RESCUING POTENTIAL DROPOUTS

IN MOROCCO

DROPOUT EARLY WARNING SYSTEM (DEWS)

Othman Bensouda Koraïchi

Stanford University

June 2023

Table of Contents

Abstract

Morocco faces a critical challenge with its student dropout rates. While dropout rate stands at **3.6%** in primary school, is escalates to **14.3%** in middle-school**,** and **10.4%** in high school as of 2019. Precise identification of students vulnerable to academic discontinuation offers an opportunity for proactive remedial intervention, enabling schools to orchestrate timely preventive measures. This study focuses on utilizing **data mining** and **machine learning** techniques to predict academic dropouts and facilitate timely intervention in middle school and high school. By leveraging a comprehensive dataset encompassing academic, demographic, and socio-economic information for **336,135 students** in the **region of Fes-Meknes** in **2015-2019**, the research aims to achieve two primary objectives: (1) **modeling** machine learning algorithms to forecast student dropout, aiding in early detection and intervention for at-risk students, and (2) **identifying** key data features that encapsulate the risk factors leading to dropout, aiding in early detection and intervention for at-risk students. Through a comparative analysis of different machine learning methodologies, the study reveals promising results, demonstrating the ability to correctly identify **84%** of potential dropouts by filtering just **19%** of the dataset using Gradient Boosted Trees. The research identifies unauthorized absences, Grade Point Average (GPA), and class rank as crucial indicators for predicting school dropout. These findings

offer valuable insights and pave the way for implementing predictive data science in the education sector, potentially mitigating dropout rates and promoting academic success in Morocco.

Introduction

Students' dropout is a serious problem for students, society, and policy makers. Across many low and middle-income countries, a sizable share of young people drop out of school before completing a full course of basic education. In Morocco, while education is compulsory until the age of 16[1], this share amounts to 3.6% in primary school, 14.3% in middle-school and 10.4% in high school, at the national level in the school year 2018-2019[2].

Dropping out before the end of high school can lead to various negative outcomes. Indeed, compared to high school graduates, dropouts have: higher rates of unemployment; lower earnings; poorer health and higher rates of mortality; higher rates of criminal behavior and incarceration; increased dependence on public assistance; and are less likely to vote[3]. The negative

[1]Dahir No. 1-19-113 of Hijra 1440 (August 9, 2019) promulgating Framework Law No. 51-17 on the Education, Training, and Scientific Research System.

[2] Conseil Supérieur de l'Éducation et de la Recherche Scientifique. "ATLAS TERRITORIAL DE L'ABANDON SCOLAIRE." https://www.csefrs.ma/wp-content/uploads/2019/12/ATLAS-TERRITORIAL-DE-LABANDON-SCOLAIRE-18-12-web.pdf

[3] Rumberger, Russell, and Sun Lim. Why Students Drop out of School: A Review of 25 Years of Research. 2008.

outcomes from dropouts generate huge social costs. Not only do governments collect fewer taxes from dropouts, but they also subsidize the poorer health, higher criminal activity, and increased public assistance of dropouts. According to the National Evaluation Authority in 2016, the cost of dropping out of school was estimated to be over 2 billion Moroccan Dirhams (> $200 million) for all three levels of education (primary, middle, and high school)[4]. Leaving school has not only a financial cost, but also a socio-economic one. As stated in the report by the Higher Council of Education and Scientific Research (CSEFRS), "the impact of education is not limited to the individual level, but also affects the social fabric of the country, its position in the world, and how it projects itself into the future. Education is therefore at the heart of building the development dynamic." Furthermore, in a 2011 study on social mobility, the Higher Planning Commission (HCP)[5] estimated that an additional year of education increased a child's chances of upward mobility by 14%. This confirms that dropping out of school has detrimental consequences, not only for the individual, but also for society as a whole. In fact, any young person who drops out of school risks falling into a vicious cycle that leads to an irreversible situation marked by illiteracy, marginalization, vulnerability, delinquency, and even violence and crime.[6]

[4] [5] [6]Attijariwafa Bank. (2017). Pole Edition et Débats: Conférences. Retrieved from https://www.attijariwafabank.com/sites/default/files/widgets/files/17-00328-book-pole_edition_et_debats-conferencesa447degedition_4.pdf

Addressing the dropout crisis requires a better understanding of why students drop out. Yet identifying the causes of dropping out is extremely difficult. Recent studies have focused on predictive and descriptive modeling using machine learning in order to identify students at risk of dropping out and understand the risk factors. Furthermore, the rise of massive online open courses (MOOCs) and other online educational environments has seen an increase in the application of data mining and machine learning techniques to educational data, particularly in the domains of educational data mining and learning analytics[7] [8] [9] [10] [11]. However, in part due to the lack of appropriate data, much less quantitative research has focused on student dropout in the traditional classroom environment, especially before college. Additionally, secondary school dropout prediction, as opposed to higher education dropout prediction, is challenging since it heavily depends on the particular context of the country and is much less exportable. Finally, to date, there have been limited large-scale studies conducted in developing countries and using machine learning, particularly on the African continent, where financial

[7] Baker, R. S., & Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining*, *1*(1), 3–17. https://doi.org/10.5281/zenodo.3554657

[8] *Long*, P., *Siemens*, G., "Penetrating the Fog: Analytics in Learning and Education, Educause Review, 46 (5), 31-40, *2011*.

[9] Siemens, George & Baker, Ryan. (2012). Learning analytics and educational data mining: Towards communication and collaboration. ACM International Conference Proceeding Series. 10.1145/2330601.2330661.

[10] Romero, C. and Ventura, S. (2013) Data Mining in Education. WIREs Data Mining and Knowledge Discovery, 3, 12-27. https://doi.org/10.1002/widm.1075

[11] Baker, R.S., Inventado, P.S. (2014). Educational Data Mining and Learning Analytics. In: Larusson, J., White, B. (eds) Learning Analytics. Springer, New York, NY. https://doi.org/10.1007/978-1-4614-3305-7_4

and technical limitations make data collection and management systems more constrained.

Hence, this research aims to contribute to the scientific literature through its focus on secondary school dropout prediction using machine learning in a developing country. This paper describes a national research project facilitated by the Moroccan Ministry of Education, whose main objective is to set up a Dropout Early Warning System (DEWS) capable of identifying middle school and high school students at risk, and to determine key data features that encapsulate the risk factors leading to dropout.

Prior Work

Early Warning Systems

In the educational domain, an EWS consists of a set of procedures and instruments for early detection of indicators of students at risk of dropping out and also involves the implementation of appropriate interventions to make them

stay in school[12]. Seidman[13] developed a slogan about student retention showing that early identification of students at risk, in addition to maintaining intensive and continuous intervention, is the key to reduce dropout levels. So, to develop and use an early warning system (EWS) is a good solution for detecting students at high risk of dropout as early as possible.

As reviewed by Marquez-Vera et al.[14] , several countries have implemented EWSs, including Mexico, which uses an Excel-based system with specific critical thresholds for absenteeism, low performance, and problematic behavior[15]. The US National High School Center has also developed an EWS based on Excel, using course performance and attendance as indicators, and Delaware Department of Education has implemented a multi-variable model in several states[16]. Additionally, Austria, Croatia, and England have focused on systematic monitoring of truancy/absenteeism and results/grades[17] [18]. While Excel-based systems may

[12] Heppen, Jessica, and Susan Bowles Therriault. ISSUE BRIEF Developing Early Warning Systems to Identify Potential High School Dropouts. 2008.

[13] Seidman, A. (1996). Retention revisited: R = E, Id+E & In, Iv. College and University, 71(4), 18-20.
[14] Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Fardoun, H. M., & Ventura, S. (2015). Early dropout prediction using data mining: A case study with high school students. Expert Systems, 32(5), 3-14. DOI: 10.1111/exsy.12135

[15] Maldonado-Ulloa P.Y., Sancén-Rodríguez A.J., Torres-Valades M., Murillo-Pazarán B. (2011) Secretaria de Educación Pública de Mexico. Programa Síguele. Sistema de Alerta Temprana, Lineamientos de Operación. 1–18.

[16] Uekawa, Kazuaki, et al. REL Mid-Atlantic Technical Assistance Brief REL MA 1. No. 2, 2010, pp. 75–85, files.eric.ed.gov/fulltext/ED565682.pdf.

