PREDICT STUDENT DROPOUT AND ACADEMIC SUCCESS

PROJECT SCOPE

This project's primary objective is to address the pressing issue of academic dropout and failure in higher education. To achieve this, the project focuses on leveraging machine learning techniques to identify students who may be at risk of encountering academic challenges at an early stage in their academic journey. The overarching goal is to provide timely interventions and support strategies to help these at-risk students succeed. The dataset has been specifically created for this project and plays a central role in enabling the development of predictive models. It includes a wide range of information available at the time of student enrollment, encompassing details related to their academic path, demographics, and socio-economic factors. The core problem at hand is a classification task, where students are classified into one of three categories: dropout, enrolled, or graduate. This classification is performed at the end of the standard course duration, allowing for the assessment of student outcomes. To implement the machine learning component for addressing academic dropout and failure within the AWS Academy Cloud Foundations and Data Engineering educational programs, you can leverage various AWS services and tools.

Here is a list of AWS tools that can be instrumental in different stages of the project.

Amazon S3 (Simple Storage Service):

Use S3 to store and manage "Predict students' dropout and academic success" dataset and model training data.

Amazon Athena:

Query data stored in S3 using SQL with Amazon Athena for analysis and preprocessing.

Amazon SageMaker:

SageMaker to train the models, perform hyperparameter tuning, and deploy endpoints for predictions of dropout or graduate or enrolled.

Amazon SageMaker Ground Truth:

For data labeling and annotation tasks, use SageMaker Ground Truth to efficiently label our dataset for training.

Amazon QuickSight:

Use QuickSight for visualizing and analyzing the dataset. Create dashboards and visual representations of key metrics, trends, and patterns within the data, aiding in understanding the factors contributing to academic outcomes.

AWS Identity and Access Management (IAM):

Use IAM to manage access to AWS services securely, creating roles and permissions for different components of your project. Amazon CloudWatch: Set up CloudWatch for monitoring and logging. It helps track the performance of your machine learning models and AWS resources.

AWS CodePipeline and AWS CodeBuild:

Set up CI/CD pipelines using CodePipeline and CodeBuild for automated model training and deployment. Whenever there are updates or improvements to the machine learning models, these pipelines automate the deployment process.

AWS Documentation:

The AWS documentation serves as a valuable resource for troubleshooting, optimizing performance, and ensuring the project adheres to AWS best practices for machine learning implementations in the context of addressing academic dropout and failure.

DOMAIN

Domain - Educational Data Mining and Learning Analytics Educational Data Mining (EDM) and Learning Analytics (LA) represent interdisciplinary fields that leverage data-driven techniques to enhance educational processes and outcomes. These fields involve the collection, analysis, and interpretation of data generated in educational settings to gain insights into student behavior, engagement, and performance. The focus is on using these insights to inform decision-making and improve the overall learning experience.

Key Characteristics:

1. Data-Driven Decision Making: EDM and LA rely on the analysis of vast datasets encompassing student demographics, academic performance, online interactions, and more. This data-driven approach enables educators and administrators to make informed decisions to enhance teaching and learning strategies.
2. Predictive Modeling: One of the key aspects is the development of predictive models that anticipate student outcomes. This includes identifying students at risk of dropping out or struggling academically, allowing for timely interventions to support them.
3. Personalized Learning: The use of data allows for the customization of learning experiences. By understanding individual student needs and preferences, educators can tailor instruction to maximize engagement and knowledge retention.

Challenges:

1. Privacy Concerns: Handling sensitive student data requires stringent privacy measures to protect individuals and comply with regulations. Balancing data accessibility with privacy is an ongoing challenge.
2. Data Integration: Educational data is often siloed across various systems and platforms. Integrating diverse datasets poses challenges but is crucial for a comprehensive understanding of student behavior.
3. Interpretation and Actionability: Deriving meaningful insights from data is essential, but it is equally important to translate these insights into actionable strategies that educators and administrators can implement effectively.

Opportunities:

1. Early Intervention: Predictive analytics can identify students at risk of dropping out early in their academic journey, enabling proactive measures to be taken to support them and improve retention rates.
2. Enhanced Teaching Strategies: Insights from data analysis can inform the development of more effective teaching methods, thereby optimizing the learning experience for students.
3. Continuous Improvement: The iterative nature of data analysis in EDM and LA allows for continuous improvement in educational practices. Regular feedback loops based on data insights contribute to the refinement of teaching and learning strategies.

