MAS 5112 – Exploratory Data Analysis Group Project

Group No: Group 8

Group Details:

Name	Reg.No
Senuri Ranaweera	2025/APST/35
Erandi Dilhara	2025/APST/37
Hansika Prabodini	2025/APST/19
Kalana Ranaweera	2024/APST/26

Predict Students' Dropout and Academic Success

Introduction

Education is a fundamental pillar of individual and societal development. However, student dropout remains a persistent challenge, affecting both learners and educational institutions. Early identification of students at risk of dropping out is crucial for implementing timely interventions that can enhance academic success and overall educational quality. This study aims to predict school dropout and academic success by identifying the major factors influencing these outcomes. By leveraging regression analysis techniques, including logistic and multivariate regression, the research seeks to develop predictive models based on key student and institutional characteristics. The findings will provide valuable insights for educators and policymakers, enabling data-driven decision-making to reduce dropout rates and improve academic performance. Ultimately, this study contributes to strengthening the education sector by fostering student retention and success.

Background

In recent years, the rise of big data in education and the advancement of machine learning have given rise to Educational Data Science, a new discipline focused on analysing student data to enhance learning outcomes and decision-making (Kiss, Nagy, Molontay, & Csabay, 2019).

Alfred Binet pioneered this focus in his early 20th-century research on elementary school students in Paris, developing methods to identify children who required specialized education. His work laid the foundation for using predictive assessments to allocate learning opportunities effectively and address student challenges, shaping modern approaches to forecasting academic success and dropout risks (Travers, 1949)

The prediction of dropout and academic success is a critical area of research in the education sector, as early identification of at-risk students enables educators to implement targeted interventions that reduce dropout rates and enhance long-term academic outcomes. Dropout is influenced by multiple factors and can follow different pathways. The expected pathway is characterized by persistent academic struggles, disengagement, and long-term risk factors, making dropout a foreseeable outcome. In contrast, the unexpected pathway involves students who, despite previous academic success and engagement, leave school due to unforeseen circumstances such as personal, financial, or psychological challenges.

Parental involvement, emotional support, and social competence are key protective factors against school dropout, students with engaged parents and strong relationships are more likely

to graduate, even when facing academic challenges. Conversely, a lack of support can increase the risk of dropout, even for high-achieving students. (Michelle M. Englund, 2008)

Educational outcomes are influenced by multiple factors beyond background risks. Some students overcome high-risk conditions and succeed, while others struggle even in supportive environments. This highlights that dropout is not solely determined by external circumstances but also by individual resilience, motivation, and school engagement (Linda S. Pagani, 2008)

The 2008 economic recession disproportionately impacted male students, leading to higher dropout rates (Sorensen, 2018). Economic hardship further intensified these issues, pushing many male students toward early workforce entry or alternative paths, ultimately increasing dropout rates. Further, as key contributing factors included poor academic performance, amotivation, and absenteeism, which created a cycle of disengagement from education (Balkis, 2018)

Studies on school dropout, particularly before the 10th grade, highlight the full mediation effect of academic achievement and the direct influence of factors such as general deviance, deviant peer affiliation, family socialization, and structural strains (Battin-Pearson, 2000)

High school grades, entrance exam scores, and year of enrolment were key factors in determining whether a medical student would graduate. Among these, high school grades and entrance exam scores were significant predictors of graduation grades, while entrance exam scores also influenced the length of study (Silvija Maslov Kruzicevic, 2012)

Perceived academic control and academic emotions, play a critical role in predicting undergraduate students' academic success. Perceived academic control—a student's belief in their ability to influence academic outcomes—enhances motivation, persistence, and performance. Meanwhile, academic emotions (e.g., enjoyment, anxiety, and frustration) directly impact engagement and learning effectiveness (Ulrike E. Nett*Ulrike E. Nett3, 2017)

Research on academic success highlights intelligence, study habits, and prior academic performance as key predictive factors. Intelligence emerges as the strongest determinant of achievement, followed by the number of study hours per week, while high school preparation has a comparatively smaller impact (May, 1923). High school GPA significantly predicts students' college performance (Amani K. Hamdan Alghamdi, 2014)

While predictive models for academic success provide valuable insights, their ability to accurately forecast college performance remains limited. Various factors, including personal

motivation, mental health, socioeconomic background, and unexpected life events, influence academic outcomes but are difficult to quantify in predictive models. (Mouw, 1993)

Significance of the Data Set

The data set[1] has been created using several disjoint databases from higher education institution related to students enrolled in different undergraduate degrees, such as agronomy, design, education, nursing, journalism, management, social service, and technologies. An important research objective of this field – Educational Data Science, is to predict dropout and improve graduation rates, in particular in STEM[2] higher education. This emphasizes the relevance and significance of analysing this data set. Further, the dataset includes information known at the time of student enrolment (academic path, demographics, and social-economic factors) and the students' academic performance at the end of the first and second semesters.