[17] Vassiliou, A. 2013. Early warning systems in Europe: practice, methods and lessons. Thematic Working Group on Early School Leaving. 1-17.

[18] Márquez-Vera, Carlos, et al. "Early Dropout Prediction Using Data Mining: A Case Study with High School Students." Expert Systems, vol. 33, no. 1, 16 Nov. 2015, pp. 107–124, https://doi.org/10.1111/exsy.12135.

not be suitable for large amounts of data, statistical techniques such as logistic regression and discriminant analysis have been used to identify factors and their contributions to student dropout[19]. However, Educational Data Mining (EDM) has emerged as a new application area to detect patterns in large educational data sets[20] [21] . In summary, the implementation of EWSs with appropriate indicators and interventions is crucial to address student dropout. While traditional statistical techniques have been used, EDM provides a new opportunity to analyze large data sets and detect patterns that may be difficult to identify otherwise.

However, detecting these indicators is challenging due to the multifactorial nature of the problem, also known as the « one thousand factors problem » [22].

It is also important to note that different countries have quite different high school systems. For example, the duration of the high school can vary a lot among countries. Due to these differences, datasets from different countries can have very different meanings and, even if they include similar features, these are

[19] Kovacic, Z. Early Prediction of Student Success: Mining Students Enrolment Data. 2017, www.semanticscholar.org/paper/Early-Prediction-of-Student-Success%3A-Mining-Data-Kovacic/e48eba98bde33586c20442d46ab9a59c411196e5.

[20] Romero, C. and Ventura, S. (2013) Data Mining in Education. WIREs Data Mining and Knowledge Discovery, 3, 12-27. https://doi.org/10.1002/widm.1075

[21] Baker, R. S., & Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining*, *1*(1), 3–17. https://doi.org/10.5281/zenodo.3554657

[22] Magaña, Marina, and Hernández Pediatra. XIII Congreso de La Sociedad Española de Medicina Del Adolescente 1a Mesa Redonda CAUSAS DEL FRACASO ESCOLAR. 2002.

describing quite different situations. For this reason, works on lower levels of education are much less general and exportable to other systems. Hence, dropout prediction for primary and secondary education is especially more challenging than college dropout prediction, for which it is usually possible to easily "translate" a system into another[35].

Early warning systems that accurately identify students at risk of dropout and support them with targeted interventions have shown results and are in widespread use in high-income contexts[23] [24] [25]. However, limited literature is available in low-income countries, though recent researches have shown applications at the national level, in countries such as Honduras and Guatemala[26]. Furthermore, the available literature on the African continent is notably scarce, but there are some noteworthy successful examples. Mnyawami et al.[27] proposed one of the few models for predicting student dropouts in Tanzanian secondary schools. The authors collected data on student demographics, academic

[23] Baker, Ryan S., et al. "Predicting K-12 Dropout." Journal of Education for Students Placed at Risk (JESPAR), vol. 25, no. 1, 1 Oct. 2019, pp. 28–54, https://doi.org/10.1080/10824669.2019.1670065.

[24] Knowles, Jared E. "Of Needles and Haystacks: Building an Accurate Statewide Dropout Early Warning System in Wisconsin." Journal of Educational Data Mining, vol. 7, no. 3, 25 July 2015, pp. 18–67, jedm.educationaldatamining.org/index.php/JEDM/article/view/JEDM082, https://doi.org/10.5281/zenodo.3554725

[25] Chung, Jae Young, and Sunbok Lee. "Dropout Early Warning Systems for High School Students Using Machine Learning." Children and Youth Services Review, vol. 96, Jan. 2019, pp. 346–353, https://doi.org/10.1016/j.childyouth.2018.11.030.

[26] Adelman, Melissa, et al. "Predicting School Dropout with Administrative Data: New Evidence from Guatemala and Honduras." Education Economics, vol. 26, no. 4, 2 Feb. 2018, pp. 356–372, https://doi.org/10.1080/09645292.2018.1433127

[27] Mnyawami, Yuda N., et al. "Enhanced Model for Predicting Student Dropouts in Developing Countries Using Automated Machine Learning Approach: A Case of Tanzanian's Secondary Schools." Applied Artificial Intelligence, vol. 36, no. 1, 7 May 2022, https://doi.org/10.1080/08839514.2022.2071406

performance, family background and school infrastructure to build a predictive model using methods such as Logistic Regression, Decision Trees, Random Forest, AdaBoost, Multilayer Perceptron Classifier, K-Nearest Neighbors, Naïve Bayes, Linear Discriminant Analysis and SGD Classifier . The model accurately predicted dropout and identified important predictors such as student age, gender, academic performance, and parental education.

N. Mduma and D. Machuve[28] proposed a machine learning model for predicting student dropout in Tanzania, Kenya, and Uganda. They found that Logistic Regression outperformed Multilayer Perceptron and Random Forest, and identified factors such as student age, and gender as important predictors of dropout.

Weybright et al.[29] aimed to predict secondary school dropout among South African adolescents using a survival analysis approach. The researchers analyzed data from a longitudinal study of 601 South African adolescents and identified risk factors for dropout, such as being male, not living with one's mother, smoking

[28] N. Mduma and D. Machuve, "Machine Learning Model for Predicting Student Dropout: A Case of Tanzania, Kenya and Uganda," 2021 IEEE AFRICON, Arusha, Tanzania, United Republic of, 2021, pp. 1-6, doi: 10.1109/AFRICON51333.2021.9570956.

[29] Weybright, Elizabeth H., et al. "Predicting Secondary School Dropout among South African Adolescents: A Survival Analysis Approach." South African Journal of Education, vol. 37, no. 2, 31 May 2017, pp. 1–11, www.scielo.org.za/scielo.php?script=sci_arttext&pid=S0256-01002017000200006, https://doi.org/10.15700/saje.v37n2a1353

cigarettes in the past month, and having lower levels of leisure-related intrinsic motivation.

Class imbalance

Predicting student failure is a difficult challenge due to both the high number of factors that can affect the low performance of students and the imbalanced nature of these types of datasets.

In this context, class imbalance means that there are usually way less dropouts than non-dropouts, which can hamper model accuracy. Mduma[30] revised machine learning algorithms to predict academic dropout in developing countries, and concluded that many researchers ignore data that is unbalanced, leading to improper results. This is because training a machine learning model for binary classification with a highly unbalanced dataset may result in poor final performance, mainly because in such a scenario the classifier would underestimate the class with a lower number of samples.

[30] Mduma, Neema. "Data Balancing Techniques for Predicting Student Dropout Using Machine Learning." Data, vol. 8, no. 3, 27 Feb. 2023, p. 49, https://doi.org/10.3390/data8030049

Nevertheless, some researchers have particularly focused on solutions to solve the data imbalance problem in dropout prediction[31] [32], using methods such as the Synthetic Minority Oversampling Technique (SMOTE)[33]. In the SMOTE algorithm, the minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any or all of the k minority class nearest neighbors. Depending on the amount of over-sampling required, neighbors from the k nearest neighbors are randomly chosen.

Orooji & Chen[32] predicted dropout in a Louisiana Public High School through imbalanced learning techniques. They applied resampling, case weighting, and cost-sensitive learning to enhance the prediction performance on the rare class. Experiments show that application of imbalanced learning methods produces good results on recall but decreases precision, whereas base classifiers without regard of imbalanced data handling gives better precision but poor recall. They concluded that overall application of imbalanced learning techniques was beneficial to the dropout prediction problem.

[31] Márquez-Vera, Carlos, et al. "Early Dropout Prediction Using Data Mining: A Case Study with High School Students." Expert Systems, vol. 33, no. 1, 16 Nov. 2015, pp. 107–124, https://doi.org/10.1111/exsy.12135

[32] Orooji, Marmar, and Jianhua Chen. "Predicting Louisiana Public High School Dropout through Imbalanced Learning Techniques." IEEE Xplore, 1 Dec. 2019, ieeexplore.ieee.org/abstract/document/8999067/. Accessed 19 July 2022.

[33] Chawla, N. V., et al. "SMOTE: Synthetic Minority Over-Sampling Technique." Journal of Artificial Intelligence Research, vol. 16, no. 16, 1 June 2002, pp. 321–357, https://doi.org/10.1613/jair.953.

