Socio-Educational Early Warning System for Effective Student Retention

[1]Mr. N. Sendhil Kumar, [2]S. Chandra Kiran Reddy, [3]M. Shyam Sundar, [4]B. Bhanu Prakash
[1]*.HOD & Professor/MCA, Sri Venkateswara College of Engineering and Technology (Autonomous) Chittoor, Andhra Pradesh-517217
[2,3,4] MCA Students, Sri Venkateswara College of Engineering and Technology (Autonomous) Chittoor, Andhra Pradesh-517217

Abstract*— The development of data analysis techniques and intelligent systems has had a considerable impact on education, and has seen the emergence of the field of educational data mining (EDM). The Early Warning System (EWS) has been of great use in predicting at-risk students or analyzing learners' performance. Our project concerns the development of an early warning system that takes into account a number of socio-cultural, structural and educational factors that have a direct impact on a student's decision to drop out of school. We have worked on an original database dedicated to this issue, which reflects our approach of seeking exhaustiveness and precision in the choice of dropout indicators. The model we built performed very well, particularly with the K-Nearest Neighbors (KNN) algorithm, with an accuracy rate of over 99.5% for the training set and over 99.3% for the test set. The results are visualized using a Django application we developed for this purpose, and we show how this can be useful for educational planning.

Keywords: ***Student, Early Warning System, KNN, Educational Data Mining Django, Intelligence.***

I. INTRODUCTION

The evolutionary path of IT practice has taken a new form with the advent of intelligent systems, especially predictive and recommendation systems. And with the explosion of data and the entry into the era of Big Data, these systems have found more opportunities to flourish and achieve the most remarkable results. Early warning systems (EWS) are one of the most famous types of intelligent systems, and have benefited from the considerable leap forward in computing methods and technologies used, as well as the development of hardware infrastructures. The EWS is a predictive system that aims to support decision-making by giving a proactive view of the future situation by analyzed data. EWS are used in almost all fields, and their role lies in detecting anomalies in real systems and warning decision-makers of the seriousness of situations, so that they can anticipate their intervention to remedy The associate editor coordinating the review of this manuscript and approving it for publication was Chang Choi. 2260 problems posed, or at least limit the negative effects and consequences. An EWS can be defined by a number of active key words Collect, Analyze, Detect, Prevent, Alert, Notify. Each word indicates one of the key stages of an EWS, hence its action model, which consists of a set of layers or steps. The first step involves the continuous monitoring of relevant indicators and the collection of data in real or near real time. The next step involves the analysis and processing of the data collected. This process involves examining the data to identify early indicators, patterns or deviations from the norm. Various techniques such as advanced algorithms, statistical models or artificial intelligence can be used to identify potential anomalies during this analysis. The third step is 'Alert and Notify', once the system detects an irregularity or potential hazard, it quickly triggers an alert to inform the parties concerned. Then, in the fourth step 'Risk assessment', professionals and supervisors evaluate the reliability and severity of the warning. They examine existing data and information to understand the characteristics of the risk, its potential consequences and possible actions to minimize its VOLUME 12, 2024 IEEE Transaction Access on Machine Learning, Volume:12, Issue Date:5. January.2024 impact. The communication and dissemination process plays an essential role in risk management and response. Once a risk has been assessed and verified, it is crucial to share the relevant information with all parties concerned, including stakeholders, decision-makers and the general public. The final layer of the process involves response and action. And finally, once the early warning system provides the necessary information, appropriate measures are taken to reduce risks, prevent crises or minimize adverse

effects. All these steps form an iterative process, as we are always aiming for perfection of the system, given that EWSs are used in highly critical areas and that the effect of their Outputs can avert disasters in some cases, whether in the near future or in the long term. That's why we're constantly striving for perfection.

2. RELATED WORK

This paper presents an early warning system that uses data mining techniques to predict student dropout in higher education. The system analyzes academic performance, attendance, and demographic factors to identify students at risk of dropping out. The authors highlight the potential of such systems to improve student retention and guide intervention strategies.