Research Problems

1. What are the most significant predictors of student dropout and academic success?
2. How do attendance, GPA, and engagement levels mediate the relationship between at-risk students and dropout likelihood?
3. How do demographic factors (e.g., gender, ethnicity, family background) impact dropout rates and academic success?
4. How do socioeconomic status, mental health, and motivation levels influence dropout rates and academic performance?
5. How do national economic indicators such as the unemployment rate, inflation rate, and GDP growth affect students' likelihood of dropping out?

[1] Predict Students' Dropout and Academic Success - UCI Machine Learning Repository

[2] Science, Technology, Engineering, and Mathematics (STEM). The concept of STEM education traces its origins to the early 2000s, when the National Science Foundation (NSF) in the United States coined the term STEM to emphasize the importance of integrating science, technology, engineering, and mathematics into educational curricula

Research Objectives

1. To identify key factors influencing student dropout and academic success by analysing academic performance, socioeconomic background, motivation levels, attendance patterns, and psychological factors.
2. To develop and evaluate predictive models that can accurately forecast student dropout risk and academic success using statistical and machine learning techniques.
3. To assess the effectiveness of existing early intervention strategies in reducing dropout rates and improving academic performance in different educational settings.
4. To examine the role of academic engagement and institutional support in mitigating dropout risk and enhancing student success.
5. To propose data-driven recommendations for educational institutions to improve student retention and academic achievement through tailored interventions and policy adjustments.
6. To enhance student retention through predictive analytics, identifying at-risk students early and enabling proactive interventions to improve retention rates.
7. To improve academic performance by leveraging student data, helping institutions understand learning progress and create targeted initiatives that enhance outcomes at both individual and institutional levels.
8. To address accessibility challenges in higher education, using demographic and socioeconomic insights to develop initiatives that promote inclusivity and equitable access to resources.
9. To classify students into dropout, enrolled, or graduate categories using predictive models based on academic, socioeconomic, and institutional factors

Data

Variable Name	Description	Level of Measurement	Numerical/ Categorical
Marital Status	1 – single 2 – married 3 – widower 4 – divorced 5 – facto union 6 – legally separated	Nominal	Categorical
Application mode	1 - 1st phase - general contingent 2 - Ordinance No. 612/93 5 - 1st phase - special contingent (Azores Island) 7 - Holders of other higher courses 10 - Ordinance No. 854-B/99 15 - International student (bachelor) 16 - 1st phase - special contingent (Madeira Island) 17 - 2nd phase - general contingent 18 - 3rd phase - general contingent 26 - Ordinance No. 533-A/99, item b2) (Different Plan) 27 - Ordinance No. 533-A/99, item b3 (Other Institution) 39 - Over 23 years old 42 - Transfer 43 - Change of course 44 - Technological specialization diploma holders 51 - Change of institution/course 53 - Short cycle diploma holders 57 - Change of institution/course (International)	Nominal	Categorical
Application order	Application order (between 0 - first choice; and 9 last choice)	Ordinal	Categorical
Course	33 - Biofuel Production Technologies 171 - Animation and Multimedia Design 8014 - Social Service (evening attendance) 9003 - Agronomy 9070 - Communication Design 9085 - Veterinary Nursing 9119 - Informatics Engineering 9130 - Equinculture 9147 - Management 9238 - Social Service 9254 - Tourism 9500 - Nursing 9556 - Oral Hygiene 9670 - Advertising and Marketing Management 9773 - Journalism and Communication 9853 - Basic Education 9991 - Management (evening attendance)	Nominal	Categorical
Daytime/evening attendance	1 – daytime 0 - evening	Nominal	Categorical
Previous qualification	1 - Secondary education 2 - Higher education - bachelor's degree 3 - Higher education - degree 4 - Higher education - master's 5 - Higher education - doctorate 6 - Frequency of higher education 9 - 12th year of schooling - not completed 10 - 11th year of schooling - not completed 12 - Other - 11th year of schooling 14 - 10th year of schooling 15 - 10th year of schooling - not completed 19 - Basic education 3rd cycle (9th/10th/11th year) or equiv. 38 - Basic education 2nd cycle (6th/7th/8th year) or equiv. 39 - Technological specialization course 40 - Higher education - degree (1st cycle) 42 - Professional higher technical course 43 - Higher education - master (2nd cycle)	Ordinal	Categorical
Previous qualification (grade)	Grade of previous qualification (between 0 and 200)	Ratio	Numerical
Nationality	1 - Portuguese; 2 - German; 6 - Spanish; 11 - Italian; 13 - Dutch; 14 - English; 17 - Lithuanian; 21 - Angolan; 22 - Cape Verdean; 24 - Guinean; 25 - Mozambican; 26	Nominal	Categorical

