%) and testing (20%) sets.
  - Data is shuffled randomly to avoid bias during model training.
- **Model Training:**
  - Four algorithms are implemented and evaluated:
    - **Support Vector Machine (SVM):** Trains on student data and outputs classification-based GPA predictions.
    - **Random Forest:** Utilizes an ensemble of decision trees for robust classification and prediction.
    - **Logistic Regression:** Models probability of academic outcomes based on feature weights.
    - **Ensemble-based Progressive Prediction (EPP):** An advanced bagging technique using Random Forest as a base estimator to combine multiple predictions progressively.
- **Performance Evaluation:**
  - Models are evaluated using Mean Squared Error (MSE) and Accuracy.
  - Lower MSE values indicate better performance; EPP showed superior accuracy and minimal error.
- **Prediction Mechanism:**
  - New student data can be uploaded for performance prediction.
  - Based on the trained EPP model, the system forecasts future GPA and classifies outcomes as "High" or "Low."
- **Course Recommendation:**
  - Depending on predicted GPA, the system recommends optimal courses from a predefined list aligned with the student's skills and interests.
- **Visualization:**
  - A bar chart visualizes the Mean Square Errors of all models, aiding in comparative analysis.
  - Graphs offer insights into the accuracy trends of each ML approach.

RESULTS:

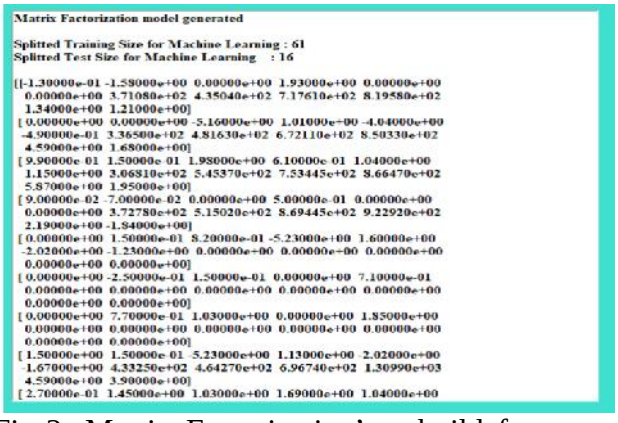

Fig-2: Matrix Factorization' to build feature vector from dataset

Fig 3: SVM MSE is 56%.

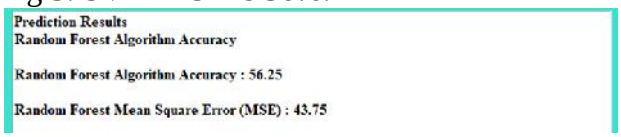

Fig: 4 Random forest got 43% MSE

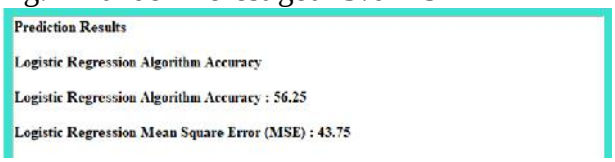

Fig: 5 Logistic regression got 43% MSE

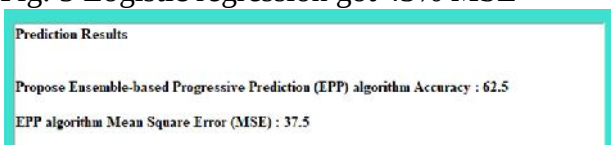

Fig: 6 EPP propose algorithm got 37% MSE

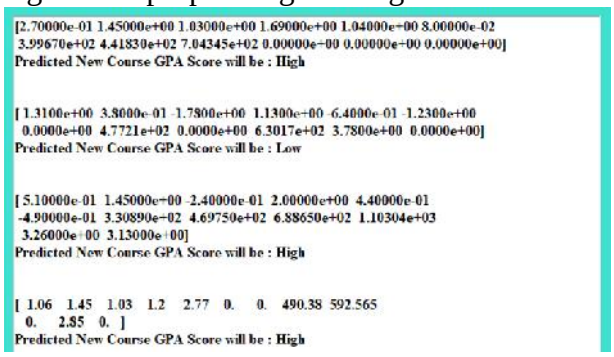

Fig: 7Student marks of ongoing subjects and this marks are converted to matrix factorization and then applied on EPP train model to predict GPA as LOW or HIGH.

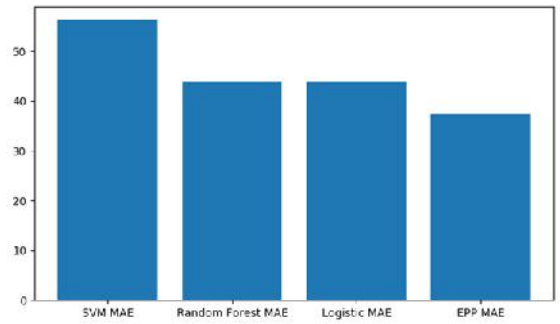

Fig: 8 : comparison algorithm.

