DOI: 10.2478/seeur-2024-0020

A DATA WAREHOUSING FRAMEWORK FOR PREDICTIVE ANALYTICS IN HIGHER EDUCATION: A FOCUS ON STUDENT AT-RISK IDENTIFICATION

Burim Ismaili PhD(c) (corresponding author)
Faculty of Contemporary Sciences and Technologies,
South East European University,
Tetovo, North Macedonia
b.ismaili@seeu.edu.mk

Adrian Besimi (co-author)
Faculty of Contemporary Sciences and Technologies,
South East European University,
Tetovo, North Macedonia
a.besimi@seeu.edu.mk

ABSTRACT

This paper will examine the development of a data warehouse aimed at improving decision-making in higher education, which focuses on the identification of students at-risk of academic failure through machine learning techniques. This research utilizes South East European University (SEEU) as a case study to show how data warehousing can integrate various student data—including demographics, academic performance, grades, attendance, and engagement—into an integrated framework that enables predictive analytics.

The overall approach allows SEEU decision-makers, administrators, and faculty to proactively identify and assist at-risk students, which improves student retention and their academic success. The results underscore the crucial role of data warehousing in enhancing student achievement and facilitating informed decision-making in higher education.

The paper concludes with concrete suggestions for leveraging data to enhance decision-making processes in a digital educational institution.

Key words: data warehouse, higher Education, data efficient, student progress, decision making

INTRODUCTION

The paper seeks to provide a comprehensive review of the existing literature and findings relating to improved decision-making through data analytics, particularly in the context of predicting the performance of at-risk students. We can use the insights from this analysis to create a recommendation system that will enhance the university's decision-making process. As a part of South East European University, an institution deeply engaged in technology and the digitization of processes, data collection, and analysis, it has always played a vital role in enhancing decision-making for top management. Data visualization has emerged as an essential instrument in education, empowering educators and administrators to identify patterns and trends and deepen their comprehension of information. Educational institutions increasingly recognize data-informed decision-making (DIDM) as a crucial approach to enhance educational standards, boost administrative efficiency, and foster sustainable development.

This research examines the importance of DIDM, especially regarding the South East European University (SEEU). Data-informed decision-making (DIDM) tools like educational data mining, learning analytics, and business intelligence can greatly enhance student recruitment efforts, monitor student achievement, and assess the performance of academic staff (Gaftandzhieva et al., 2023). Various research efforts have employed machine learning and deep learning techniques to forecast the performance of students at risk, highlighting the advantages of leveraging data analytics, prediction, and visualization to improve decision-making processes. As Teng et al. (2022) mentioned, machine learning algorithms and AI models improve decision-making by examining extensive datasets, resulting in more effective policy and strategy development.

Data warehousing (DW) is essential for handling and examining educational data, equipping institutions with the resources necessary to make data-driven decisions based on thorough insights. In the field of education, the significance of data warehousing is rising as schools and universities encounter escalating volumes of data from multiple sources, including student information systems, learning management systems, and administrative databases. A data warehouse (DW) serves as a centralized hub that stores, manages, and analyzes substantial amounts of data collected from various sources. Its main objective is to support business intelligence (BI) efforts, empowering organizations to extract insights from both historical and up-to-date data. By bringing together data from various sources, a data warehouse acts as a singular source of truth, guaranteeing that decision-makers can access precise and consistent information for analysis and reporting. (What Is a Data Warehouse? | Definition, Components, Architecture | SAP, n.d.)

This research explores the integration of various sources of data from different in-house datasets, such as academic records, demographic information, and behavioral data, to create accurate prediction models using a custom-developed data warehouse (DW).

We organize the paper as follows: Section 2 examines the foundational information and current data warehouse architectures specifically designed for educational data. Section 3 outlines the architecture model of the data warehouse, explaining the implementation of the schema design and the process of data integration. Section 4 delineates the analytical techniques employed to pinpoint students at risk, along with the essential algorithms we will employ in our study and their efficacy. The final section delves into the significance of a data warehouse in managing student data, enhancing the decision-making process, and highlighting future opportunities for at-risk student prediction in higher educational institutions. The paper goes further on these research endeavors.

BACKGROUND INFORMATION

The first part of the paper looks at important findings from previous research on data warehousing. The paper concentrates on utilizing data warehousing in schools to aid in decision-making regarding student performance, at-risk students, dropping out, and academic success. Presented below is a comprehensive analysis of this subject.

REVIEW OF EXISTING DATA WAREHOUSE ARCHITECTURES TAILORED TO EDUCATIONAL DATA.

Data warehousing has become crucial for improving decision-making in higher education, particularly for predicting students at-risk. Leveraging large-scale data analytics and predictive modeling, educational institutions can identify students who need additional support, which helps improve retention rates and academic performance. The review of current data warehouse concepts designed for educational data reveals various approaches aimed at improving the processing and analysis of data within educational institutions. These systems aim to streamline decision-making processes, improve learning outcomes, and guarantee data security. In educational contexts, these architectures facilitate the integration of student information, academic records, and administrative data into a single repository, thereby enhancing analysis and reporting (Kimball & Ross, 2013). The following sections outline key aspects of these architectures.