Baker et al.[34] used over-sampling to predict dropout for K-12 students in Texas. In other words, data points in the rarer category (in this case, students who dropped out) were duplicated several times. They selected the number of duplications that best equalized the number of data points in the rarer and more common categories. Re-sampling was only used on the training sets, and all calculations of model goodness took place in unmodified test sets.

Some researchers, such as Del Bonifro et al.[35], used under-sampling. They predicted drop out for first-year undergraduate students with a very unbalanced dataset (<13% dropout). As a solution, they randomly selected half of the negative samples (i.e., the students who effectively dropped) and used it in the training set. An equal number of instances of the other class was randomly sampled from the dataset and added to the training set. In doing so, they obtained a balanced training set, which was used to train the supervised models. The remaining samples constituted an unbalanced test set which they used to measure the performance of the trained models.

[34] Baker, Ryan S., et al. "Predicting K-12 Dropout." Journal of Education for Students Placed at Risk (JESPAR), vol. 25, no. 1, 1 Oct. 2019, pp. 28–54, https://doi.org/10.1080/10824669.2019.1670065

[35] Del Bonifro, F., Gabbrielli, M., Lisanti, G., Zingaro, S.P. (2020). Student Dropout Prediction. In: Bittencourt, I., Cukurova, M., Muldner, K., Luckin, R., Millán, E. (eds) Artificial Intelligence in Education. AIED 2020. Lecture Notes in Computer Science(), vol 12163. Springer, Cham. https://doi.org/10.1007/978-3-030-52237-7_11

Predictors

One particular caveat of the dropout prediction problem in traditional classrooms is the available data. Usual features mainly include academic data, but prior work suggests that other types of data, such as health data, socioeconomic data, social data, and even personality/motivation data could greatly enhance the models. While there no consistent agreement among different studies regarding the best predictors[36], Allensworth & Easton[37] indicate that the most powerful predictors of whether a student will complete high school include course performance and attendance during the first year of high school. Therefore, they recommend to systematically collect student attendance and course performance data to develop an effective early warning system that can also be tailored to local contexts.

Academic and demographic features are the most widely used in literature, mainly because they are easier to collect and to interpret. Moreover, it is worth noting that researchers sometimes augment their data with external datasets. Adelman et al.[26] looked at dropout prediction in Guatemala and Honduras in the transition from Primary to Lower Secondary School. Their analysis augmented the

[36] Dekker, Gerben, et al. Predicting Students Drop Out: A Case Study. 2009

[37] Allensworth, Elaine, and John Easton. What Matters for Staying On-Track and Graduating in Chicago Public High Schools a Close Look at Course Grades, Failures, and Attendance in the Freshman Year. 2007.

most basic administrative data with additional data to improve the quality of prediction (in the case of Guatemala, with periodic national exam data for primary students; in the case of Honduras, with household survey and census data). However, complexities arise when academic data is enriched with external data. For example, Adelman et al.[26] mentioned that their augmentation could affect the consistency of the prediction from year to year, as the availability of household survey and other data varies over time. Furthermore, it could also make the approach more difficult to replicate for practitioners within ministries of education, who may not be familiar with accessing and analyzing these other data sources.

Geryk et al.[38] used data derived from students social behaviour to enrich their dataset and improve their dropout predictions. Their social data described social dependencies gathered from e-mail and discussion board conversations, among other sources. Their work showed that the use of social behaviour data results in significant increase of the prediction accuracy.

Marquez Vera et al.[14] worked on a very rich dataset that also included social data (e.g. number of friends, study habits), health data (e.g. smoking, alcohol

[38] Bayer, Jaroslav; Bydzovska, Hana; Geryk, Jan; Obsivac, Tomas; Popelinsky, Lubomir. *International Educational Data Mining Society*, Paper presented at the International Conference on Educational Data Mining (EDM) (5th, Chania, Greece, Jun 19-21, 2012)

consumption, disabilities, time in 1000m race) and even personality/motivation data (e.g. type of personality, level of boredom in class, level of motivation) . These factors enhanced predictions, but they are rarely used in common applications mainly because of the complexity to collect them.

Feature importance & Selection

Another challenge of dropout prediction is to understand which factors contribute to the model. Baranyi et al.[39] used Permutation Importance and SHapley Additive exPlanations to this end. Permutation importance can be measured by looking at how much the performance of a trained model decreases when the values of a feature are permuted (shuffled), ceteris paribus. Baranyi et al.[39] first trained a model and calculated its performance. Then, they shuffled the values of one feature, and without re-training the model, they again calculated its performance on the shuffled data and investigated how much the performance worsened. They compared the results of Permutation with SHapley Additive exPlanations (SHAP), which is a state-of-the-art machine learning model explainability tool that is based on the Shapley value, a cooperative game theoretical concept that quantifies the contribution of each player in a coalition.

[39] Baranyi, M., Nagy, M., & Molontay, R. (2020). Interpretable deep learning for university dropout prediction. Proceedings of the 21st Annual Conference on Information Technology Education (pp. 13–19)

Other methods of variable selection, such as LASSO (least absolute shrinkage and selection operator) have been used by Vallabhaneni[40] to predict dropout for students at Kansas State University. Essentially, The LASSO method regularizes model parameters by shrinking the regression coefficients, reducing some of them to zero. The feature selection phase occurs after the shrinkage, where every non-zero value is selected to be used in the model.

Finally, researchers such as Song et al.[41] and Realinho et al.[42] use the built-in feature importance of certain decision tree libraries such as LightGBM, XGBoost, Catboost, and Scikit-learn Random Forest. In this case, the default importance represents the total "gain" of all splits which use the feature, and the gain measures the improvement in accuracy brought by a feature to the branches it is on. In other words, it measures how much the model's accuracy improves when it splits on a certain feature.

Models

Machine learning models have shown promise in predicting student dropout, aiding in the early detection of students who might be veering off their

[40] Vallabhaneni, Teja. Prediction of University Student Attrition Rate Using Ridge and Lasso Regression. 2019.
[41] Song, Zihan, et al. “All-Year Dropout Prediction Modeling and Analysis for University Students.” Applied Sciences, vol. 13, no. 2, 14 Jan. 2023, p. 1143, https://doi.org/10.3390/app13021143.

[42] Realinho, Valentim, et al. “Predicting Student Dropout and Academic Success.” Data, vol. 7, no. 11, 1 Nov. 2022, p. 146, www.mdpi.com/2306-5729/7/11/146, https://doi.org/10.3390/data7110146.

educational path. However, the choice of the appropriate ML model is crucial and can greatly impact the accuracy of these predictions. There are many different types of ML models available, each with its own strengths and weaknesses. These include decision trees, logistic regression, support vector machines, neural networks, ensemble methods like random forests or gradient boosting, and more. The performance of these models can vary depending on the nature and quality of the data at hand.

In the context of student dropout prediction, it is crucial to choose a model that can handle the kind of data typically available in educational settings. As mentioned previously, this data may include demographic information, academic performance, attendance records, and behavioral indicators, among other features. Some models may handle categorical data better, while others might be more suited to continuous or time-series data. Furthermore, some models can better handle missing or imbalanced data, a common occurrence in educational datasets.

Moreover, the interpretability of the model is another significant factor. Education professionals and policymakers who will use the results of the prediction need to understand the factors contributing to the risk of dropout. Some ML models, like decision trees and logistic regression, provide clearer insights into which features

are most influential in the predictions. Other models, like neural networks, may offer better prediction performance but are often seen as "black boxes" due to their complex internal workings. Finally, model choice also affects the feasibility and efficiency of deployment. Some models are more computationally intensive than others, which can be a concern for institutions with limited computational resources.

Oqaidi et al. [43] identified the most commonly used machine learning algorithms in the literature, namely (LR) Logistic Regression, (DT) Decision Tree, (RF) Random Forest, (ANN) Artificial Neural Networks, (kNN) k-Nearest Neighbors, (SVM) Support Vector Machine, (NB) Naive Bayes, and other less used algorithms like Linear regression, AdaBoost, Enhanced ML Algorithm, Linear Discriminant Analysis or eXtremeGBoost.