Stakeholders:

1. Educators and Administrators: Benefit from insights to optimize teaching methods, allocate resources effectively, and implement interventions for at-risk students.
2. Students: Gain from personalized learning experiences and support systems that increase the likelihood of academic success.
3. Institutions: Improve overall retention rates, enhance institutional reputation, and contribute to the success of their students.
4. Policy Makers: Use data-driven insights to inform educational policies and initiatives aimed at reducing dropout rates and improving the quality of education.

LITERATURE REVIEW

1. Predicting Student Dropout and Academic Success (MDPI). [1] The research paper from the Polytechnic Institute of Portalegre focused on predicting student dropout and academic success using a dataset that includes 4,424 records with 35 attributes related to students enrolled in various undergraduate degrees. The dataset was notable for its imbalance, with 50% of the records representing graduates, 32% dropouts, and 18% still enrolled. This imbalance necessitated careful consideration in data handling and analysis. To address this, the study proposed using data-level approaches like the Synthetic Minority Over-Sampling Technique (SMOTE) or the Adaptive Synthetic Sampling Approach (ADASYN), or algorithm-level approaches like Balanced Random Forest or Easy Ensemble. The dataset adhered to privacy and data protection standards, including the General Data Protection Regulation (GDPR), and complied

with the FAIR principles for scientific data management. It's designed to be useful for researchers conducting comparative studies on student academic performance and for training in machine learning

2. Classification and prediction of student performance data using various machine learning algorithms [2] The study "Classification and Prediction of Student Performance Data Using Various Machine Learning Algorithms" focuses on forecasting student performance in higher education. It emphasizes the importance of predicting and classifying students based on their talents to enhance their future academic achievements. Utilizing data mining techniques, the study analyzes massive data to uncover hidden patterns that can be valuable for categorization and prediction. The research employs machine learning algorithms like Naive Bayes, ID3, C4.5, and SVM to analyze a dataset from the UCI machinery, consisting of 33 attributes and 649 instances. This data set, contributed by the University of Minho, Portugal, is used to compare the accuracy of these algorithms in classifying student performance. The findings suggest that a student's prior academic performance can be a significant predictor of their future success. Educational data mining assists in identifying key disciplines that are interdependent, helping students focus on crucial subjects for their academic and future career success. The research concludes that SVM is the most accurate algorithm for classifying student performance data, highlighting its potential in improving the quality of educational instruction and assessment. This type of analysis is critical for educational institutions to lower failure rates and enhance the overall educational experience.

3. Educational Data Mining Techniques for Student Performance Prediction: Method Review and Comparison Analysis [3] The research paper titled "Educational Data Mining Techniques for Student Performance Prediction: Method Review and Comparison Analysis" provides a systematic review of student performance prediction (SPP) from a machine learning and data mining perspective. SPP aims to evaluate a student's grade before course enrollment or exam participation, contributing to personalized education and attracting significant attention in artificial intelligence and educational data mining. The paper outlines five stages of SPP: data collection, problem formulation, model development, performance prediction, and practical application. Various machine learning methods like decision trees, neighborhood methods, linear regression, and neural networks are employed in SPP. The ultimate goal of these studies is to improve student learning performance and reduce educational costs. SPP models analyze student and course features to predict grades, offering solutions for diverse educational scenarios. Once a model is learned, it can predict grades for new students in new courses, with different strategies used in current studies. Ultimately, SPP aims to provide explainable patterns that help educators and other stakeholders improve educational tasks, identify at-risk students, and enhance overall learning outcomes

4. Educational Data Mining to Predict Bachelor Students' Success [4] The research paper "Educational Data Mining to Predict Bachelors Students’ Success" aims to identify the most relevant attributes for academic success using educational data mining (EDM) techniques applied to historical data from a Portuguese business school. The study developed two predictive models to assess academic success at enrolment and the end of the first academic year, employing the SEMMA (Sample, Explore, Modify, Model, and Assess) methodology. The study analyzed data spanning

from 2007/2008 to 2017/2018, including socio-economic, socio-demographic, and academic information of students. The study partitioned each dataset into training, validation, and test sets and utilized logistic regression for feature selection to feed the most relevant features into the learning algorithms. The models used included decision trees, KNN, neural networks, and SVM. The best classifier for entry-level academic success was a random forest with 69% accuracy, while an MLP artificial neural network achieved 85% accuracy for the end of the first academic year. The study concluded that grades and student engagement are crucial for academic success. It also found that the random forest model was the most effective in classifying students' academic outcomes at enrolment, particularly in predicting true negatives, while SVM and ANN were slightly better at predicting true positives

