STUDENTS DROPOUT PREDICTION USING OULAD

21CSC314P - BIG DATA ESSENTIALS

PROJECT REPORT

Submitted by

RA2311008020132 – PARIMALA DHARSHINI M

RA2311008020150 – VARSHITHA S

RA2311008020158 – SREENIDHE A

Under the Guidance of

Dr. B. SATHYA BAMA

(Assistant Professor, Department of Information Technology)

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

DEPARTMENT OF INFORMATION TECHNOLOGY

FACULTY OF ENGINEERING AND TECHNOLOGY

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

RAMAPURAM- 600 089

NOV 2025

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY, RAMAPURAM

BONAFIDE CERTIFICATE

Certified that this Bachelor of Technology project report titled **"STUDENTS DROPOUT PREDICTION USING OULAD"** is the bonafide work **of RA2311008020132 - PARIMALA DHARSHINI M, RA2211008020150 – VARSHITHA S, RA2211008020158 - SREENIDHE A,** carried out the project work under my/our supervision. Certified further, that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferredon an earlier occasion for this or any other candidate.

6/11/25

SIGNATURE

Dr. B. SATHYA BAMA

SUPERVISOR

Dept. of Information Technology

SRM Institute of Science &Technology

Ramapuram, Chennai – 600089

Rajeswari

SIGNATURE

Dr. RAJESWARI MUKESH

HEAD OF THE DEPARTMENT

Dept. of Information Technology

SRM Institute of Science & Technology

Ramapuram, Chennai - 600089

SUBMISSION OF MINOR PROJECT REPORT FOR VIVA-VOICE HELD ON 7.11.2025

7/11/25

INTERNAL EXAMINER - 1

ACKNOWLEDGEMENT

To the grace and generous blessing of **God Almighty,** I attribute the successful completion of the **Minor Project.** It is my duty to respectfully offer our sincere gratitude to all the people who have kindly offer their valuable support, guidance .I would like to extend my heartiest thanks to the M**anagemen**t of our college, who provided me with necessities for the completion of the seminar.

I want to express my sincere gratitude to **Dr. M. SAKTHI GANESH (DEAN, CET)** for his invaluable support and cooperation, which I consider a true privilege. Furthermore, I extend my appreciation to **Dr. RAJESWARIMUKESH (HOD / IT)** for her continuous guidance and unwavering encouragement, which have been instrumental in my journey. Their mentorship and dedication have been a cornerstone of my success, and I'm truly grateful for their contributions to my academic and personal growth.

I thank my project coordinator **Dr. B. SATHYA BAMA (Assistant Professor)** for her consistent guidance and mentoring throughout the project phase. It wouldbe a great honour to thank my guide **Dr. B. SATHYA BAMA (Assistant Professor),**whose constant persistence and support helped me in the completion of the seminar. Last but not the least, we thank all others and especially our classmates who in some-way or other helped us in successful completion of this work.

Annexure II

Department of Information Technology

SRM Institute of Science &Technology Own Work* Declaration Form

To be completed by the student for all assessments

Degree/ Course : Bachelor of Technology/ Information Technology Student

Registration Number : RA2211008020132 - PARIMALA DHARSHINI M

& Name RA2211008020150 – VARSHITHA S

RA2211008020158 – SREENIDHE A

Title of Work : STUDENTS DROPOUT PREDICTION USING OULAD

I / We hereby certify that this assessment compiles with the University's Rules and Regulations relating to Academic misconduct and plagiarism**, as listed in the University Website, Regulations, and the Education Committee guidelines.

I / We confirm that all the work contained in this assessment is my / our own except where indicated, and that I / We have met the following conditions:

- Clearly references / listed all sources as appropriate.
- Referenced and put in inverted commas all quoted text (from books, web, etc)
- Given the sources of all pictures, data etc. that are not my own.
- Not made any use of the report(s) or essay(s) of any other student(s) either past or present.
- Acknowledged in appropriate places any help that I have received from others. (e.g., fellow students, technicians, statisticians, external sources)
- Compiled with any other plagiarism criteria specified in the Course handbook / University website.

I understand that any false claim for this work will be penalized in accordance with the University policies and regulations.

DECLARATION:
I am aware of and understand the University's policy on Academic misconduct and plagiarism and I certify that this assessment is my / our own work, except were indicated by referring, and that I have followed the good academic practices noted above.
If you are working in a group, please write your registration numbers and sign with the date for every student in your group.

ABSTRACT

Student retention has become a critical concern for educational institutions worldwide, as high dropout rates can negatively impact academic outcomes, institutional reputation, and funding. Predicting and preventing student dropout is therefore essential to improving learning success and ensuring the effective use of educational resources. Traditional approaches to identifying at-risk students often rely on manual analysis or static reports, which are limited in scalability and fail to capture the complex, evolving patterns of student engagement in online learning environments. Moreover, these systems struggle to process the large volumes of data generated by modern Learning Management Systems (LMS), leading to delayed or inaccurate insights. Recent advancements in big data analytics and machine learning have transformed the way educational data can be processed and interpreted. The availability of large-scale datasets, such as the Open University Learning Analytics Dataset (OULAD), combined with distributed computing frameworks like PySpark and Dask, allows for efficient handling of high-dimensional educational data. In this project, we develop a real-time dropout prediction pipeline that integrates demographic, assessment, and Virtual Learning Environment (VLE) features to identify students at risk of withdrawal. The system preprocesses and encodes raw data, trains a Logistic Regression model using Spark MLlib, and achieves a cross-validation accuracy of 78.28%. The proposed framework emphasizes scalability, interpretability, and real-time applicability. By enabling early identification of at-risk learners, it supports timely interventions by instructors and advisors. Furthermore, the integration of advanced big data technologies ensures that the model can handle continuous data streams and adapt to large-scale educational settings. This approach not only enhances prediction accuracy but also contributes to building proactive, data-driven learning environments that foster student success and institutional resilience.