Table 1. Key aspects of existing (DW) architectures tailored to educational data

Data Collection and Types	Predictive Techniques	Application and Benefits
Data warehouses **aggregate information from various sources**, including academic performance, socio-demographics, and learning behaviors, particularly from learning management systems (Li et al., 2023). Students are classified into **risk categories** (Low, Medium, High) based on predictive analytics, which helps in tailoring support strategies (Kamal & Ahuja, 2017).	Techniques such as **decision trees, neural networks**, and **Bayesian networks** are commonly employed to predict student performance (Li et al., 2023) and (Dhankhar et al., 2021). Evaluation of prediction accuracy often utilizes metrics like **classification accuracy, recall**, and **sensitivity** (Li et al., 2023).	Data warehouses support the design of **personalized interventions, enhancing the educational experience** and **reducing dropout rates** (Kurniawan & Halim, 2013) and (Fei-fei, 2007). Systems developed using data warehouse techniques can **proactively identify potential crises among students**, allowing for timely support (Fei-fei, 2007).

The above studies demonstrate the increasing use of data analytics and data warehouses to identify at-risk students. While not all studies specifically address data warehouse architectures for educational data, they all emphasize the use of data mining and learning analytics to enhance performance. In order to provide accurate prediction, machine learning techniques such as decision trees and Bayesian models often use academic and demographic data for this purpose. Additionally, other research looks into data warehouse models that combine various datasets to assist administrators and teachers in assessing student performance and acting promptly on time. These methods demonstrate clearly how predictive analytics can help educational institutions to make better decisions and take early action to improve student success and academic performance. As a famous quote from Albert Einstein states, “Intellectuals solve problems; geniuses prevent them,” so it is of utmost importance for educational institutions to prevent on time in order to make better decision-making and have successful academic performance.

These architectures present significant advantages for educational institutions; the main challenges remain in the implementation and integration of these systems, particularly regarding data quality and user training. While focusing on predictive analytics is crucial, it's equally important to consider the ethical implications of data usage in education, ensuring the preservation of **student data privacy** and the equity of interventions.

PREVIOUS RESEARCH ON IDENTIFYING AT-RISK STUDENTS

For educational institutions to provide on-time interventions and support, it is essential to identify students who are at-risk. The use of several machine learning techniques has made it easier to identify students who are at risk. By using a variety of data sources, these techniques increase predictive accuracy and allow for prompt actions that can greatly benefit student results and academic performance. The studies that follow specifically address this issue.

Numerous studies have suggested data-driven techniques for identifying at-risk students. Duarte et al. (2014) presented a method that solely **utilizes academic data**, demonstrating its feasibility through a case study. In their 2015 study, He et al. examined Massive Open Online Courses (MOOCs), closely examining **student participation behaviors** such as watching lectures and attempting assignments. In a different study, Pei and Xing (2022) built an interpretable pipeline to **detect at-risk students**, highlighting the importance of model explainability. Osborne and Lang's research (2023) similarly discovered elements in **Learning Management System data** without detailing them. Further research is essential to thoroughly understanding the key components in various educational situations.

Furthermore, examining the influence of non-academic elements is crucial in identifying students who are at-risk. The growing transition to an online education environment requires the identification of at-risk students in digital learning environments, which will remain critically important. Gupta et al. (2022) underscore the potential for **early identification** as early as the third week of a course, emphasizing the need for prompt interventions. Vasconcelos et al. (2023) have underscored the importance of relational models in early warning systems to prevent school dropouts. These findings highlight the imperative to combine various data sources, particularly with the expansion of online education, and to investigate relational and behavioral insights for enhanced intervention approaches.

PREDICTING ACADEMIC PERFORMANCE

One of the major challenges that higher education institutions are facing has been predicting future academic performance. Research in the critical area of predicting academic performance aims to improve educational outcomes by identifying students at-risk and enabling timely interventions. This should assist educators in tailoring their teaching methodologies and offering supplementary support to students at-risk, while also predicting timely graduation for students with positive success. A variety of studies looked into various methodologies for developing predictive models of academic performance. Rubiano and Garcia (2016) investigated **data mining techniques** for model development, highlighting the importance of selecting appropriate algorithms. Pardo et al. (2016) employed a **recursive partitioning technique** to develop a predictive model, highlighting the necessity of ensuring these models are applicable for educators. Identifying the **key factors** that influence academic performance is crucial for developing precise predictive models. Aman et al. (2019) indicate that various factors serve as a predictive function at distinct phases of a student's academic experience, underscoring the significance of incorporating the **longitudinal dimension** of learning into models. Recent findings indicate that **age and entry qualifications** significantly influence students' academic self-belief and sense of control, which in turn affects their academic performance (Ofori & Charlton, 2002).

It is important to understand that despite personal factors such as age, entry qualifications, and non-academic skills showing significant importance, the **learning environment** and **socio-cultural context** play an essential role as well. Additional investigation is required to comprehend the precise mechanisms by which these factors affect academic performance and to create targeted interventions that focus on enhancing student success. Furthermore, Asiah et al. (2019) concentrated on recent research on predictive models, **emphasizing the crucial role** of prediction in enhancing educational outcomes for both **educators** and **students**.