[43] Oqaidi, K., Aouhassi, S., & Mansouri, K. (2022). Towards a Students' Dropout Prediction Model in Higher Education Institutions Using Machine Learning Algorithms. *International Journal of Emerging Technologies in Learning (iJET)*, *17*(18), pp. 103–117. https://doi.org/10.3991/ijet.v17i18.25567

Decision trees[44] [45] [50] [46] [47]and logistic regression[23] [24] [27] [28] [48] [49] [50] [51] are often used as a baseline, since they usually yield convincing result while being simple and interpretable. Common methods also involve SVM [46] [51] [52] [53] , KNN [48] [54], or Naïve Bayes [46] [51] [53] [54].

However, a variety of researches employ more complex methods that generally generate better results, such as Random Forest [48] **Error! Bookmark not defined.**

[44] K. Limsathitwong, K. Tiwatthanont and T. Yatsungnoen, "Dropout prediction system to reduce discontinue study rate of information technology students," *2018 5th International Conference on Business and Industrial Research (ICBIR)*, Bangkok, Thailand, 2018, pp. 110-114, doi: 10.1109/ICBIR.2018.8391176.

[45] Kemper, Lorenz, et al. "Predicting Student Dropout: A Machine Learning Approach." European Journal of Higher Education, vol. 10, no. 1, 2 Jan. 2020, pp. 28–47, https://doi.org/10.1080/21568235.2020.1718520.

[46] Zhang, Ying & Oussena, Samia & Clark, Tony & Kim, Hyeonsook. (2010). Use Data Mining to Improve Student Retention in Higher Education - A Case Study.. 190-197.

[47] M. Orooji and J. Chen, "Predicting Louisiana Public High School Dropout through Imbalanced Learning Techniques," *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, 2019, pp. 456-461, doi: 10.1109/ICMLA.2019.00085.

[48] Aulck, L.; Velagapudi, N.; Blumenstock, J.; West, J. Predicting Student Dropout in Higher Education. ICML Workshop on #Data4Good: Machine Learning in Social Good Applications 2016. *arXiv* **2017**

[49] Saranya, A.; Rajeswari, J. Enhanced Prediction of Student Dropouts Using Fuzzy Inference System and Logistic Regression. *ICTACT J. Soft Comput.* **2016**, *6*, 1157–1162

[50] Jadrić, Mario, and Željko Garača. STUDENT DROPOUT ANALYSIS with APPLICATION of DATA MINING METHODS Maja Ćukušić. 2010.

[51] Rovira, S.; Puertas, E.; Igual, L. Data-driven System to Predict Academic Grades and Dropout. *PLoS ONE* **2017**, *12*, e0171207

[52] Fernandez-Garcia, Antonio Jesus, et al. "A Real-Life Machine Learning Experience for Predicting University Dropout at Different Stages Using Academic Data." IEEE Access, vol. 9, 2021, pp. 133076–133090, https://doi.org/10.1109/access.2021.3115851

[53] Ara, Nicolae-Bogdan, et al. High-School Dropout Prediction Using Machine Learning a Danish Large-Scale Study. 2015.

[54] Hutagaol, Nindhia, and Suharjito Suharjito. "Predictive Modelling of Student Dropout Using Ensemble Classifier Method in Higher Education." Advances in Science, Technology and Engineering Systems Journal, vol. 4, no. 4, 2019, pp. 206–211, https://doi.org/10.25046/aj040425

[55] [56], Gradient Boosted Trees [41] [42] **Error! Bookmark not defined.**, or Neural Networks [50] [57] [58] [59].

Finally, a few studies have used a survival analysis approach [60] [61] [62] [63], which analyzes time until dropout by estimating the probability of dropout at a given time using techniques such as Kaplan-Meier estimation or Cox regression.

Data

[55] Sales, A.; Balby, L.; Cajueiro, A. Exploiting Academic Records for Predicting Student Drop Out: A case study in Brazilian higher education. *J. Inf. Data Manag.* **2016**, *7*, 166–180.

[56] Halland, R.; Igel, C.; Alstrup, S. High-School Dropout Prediction Using Machine Learning: A Danish Large-scale Study. In Proceedings of the 23rd European Symposium on Artificial Neural Networks, Bruges, Belgium, 22–23 April 2015; pp. 22–24.

[57] Siri, A.; Siri, A. Predicting Students' Dropout at University Using Artificial Neural Networks. *Ital. J. Sociol. Educ.* **2015**, *7*, 225–247.

[58] Oancea, B.; Dragoescu, R.; Ciucu, S. Predicting Students' Results in Higher Education Using Neural Networks. In Proceedings of the International Conference on Applied Information and Communication Technologies, Baku, Azerbaijan, 23–25 October 2013; pp. 190–193.

[59] Kiss, Botond, et al. "Predicting Dropout Using High School and First-Semester Academic Achievement Measures." IEEE Xplore, 1 Nov. 2019, ieeexplore.ieee.org/abstract/document/9040158/. Accessed 19 July 2022.

[60] Masci, Chiara, et al. "SURVIVAL MODELS for PREDICTING STUDENT DROPOUT at UNIVERSITY across TIME." Education and New Developments 2022 – Volume I, 17 June 2022, end-educationconference.org/wp-content/uploads/2022/07/2022v1end043.pdf, https://doi.org/10.36315/2022v1end043

[61] Ameri, Sattar, et al. "Survival Analysis Based Framework for Early Prediction of Student Dropouts." Proceedings of the 25th ACM International on Conference on Information and Knowledge Management - CIKM '16, 2016, https://doi.org/10.1145/2983323.2983351. Accessed 30 Nov. 2019.

[62] Weybright EH, Caldwell LL, Xie HJ, Wegner L, Smith EA. Predicting secondary school dropout among South African adolescents: A survival analysis approach. S Afr J Educ. 2017 May;37(2):1353. doi: 10.15700/saje.v37n2a1353. PMID: 30287979; PMCID: PMC6168088.

[63] Survival Analysis based Framework for Early Prediction of Student Dropouts Authors: Sattar Ameri , Mahtab J. Fard , Ratna B. Chinnam , Chandan K. Reddy

In 2013, the Moroccan government launched MASSAR, an education management information system which aimed to provide every student in the country with a unique identification number. Using a dedicated website or mobile application, teachers and school directors enter data such as age, gender, and performance for each student. This unique ID number stays with the student throughout their entire schooling, allowing them to be tracked even if they move or migrate to a different school. In addition to student information, MASSAR also includes data on schools, such as the municipality, number of classes per level, and teachers per class. As part of our research, we were granted access to anonymous MASSAR data by the Moroccan Ministry of Education, augmented with administrative data, for the region of Fes - Meknes from 2015 to 2019.

Our dataset contains data for **336,135 students** in **middle-school (Public)** and **high school (Public)**, enrolled in the academic year **2017-2018**. Variables are collected from the academic year **2015-2016** to **2018-2019** at the **semester level,** and include academic performance, absences, special needs (e.g. handicap), socio-economic indicators (e.g. mother/father profession), demographic variables, and school characteristics. The full list can be found in **Appendix A**. As expected, the dataset exhibited class imbalance, as evidenced by a dropout rate of **10.5%,** as shown on **Figure 1**.

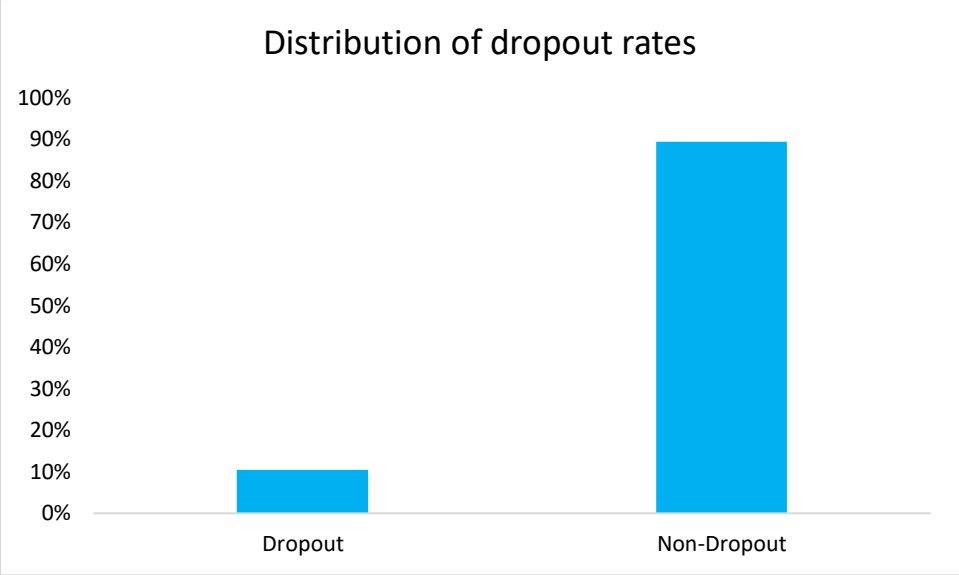

Figure 1 : Distribution of dropout rates

Methods

Objective

We treat this problem as a binary classification problem. The task is to predict whether a student enrolled in the academic year 2017-2018 will dropout for the subsequent academic year. Our main goal is to flag the maximum number of potential dropouts in order to rightly intervene before they make this critical decision. However, inaccurately flagging non-dropouts as dropouts could lead to increased costs and complexity that we would also like to avoid, as far as possible. Finally, we believe that false negatives (inaccurately flagging a dropout as a non-dropout) should cost much more than false positives (inaccurately flagging a non-