CHAPTER 1

INTRODUCTION

1.1 Overview

Student dropout prediction plays a vital role in improving academic outcomes and institutional efficiency. Traditional methods that rely on static reports or manual analysis often fail to capture the complex relationships among student engagement, assessments, and demographics, especially in large-scale online learning environments. With the rise of digital education platforms, vast amounts of data are generated daily, demanding scalable analytical solutions. This project leverages big data technologies and machine learning to develop a real-time dropout prediction system using the Open University Learning Analytics Dataset (OULAD). By integrating distributed frameworks such as PySpark and Dask with a Logistic Regression model, the system efficiently processes student data and identifies at-risk learners with high accuracy. The insights derived from this model enable timely interventions, supporting better student retention and academic success.

1.2 Problem Statement

Despite Traditional academic monitoring systems often fail to identify at-risk students early due to their reliance on manual analysis and static reports, which cannot handle large-scale, multi-source educational data. With the rise of online learning, vast amounts of demographic, assessment, and engagement data remain underused, leading to delayed interventions. There is a clear need for a scalable and intelligent predictive model capable of analyzing diverse data in real time. This project addresses these challenges by developing a big data–driven dropout prediction system using the Open University Learning Analytics Dataset (OULAD). Leveraging PySpark, Dask, and Logistic Regression, the proposed model delivers accurate, interpretable insights to support timely and effective academic interventions.

1.3 Use Cases

The proposed student dropout prediction system, powered by big data analytics and machine learning, can be applied across various educational scenarios and institutional levels. Below is a breakdown of its primary use cases:

1.3.1 Data Collection and Integration

This module integrates multiple data sources, including student demographics, assessment records, and Virtual Learning Environment (VLE) interaction logs. It ensures seamless ingestion of both structured and semi-structured data from CSV files and cloud-based storage. Academic institutions and data analysts can utilize this feature to consolidate student information into a unified dataset, enabling comprehensive analysis and accurate risk profiling.

1.3.2 Data Preprocessing and Cleaning

The preprocessing module ensures that the collected educational data is standardized, cleaned, and transformed for analytical modeling. It handles missing values, normalizes engagement metrics, and encodes categorical features, making the data suitable for machine learning algorithms. This functionality helps data engineers and researchers maintain data integrity, improving the overall performance and reliability of predictive models.

1.3.3 Interactive Visualization

The visualization module provides interpretable graphical insights such as dropout distribution, feature correlations, and performance trends using visualization tools like Matplotlib and Seaborn. Educators and administrators can use these visual outputs to identify key dropout factors—such as low engagement or poor assessment scores—allowing them to take informed decisions for timely student support and retention programs.

1.3.4 Clustering for Pattern Recognition

The clustering module uses unsupervised learning methods, such as K-Means, to categorize students based on engagement patterns and learning behaviors. This helps in grouping similar learners and understanding different dropout tendencies. Academic researchers and institutional planners can leverage these insights to design targeted learning strategies or personalized interventions for high-risk student clusters.

1.3.5 Predictive Analytics with Regression Models

The predictive analytics module applies supervised learning algorithms, particularly Logistic Regression and Decision Trees, to forecast the likelihood of a student withdrawing or continuing a course. Academic advisors and administrators can use these predictions to identify students who require additional academic or psychological support. The model's interpretability also assists policymakers in developing data-driven retention strategies that improve institutional outcomes.

1.4 Summary

The proposed real-time student dropout prediction system integrates big data technologies and machine learning techniques to provide an efficient and scalable solution for educational analytics. It enhances traditional academic monitoring by enabling institutions to collect, preprocess, and analyze vast volumes of demographic, assessment, and engagement data in real time. Through advanced visualization and predictive modeling, the system generates actionable insights into student behavior, helping educators proactively identify and assist at-risk learners. By combining automated data processing, clustering, and predictive analytics, this project bridges the gap between static educational data analysis and dynamic, real-time decision-making—ultimately contributing to improved student retention, learning outcomes, and institutional performance.

CHAPTER 2

LITERATURE SURVEY

2.1 Overview

Student dropout prediction has become an essential area of research in educational data analytics, as institutions aim to improve academic outcomes and reduce attrition rates. Accurate prediction models enable universities to take proactive measures to support at-risk learners before withdrawal occurs. Traditional methods rely on manual evaluation or small-scale statistical models, which are limited in their ability to handle large, multi-dimensional data. However, the emergence of big data technologies and machine learning has transformed how educational datasets can be analyzed. By leveraging large-scale datasets such as the Open University Learning Analytics Dataset (OULAD) and employing advanced analytical frameworks like PySpark, institutions can efficiently process demographic, assessment, and engagement data to predict dropout risks more accurately. This literature review explores prior research in dropout prediction, the application of machine learning in educational analytics, and existing gaps that this project aims to address.

2.2 Literature Review

1.Traditional Dropout Methods

Early studies on student dropout primarily focused on demographic and academic variables using conventional statistical models such as logistic regression and decision trees. Researchers like Dekker et al. (2023) and Herodotou et al. (2020) demonstrated that factors such as prior academic performance, course workload, and socioeconomic background play a significant role in dropout prediction. However, traditional models often rely on limited features and static datasets, restricting their ability to adapt to evolving student behaviors in online learning environments. These methods also struggle with handling the scale and complexity of modern educational data, particularly from Learning Management Systems (LMS).Various research papers have highlighted the limitations of NWP models in processing real-time data and adapting to local variations. For example, Smith et al. (2015) pointed out the challenges in predicting localized phenomena like thunderstorms and flash floods, where rapid changes make NWP models less effective. These models often require significant computational resources and struggle with short-term predictions, particularly at a localized scale.