Despite the progress, several challenges remain in the field of academic performance prediction. Helal et al. (2018) emphasize that students' diverse backgrounds and varying course selections pose significant challenges. Models that consider **student heterogeneity** and **sub-populations** tend to perform better. There is a need for more detailed reporting of methods and results, as well as efforts to validate and replicate findings to improve research quality (Hellas et al., 2018). Lastly, Ouatik et al. (2022) inform us that the increasing volume of educational data urges the use of Big Data technologies to handle data processing efficiently without compromising the accuracy of predictions. Different machine learning models and a wide range of predictive features enhance the difficult task of predicting academic performance. Continuous research is crucial to address current obstacles and enhance the accuracy and significance of predictive models within higher educational environments, despite the achievements.

DATA WAREHOUSING ARCHITECTURE

This section of the paper will demonstrate the high-level architecture of the data warehouse (DW) with a focus on choosing the right design schema and the process of data integration from various in-house datasets.

HIGH-LEVEL ARCHITECTURE

The initial step in the development and implementation of the project involved researching the datasets available at our university and determining how to access the ones required for it. Accessing the data remotely from outside the university campus presented the first challenge. We collaborated with the IT Office to address this issue, which focused on data security and privacy. To resolve this, we worked closely with the head of the IT Office to identify a solution that balanced accessibility, security, and privacy. In the end, we end up with the below list of available datasets that were essential for this research.

Table2. Available in-house datasets

NR	KEY	QUERY PARAMETERS		DESCRIPTION
1	Applicant01	TermCode	DegreeLevel	Applicants
2	Applicant02	TermCode	DegreeLevel	Accepted applicants

3	**TeacherAtt**	TermCode	WeekNo		Teaching staff
4	**StdBalance**	NULL			Balance
5	**StdPayInv**	TermCode			Invoice and payment
6	StdCeremony	TermCode			Student ceremony
7	**StdCert01**	NULL			Issued certificates
8	StdCert02	NULL			Certificates II
9	**StdCohort**	TermCode			Student cohort
10	StdDiploma	NULL			Issued diplomas
11	**Engagement**	TermCode			Engagements
12	StdEnroll	TermCode	DegreeLevel	FacultyCode	Student enrollment
13	StdEnrollActive	TermCode			Active students
14	StdF_Aid	TermCode			Financial aids
15	**StdGPA**	NULL			Student GPA
16	**StdGradesAvg**	TermCode			Average grades
17	**StdGradesFin**	SessionID			Finalized student grades
18	StdGradesinStd	SessionID			Student grades
19	StdListAll	NULL			Students list
20	**StdProgress**	TermCode			Student progress
21	**Roster**	TermCode			Roster
22	**ScheduleActivity**	TermCode			Schedule activities
23	**StdRoster**	TermCode			Student roster
24	StdList	NULL			List of students
25	StdAdress	TermCode			Student address
26	Transcriptofstd	NULL	DegreeLevel	FacultyCode	Student transcript
27	**StudentAtt**	Year	WeekNo		Week number
28	TeacherAtt	Year	WeekNo		Week number

The table above presents 28 identified in-house datasets, from which we selected those most suitable for our project. Following a thorough analysis of the datasets and based on our research, we decided to utilize the following datasets: **TeacherAtt, StdBalance, StdPayInv, StdCert01, StdCohort, Engagement, StdGPA, StdGradesAvg, StdGradesFin, StdProgress, Roster, ScheduleActivity, StdRoster,** and **StudentAtt.**

There are two main data integration processes: **ETL** (Extract, Transform, Load) and **ELT** (Extract, Load, Transform) used to prepare data for analysis and decision-making. They involve extracting data from multiple sources, cleaning and processing it, and storing it in a target system. The primary difference lies in the sequence of transformation and loading. Since our institution uses traditional data warehouses and works primarily with structured datasets, the above datasets need to be cleaned and transformed before analysis, aiming to analyze student performance by integrating grades, attendance, and demographic information. We would extract the relevant data, clean it by removing missing values, and load it into a structured data warehouse to enable accurate prediction and enhanced reporting. For the purpose of our research, we chose to go with the **ETL (Extract, Transform, Load) process**.

According to Kimball & Ross (2013), the ETL process, a traditional method in data warehousing and analytics, serves as an essential framework in the field of data management.

The main steps to successfully perform data transormation is as follows:

1. First, data is **extracted** from various source systems such as databases, APIs, and spreadsheets.
2. Next step the data undergoes **transformation** in a staging area where it is cleaned, standardized, and converted into the desired format or structure. This step ensures that only clean and standardized data is moved forward.
3. Finally, the transformed data is **loaded** into a target system, typically a data warehouse.

This approach is characterized by performing transformations before loading, which is particularly effective for structured data environments. The advantages of this method include ensuring high data quality and optimization for structured data and relational databases. However, it poses certain challenges, including slower processing for large datasets due to pre-loading transformations and the need for a separate staging area for data transformation (Inmon, 2005). However, our research did not involve working with very large datasets, so this was not a problem.