dropout as a dropout). Moreover, our predictions simulate real-life settings : for all students present in the second semester of the academic year 2017-2018, we predict whether they will drop out in the following academic year. This method aims to emulate scenarios wherein educational administrators conduct retrospectives of students' annual progress, scrutinizing their academic achievements, conduct, and incidences of absenteeism upon conclusion of the second semester.

To account for our belief that false negatives are more costly than false positives, we place more weight on recall than on precision. One way to do this is to use an Fβ score variant that places more weight on recall, such as the F2 score. In our research, **Precision** is the fraction correctly flagged dropouts among all flagged students, and **Recall** is the fraction of flagged true dropouts among all actual dropouts. Note that, even though we will calculate accuracy, we believe that it is not the greatest metric because our dataset is heavily imbalanced (almost 90% are non-dropouts). Hence, one could get a high accuracy just by flagging all students as non-dropouts, which does not bring real value to our model.

Based on our review of related work, we have decided to try 5 different algorithms : Logistic Regression, Random Forest, XGBoost, LightGBM, and Catboost. Note that we refrain from using Neural Networks even if they usually perform well. This

choice stems from two simple facts. First, existing research has indicated that gradient boosted trees generally achieve comparable or superior performance compared to Neural Networks, even outside the student dropout prediction realm [64]. Then, we believe that it is also important to keep the context in mind. Ultimately, this predictive model is aimed for school administrators with no prior knowledge in machine learning. In this regard, we contend that while approaches like gradient boosted trees may be perceived as complex when explained to a non-technical audience, the concept of decision trees is comparatively more straightforward and visual than neural networks. This simplicity is pivotal for widespread adoption of the model and demystifying its implementation. This is further highlighted by Bowers[65], who defines accessibility as one of the core components of an Early Warning System. According to Bowers[65] the algorithm should be accessed, examined, and understood. In this context, accessible is the opposite of proprietary, hidden, or machine learned algorithms that obfuscate how the prediction takes place, but instead is open, public, and understandable.

Data Preprocessing

[64] Giannakas, Filippos, et al. "XGBoost and Deep Neural Network Comparison: The Case of Teams' Performance." Intelligent Tutoring Systems, 2021, pp. 343–349, https://doi.org/10.1007/978-3-030-80421-3_37.

[65] Bowers, A. J. (2021). Early warning systems and indicators of dropping out of upper secondary school: The emerging role of digital technologies. Teachers College, Columbia University. www.oecd-llibrary.org, www.oecd-ilibrary.org/sites/c8e57e15-en/index.html?itemId=/content/component/c8e57e15-en

Data cleaning and preprocessing involved a multitude of steps before being able to use the data. The main steps include, but are not limited to :

- Creating the label : The dropout label was not directly available, but was created using the strategy outlined in the initial report describing the objective of MASSAR[2].
- Correcting for obvious outliers that fall outside the range of reasonable values .
- Converting each variable to an appropriate type.
- Generating new categories for categorical variables when appropriate.
- Creating lagged variables to get 1 row per student and keeping information for each semester as columns.
- Filling missing values with the median for numeric variables, and with the mode for categorical variables.
- Engineering features such as class rank (percentile), percentage of female students in the class.
- Treating columns that mixed Arabic and French. For example, mother/father profession are information directly filled by parents. It can include spelling mistakes, different languages, and different words with similar meanings, such as housewife or unemployed. For this, we used GPT-3 with Few-Shot Learning Prompting in order to create clusters of jobs.

Algorithms

Logistic regression

Logistic Regression is a widely used statistical model in machine learning for binary classification tasks. It estimates the probability of a binary outcome based on input variables. It assumes that the log odds of the output variable are a linear function of the input variables. The model uses a sigmoid function to transform the linear combination of inputs into a probability between 0 and 1. The coefficients of the model are estimated using maximum likelihood estimation, which maximizes the likelihood of the observed data. Logistic Regression is advantageous due to its simplicity, interpretability, and ability to handle different types of variables. However, it assumes a linear relationship between inputs and the log odds, which may not always be suitable for complex data.

Random forest

Random Forest is an ensemble learning technique used in machine learning for both classification and regression tasks. It is based on the idea of combining multiple decision trees to improve the accuracy and reduce the risk of overfitting. In Random Forest, a large number of decision trees are created, each with a random subset of features and a random subset of training data. Each tree is trained independently using a random subset of the input variables, and the final

result is a combination of the predictions of all the trees. To make a prediction using Random Forest, the input data is passed through all the trees in the forest, and the output of each tree is considered. The final prediction is then made by taking the majority vote (in classification tasks) of the predictions from all the trees. Random Forest has several advantages over a single decision tree. It is less prone to overfitting as it combines the predictions of many trees, each with different subsets of features and training data. It is also more accurate than a single decision tree, as it reduces the variance and bias of the model. Additionally, it can handle both categorical and numerical data and is robust to missing data.

However, Random Forest can be computationally expensive, especially for large datasets with many features. It also requires careful selection of hyperparameters, such as the number of trees, the maximum depth of each tree, and the size of the feature subset.

Gradient boosting (XGBoost, LightGBM, Catboost)

Gradient Boosting is a type of ensemble learning that combines multiple weak models into a single strong model. In Gradient Boosting, a series of decision trees are constructed sequentially, each tree learning from the errors made by the previous tree. The first decision tree is trained on the original dataset, and the subsequent trees are trained on the residuals (or errors) of the previous trees.

The process of building the model involves minimizing a loss function, typically using gradient descent optimization. At each step, the model calculates the gradient of the loss function with respect to the output of the previous tree and trains the next tree to approximate this gradient. The output of each tree is then added to the previous output, with a learning rate (a hyperparameter that controls the contribution of each tree) to prevent overfitting. The final prediction of the Gradient Boosting model is the sum of the predictions of all the trees.

Gradient Boosting has several advantages over other machine learning models, including high accuracy, robustness to outliers, and ability to handle both numerical and categorical data. However, it can be sensitive to hyperparameters and prone to overfitting, especially with a large number of trees. To prevent overfitting, techniques such as early stopping, regularization, and cross-validation are used. XGBoost[66], CatBoost[67], and LightGBM[68] are three popular gradient boosting frameworks used for supervised learning tasks such as classification. They are designed to improve the accuracy and speed of traditional gradient boosting methods.

Results

[66] Choi, DK. Data-Driven Materials Modeling with XGBoost Algorithm and Statistical Inference Analysis for Prediction of Fatigue Strength of Steels. Int. J. Precis. Eng. Manuf. 20, 129–138 (2019). https://doi.org/10.1007/s12541-019-00048-6

[67] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. arXiv preprint arXiv:1706.09516.

[68] Ke, Guolin, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. 2017.

Firstly, we divide the entire dataset into training (80%), validation (10%), and testing (10%) datasets using the train_test_split function from sklearn.model_selection. This separation allows us to not only train our model but also validate and test its performance on unseen data.

As mentioned earlier, the class imbalance can bias our machine learning model towards the majority class, and using the imbalance ratio as a parameter can help us tackle this issue. Thus, for all models, we test both the Synthetic Minority Oversampling Technique (SMOTE) and adjusting class weights to mitigate class imbalance. Note that we only apply SMOTE on the training set after the split, since using SMOTE before the split could have resulted in data leakage, which is a highly undesirable outcome. While Logistic Regression performed similarly with both methods, all tree methods performed significantly better with class weights adjustments compared to SMOTE.