	- Santomean; 32 - Turkish; 41 - Brazilian; 62 - Romanian; 100 - Moldova (Republic of); 101 - Mexican; 103 - Ukrainian; 105 - Russian; 108 - Cuban; 109 - Colombian		
Mother's qualification	1 - Secondary Education - 12th Year of Schooling or Eq. 2 - Higher Education - Bachelor's Degree 3 - Higher Education - Degree 4 - Higher Education - Master's 5 - Higher Education - Doctorate 6 - Frequency of Higher Education 9 - 12th Year of Schooling - Not Completed 10 - 11th Year of Schooling - Not Completed 11 - 7th Year (Old) 12 - Other - 11th Year of Schooling 14 - 10th Year of Schooling 18 - General commerce course 19 - Basic Education 3rd Cycle (9th/10th/11th Year) or Equiv. 22 - Technical-professional course 26 - 7th year of schooling 27 - 2nd cycle of the general high school course 29 - 9th Year of Schooling - Not Completed 30 - 8th year of schooling 34 - Unknown 35 - Can't read or write 36 - Can read without having a 4th year of schooling 37 - Basic education 1st cycle (4th/5th year) or equiv. 38 - Basic Education 2nd Cycle (6th/7th/8th Year) or Equiv. 39 - Technological specialization course 40 - Higher education - degree (1st cycle) 41 - Specialized higher studies course 42 - Professional higher technical course 43 - Higher Education - Master (2nd cycle) 44 - Higher Education - Doctorate (3rd cycle)	Ordinal	Categorical
Father's qualification	1 - Secondary Education - 12th Year of Schooling or Eq. 2 - Higher Education - Bachelor's Degree 3 - Higher Education - Degree 4 - Higher Education - Master's 5 - Higher Education - Doctorate 6 - Frequency of Higher Education 9 - 12th Year of Schooling - Not Completed 10 - 11th Year of Schooling - Not Completed 11 - 7th Year (Old) 12 - Other - 11th Year of Schooling 13 - 2nd year complementary high school course 14 - 10th Year of Schooling 18 - General commerce course 19 - Basic Education 3rd Cycle (9th/10th/11th Year) or Equiv. 20 - Complementary High School Course 22 - Technical-professional course 25 - Complementary High School Course - not concluded 26 - 7th year of schooling 27 - 2nd cycle of the general high school course 29 - 9th Year of Schooling - Not Completed 30 - 8th year of schooling 31 - General Course of Administration and Commerce 33 - Supplementary Accounting and Administration 34 - Unknown 35 - Can't read or write 36 - Can read without having a 4th year of schooling 37 - Basic education 1st cycle (4th/5th year) or equiv. 38 - Basic Education 2nd Cycle (6th/7th/8th Year) or Equiv. 39 - Technological specialization course 40 - Higher education - degree (1st	Ordinal	Categorical

	cycle) 41 - Specialized higher studies course 42 - Professional higher technical course 43 - Higher Education - Master (2nd cycle) 44 - Higher Education - Doctorate (3rd cycle)		
Mother's occupation	0 - Student 1 - Representatives of the Legislative Power and Executive Bodies, Directors, Directors and Executive Managers 2 - Specialists in Intellectual and Scientific Activities 3 - Intermediate Level Technicians and Professions 4 - Administrative staff 5 - Personal Services, Security and Safety Workers and Sellers 6 - Farmers and Skilled Workers in Agriculture, Fisheries and Forestry 7 - Skilled Workers in Industry, Construction and Craftsmen 8 - Installation and Machine Operators and Assembly Workers 9 - Unskilled Workers 10 - Armed Forces Professions 90 - Other Situation 99 - (blank) 122 - Health professionals 123 - teachers 125 - Specialists in information and communication technologies (ICT) 131 - Intermediate level science and engineering technicians and professions 132 - Technicians and professionals, of intermediate level of health 134 - Intermediate level technicians from legal, social, sports, cultural and similar services 141 - Office workers, secretaries in general and data processing operators 143 - Data, accounting, statistical, financial services and registry-related operators 144 - Other administrative support staff 151 - personal service workers 152 - sellers 153 - Personal care workers and the like 171 - Skilled construction workers and the like, except electricians 173 - Skilled workers in printing, precision instrument manufacturing, jewelers, artisans and the like 175 - Workers in food processing, woodworking, clothing and other industries and crafts 191 - cleaning workers 192 - Unskilled workers in agriculture, animal production, fisheries and forestry 193 - Unskilled workers in extractive industry, construction, manufacturing and transport 194 - Meal preparation assistants	Nominal	Categorical
Father's occupation	0 - Student 1 - Representatives of the Legislative Power and Executive Bodies, Directors, Directors and Executive Managers 2 - Specialists in Intellectual and Scientific Activities 3 - Intermediate Level Technicians and Professions 4 - Administrative staff 5 - Personal Services, Security and Safety Workers and Sellers 6 - Farmers and Skilled Workers in Agriculture, Fisheries and Forestry 7 - Skilled Workers in Industry, Construction and Craftsmen 8 - Installation and Machine Operators and Assembly Workers 9 - Unskilled Workers 10 - Armed Forces Professions 90 - Other Situation 99 -	Nominal	Categorical