SCHEMA DESIGN IN DATA WAREHOUSING

Schema design, which defines the organization, storage, and retrieval of information for further analysis, must be carefully considered when building a data warehouse. An effectively structured schema guarantees peak performance, scalability, and user friendliness. Schema design is essential for ensuring that the data warehouse meets the analytical needs of stakeholders and supports advanced analytics, like predicting at-risk students and enhancing decision-making in educational institutions. In the context of educational data warehousing that aims to predict students at-risk, this section describes the essential components and methodologies for schema design.

Table3. Key Schema types available

Star Schema	Snowflake Schema	Galaxy Schema	Data Vault Schema
A central fact table connected to dimension tables in a star-like pattern.	An extension of the star schema, where dimension tables are normalized into multiple related tables.	Multiple fact tables share common dimension tables, often used for multidimensional data analysis.	Consists of hubs (core entities), links (relationships), and satellites (descriptive attributes).
A fact table stores metrics like attendance percentages, GPA, and engagement scores, while dimensions include student demographics, course details, and time.	The student demographics dimension is split into separate tables for age, gender, and socioeconomic data.	Separate fact tables for grades and payment history share dimensions like student and time.	Hubs store student IDs, links represent relationships like enrollment, and satellites hold descriptive attributes like grades and attendance.

Star and snowflake schemas are essential for analyzing student performance data, particularly in educational data warehouses. These schemas enable effective data organization and querying, which is crucial for deriving insights from extensive datasets. The subsequent sections emphasize their applications and effectiveness regarding the analysis of student performance. **Star Schema** consists of a central fact table (e.g., student grades) linked to dimension tables (e.g., student demographics, courses). Queries involving fewer joins are faster, making it suitable for straightforward analyses like average grades by course or demographic group (Cheng & Schneider, 2014). A star schema can efficiently support dashboards that visualize student performance metrics, allowing educators to quickly identify trends.

On the other hand, **Snowflake Schema** normalizes dimension tables, which can reduce redundancy (e.g., separating course details into multiple related tables). This schema allows for easier integration of new dimensions, such as additional performance metrics or student attributes, without significant restructuring (Akid et al., 2022). It is beneficial for complex queries that require detailed analysis across multiple dimensions, such as performance comparisons across different academic years or programs. Star schemas demonstrate superior performance for straightforward queries, whereas snowflake schemas provide enhanced adaptability for changing data requirements. This trade-off underscores the necessity of selecting the suitable schema in accordance with particular analytical needs and the intricacies of the data (Iqbal et al., 2019).

For our research, we decided to go with the **Star Schema design** because of the characteristics of our datasets and the particular analytical requirements of our project. The Star Schema features a clear and straightforward structure that aligns effectively with our goals, facilitating very efficient querying and simple reporting. Linking to descriptive dimension tables, the central fact table establishes an efficient framework for analyzing metrics such as student attendance, roster, GPA, and financial status. This design effectively accommodates structured data, making it particularly suitable for the integration and analysis of datasets like demographics, course details, grades, and academic performance. Utilizing the Star Schema allows us to develop a scalable and user-friendly solution that aligns with our objectives and promotes the development of actionable insights.

Fact and dimension tables tailored to student performance data. Based on the aforementioned datasets and an analysis of their structure and attributes, we have designed fact and dimension tables tailored to student performance data for this research. Here is the initial proposal framework for the **Student Performance Schema**:

Fact Table

Fact_StudentPerformance. The fact table will contain measurable data points central to the analysis.

Attribute	**Description**
Student_ID (FK)	Unique identifier for the student.
Course_ID (FK)	Unique identifier for the course.
TermCode	Academic term or session. (eg. 1232,1239)
GPA	Grade Point Average.
Attendance_Percentage	Percentage of classes attended during semester

Balance	Financial balance status of the student.
Grades_Final	Finalized grades for the term.
Progress_Status	Progress indicators (e.g., completed, on track).

Dimension Tables

Dimension tables provide descriptive attributes to contextualize the data in the fact table.

1. **Dim_Student:**
   Attributes: Student_ID (PK), Name, Age, Gender, Cohort, Demographic Details.
2. **Dim_Course:**
   Attributes: Course_ID (PK), Course_Name, Instructor_ID, Schedule.
3. **Dim_Teacher:**
   Attributes: Teacher_ID (PK), Name, Attendance (from TeacherAtt dataset).
4. **Dim_Time:**
   Attributes: Time_ID (PK), Academic_Year, Semester, Week.
5. **Dim_Financial:**
   Attributes: Financial_ID (PK), Balance, Payments, Invoices.
6. **Dim_Attendance:**
   Attributes: Attendance_ID (PK), Total Classes, Present percentage.

The table below displays the Star Schema diagram pertaining to student performance data.