To find the best combination of hyperparameters, we use Grid Search Cross Validation (GridSearchCV). This technique performs an exhaustive search over the specified parameter values for an estimator. It divides the validation set into a certain number of 'folds' (5 in this case), trains the model on these, and then evaluates it. This process is repeated for all the combinations of hyperparameters,

and the one that gives the best performance (based on a specific scoring metric, in this case, F2 score) is chosen.

Finally, after we've found the best hyperparameters, we create a new model using these parameters and fit it on the training data. This is our final model. We then use this model to predict the dropout status for our test data and evaluate the model's performance using a classification report.. The classification report provides key metrics such as accuracy, precision, recall, F2 score, macro-average precision, and macro-average recall.

As a reminder, the following are formulas for the metrics measured :

Accuracy : Number of correct predictions / Total number of predictions

Precision : TP / (TP + FP)

Recall : TP / (TP + FN)

F2 Score : 5 * (precision * recall) / (4 * precision + recall)

Macro-average precision : Arithmetic mean of individual classes' precision

Macro-average recall : Arithmetic mean of individual classes' recall

Figure 2 summarizes the best results obtained for each model.

Model Name	Accuracy	Macro average precision	Macro average Recall	Precision	Recall	F2 score
Logistic Regression	83.1%	0.66	0.81	0.35	0.79	0.63
Random Forest	90.3%	0.74	0.83	0.51	0.73	0.67
XGBoost	88.3%	0.72	0.85	0.45	0.81	0.70
LightGBM	86.6%	0.70	0.85	0.41	0.84	0.70
CatBoost	89.8%	0.73	0.86	0.49	0.81	0.72

Figure 2 : Prediction metrics results

Results show that **Logistic Regression** does fairly well in terms of recall (0.79), meaning it correctly identifies a high percentage of actual student dropouts, but that its precision is low (0.35), indicating that it also predicts many false positives. This means it might wrongly identify quite a few students as probable dropouts.

Random Forest has the highest accuracy (90.3%), a reasonable recall (0.73), and significantly better precision (0.51) than the Logistic Regression model. This suggests that it balances both correctly identifying student dropouts and reducing false alarms.

XGBoost has a good balance between precision (0.45) and recall (0.81), and also has a high F2 score (0.70), indicating a balance between precision and recall, with

more weight on recall. Thus, it might be a good choice when identifying as many dropouts as possible is critical, even if it means having more false positives.

LightGBM's recall is the highest among all models (0.84), suggesting it's most effective at identifying actual student dropouts. However, its precision (0.41) is relatively low, implying it may incorrectly flag many non-dropout students.

Finally, **CatBoost** has a strong balance between accuracy, recall, and precision, with the highest F2 score (0.72) among the models. Catboost could arguably be our best choice since we're looking for an overall balanced model that correctly identifies dropouts (high recall) while keeping false positives reasonably low (moderate precision).

As mentioned previously, while **Random Forest** has the highest accuracy, the latter is not a great metric in the context of class imbalance, since flagging everyone as non-dropouts would result in a 89.5% accuracy. In these settings, it is more important to focus on Recall and F2-Score, since we focus on correctly identifying potential dropouts while keeping a moderate precision.

In our case, an average precision is not problematic. If we take LightGBM for example, we can correctly identify 84% of potential dropouts while only filtering

19% of the dataset. Note that a completely perfect model would flag 10.5% of the dataset. Furthermore, it is worth noting the temporal aspect of our prediction : we precisely predict whether the student will drop out in T+1, but it is also possible that the student drops out in T+2, T+3.. T+N, which means that the students appearing as misidentified might actually be at risk for later years. This further means that moderate precision is not problematic in this case.

Since this predictive model aims to emulate scenarios wherein educational administrators conduct retrospectives of students' annual progress upon conclusion of the second semester, it could also be coupled with a student dashboard which displays the evolution of grades, absences, and other indicators that can help school administrators to judge the gravity of the situation at the individual level.

Ultimately, beyond the realm of predictions, we aim to discern the pivotal elements that could potentially influence the prediction of student dropout. **Figure 3** shows the feature importance of LightGBM, which is our model with the highest recall and a moderate precision. This plot provides insight into which features have the most impact on the predictions made by the model.

Note that we normalize feature importance, so that it sums to 1, and we sum up the importance variables belonging to a similar group (e.g. lagged variables, units of absences and days of absences).

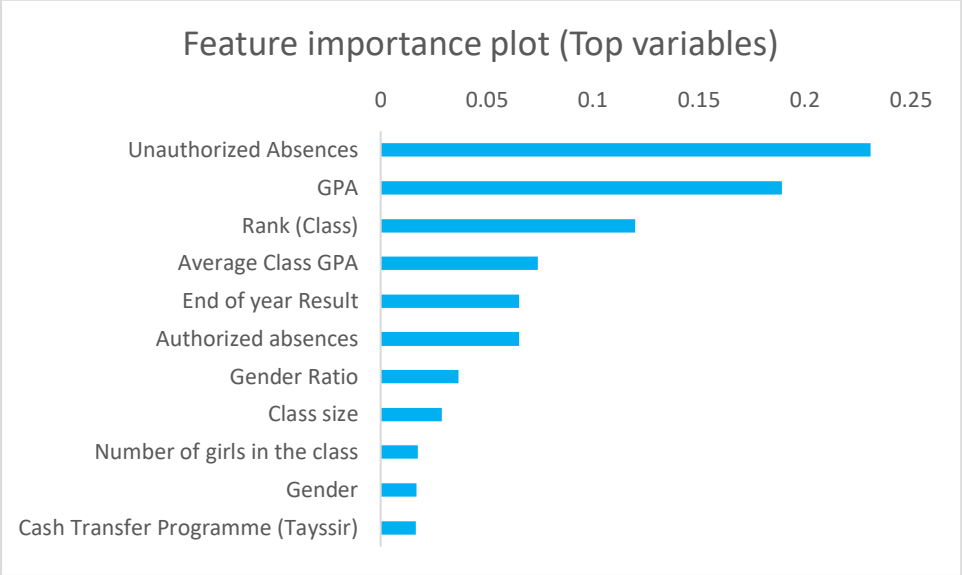

Figure 3 : LightGBM variable importance (top variables)

The feature importance plot perfectly corresponds to what prior work has shown. Academic performance (GPA, class rank, end of year result), as well as absences and gender, play a significant role in the prediction of dropout. Perhaps more interestingly, the gender ratio (% of girls), class size, and Tayssir (Cash Transfer Programme to combat absenteeism) are also among the most important variables. School characteristics, which are not present on this plot considering their low importance, do not seem to play a huge role compared to the variables

outlined in **Figure 3**. An interesting point to mention is that unauthorized absences are a much better predictor than authorized absences, which completely makes sense from a contextual point of view. It is worth noting that while the feature importance plots gives great insights, the study is still unable to draw any causal effect. This means that we acknowledge the importance of these predictors, but that we cannot guarantee that they directly cause dropout.

Conclusion and future work

In this research, we developed an Early Warning System to accurately predict student dropout for middle school and high school students in Morocco, and to shed light on important variables that could support student dropout prediction.

The primary constraint of the model was the absence of certain features, notably student behavior, which could have considerably amplified the model's effectiveness. Furthermore, the imbalance in classes posed a significant challenge. Subsequent investigations could direct their attention to predicting dropout rates in primary schools, where data imbalance is markedly more pronounced (~3%). Despite the study’s success in revealing insights regarding key predictive factors, it falls short of establishing causal relationships.

Finally, while the focus has been on the specifics of machine learning models and their accuracy, this step is only one small part of the vast system in education organisations that can be leveraged to help support student success, as exemplified in the work in the Chicago context in the United States [65]. The Chicago on-track indicator, recognized for its reliability and accuracy, serves as a cross-sectional early warning sign for potential student dropouts. Over the past twenty years, Chicago has witnessed a remarkable surge in graduation rates, escalating from 52.4% in 1998 to more than 90% by 2019. Yet, as highlighted by Chicago-based researchers, the mere identification of accurate on-track indicators and their incorporation into an early warning system doesn't directly result in improvements, as it addresses only a fraction of the complexities associated with school dropouts. Instead, the early warning system forms a minor part of a comprehensive array of systems, which when coupled with active educational measures, supplies valuable data that assists educators in customizing interventions for students.