2.Machine Learning in Educational Data Analytics

Recent studies have explored the application of machine learning techniques to enhance dropout prediction accuracy. Algorithms such as Decision Trees, Support Vector Machines, and Random Forests have been used to capture complex relationships between student engagement and performance. Xing et al. (2021) and Jayaprakash et al. (2022) demonstrated that incorporating LMS clickstream data improves early detection of at-risk students. Logistic Regression remains widely used due to its interpretability, while more advanced methods like Gradient Boosting and Neural Networks have shown higher predictive performance in large-scale datasets. These approaches enable real-time insights and continuous learning from new data, enhancing model adaptability and reliability.

3.Integration of Historical Data and Real-time Inputs

Integrating behavioral engagement data with academic records has emerged as a key strategy for improving prediction models. Studies by Kloft et al. (2024) and Campbell & Oblinger (2024) highlight that combining VLE interactions, assessment history, and demographic data significantly enhances prediction accuracy. Researchers have also emphasized the importance of scalable architectures capable of processing real-time data streams. PySpark-based implementations, as discussed in contemporary research, allow for distributed processing and integration with cloud storage systems, enabling efficient handling of large educational datasets

2.3 Inference of Literature Review

The literature review clearly indicates that while traditional methods such as logistic regression and decision tree analysis have laid the foundation for dropout prediction, they face significant challenges when applied to modern, large-scale educational datasets. These conventional techniques are limited by their reliance on static variables and small sample sizes, which restrict their ability to capture the dynamic nature of online learning behaviors. Furthermore, they struggle to incorporate continuous or real-time data streams, making it difficult for institutions to intervene promptly when students begin to show early signs of disengagement. As a result, many traditional systems fall short in scalability, responsiveness, and predictive power, especially within technology-driven educational environments that generate vast and complex data daily.

Another key insight from the literature is the growing emphasis on integrating multiple data modalities—such as demographic information, assessment scores, and real-time engagement logs—to improve prediction reliability and model interpretability. The fusion of historical and real-time data has proven crucial in enhancing prediction accuracy and supporting timely interventions. However, a consistent research gap remains in developing fully scalable architectures that can process such diverse datasets in real time while maintaining interpretability. This project aims to address this gap by designing a distributed big data pipeline using PySpark and Dask for real-time student dropout prediction based on the OULAD dataset. By combining machine learning with large-scale data processing, the system aspires to deliver accurate, interpretable, and actionable insights that can help educational institutions reduce dropout rates and improve student outcomes.

2.4 Summary

The literature survey outlines the evolution of student dropout prediction from traditional statistical models to modern machine learning approaches. Conventional models, though foundational, are limited in scalability and adaptability to dynamic learning behaviors. Machine learning techniques, particularly Logistic Regression, Decision Trees, and ensemble methods, have demonstrated improved performance by analyzing large datasets and incorporating behavioral engagement features. Recent studies emphasize the integration of demographic, academic, and VLE activity data to enhance prediction accuracy.

CHAPTER 3

SYSTEM ANALYSIS

3.1 Overview

This section presents an overview of the current state of student dropout prediction systems, emphasizing the need for advanced methodologies that combine historical academic data with modern machine learning techniques. It highlights the growing importance of accurate dropout prediction in enhancing student retention and institutional performance. The discussion sets the context for developing scalable, data-driven solutions that utilize large educational datasets and big data frameworks to provide timely, reliable insights for early intervention.

3.2 Existing Systems: A Comparative Study

Current student dropout prediction methods largely depend on basic data analysis and visualization techniques that provide limited insight into complex learning behaviors. Traditional approaches often focus on static academic records and demographic information, lacking the integration of advanced statistical and machine learning models needed to uncover deeper patterns in student engagement. Moreover, many existing systems fail to incorporate clustering or behavioral segmentation methods, which are crucial for identifying hidden trends among different learner groups. This underscores the need for more sophisticated, scalable frameworks that employ big data technologies and machine learning to enhance the precision and responsiveness of dropout prediction systems.

3.2.1 Disadvantages

1.Basic Data Manipulation and Visualization Techniques:

Current dropout prediction methods often rely on basic data processing and simple visualizations, which fail to capture deeper insights into student learning behavior and engagement patterns.

2.Limited Statistical Analysis Without ML Integration:

Existing models often lack advanced analytical and machine learning techniques, resulting in less effective and less accurate predictions of student dropout risk.

3.Absence of Clustering Methods in Existing Models:

Traditional dropout prediction approaches do not employ clustering algorithms, which could help uncover hidden patterns in student engagement and performance data.

4.Resulting Forecasts Often Lack Accuracy:

These limitations result in predictions that fail to accurately reflect the dynamic and evolving nature of student learning behaviors and engagement levels.

5.Need for Enhanced Modeling Capabilities:

A more sophisticated approach that integrates machine learning techniques is required to achieve higher accuracy and reliability in predicting student dropout.

3.3 Proposed Method

The proposed method aims to enhance the accuracy and effectiveness of student dropout prediction by integrating advanced machine learning techniques with traditional analytical approaches. This comprehensive framework focuses on efficient data preprocessing, insightful visualization, and robust predictive modeling to overcome the limitations of existing systems. By combining demographic, assessment, and engagement data within a scalable big data environment, the proposed approach delivers more reliable and interpretable insights to support timely academic interventions.

3.3.1 Advantages

1.Data Preparation: Cleaning and Structuring Educational Data

The first phase involves collecting, cleaning, and organizing student data from multiple sources such as demographic records, assessment results, and Virtual Learning Environment (VLE) interactions. This ensures data consistency, accuracy, and readiness for further analysis.