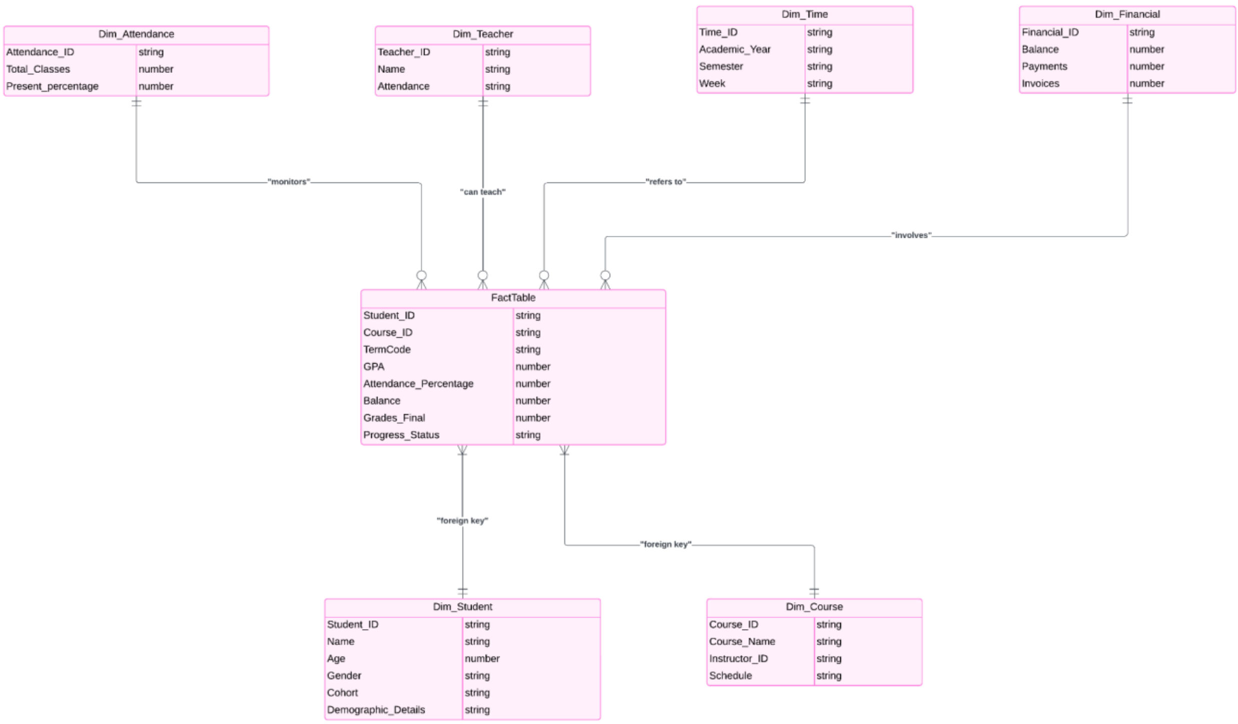

Figure1. Star Schema Diagram

This is a representation of the Star Schema diagram related to student performance data. The primary fact table, Fact_StudentPerformance, links to multiple dimension tables, including Dim_Student, Dim_Course, Dim_Teacher, Dim_Time, Dim_Financil and Dim_Attendace, creating a well-structured and effective schema for querying and analysis.

DATA INTEGRATION AND ETL PROCESSES

Methods for extracting, transforming, and loading (ETL) data from various sources. We initially used JSON scripts to save and transfer data objects with attribute-value pairs and arrays in CSV format, which we then integrated into Cloud Solutions technology. However, this method posed challenges with data updates, as the integration of CSV files with Cloud Solutions technology proved inefficient. After exploring alternative solutions, we ultimately decided to use the KNIME **Analytics platform** for transforming and storing data from the datasets. We successfully integrated this approach with Cloud Solutions technology, which features an automated data update process based on the extracted dimensions. The user can choose between auto-update and a specific time for daily or date-based repetition. The method below describes the steps that we undertake to achieve data integration and transformation. First of all, our data is gathered from various datasets described in the table above (e.g., our datasets consist of extracted csv files downloaded in a folder that is connected to the KNIME Analytics platform).

To protect the privacy and security of personal data, it is essential to make sure that regional and international data privacy laws, like the General Data Protection Regulation GDPR, are followed. A commonly employed technique to accomplish this is **data anonymization**, which involves altering personal data to eliminate or obscure identifiable components while maintaining its usefulness for analysis. This method is widely used to protect privacy and commitment to regulations. What follows describes the construction of an ETL pipeline for pulling datasets from several in-house databases within SEEU repositories, transforming the data into valuable information for future analyses, and loading it into a custom-built data warehouse required for our research. Data is collected from 28 various databases, crucial for the higher education institution (in particular, SEEU) to generate real-time insights that enhance the decision-making process.