	(blank) 101 - Armed Forces Officers 102 - Armed Forces Sergeants 103 - Other Armed Forces personnel 112 - Directors of administrative and commercial services 114 - Hotel, catering, trade and other services directors 121 - Specialists in the physical sciences, mathematics, engineering and related techniques 122 - Health professionals 123 - teachers 124 - Specialists in finance, accounting, administrative organization, public and commercial relations 131 - Intermediate level science and engineering technicians and professions 132 - Technicians and professionals, of intermediate level of health 134 - Intermediate level technicians from legal, social, sports, cultural and similar services 135 - Information and communication technology technicians 141 - Office workers, secretaries in general and data processing operators 143 - Data, accounting, statistical, financial services and registry-related operators 144 - Other administrative support staff 151 - personal service workers 152 - sellers 153 - Personal care workers and the like 154 - Protection and security services personnel 161 - Market-oriented farmers and skilled agricultural and animal production workers 163 - Farmers, livestock keepers, fishermen, hunters and gatherers, subsistence 171 - Skilled construction workers and the like, except electricians 172 - Skilled workers in metallurgy, metalworking and similar 174 - Skilled workers in electricity and electronics 175 - Workers in food processing, woodworking, clothing and other industries and crafts 181 - Fixed plant and machine operators 182 - assembly workers 183 - Vehicle drivers and mobile equipment operators 192 - Unskilled workers in agriculture, animal production, fisheries and forestry 193 - Unskilled workers in extractive industry, construction, manufacturing and transport 194 - Meal preparation assistants 195 - Street vendors (except food) and street service providers		
Admission grade	Admission grade (between 0 and 200)	Ratio	Numerical
Displaced	1 – yes 0 – no	Nominal	Categorical
Educational special needs	1 – yes 0 – no	Nominal	Categorical
Debtor	1 – yes 0 – no	Nominal	Categorical
Tuition fees up to date	1 – yes 0 – no	Nominal	Categorical
Gender	1 – male 0 – female	Nominal	Categorical
Scholarship holder	1 – yes 0 – no	Nominal	Categorical
Age at enrolment	Age of student at enrolment	Ratio	Numerical
International	1 – yes 0 – no	Nominal	Categorical
Curricular units 1stsem (credited)	Number of curricular units credited in the 1st semester	Ratio	Numerical

Curricular units 1st sem (enrolled)	Number of curricular units enrolled in the 1st semester	Ratio	Numerical
Curricular units 1st sem (evaluations)	Number of evaluations to curricular units in the 1st semester	Ratio	Numerical
Curricular units 1st sem (approved)	Number of curricular units approved in the 1st semester	Ratio	Numerical
Curricular units 1st sem (grade)	Grade average in the 1st semester (between 0 and 20)	Ratio	Numerical
Curricular units 1st sem (without evaluations)	Number of curricular units without evalutions in the 1st semester	Ratio	Numerical
Curricular units 2nd sem (credited)	Number of curricular units credited in the 2nd semester	Ratio	Numerical
Curricular units 2nd sem (enrolled)	Number of curricular units enrolled in the 2nd semester	Ratio	Numerical
Curricular units 2nd sem (approved)	Number of curricular units approved in the 2nd semester	Ratio	Numerical
Curricular units 2nd sem (grade)	Grade average in the 2nd semester (between 0 and 20)	Ratio	Numerical
Curricular units 2nd sem (without evaluations)	Number of curricular units without evalutions in the 1st semester	Ratio	Numerical
Unemployment rate	Unemployment rate (%)	Ratio	Numerical
Inflation rate	Inflation rate (%)	Ratio	Numerical
GDP	GDP	Ratio	Numerical
Target	Target. The problem is formulated as a three category classification task (dropout, enrolled, and graduate) at the end of the normal duration of the course	Nominal	categorical