5. Educational Data Mining: Prediction of Students' Academic Performance [5] This research study aimed to develop a new model based on machine learning algorithms to predict undergraduate students' final exam grades using only their midterm exam grades, faculty, and department data, excluding demographic and socio-economic data. Data were sourced from a Turkish State University's Student Information System, focusing on the 2019–2020 fall semester, involving 1854 students who took the Turkish Language-I course. The study employed machine learning classification algorithms like Random Forest (RF), Neural Networks (NN), Logistic Regression (LR), Support Vector Machines (SVM), Naïve Bayes (NB), and k-Nearest Neighbour (kNN), with performance evaluated through tenfold cross-validation. The study's results indicated that the proposed model achieved a classification accuracy between 70–75%, demonstrating that midterm exam grades are significant predictors of final exam grades. The research highlighted the effectiveness of these algorithms in predicting academic performance, emphasizing that such predictions could be made using only midterm exam grades, department data, and faculty data. The findings suggest the potential of these machine learning models to aid in the development of educational policies and interventions aimed at reducing the number of potentially unsuccessful students.

References:

[1]. Realinho, Valentim, Jorge Machado, Luís Baptista, and Mónica V. Martins. 2022."Predicting Student Dropout and Academic Success" Data 7, no. 11: 146.https://doi.org/10.3390/data7110146
[2]. https://doi.org/10.1016/j.matpr.2021.07.382
[3]. https://doi.org/10.3389/fpsyg.2021.698490
[4]. https://doi.org/10.28991/ESJ-2023-SIED2-013
[5]. Yağcı, M. Educational data mining: prediction of students' academic performance using machine learning algorithms. Smart Learn. Environ. 9, 11 (2022).https://doi.org/10.1186/s40561-022-00192-

DATA SOURCE

Predict Student Dropout and Academic Success
- https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

This dataset serves as the foundation for constructing classification models designed to predict students likelihood of dropout or academic success. The classification task involves three categories: dropout, enrolled, and graduate, emphasizing a specific challenge or imbalance in class distribution. The dataset is acquired from various disjoint databases within a higher education institution, focusing on students enrolled in diverse undergraduate degrees, such as agronomy, design, education, nursing, journalism, management, social service, and technologies. It incorporates data on students academic performance at the conclusion of the first and second semesters.
• Contains details about students, courses, academic performance and enrolment history.
• Information about students demographics including age, gender, nationality, and marital status.
• Information about the academic path of students including previous qualifications, admission grades and application details.
• It encompasses details available at the time of student enrolment including academic path, demographics and socio-economic factors

Challenges and Considerations

Data Quality:

Conducted a thorough data quality assessment, identifying and rectifying inconsistencies and inaccuracies in disjoint databases to ensure reliable results.

Data Imbalance:

Ensured a balanced representation for model training. Measures were taken to stabilize the use of weighted classes to prevent bias and enhance the model's ability to generalize across all categories.

Data Preprocessing:

Conducted extensive preprocessing to handle anomalies, outliers, and missing values. This involved a systematic approach to cleaning and organizing the data, ensuring that it was suitable for analysis.

Privacy and Ethics:

Ensured compliance with privacy regulations and ethical guidelines when using sensitive demographic and socio-economic data.

DOMAIN SPECIFIC CHALLENGES

The following are the domain specific challenges for Educational Data Mining and Learning Analytics:

1. Data Privacy and Ethics: When it comes to data, we need to be careful, about protecting students' sensitive information. It's crucial to follow data privacy regulations and maintain practices in handling this kind of data.
2. Bias and Fairness: We should always be mindful of any bias in the data especially when using it to predict student outcomes. Bias can greatly affect the fairness of our models. This leads to results that lack equity.
3. Data Completeness: Educational datasets can vary in terms of quality and completeness. It's important to address any inaccurate data as they can significantly impact the effectiveness of models. That's why thorough preprocessing and cleaning are necessary.
4. Interpretable Models for Educational Stakeholders: In settings in which everyone is not familiar with machine learning models, such as teachers or administrators, it is important to ensure these stakeholders fully understand the insights provided by the models and can act based on them it's crucial that the models are interpretable and provide insights.

KEY PERFORAMANCE INDICATORS (KPIs):

The following are the KPIs that are used to find the measure of success:

1. Accuracy of Predictions: Evaluate how well the model accurately classifies students into categories such as dropout, enrolled or graduate.
2. Precision: Assess the model's ability to correctly identify students who're at risk without generating many false positive results.
3. Recall: Evaluate the model's effectiveness in identifying all students who are at risk without generating many false negative results.
4. Metrics for Addressing Class Imbalance: Utilize metrics like F1 score, which consider both precision and recall tackling the challenges posed by imbalanced class distributions.