The streaming of ETL begins with the **acquisition of data** using a customized **API** and the **extraction** of raw data in **CSV format** from the official University Management System (UMS). Data flows necessitate a variety of configurations to clean and organize the data, facilitate its transformation and aggregation into larger files through procedures like compression and encryption, and prepare it for integration with tools. Analytical tools then clean up and prepare the data for processing. After data collection, the KNIME Analytics platform **filters**, **cleans**, **sorts**, **normalizes**, and **anonymizes** the data, **transforming** it into a suitable format for further processing in an in-house custom-made data warehouse. This step guarantees the forwarding of only clean and standardized data. Streaming ETL pipelines facilitate the efficient ingestion, processing, and continuous analysis of huge data quantities, facilitating quicker accessibility for decision-makers. Finally, we **load** the transformed data into a target system, usually our custom-made data warehouse. Below is the ETL Pipeline Diagram:

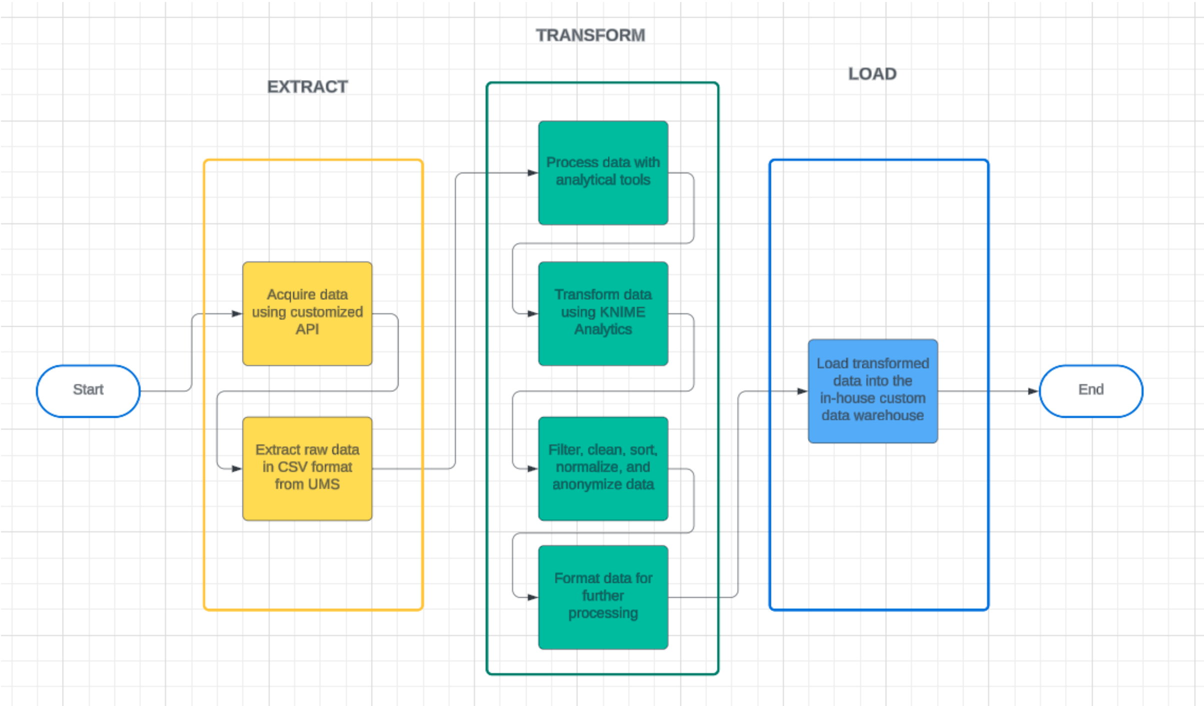

Figure2. ETL Pipeline Diagram

METHODS FOR PREDICTING AT-RISK STUDENTS

Identifying at-risk students through analytical methods is crucial for timely intervention and support. Various predictive models leverage learning analytics, machine learning, and data mining techniques to enhance the accuracy of these predictions.

The following sections of the research outline the key methodologies and findings from recent studies. **Previous grades** and **early semester performance** are significant predictors of student success (Gonzalez-Nucamendi et al., 2023) and (Osborne & Lang, 2023). On the other hand, **age**, **financial status**, and **socio-demographics** also play a role in dropout rates (Gonzalez-Nucamendi et al., 2023). Machine Learning Techniques like Random Forest, neural networks, and probabilistic logistic regression have shown high accuracy rates (up to 92.81%) in predicting at-risk students (Gonzalez-Nucamendi et al., 2023), (Osborne & Lang, 2023) and (Nimy et al., 2023). **Clustering** and **Classification** methods help in identifying patterns and categorizing students based on risk levels (Khalifa et al., 2023).

Common metrics for evaluating prediction models include **classification accuracy**, **sensitivity**, and **true positive rates**, which help in assessing the effectiveness of the interventions (Li et al., 2023). While these analytical methods demonstrate potential in identifying at-risk students, they still face significant challenges in effectively tailoring interventions to individual needs, underscoring the need for continuous refinement of predictive models and strategies. Machine learning algorithms significantly enhance the predictive models for academic performance by analyzing diverse datasets and identifying key factors influencing student outcomes. These algorithms, such as Decision Trees, Support Vector Machines (SVM), and linear regression, utilize historical data to forecast future performance, enabling targeted interventions to improve higher educational results.