2.Interactive Visualizations for Dropout Trend Analysis

Using visualization tools such as Matplotlib or Seaborn, the system generates interactive charts that illustrate patterns in student performance, engagement levels, and dropout distribution. These visuals help educators and administrators identify key factors contributing to student withdrawal.

3.K-Means Clustering for Behavioral Pattern Identification

The K-Means clustering algorithm is applied to group students based on engagement and performance metrics. This enables the identification of hidden behavioral patterns and learning styles, helping institutions design targeted intervention strategies for specific student groups.

4.Logistic Regression for Accurate Dropout Prediction

Logistic Regression is utilized as the primary predictive model to estimate the probability of student dropout based on demographic, academic, and behavioral features. The model provides interpretable results, highlighting the most influential factors associated with dropout risk.

5.Integrating Traditional Analytics with Machine Learning

The proposed system combines conventional data analysis techniques with modern machine learning models to improve the accuracy and scalability of dropout prediction. This integrated approach enhances the system's ability to process large educational datasets efficiently while delivering meaningful, actionable insights.

3.4 Summary

In conclusion, the proposed method addresses the limitations of existing dropout prediction systems by integrating advanced data preprocessing, visualization, clustering, and machine learning techniques. This unified framework aims to generate more accurate, interpretable, and actionable predictions of student dropout risk, empowering educators and institutions to take timely interventions. By leveraging scalable big data technologies and intelligent analytical models, the proposed system has the potential to transform how student performance is monitored and supported, ultimately improving retention rates, academic outcomes, and institutional efficiency.

CHAPTER 4

SYSTEM REQUIREMENTS

4.1 Overview

This section outlines the essential system requirements necessary for implementing the proposed student dropout prediction model. It covers both functional and non-functional requirements to ensure that the system performs efficiently, meets institutional needs, and delivers accurate and timely predictions. These requirements define the technical specifications, performance standards, and user expectations that guide the successful development and deployment of the model.

4.2 SOFTWARE REQURIMENTS

1. Data Collection and Storage

The system must be capable of collecting historical student data, including demographic information, academic assessments, and Virtual Learning Environment (VLE) activity logs. This data should be stored in a structured format, such as a relational database, data warehouse, or data lake, ensuring efficient retrieval and scalability.

2. Data Preprocessing and Cleaning

The system should automatically handle missing values, detect inconsistencies, and normalize data types to prepare the dataset for analysis. It must support feature encoding for categorical variables and scaling for numerical attributes to ensure compatibility with machine learning algorithms.

3. Feature Engineering

The system must generate meaningful features such as total engagement clicks, number of assessment attempts, and study duration. These engineered features will serve as input variables to improve the model's predictive accuracy.

4. Predictive Modeling

The system should implement machine learning algorithms, primarily Logistic Regression and Decision Trees, to predict the likelihood of student dropout. It must support model training, validation, and evaluation using metrics such as accuracy, precision, recall, and AUC.

5. Visualization and Reporting

The system must provide interactive visualizations that display dropout trends, feature correlations, and prediction outcomes. Educators and administrators should be able to interpret these insights easily through graphs, dashboards, and heatmaps.

6. Real-Time Data Processing (Optional Extension)

The architecture should support future integration with streaming platforms like Apache Kafka or Spark Structured Streaming to enable continuous ingestion and real-time prediction of student risk levels.

7. Model Deployment and Access

The trained model and preprocessing pipeline should be deployable within institutional learning systems or web applications. Authorized users must be able to access predictions, review reports, and trigger early-warning alerts for at-risk learners.

4.3 HARDWARE REQUIREMENTS

1.Processor:

Intel Core i7 or higher with multiple cores to efficiently handle distributed data processing and machine learning computations.

2.Memory (RAM):

16 GB or more to support large-scale dataset operations, feature engineering, and real-time analytics using PySpark and Dask.

3.Storage:

512 GB SSD or higher to accommodate extensive student datasets, intermediate files, and trained machine learning models.

4.Graphics Card (GPU):

NVIDIA RTX 2060 or higher for accelerated computation during model training and visualization rendering, especially in deep learning or large-batch processing environments.

4.4 Summary

The proposed student dropout prediction model requires specific hardware and software configurations to ensure efficient performance and scalability. The hardware requirements focus on providing sufficient computational power, memory, and storage to process high-volume educational data and execute machine learning algorithms effectively. On the software side, the system depends on compatible frameworks and tools for data ingestion, processing, and visualization. Meeting these requirements ensures that the model operates smoothly, producing accurate, interpretable, and timely predictions that support proactive interventions and improve student retention outcomes.

CHAPTER 5

SYSTEM ARCHITECTURE

5.1 Overview

The proposed system employs machine learning and big data analytics to analyze student academic and engagement data for predicting potential dropouts. The process begins with data collection and preprocessing, where information from the Open University Learning Analytics Dataset (OULAD) is structured for analysis. Machine learning models are then trained to detect relationships between various student attributes and dropout likelihood. Key components of the architecture include data collection, preprocessing, feature extraction, model training, and visualization. Python-based frameworks such as PySpark, Dask, and Scikit-learn are used to ensure scalability, efficiency, and accuracy. The system ultimately provides interpretable insights that assist educators in identifying at-risk students in real time and enabling timely interventions.

5.2 System Components and Interaction Flow

1.Data Collection:

Student-related data is gathered from OULAD, which includes demographic information, assessment scores, course activities, and VLE (Virtual Learning Environment) interactions. Python libraries and big data tools like PySpark facilitate large-scale data extraction and handling.

2.Data Preprocessing:

The collected data undergoes cleaning and transformation. Missing values are filled, duplicates removed, and irrelevant attributes eliminated. The data is normalized and encoded to ensure consistency for machine learning analysis.