Ultimately, the technology of early warning systems and indicators constitutes a minor, albeit significant, segment of the extensive system of data utilization in educational institutions. This system necessitates active community engagement, , and an unwavering commitment to promoting student success via individualized interventions. The creation of methods and instruments that convert data into actionable insights represents merely a stride in the direction of impactful actions

that foster learning enhancement and student success. However, it is an encouraging avenue that, with the assistance of digitization and breakthroughs in data mining and analytics, should feasibly be sustainable in the impending future of Morocco.

References

- Dahir No. 1-19-113 of Hijra 1440 (August 9, 2019) promulgating Framework Law No. 51-17 on the Education, Training, and Scientific Research System.

- Conseil Supérieur de l'Éducation et de la Recherche Scientifique. "ATLAS TERRITORIAL DE L'ABANDON SCOLAIRE." https://www.csefrs.ma/wp-content/uploads/2019/12/ATLAS-TERRITORIAL-DE-LABANDON-SCOLAIRE-18-12-web.pdf

- Rumberger, Russell, and Sun Lim. Why Students Drop out of School: A Review of 25 Years of Research. 2008.

- Attijariwafa Bank. (2017). Pole Edition et Débats: Conférences. Retrieved from https://www.attijariwafabank.com/sites/default/files/widgets/files/17-00328-book-pole_edition_et_debats-conferencesa447degedition_4.pdf

- Baker, R. S., & Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining*, *1*(1), 3–17. https://doi.org/10.5281/zenodo.3554657

- *Long*, P., *Siemens*, G., "Penetrating the Fog: Analytics in Learning and Education, Educause Review, 46 (5), 31-40, *2011*.

- Siemens, George & Baker, Ryan. (2012). Learning analytics and educational data mining: Towards communication and collaboration. ACM International Conference Proceeding Series. 10.1145/2330601.2330661.

- Romero, C. and Ventura, S. (2013) Data Mining in Education. WIREs Data Mining and Knowledge Discovery, 3, 12-27. https://doi.org/10.1002/widm.1075

- Baker, R.S., Inventado, P.S. (2014). Educational Data Mining and Learning Analytics. In: Larusson, J., White, B. (eds) Learning Analytics. Springer, New York, NY. https://doi.org/10.1007/978-1-4614-3305-7_4

- Heppen, Jessica, and Susan Bowles Therriault. ISSUE BRIEF Developing Early Warning Systems to Identify Potential High School Dropouts. 2008.

- Seidman, A. (1996). Retention revisited: R = E, Id+E & In, Iv. College and University, 71(4), 18-20.

- Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Fardoun, H. M., & Ventura, S. (2015). Early dropout prediction using data mining: A case study with high school students. Expert Systems, 32(5), 3-14. DOI: 10.1111/exsy.12135

- Maldonado-Ulloa P.Y., Sancén-Rodríguez A.J., Torres-Valades M., Murillo-Pazarán B. (2011) Secretaria de Educación Pública de Mexico. Programa Síguele. Sistema de Alerta Temprana, Lineamientos de Operación. 1–18.

- Uekawa, Kazuaki, et al. REL Mid-Atlantic Technical Assistance Brief REL MA 1. No. 2, 2010, pp. 75–85, files.eric.ed.gov/fulltext/ED565682.pdf.

- Vassiliou, A. 2013. Early warning systems in Europe: practice, methods and lessons. Thematic Working Group on Early School Leaving. 1-17.

- Márquez-Vera, Carlos, et al. “Early Dropout Prediction Using Data Mining: A Case Study with High School Students.” Expert Systems, vol. 33, no. 1, 16 Nov. 2015, pp. 107–124, https://doi.org/10.1111/exsy.12135.

- Kovacic, Z. Early Prediction of Student Success: Mining Students Enrolment Data. 2017, www.semanticscholar.org/paper/Early-Prediction-of-Student-Success%3A-Mining-Data-Kovacic/e48eba98bde33586c20442d46ab9a59c411196e5.

- Romero, C. and Ventura, S. (2013) Data Mining in Education. WIREs Data Mining and Knowledge Discovery, 3, 12-27. https://doi.org/10.1002/widm.1075

- Baker, R. S., & Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining*, *1*(1), 3–17. https://doi.org/10.5281/zenodo.3554657

- Magaña, Marina, and Hernández Pediatra. XIII Congreso de La Sociedad Española de Medicina Del Adolescente 1a Mesa Redonda CAUSAS DEL FRACASO ESCOLAR. 2002.

- Baker, Ryan S., et al. “Predicting K-12 Dropout.” Journal of Education for Students Placed at Risk (JESPAR), vol. 25, no. 1, 1 Oct. 2019, pp. 28–54, https://doi.org/10.1080/10824669.2019.1670065.

- Knowles, Jared E. “Of Needles and Haystacks: Building an Accurate Statewide Dropout Early Warning System in Wisconsin.” Journal of Educational Data Mining, vol. 7, no. 3, 25 July 2015, pp. 18–67, jedm.educationaldatamining.org/index.php/JEDM/article/view/JEDM082, https://doi.org/10.5281/zenodo.3554725

- Chung, Jae Young, and Sunbok Lee. “Dropout Early Warning Systems for High School Students Using Machine Learning.” Children and Youth Services Review, vol. 96, Jan. 2019, pp. 346–353, https://doi.org/10.1016/j.childyouth.2018.11.030.

- Adelman, Melissa, et al. “Predicting School Dropout with Administrative Data: New Evidence from Guatemala and Honduras.” Education Economics, vol. 26, no. 4, 2 Feb. 2018, pp. 356–372, https://doi.org/10.1080/09645292.2018.1433127

- Mnyawami, Yuda N., et al. “Enhanced Model for Predicting Student Dropouts in Developing Countries Using Automated Machine Learning Approach: A Case of Tanzanian’s Secondary Schools.” Applied Artificial Intelligence, vol. 36, no. 1, 7 May 2022, https://doi.org/10.1080/08839514.2022.2071406

- N. Mduma and D. Machuve, "Machine Learning Model for Predicting Student Dropout: A Case of Tanzania, Kenya and Uganda," 2021 IEEE AFRICON, Arusha, Tanzania, United Republic of, 2021, pp. 1-6, doi: 10.1109/AFRICON51333.2021.9570956.

- Weybright, Elizabeth H., et al. “Predicting Secondary School Dropout among South African Adolescents: A Survival Analysis Approach.” South African Journal of Education, vol. 37, no. 2, 31 May 2017, pp. 1–11, www.scielo.org.za/scielo.php?script=sci_arttext&pid=S0256-01002017000200006, https://doi.org/10.15700/saje.v37n2a1353

- Mduma, Neema. “Data Balancing Techniques for Predicting Student Dropout Using Machine Learning.” Data, vol. 8, no. 3, 27 Feb. 2023, p. 49, https://doi.org/10.3390/data8030049

- Márquez-Vera, Carlos, et al. “Early Dropout Prediction Using Data Mining: A Case Study with High School Students.” Expert Systems, vol. 33, no. 1, 16 Nov. 2015, pp. 107–124, https://doi.org/10.1111/exsy.12135

- Orooji, Marmar, and Jianhua Chen. “Predicting Louisiana Public High School Dropout through Imbalanced Learning Techniques.” IEEE Xplore, 1 Dec. 2019, ieeexplore.ieee.org/abstract/document/8999067/.

- Chawla, N. V., et al. “SMOTE: Synthetic Minority Over-Sampling Technique.” Journal of Artificial Intelligence Research, vol. 16, no. 16, 1 June 2002, pp. 321–357, https://doi.org/10.1613/jair.953.