Key algorithms that we will use in our research and their effectiveness. Identified as the most effective algorithm in several studies, **Decision Trees** provide clear visualizations of decision-making processes, allowing educators to understand the factors leading to student success or failure (Owolabi & Obadaki, 2024). **Support Vector Machines (SVM)** achieve high accuracy rates (up to 96%) and excel in classifying complex datasets, making them suitable for predicting academic performance based on various input features (Ahmed, 2024). Used to analyze relationships between student performance and factors like attendance and study environment, the **Linear regression** models showed high accuracy (88.3%) in predicting student outcomes (Ogundele et al., 2024). While machine learning offers robust predictive capabilities, it is essential to consider the potential limitations, such as data quality and the need for continuous model updates to reflect changing educational dynamics. Identifying students at-risk is essential for higher education institutions to offer on-time assistance and enhance student success. Key performance indicators (KPIs), such as GPA, attendance rate, and engagement measures, frequently identify at-risk students. The following review examines the effectiveness of these KPIs based on recent findings from the research.

Tucker and McKnight (2017) emphasized that the high school **GPA** serves as a robust indicator of future success in college, frequently surpassing standardized test scores such as the American College Test (ACT). First semester GPA is a crucial predictor of future academic achievement, highlighting the role of initial performance in identifying potentially at-risk students. However, despite the absence of explicit detail in the provided summaries, **attendance data** is widely considered a crucial element in forecasting student performance. Consistent attendance frequently aligns with improved academic performance and reduced dropout rates. As noted by Helal et al. (2018) **engagement** of students with online learning activities, such as Learning Management Systems (LMS), serves as a significant indicator of academic performance. Increased levels of engagement correlate with improved academic results. Finally, the **socio-demographic** characteristics and study modes (such as full-time versus part-time) play a significant role in shaping engagement and, as a result, academic performance. Adjusting predictive models to consider these variations can enhance their precision. Utilizing data mining and machine learning techniques can improve the accuracy of these forecasts, allowing institutions to offer on-time assistance to students who are at risk. Our hypotheses and research goal align with this, as we aim to develop a data warehouse that enhances decision-making in higher education by identifying students at risk of academic failure through machine learning techniques.

CONCLUSION

In summary, data warehousing plays a fundamental role in enhancing decision-making and overseeing student data in higher education. Integrating diverse datasets like grades, attendance, and financial records creates a unified platform through a data warehouse, ensuring data consistency, accessibility, and analytical precision. This integration allows institutions to obtain detailed insights into student performance, promoting informed decision-making across all levels of administration (Rubiano & Garcia, 2016). Data warehousing plays a crucial role in the advancement of predictive models, utilizing machine learning algorithms to more accurately and promptly identify at-risk students than conventional approaches (Pardo et al., 2016). Researchers have identified important indicators such as **GPA**, **attendance**, and **demographic factors** as crucial predictors of student at-risk, highlighting the crucial role of data warehouses in facilitating proactive and tailored interventions (Aman et al., 2019).

Furthermore, the potential advancements offered by data warehousing are extensive. Utilizing advanced predictive analytics can significantly improve early warning systems by delivering real-time insights into student engagement and performance, enabling institutions to take prompt support actions (Gupta et al., 2022). The ability of data warehouses to scale allows them to handle the increasing intricacies of data sources, including learning management systems and behavioral analytics, thereby providing a comprehensive perspective on student requirements (Vasconcelos et al., 2023).

Moreover, these systems enhance the investigation of novel indicators related to academic achievement and attrition rates, allowing for the ongoing improvement of educational approaches. In regard to challenges such as data quality and privacy concerns, following rigorous ETL processes and adhering to regulations like GDPR can help mitigate risks and promote ethical data usage. Over time, data warehouses enable institutions to tackle current issues while also facilitating strategic planning for the changing dynamics of education. The integration of structured data management and predictive technologies revolutionizes educational settings, enhancing student achievement and institutional effectiveness in an era defined by data-driven decision-making. Policy development and strategic planning can benefit from predictive insights, allowing administrators to inform policies, allocate resources effectively, and optimize course offerings in response to anticipated student demand or performance trends.

Through ongoing improvements in its functionalities, a data warehouse enables educational institutions to effectively tackle challenges, promote student achievement, and maintain flexibility in response to changing educational needs.