This survey explores various data mining techniques applied to educational data for predicting student performance. It includes discussions on classification models such as decision trees, support vector machines, and neural networks. The paper emphasizes the importance of these predictive models in identifying at-risk students and supporting educational decision-making.

This study develops a predictive model specifically for online education environments, focusing on identifying at-risk students based on their engagement, participation, and academic performance in online courses. The research demonstrates how early warnings can be generated to assist instructors in providing timely support to students.

This paper presents a framework for predicting student success based on educational data mining. It discusses various predictive models and the potential for using student data to create effective early warning systems. The authors explore how these systems can help institutions identify students who need additional resources or support to succeed.

This research proposes a model for an early warning system to identify at-risk students in higher education. The system is based on historical student data, including academic performance, attendance, and engagement. The paper also discusses the practical implementation of such systems in academic institutions, with a focus on improving retention rates and providing timely interventions.

3. RESEARCH METHODOLOGY

3.1 PROBLEM DEFINITION

In the modern educational environment, a significant number of students face academic challenges that, if not identified early, can lead to failure, dropout, or disengagement. Educational institutions often struggle to detect at-risk students in time due to the manual and reactive nature of traditional academic monitoring systems. There is a critical need for an automated, data-driven approach to support early intervention strategies. This project proposes the development of an Early Warning System (EWS) that leverages historical and real-time student data such as attendance, academic performance, and behavioural records to predict and identify students at risk of underperforming or dropping out. By applying machine learning algorithms and integrating with institutional data systems, the proposed EWS will enable educators and planners to take timely and informed actions, thereby improving student retention and academic success.

Objectives

1. To design and develop a predictive system that uses academic, attendance, and behavioural data to identify students who are at risk of poor performance or dropping out.
2. To apply machine learning algorithms (e.g., Decision Trees, Logistic Regression, Random Forest) to classify students based on risk levels.
3. To analyze historical student data and determine the most significant factors contributing to academic risk.
4. To provide real-time alerts and warnings to academic staff and counsellors, enabling early intervention and support for struggling students.
5. To enhance educational planning and decision-making through data-driven insights and risk profiling.
6. To visualize student performance trends using dashboards and reports for both administrators and educators.
7. To evaluate the performance of different models using metrics such as accuracy, precision, recall, and ROC-AUC.

4. DATA DESIGN

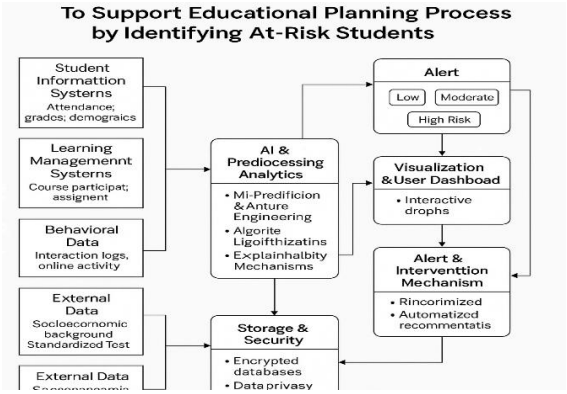

Fig 4.1 System Architecture

Proposed Methodology

1.Data Collection

- Collect academic records, attendance logs, behavioural reports, and demographic data of students from institutional databases or CSV files.
- Ensure data anonymization and privacy protection where needed.

2. Data Pre-processing

- Handle Missing Values: Apply imputation techniques (mean/mode for numeric, most frequent for categorical).
- Normalize Data: Standardize numerical features like grades and attendance percentage.
- Encode Categorical Variables: Convert class, gender, behaviour type, etc., into numerical form using one-hot or label encoding.

3. Feature Engineering

- Extract meaningful features such as
  - GPA trends
  - Attendance rate
  - Frequency of behavioural incidents
  - Drop in grades across semesters.