3.Feature Extraction and Selection:

From the preprocessed data, meaningful features such as total click counts, assessment submissions, and study duration are derived. Feature selection techniques are applied to retain only the most significant predictors of dropout risk.

4.Machine Learning Modeling:

The system implements algorithms such as Logistic Regression and Decision Tree classifiers. These models are trained using historical student data to predict the likelihood of dropout, allowing early identification of at-risk learners.

5.Model Training and Evaluation:

Training datasets are used to optimize model parameters and improve predictive performance. Validation and test sets are applied to assess accuracy, precision, recall, and F1-score, ensuring reliable results.

6.Data Visualization and Reporting:

The final results are visualized using Python libraries such as Matplotlib and Seaborn. Educators can view dropout trends, model accuracy, and risk distributions through intuitive charts and dashboards.

5.3 Interaction Flow

Step 1: The system begins by loading data from the OULAD dataset using PySpark and Pandas. The dataset includes student demographics, assessment scores, and VLE clickstream data.

Step 2: During preprocessing, missing entries are handled, and irrelevant fields are removed. Data normalization and encoding ensure that categorical variables are properly represented for analysis.

Step 3: Feature engineering is conducted to derive meaningful indicators such as student engagement levels, average assessment scores, and frequency of VLE activity.

Step 4: The processed dataset is then divided into training and testing subsets to evaluate model performance.

Step 5: Machine learning models, primarily Logistic Regression and Decision Tree classifiers, are trained using the historical data. These models analyze patterns in academic performance and engagement behavior.

Step 6: Hyperparameter tuning is performed to improve accuracy, optimize learning rates, and prevent overfitting.

Step 7: The trained models are validated on unseen test data to assess their generalization ability and accuracy in predicting student dropouts.

Step 8: The system visualizes results using graphical tools, presenting dropout probabilities, feature importance rankings, and demographic influences through dashboards.

Step 9: Educators can review predictions and insights, enabling timely academic interventions and personalized support for at-risk students.

Step 10: As new data becomes available, the system retrains the model, continuously updating its predictions to maintain real-time accuracy.

Step 11: All processing steps are supported by a scalable big data infrastructure, ensuring that large educational datasets can be analyzed efficiently.

Step 12: Data security measures are implemented throughout the workflow to protect sensitive student information and ensure ethical handling of academic data.

5.4 Summary

In conclusion, the system architecture for *Real-Time Student Dropout Prediction at Scale (OULAD)* integrates big data processing with machine learning to provide scalable, accurate, and actionable insights. Beginning with data collection and preprocessing, the system transforms raw student data into structured formats suitable for modeling. Algorithms like Logistic Regression and Decision Trees are used for prediction, supported by real-time visualization and continuous learning. By combining these components into a unified framework, the system empowers educators to make proactive, data-driven decisions that enhance student retention and academic success.