X-axis represents algorithm name and y-axis represents MSE (mean square error). From above graph we can see propose algorithm got less MSE error and has high accuracy compare to other algorithms.

CONCLUSION

Colleges can guarantee higher graduation rates by using an efficient machine learning technology that automatically predicts the academic performance of learners as a bilayered model and ensemble-based EPP that predicts GPA results and suggested the correct courses to be taken. These approaches solve model limitations like poor adaptability and statically preset predictions. EPP has demonstrated greater accuracy with the least mean square error as compared to other models. This framework provides educators with valuable intelligence regarding the learner's academic progression, enabling them to tailor engagement strategies and pathways for the learner, facilitating the institution in achieving elevated standards of academic performance, and aiding the learner in succeeding.

REFERENCES:

[1] The White House, “Making college affordable,” https:// www.whitehouse.gov/issues/education/higher -education/ making-college-affordable, 2016.

[2] Complete College America, “Four-year myth: Making college more affordable,” http://completecollege.org/wp-content/uploads/2014/ 11/4-Year-Myth.pdf, 2014.

[3] H. Cen, K. Koedinger, and B. Junker, “Learning factors analysis–a general method for cognitive model evaluation and improvement,” in International Conference on Intelligent Tutoring Systems. Springer, 2006, pp. 164–175.

[4] M. Feng, N. Heffernan, and K. Koedinger, “Addressing the assessment challenge with an online system that tutors as it assesses,” User Modeling and User-Adapted Interaction, vol. 19, no. 3, pp. 243–266, 2009.

[5] H.-F. Yu, H.-Y. Lo, H.-P. Hsieh, J.-K. Lou, T. G. McKenzie, J.-W. Chou, P.-H.

Chung, C.-H. Ho, C.-F. Chang, Y.-H. Wei et al., "Feature engineering and classifier ensemble for kdd cup 2010," in Proceedings of the KDD Cup 2010 Workshop, 2010, pp. 1–16.

[6] Z. A. Pardos and N. T. Heffernan, "Using hmms and bagged decision trees to leverage rich features of user and skill from an intelligent tutoring system dataset," Journal of Machine Learning Research W & CP, 2010.

[7] Y. Meier, J. Xu, O. Atan, and M. van der Schaar, "Personalized grade prediction: A data mining approach," in Data Mining (ICDM), 2015 IEEE International Conference on. IEEE, 2015, pp. 907–912.

[8] C. G. Brinton and M. Chiang, "Mooc performance prediction via clickstream data and social learning networks," in 2015 IEEE Conference on Computer Communications (INFOCOM). IEEE, 2015, pp. 2299– 2307.

[9] KDD Cup, "Educational data minding challenge," https://pslcdatashop. web.cmu.edu/KDDCup/, 2010.

[10] Y. Jiang, R. S. Baker, L. Paquette, M. San Pedro, and N. T. Heffernan, "Learning, moment-by-moment and over the long term," in International Conference on Artificial Intelligence in Education. Springer, 2015, pp. 654–657.

[11] C. Marquez-Vera, C. Romero, and S. Ventura, "Predicting school failure using data mining," in Educational Data Mining 2011, 2010. [12] Y.-h. Wang and H.-C. Liao, "Data mining for adaptive learning in a tesl-based e-learning system," Expert Systems with Applications, vol. 38, no. 6, pp. 6480–6485, 2011.

[13] N. Thai-Nghe, L. Drumond, T. Horvath, L. Schmidt-Thieme ´ et al., "Multi-relational factorization models for predicting student performance," in Proc. of the KDD Workshop on Knowledge Discovery in Educational Data. Citeseer, 2011.

[14] A. Toscher and M. Jahrer, "Collaborative filtering applied to educational data mining," KDD cup, 2010.

[15] R. Bekele and W. Menzel, "A bayesian approach to predict performance of a student (bapps): A case with ethiopian students," algorithms, vol. 22, no. 23, p. 24, 2005.

[16] N. Thai-Nghe, T. Horvath, and L. Schmidt-Thieme, "Factorization mod- ´ els for forecasting student performance," in Educational Data Mining 2011, 2010.

[17] Y. Meier, J. Xu, O. Atan, and M. van der Schaar, "Predicting grades," IEEE Transactions on Signal Processing, vol. 64, no. 4, pp. 959–972, Feb 2016.

[18] N. Cesa-Bianchi and G. Lugosi, Prediction, learning, and games. Cambridge university press, 2006.

[19] Y. Koren, R. Bell, C. Volinsky et al., "Matrix factorization techniques for recommender systems," Computer, vol. 42, no. 8, pp. 30–37, 2009.

[20] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorization," in NIPS, vol. 20, 2011, pp. 1–8.