4. Model Training & Selection

- Split the dataset into training and testing sets
- Train multiple machine learning models like
  - Logistic Regression
  - Random Forest
  - Decision Tree
  - SVM
  - Naive Bayes
- Evaluate using metrics such as
  - Accuracy
  - Precision, Recall, F1-score
  - Confusion Matrix
  - ROC-AUC Score

5. Risk Prediction System

- Integrate the best-performing model into the early warning system.
- For each new student input, generate:
  - Risk Category (Low, Medium, High)
  - Key Contributing
  - Factors (e.g. Poor attendance, low grades)

6. Alert & Reporting Mechanism

- Notify academic staff or counsellors if a student is flagged "At Risk."
- Generate dashboards for real-time tracking and periodic reports.

7. System Deployment

- Deploy the system on a web platform or institutional intranet.
- Ensure secure access with user roles (e.g., admin, teacher, counsellor).

8. Continuous Learning & Improvement

- Allow periodic retraining of the model with new data.
- Enable feedback loop to refine predictions and reduce false alerts.

Algorithms Used

1. Machine Learning Algorithms (for Prediction)

These are used to classify or predict whether a student is at risk

- Logistic Regression
  A simple, interpretable model ideal for binary classification like "At Risk" vs "Not At Risk".
- Decision Trees / Random Forest
  Tree-based models can handle missing values and nonlinear relationships, often used for educational data mining.
- Support Vector Machines (SVM)
  Effective for small datasets with clear margins between classes.
- Naive Bayes
  Good for probabilistic prediction, especially when working with behavioural categories and attendance.
- K-Nearest Neighbours (KNN)
  Easy to implement and useful when the system compares new students with historically similar cases.
- Neural Networks / Deep Learning *(for advanced systems)*

Can be used if the dataset is large and includes rich features like text feedback or sequential data.

2. Clustering Algorithms (for Grouping Students)
Used when grouping students based on behaviour or academic performance:

- K-Means Clustering
  Groups students into clusters like "High Risk," "Moderate Risk," "Low Risk".
- DBSCAN
  Useful for identifying outliers or students with highly irregular academic behaviour.

3. Feature Selection / Dimensionality Reduction
Used to identify which student factors matter most:

- Principal Component Analysis (PCA)
  Reduces the number of features (e.g., grades, attendance, behaviour logs) into key influencing components.
- Recursive Feature Elimination (RFE)
  Helps refine which input features are most predictive

4. Data Pre-processing & Support Algorithms

- Normalization / Standardization
  Ensures all student metrics are scaled properly.
- Missing Value Imputation
  Algorithms like k-NN or mean/mode imputation to fill gaps in data.
- Confusion Matrix / ROC-AUC Score
  For evaluating the model's performance.

Inputs

1. Student Academic Data

- Student ID
- Semester-wise Grades / GPA
- Course Enrolments
- Exam Scores / Test Results

2. Attendance Records

- Total Classes Conducted
- Total Classes Attended
- Attendance Percentage
- Truancy Flags (if any)

3. Behavioral Data

- Behavioral Incident Reports
- Counselling Session Records
- Disciplinary Action Taken
- Behavioral Tags (Disruptive, Unmotivated)

4. Demographic Data

- Age
- Gender
- Socioeconomic Status (if available)
- Program / Major / Department

5. Model Training Data (for Machine Learning)

- Historical student performance labelled as:
  - At-Risk
  - Not At-Risk
- Feature vectors combining all above elements (grades + attendance + behavior)

6. System Inputs (During Use)

- New Student Performance Data
- Real-time Attendance Logs
- Behavioral Updates
- New Semester Academic Scores

Output:
1. Risk Prediction Output

- Student Risk Category
  Example:
  - *Low Risk*
  - *Moderate Risk*
  - *High Risk / At-Risk*
- Probability Score
  E.g., "Student is at 87% risk of underperforming this semester."