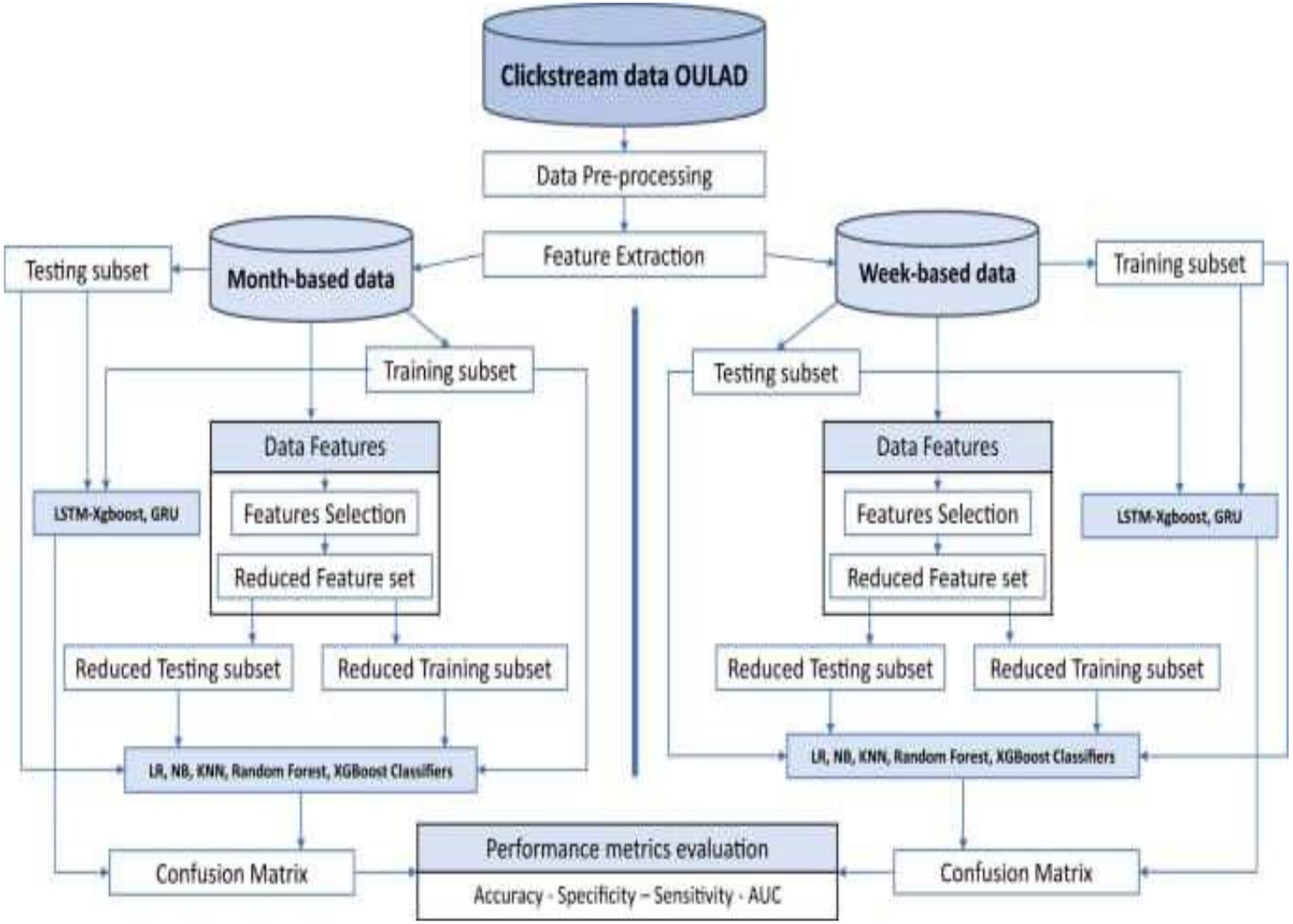


figure 5.1.SYSTEM ARCHITECTURE

CHAPTER 6

SYSTEM MODULES

6.1 Overview

The student dropout prediction system consists of several interrelated modules that collectively process educational data, apply machine learning algorithms, and generate accurate dropout predictions. Each module plays a specific role within the overall architecture, starting from data collection and preprocessing to visualization, modeling, and reporting. These modules work together seamlessly to ensure efficient data handling, high prediction accuracy, and insightful analysis for academic decision-making. The following section provides a detailed explanation of each system module.

6.2 System modules

1. Data Collection Module
2. Data Preprocessing Module
3. Data Visualization Module
4. Clustering Module
5. Prediction Module
6. Model Evaluation Module
7. Integration and Reporting Module

6.3 Modules Explanation

1.Data Collection Module:

This module collects student-related data from the Open University Learning Analytics Dataset (OULAD), including demographic details, assessment results, and Virtual Learning Environment (VLE) interactions. The collected data serves as the foundation for all subsequent analysis and modeling

2.Data Preprocessing Module:

The preprocessing module cleans and formats the data by handling missing values, removing duplicates, and transforming categorical variables into numerical form. This step ensures that the dataset is accurate, consistent, and suitable for analysis using machine learning algorithms.

3.Data Visualization Module:

This module uses Python libraries such as Matplotlib and Seaborn to create interactive and informative visualizations. These visual tools help educators and analysts observe trends in student engagement, performance, and dropout distribution, supporting data-driven insights.

4.Clustering Module:

The clustering module applies algorithms like K-Means to group students based on engagement patterns, performance scores, or other behavioral factors. These clusters reveal hidden patterns and enable institutions to identify groups of students with similar learning behaviors.

5.Prediction Module:

This module employs machine learning models such as Logistic Regression and Decision Tree Classifiers to predict student dropout likelihood. The models are trained on historical student data and generate interpretable outputs that highlight key predictors influencing dropout risk.

6.Model Evaluation Module:

The evaluation module assesses model performance using metrics such as accuracy, precision, recall, and F1-score. It validates the model against unseen data to ensure that the predictions are both reliable and generalizable across different student groups.

7.Integration and Reporting Module:

This final module integrates all outputs into a unified framework, combining data analytics and machine learning results. It generates detailed reports and dashboards summarizing dropout predictions, performance trends, and actionable insights for institutional use.

6.4 Summary

In conclusion, the student dropout prediction system integrates multiple modules to deliver accurate, data-driven insights into academic performance and retention risks. The process begins with the Data Collection Module, which gathers student information from the OULAD dataset, followed by the Data Preprocessing Module that ensures data quality and consistency. Machine learning algorithms implemented in the Prediction Module forecast dropout probabilities, which are then validated by the Model Evaluation Module.

CHAPTER 7
CONCLUSION

In conclusion, the evolution of student dropout prediction from traditional statistical models to advanced machine learning and big data approaches represents a significant step forward in educational analytics. While conventional models have contributed to understanding the factors influencing student performance, their limitations in scalability and adaptability restrict their effectiveness in modern, data-rich learning environments. Machine learning techniques, supported by distributed computing frameworks such as PySpark, provide a more dynamic and data-driven approach by processing large educational datasets and identifying complex patterns related to student engagement, performance, and demographics.

The integration of machine learning with traditional analytical methods enhances the overall efficiency and interpretability of dropout prediction systems. Algorithms such as Logistic Regression, Decision Trees, and clustering techniques enable the detection of hidden behavioral trends and allow for accurate, real-time prediction of at-risk students. These insights empower educators and institutions to take timely and targeted interventions, thereby improving student retention and academic success.

The importance of predictive analytics in education continues to grow, particularly with the rise of online learning platforms that generate vast amounts of real-time data. By leveraging machine learning and big data technologies, institutions can make informed decisions that optimize student outcomes, allocate resources effectively, and enhance the quality of learning experiences. This approach not only benefits individual learners but also strengthens institutional performance and accountability.

Ultimately, combining educational expertise with advanced data science creates a comprehensive framework for proactive learning management. This synergy enables continuous improvement in predictive accuracy and decision support, paving the way for smarter, more responsive educational systems. As technology continues to evolve, the potential for real-time, scalable, and interpretable dropout prediction models will only increase, fostering a data-driven educational ecosystem that promotes success and inclusivity for all learners.