This dataset is supported by program SATDAP - Capacitação da Administração Pública[3] under grant POCI-05-5762-FSE-000191, Portugal

[3] Capacity Building of Public Administration

Research Methodology

1. Exploratory Data Analysis (EDA)

Before developing predictive models, an exploratory data analysis (EDA) was conducted to understand the dataset, identify trends, detect anomalies, and ensure data quality. The EDA process included the following steps:

2. Choose the Type of Regression Model
   (a) Linear Regression: If predicting continuous variables (e.g., GPA as a measure of academic success).
   (b) Logistic Regression: If predicting binary outcomes (e.g., Dropout = 1, Non-dropout = 0).
   (c) Multiple Regression: If using multiple predictors (e.g., attendance, financial background, and academic performance together).

Model Development and Interpretation

   (a) Define the dependent variable (Y):

- For dropout prediction: Dropout status (Yes/No)
- For academic success: GPA or final grades

   (b) Define independent variables (X1, X2, X3...Xn)

- Attendance rate
- Socioeconomic status
- Academic engagement
- Psychological factors (e.g., motivation)

3. Detecting and Removing Duplicates: Duplicate records were identified and removed to prevent data bias.

Outlier Detection: Boxplots were used to identify extreme values, especially in GPA and attendance rates. Outliers were analysed to determine whether they were valid data points or required removal.

4. Univariate Analysis

Numerical Variables: The distribution of key numerical variables (GPA, attendance rate, age at enrollment) was analysed using histograms and boxplots to detect skewness, normality, or potential transformation needs.

Categorical Variables: Frequency distributions of categorical variables such as gender, course selection, and scholarship status were visualized using bar charts to identify any imbalances.

5. Bivariate Analysis

(a) Correlation Analysis: A heatmap was generated to analyze correlations between predictor variables (e.g., socioeconomic background, academic performance, attendance) and dropout status.

(b) Comparative Analysis: Boxplots and bar charts were used to compare dropout rates across different socioeconomic groups, GPA levels, and institutional support factors.

(c) Chi-Square Test for Categorical Variables: Statistical tests were performed to determine whether categorical features such as scholarship status and parental education had a significant association with dropout risk.

6. Multivariate Analysis

Factor Interaction Analysis: Multivariate regression techniques were applied to examine the combined effects of multiple factors, such as attendance and socioeconomic status, on student dropout.

Principal Component Analysis (PCA): A dimensionality reduction technique was used to identify the most influential variables contributing to dropout and academic success.

7. Hypothesis Testing – Validating Findings

(a) Null Hypothesis (H_0): There is no significant relationship between predictor variables and dropout/academic success.

(b) Alternative Hypothesis (H_1): There is a significant relationship between predictor variables and dropout/academic success.

References

Amani K. Hamdan Alghamdi, A. A.-H. (2014). The Accuracy of Predicting University Students' Academic Success. *Alghamdi, A. K. H., & Al-Hattami, A. A. (2014). TJournal of Saudi Educational and Psychological Association*.

Balkis, M. (2018). Academic amotivation and intention to school dropout: the mediation role of academic achievement and absenteeism.

Battin-Pearson, S. N. (2000). Predictors of early high school dropout: A test of five theories. *American Psychological Association*.

Kiss, B., Nagy, M., Molontay, R., & Csabay, B. (2019). Predicting Dropout Using High School and First-semester Academic Achievement Measures. *2019 17th International Conference on Emerging eLearning Technologies and Applications (ICETA)*.

Linda S. Pagani, F. V. (2008). When Predictions Fail: The Case of Unexpected Pathways Toward High School Dropout.

May, M. A. (1923). Predicting academic success. *Journal of Educational Psychology*.

Michelle M. Englund, B. E. (2008). Exceptions to High School Dropout Predictions in a Low-Income Sample: Do Adults Make a Difference?

Mouw, J. T. (1993). Prediction of academic success: A review of the literature and some recommendations. *College Student Journal*.

Silvija Maslov Kruzicevic, K. J. (2012). Predictors of Attrition and Academic Success of Medical Students: A 30-Year Retrospective Study.

Sorensen, L. C. (2018). "Big Data" in Educational Administration: An Application for Predicting School Dropout Risk.

Travers, R. M. (1949). Significant research on the prediction of academic success.

Ulrike E. Nett*Ulrike E. Nett3, R. S. (2017). Perceived Academic Control and Academic Emotions Predict Undergraduate University Student Success: Examining Effects on Dropout Intention and Achievement.