REFERENCES

1. Ahmed, Amer. (2024). 1. Predictive Model for Student's Academic Performance Using Machine Learning Techniques. International Journal for Science Technology and Engineering
2. Alejandro, A., Vaisman., Florencia, Besteiro., Maximiliano, Valverde. (2019). 5. Modelling and Querying Star and Snowflake Warehouses Using Graph Databases.
3. Aman, F., Rauf, A., Ali, R., Iqbal, F., & Khan, M. (2019). A predictive model for predicting students' academic performance.
4. Amani, Khalifa., Fatma, BenSaid., Yessine, Hadj, Kacem., Zouhaier, Jridi. (2023). 4. At-Risk Students Identification based on Machine Learning Approach: A Case Study of Computer Science Bachelor Student in Tunisia.
5. Amita, Dhankhar., Kamna, Solanki., Sandeep, Dalal., Omdev. (2021). Predicting Students Performance Using Educational Data Mining and Learning Analytics: A Systematic Literature Review.
6. Andres, Gonzalez-Nucamendi., Julieta, Noguez., Luis, Neri., V., Robledo-Rella., R., M., G., García-Castelán. (2023). 2. Predictive analytics study to determine undergraduate students at risk of dropout. Frontiers in Education,
7. Asiah, M., Zulkarnaen, K. N., Safaai, D., & Shahbudin, H. M. (2019). A review on predictive modeling technique for student academic performance monitoring.
8. Akid, H., Frey, G., Ayed, M., & Lachiche, N. (2022). Performance of NoSQL graph implementations of star vs. snowflake schemas. IEEE Access
9. Duarte, R., Ramos-Pires, A., & others. (2014). Identifying at-risk students in higher education.
10. E., N., M., Nimy., Moeketsi, Mosia., Colin, Chibaya. (2023). 5. Identifying At-Risk Students for Early Intervention – a Probabilistic Machine Learning Approach. Social Science Research Network,
11. Gaftandzhieva, S., Hussain, S., Hilcenko, S., Doneva, R., & Boykova, K. (2023). Data-driven decision making in higher education institutions: state-of-play.
12. Gupta, A., Garg, D., & Kumar, P. (2022). An ensembling model for early identification of at-risk students in higher education.
13. He, J., Bailey, J., Rubinstein, B., & Zhang, R. (2015). Identifying at-risk students in massive open online courses.
14. Helal, S., Li, J., Liu, L., Ebrahimie, E., Dawson, S., Murray, D., & Long, Q. (2018). Predicting academic performance by considering student heterogeneity.
15. Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., & Liao, S. (2018). Predicting academic performance: a systematic literature review. Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education.
16. Inmon, W. H. (2005). Building the Data Warehouse (4th ed.). Wiley.
17. Iqbal, M., Mustafa, G., Sarwar, N., Wajid, S., Nasir, J., & Siddque, S. (2019). A Review of Star Schema and Snowflakes Schema.
18. Isreal, M., Ogundele., Olutosin, Taiwo., Asegunloluwa, Eunice, Babalola., Olumide, C, Ayeni. (2024). 4. Prediction of Student Academic Performance Based on Machine Learning Model.
19. J., B., Osborne., A.-S., Lang. (2023). 3. Predictive Identification of At-Risk Students: Using Learning Management System Data.

20. Kam, Cheong, Li., Billy, Wong., Hon, Tung, Chan. (2023). 1. Prediction of At-Risk Students Using Learning Analytics: A Literature Review. Communications in computer and information science
21. Kam, Cheong, Li., Billy, Wong., Hon, Tung, Chan. (2023). 1. Prediction of At-Risk Students Using Learning Analytics: A Literature Review. Communications in computer and information science
22. Kam, Cheong, Li., Billy, Wong., Hon, Tung, Chan. (2023). Prediction of At-Risk Students Using Learning Analytics: A Literature Review. Communications in computer and information science.
23. Kimball, R., & Ross, M. (2013). The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling (3rd ed.). Wiley.
24. M., Zafar, Iqbal., Ghulam, Mustafa., Nadeem, Sarwar., Syed, Hamza, Wajid., Junaid, Nasir., Shaista, Siddque. (2019). 3. A Review of Star Schema and Snowflakes Schema.
25. Manish, A., Bhide., Srinivas, K., Mittapalli., Sriram, Padmanabhan. (2011). 4. Star and snowflake schemas in extract, transform, load processes.
26. Ofori, R., & Charlton, J. P. (2002). A path model of factors influencing the academic performance of nursing students.
27. Osborne, J.B., & Lang, A.S.I.D. (2023). Predictive Identification of At-Risk Students: Using Learning Management System Data.
28. Ouatik, F., Erritali, M., Ouatik, F., & Jourhmane, M. (2022). Predicting Student Success Using Big Data and Machine Learning Algorithms. Int. J. Emerg. Technol. Learn.
29. Pardo, A., Mirriahi, N., Martinez-Maldonado, R., Yacef, K., & Kay, J. (2016). Generating actionable predictive models of academic performance.
30. Pei, B., & Xing, W. (2022). An interpretable pipeline for identifying at-risk students. Journal of Educational Computing Research, 60(6), 743-763.
31. Preet, Kamal., Sachin, Ahuja. (2017). A Review on Prediction of Academic Performance of Students At-Risk Using Data Mining Techniques.
32. Shen, Fei-fei. (2007). An explorer of the student crisis forecast and support system in higher education based on data warehouse. Journal of Chaohu College.
33. Teng, Y., Zhang, J., & Sun, T. (2022). Data-driven decision-making model based on artificial intelligence in higher education system of colleges and universities.
34. Tucker, L., & McKnight, O. (2019). Assessing the Validity of College Success Indicators for the At-Risk Student: Toward Developing a Best-Practice Model. Journal of College Student Retention: Research, Theory & Practice
35. Vasconcelos, A.N., Freires, L.A., Loureto, G.D.L., & others. (2023). Advancing school dropout early warning systems: the IAFREE relational model for identifying at-risk students.
36. What is a data warehouse? | Definition, components, architecture | SAP. (n.d.). SAP.https://www.sap.com/products/technology-platform/datasphere/what-is-a-data-warehouse.html
37. Xun, Cheng., Peter, Schneider. (2014). 2. Star and snowflake join query performance.
38. Yohannes, Kurniawan., Erwin, Halim. (2013). Use data warehouse and data mining to predict student academic performance in schools: A case study (perspective application and benefits).