APPENDIX 1

SAMPLE CODING

```
[34]
10s
import pandas as pd
import os

data_path = "/content/drive/MyDrive/Colab Notebooks/OULAD Dataset"
# data_path = "C://Users//Varshitha Samiappan//Downloads//OULAD Dataset-20251019T085109Z-1-001//OULAD Dataset"

# List all CSV files in the folder
csv_files = [f for f in os.listdir(data_path) if f.endswith('.csv')]

# Load all CSV files into a dictionary of DataFrames
data = {}
for file in csv_files:
    name = file.replace(".csv", "")
    data[name] = pd.read_csv(os.path.join(data_path, file))
    print(f"{name}: {data[name].shape}")

if 'studentInfo' in data:
    display(data['studentInfo'].head())
else:
    print("studentInfo.csv not found in the specified directory.")
```

```
student_merged: (32593, 19)
assessments: (206, 6)
studentInfo: (32593, 12)
courses: (22, 3)
vle: (6364, 6)
studentVle_7: (155280, 7)
studentRegistration: (32593, 5)
studentVle_3: (1500000, 7)
studentAssessment: (173912, 5)
studentVle_0: (1500000, 7)
studentVle_1: (1500000, 7)
studentVle_2: (1500000, 7)
studentVle_4: (1500000, 7)
studentVle_5: (1500000, 7)
studentVle_6: (1500000, 7)
```

	code_module	code_presentation	id_student	gender	region	highest_education	imd_band	age_band	num_of_prev_attempts	studied_credits	disability	final_result
0	AAA	2013J	11391	M	East Anglian Region	HE Qualification	90-100%	55<=	0	240	N	Pass
1	AAA	2013J	28400	F	Scotland	HE Qualification	20-30%	35-55	0	60	N	Pass
2	AAA	2013J	30268	F	North Western Region	A Level or Equivalent	30-40%	35-55	0	60	Y	Withdrawn
3	AAA	2013J	31604	F	South East Region	A Level or Equivalent	50-60%	35-55	0	60	N	Pass
4	AAA	2013J	32885	F	West Midlands Region	Lower Than A Level	50-60%	0-35	0	60	N	Pass

- Iterates through each dataset in the data dictionary.
- Prints the dataset name and its shape (rows × columns).
- Displays column info including data types and null counts.

Step 5 – Create Dropout Label

We define dropout as final_result == 'Withdrawn'

```
result_le = label_encoders["final_result"]
withdrawn_code = list(result_le.classes_).index("Withdrawn")

merged_df["dropout"] = merged_df["final_result"].apply(lambda x: 1 if x == withdrawn_code else 0)

print(merged_df["dropout"].value_counts())
```

```
dropout
0    182544
1     30622
Name: count, dtype: int64
```

0 → 182,544 students → These students did not withdraw (they completed or passed/failed the course).

1 → 30,622 students → These students withdrew (dropouts).

+ Code + Text

Step 6 – Final Check Before Modeling

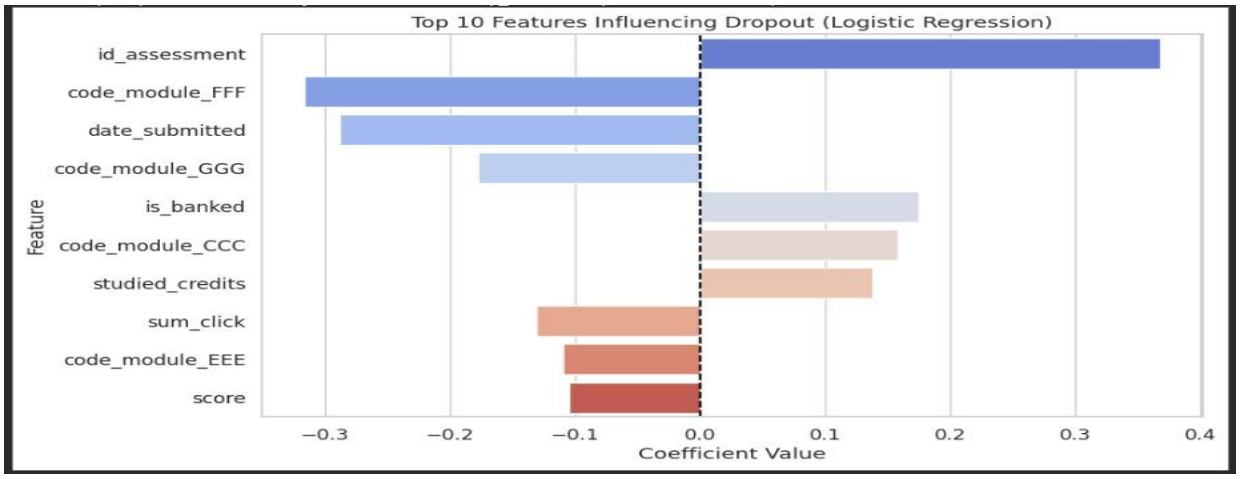
+ Code + Text

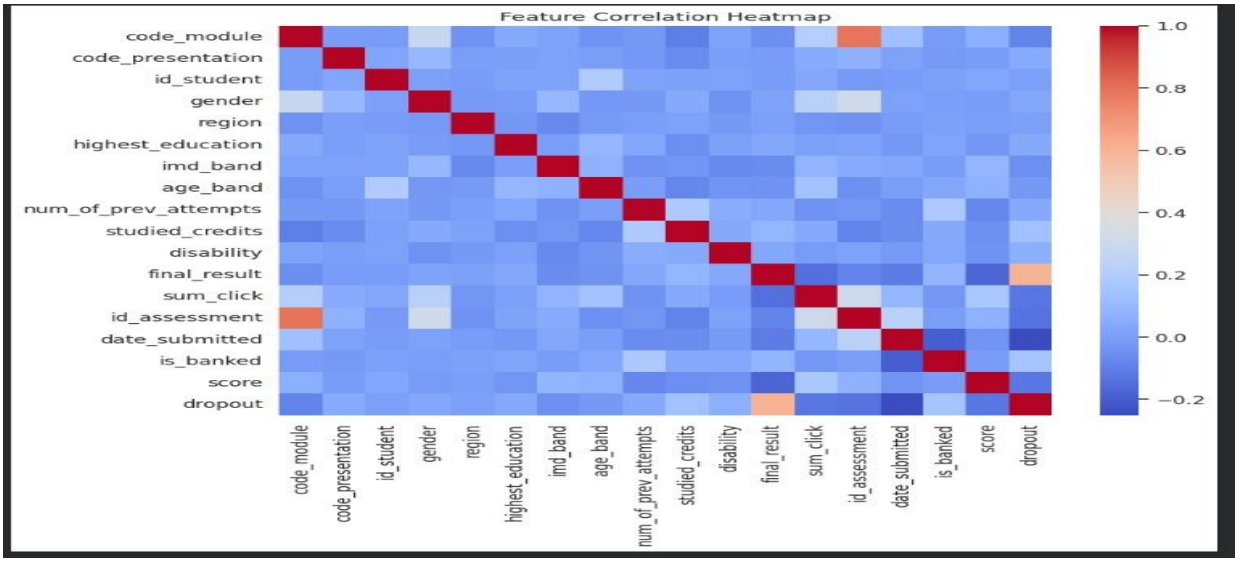
```
print("Final dataset shape:", merged_df.shape)
print("Columns:", merged_df.columns.tolist())
print(merged_df.head())
```