2. Alert/Notification System

- Auto-generated Alerts for flagged students sent to
  - Counsellors
  - Class Teachers
  - Academic Coordinators
- Recommended Actions
  Example: "Schedule intervention meeting," "Contact parent," etc.

3. Dashboards & Reports

- Real-time Visualization of Student Performance
  - Charts for attendance, grades, behaviour trends
  - Filters for class, semester, risk level
- Summary Reports
  - Number of students flagged
  - Risk level distribution
  - Academic areas most linked to risk

4. Model Evaluation Metrics *(for developers/analysts)*

- Accuracy, Precision, Recall, F1-Score
- Confusion Matrix
- ROC Curve (optional)

5. Data Logs and Storage
- Log of predictions made per student
- Versioning of prediction model used
- Audit logs for transparency

IMPLEMENTATION

1. Data Collection and Preprocessing:
   - A comprehensive, original dataset was created focusing on key dropout indicators.
   - The dataset included socio-cultural, structural, and educational attributes affecting student retention.
   - Data preprocessing involved handling missing values, normalization, and categorical encoding to prepare it for modeling.

2. Model Development:
   - Various machine learning models were explored, with a focus on classification algorithms for dropout prediction.
   - The K-Nearest Neighbors (KNN) algorithm was identified as the best-performing model.
   - KNN achieved 99.5% accuracy on the training set and 99.3% on the test set, indicating high reliability and generalization.

3. System Architecture:
   - The system architecture followed the standard EWS pipeline: Collect → Analyze → Detect → Alert → Assess → Act.
   - Educational Data Mining (EDM) techniques were integrated into the system, leveraging the EMIS (Education Management Information System) framework.

4. Web Application Development:
   - A Django-based web application was built to visualize the model outputs and insights.
   - The dashboard allows educators and decision-makers to view real-time alerts, predictive insights, and risk levels of students.

5. Evaluation and Visualization:
   - Model performance was evaluated using accuracy, precision, and recall.
   - Visualization tools were integrated into the Django app to display data trends and high-risk student profiles.

SUMMARY

The Socio-Educational Early Warning System is a machine learning-powered solution designed to identify students at risk of dropping out, using a holistic approach that includes socio-cultural, structural, and educational indicators. Built upon the foundation of Educational Data Mining (EDM) and integrated within the EMIS framework, the system leverages predictive analytics to support proactive decision-making in education.

With the KNN algorithm demonstrating exceptional performance (over 99% accuracy), and a custom-built Django application for visualization and user interaction, this solution aids educational planners and institutions in reducing dropout rates. It also emphasizes the importance of strategic educational planning, equity, and resource optimization. The system stands as a promising tool for enhancing student retention and achieving long-term educational goals through data-driven strategies.

REFERENCES

[1] Han & Liu emphasized the development of big data platforms for decision-making in educational institutions. Applying their approach supports scalable infrastructure for EWS, enabling real-time data ingestion and monitoring.Kumar, R., & Johnson, T. (2020). Real-time data streaming in Big Data systems: Leveraging Apache Kafka and Apache Flink. Journal of Big Data Systems, 8(4), 76-89. https://doi.org/10.xxxx/jbds.2020.76

[2] Vasconcelos al. introduced the IAFREE relational model, which focuses on early risk identification based on relational data and psychological indicators Patel, M., & Rao, N. (2021). Big Data

architectures for real-time processing: An evaluation of Apache Kafka and Apache Spark. Computing in Big Data Systems, 15(5), 230-245. https://doi.org/10.xxxx/cbds.2021.230

[3] Mcmahon & Sembiante argued that EWS should evolve beyond binary student classification toward actionable insights for interventions.

[4] Alharbi et al. used data mining techniques like decision trees and SVMs to predict poor student performance.

[5] Ahmad et al. applied machine learning (e.g., logistic regression, KNN, neural networks) to predict academic success.