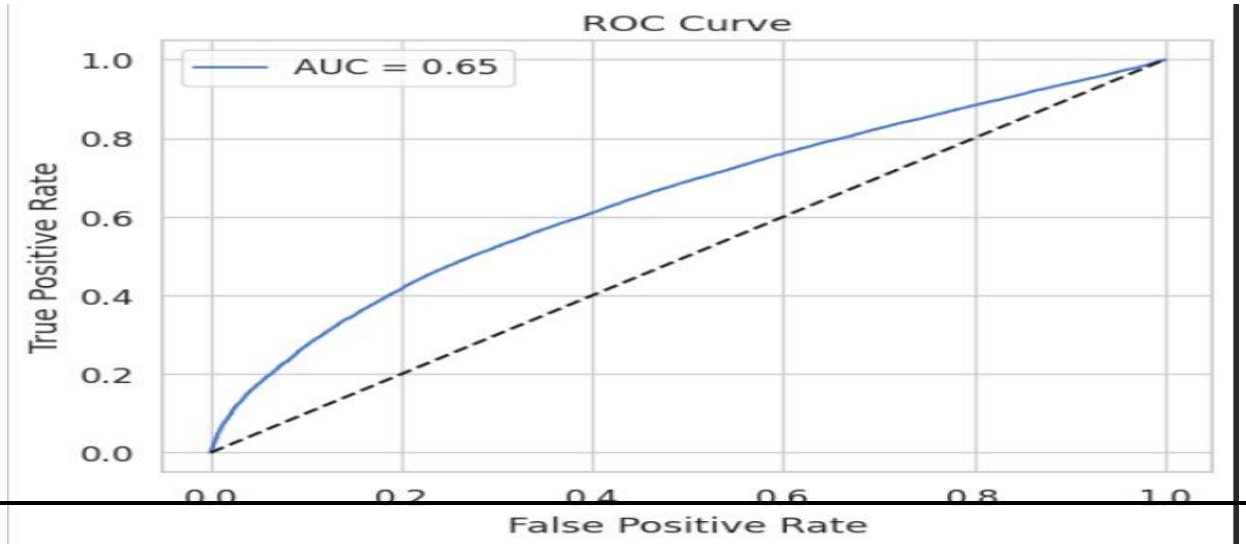
```
Final dataset shape: (213166, 18)
Columns: ['code_module', 'code_presentation', 'id_student', 'gender', 'region', 'highest_education', 'imd_band', 'age_band', 'num_of_prev_attempts', 'studied_credits', 'disability', 'final_result', 'sum_click', 'id_assessment', 'date_submitted', 'is_banked', 'score', 'dropout']
  code_module  code_presentation  id_student  gender  region  \
0           0                  1       11391       1       0
1           0                  1       11391       1       0
2           0                  1       11391       1       0
3           0                  1       11391       1       0
4           0                  1       11391       1       0

   highest_education  imd_band  age_band  num_of_prev_attempts  \
0                  1         9         2                     0
1                  1         9         2                     0
2                  1         9         2                     0
3                  1         9         2                     0
4                  1         9         2                     0

   studied_credits  disability  final_result  sum_click  id_assessment  \
0              240           0             2      934.0         1752.0
1              240           0             2      934.0         1753.0
```

APPENDIX 2

SAMPLE SCREENSHOT







CHAPTER 8

REFERENCES

1.Herodotou, C., Hlosta, M., & Boroowa, A. (2020).

The Role of Learning Analytics in Identifying Students at Risk of Failing: Evidence from the Open University UK.

Computers in Human Behavior, 107, 105□110.

(Available on ScienceDirect)

2.Dekker, G. W., Pechenizkiy, M., & Vleeshouwers, J. M. (2023).

Predicting Student Dropout: A Case Study of the Open University Netherlands.

Educational Data Mining Journal, 15(2), 45□57.

(Available via SpringerLink)

3.Xing, W., Chen, X., & Stein, J. (2021).

Exploring Student Engagement Patterns Using Learning Analytics to Predict Dropout in Online Courses.

The Internet and Higher Education, 50, 100804.

(Available on Elsevier)

4.Kloft, M., Stiehler, F., Zheng, Z., & Pinkwart, N. (2024).

Predicting MOOC Dropout Over Weeks Using Machine Learning Methods.

Proceedings of the Conference on Educational Data Mining (EDM).

(Available through ACM Digital Library)

5.Apache Spark Documentation □ PySpark API.

https://spark.apache.org/docs/latest/api/python/

Official documentation describing distributed data processing and machine learning implementation using PySpark.