- Baker, Ryan S., et al. “Predicting K-12 Dropout.” Journal of Education for Students Placed at Risk (JESPAR), vol. 25, no. 1, 1 Oct. 2019, pp. 28–54, https://doi.org/10.1080/10824669.2019.1670065

- Del Bonifro, F., Gabbrielli, M., Lisanti, G., Zingaro, S.P. (2020). Student Dropout Prediction. In: Bittencourt, I., Cukurova, M., Muldner, K., Luckin, R., Millán, E. (eds) Artificial Intelligence in Education. AIED 2020. Lecture Notes in Computer Science(), vol 12163. Springer, Cham. https://doi.org/10.1007/978-3-030-52237-7_11

- Dekker, Gerben, et al. Predicting Students Drop Out: A Case Study. 2009

- Allensworth, Elaine, and John Easton. What Matters for Staying On-Track and Graduating in Chicago Public High Schools a Close Look at Course Grades, Failures, and Attendance in the Freshman Year. 2007.

- Bayer, Jaroslav; Bydzovska, Hana; Geryk, Jan; Obsivac, Tomas; Popelinsky, Lubomir. *International Educational Data Mining Society*, Paper presented at the International Conference on Educational Data Mining (EDM) (5th, Chania, Greece, Jun 19-21, 2012)

- Baranyi, M., Nagy, M., & Molontay, R. (2020). Interpretable deep learning for university dropout prediction. Proceedings of the 21st Annual Conference on Information Technology Education (pp. 13–19)

- [1] Vallabhaneni, Teja. Prediction of University Student Attrition Rate Using Ridge and Lasso Regression. 2019.

- Song, Zihan, et al. “All-Year Dropout Prediction Modeling and Analysis for University Students.” Applied Sciences, vol. 13, no. 2, 14 Jan. 2023, p. 1143, https://doi.org/10.3390/app13021143.

- Realinho, Valentim, et al. “Predicting Student Dropout and Academic Success.” Data, vol. 7, no. 11, 1 Nov. 2022, p. 146, www.mdpi.com/2306-5729/7/11/146, https://doi.org/10.3390/data7110146.

- Oqaidi, K., Aouhassi, S., & Mansouri, K. (2022). Towards a Students’ Dropout Prediction Model in Higher Education Institutions Using Machine Learning Algorithms. *International Journal of Emerging Technologies in Learning (iJET)*, *17*(18), pp. 103–117. https://doi.org/10.3991/ijet.v17i18.25567

- K. Limsathitwong, K. Tiwatthanont and T. Yatsungnoen, "Dropout prediction system to reduce discontinue study rate of information technology students," *2018 5th International Conference on Business and Industrial Research (ICBIR)*, Bangkok, Thailand, 2018, pp. 110-114, doi: 10.1109/ICBIR.2018.8391176.

- Kemper, Lorenz, et al. “Predicting Student Dropout: A Machine Learning Approach.” European Journal of Higher Education, vol. 10, no. 1, 2 Jan. 2020, pp. 28–47, https://doi.org/10.1080/21568235.2020.1718520.

- Zhang, Ying & Oussena, Samia & Clark, Tony & Kim, Hyeonsook. (2010). Use Data Mining to Improve Student Retention in Higher Education - A Case Study.. 190-197.

- M. Orooji and J. Chen, "Predicting Louisiana Public High School Dropout through Imbalanced Learning Techniques," *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, 2019, pp. 456-461, doi: 10.1109/ICMLA.2019.00085.

- Aulck, L.; Velagapudi, N.; Blumenstock, J.; West, J. Predicting Student Dropout in Higher Education. ICML Workshop on #Data4Good: Machine Learning in Social Good Applications 2016. *arXiv* **2017**

- Saranya, A.; Rajeswari, J. Enhanced Prediction of Student Dropouts Using Fuzzy Inference System and Logistic Regression. *ICTACT J. Soft Comput.* **2016**, *6*, 1157–1162

- Jadrić, Mario, and Željko Garača. STUDENT DROPOUT ANALYSIS with APPLICATION of DATA MINING METHODS Maja Ćukušić. 2010.

- Rovira, S.; Puertas, E.; Igual, L. Data-driven System to Predict Academic Grades and Dropout. *PLoS ONE* **2017**, *12*, e0171207

- Fernandez-Garcia, Antonio Jesus, et al. "A Real-Life Machine Learning Experience for Predicting University Dropout at Different Stages Using Academic Data." IEEE Access, vol. 9, 2021, pp. 133076–133090, https://doi.org/10.1109/access.2021.3115851

- Ara, Nicolae-Bogdan, et al. High-School Dropout Prediction Using Machine Learning a Danish Large-Scale Study. 2015.

- Hutagaol, Nindhia, and Suharjito Suharjito. "Predictive Modelling of Student Dropout Using Ensemble Classifier Method in Higher Education." Advances in Science, Technology and Engineering Systems Journal, vol. 4, no. 4, 2019, pp. 206–211, https://doi.org/10.25046/aj040425

- Sales, A.; Balby, L.; Cajueiro, A. Exploiting Academic Records for Predicting Student Drop Out: A case study in Brazilian higher education. *J. Inf. Data Manag.* **2016**, *7*, 166–180.

- Halland, R.; Igel, C.; Alstrup, S. High-School Dropout Prediction Using Machine Learning: A Danish Large-scale Study. In Proceedings of the 23rd European Symposium on Artificial Neural Networks, Bruges, Belgium, 22–23 April 2015; pp. 22–24.

- Siri, A.; Siri, A. Predicting Students' Dropout at University Using Artificial Neural Networks. *Ital. J. Sociol. Educ.* **2015**, *7*, 225–247.

- Oancea, B.; Dragoescu, R.; Ciucu, S. Predicting Students' Results in Higher Education Using Neural Networks. In Proceedings of the International Conference on Applied Information and Communication Technologies, Baku, Azerbaijan, 23–25 October 2013; pp. 190–193.

- Kiss, Botond, et al. "Predicting Dropout Using High School and First-Semester Academic Achievement Measures." IEEE Xplore, 1 Nov. 2019, ieeexplore.ieee.org/abstract/document/9040158/. Accessed 19 July 2022.

- Masci, Chiara, et al. “SURVIVAL MODELS for PREDICTING STUDENT DROPOUT at UNIVERSITY across TIME.” Education and New Developments 2022 – Volume I, 17 June 2022, end-educationconference.org/wp-content/uploads/2022/07/2022v1end043.pdf, https://doi.org/10.36315/2022v1end043

- Ameri, Sattar, et al. “Survival Analysis Based Framework for Early Prediction of Student Dropouts.” Proceedings of the 25th ACM International on Conference on Information and Knowledge Management - CIKM ’16, 2016, https://doi.org/10.1145/2983323.2983351. Accessed 30 Nov. 2019.

- Weybright EH, Caldwell LL, Xie HJ, Wegner L, Smith EA. Predicting secondary school dropout among South African adolescents: A survival analysis approach. S Afr J Educ. 2017 May;37(2):1353. doi: 10.15700/saje.v37n2a1353. PMID: 30287979; PMCID: PMC6168088.

- Survival Analysis based Framework for Early Prediction of Student Dropouts Authors: Sattar Ameri , Mahtab J. Fard , Ratna B. Chinnam , Chandan K. Reddy

- Giannakas, Filippos, et al. “XGBoost and Deep Neural Network Comparison: The Case of Teams’ Performance.” Intelligent Tutoring Systems, 2021, pp. 343–349, https://doi.org/10.1007/978-3-030-80421-3_37.

- Bowers, A. J. (2021). Early warning systems and indicators of dropping out of upper secondary school: The emerging role of digital technologies. Teachers College, Columbia University. www.oecd-ilibrary.org, www.oecd-ilibrary.org/sites/c8e57e15-en/index.html?itemId=/content/component/c8e57e15-en

- Choi, DK. Data-Driven Materials Modeling with XGBoost Algorithm and Statistical Inference Analysis for Prediction of Fatigue Strength of Steels. Int. J. Precis. Eng. Manuf. 20, 129–138 (2019). https://doi.org/10.1007/s12541-019-00048-6

- Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. arXiv preprint arXiv:1706.09516.

- Ke, Guolin, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. 2017.

Appendix

Appendix A : Features

Variable	Description
Student ID (Cat)	Student ID
Year ID (Cat)	Year ID (4 years)
Level ID (Cat)	ID corresponding to the current level of the student.
School ID (Cat)	School ID
Class ID (Cat)	Class ID
End of Year result (Cat)	Result ID (e.g. pass, fail..)
GPA (Num)	Mean grade for tests throughout the semester
Session ID (Cat)	1 for 1st semester, 2 for 2nd semester
Class Rank (Num)	Rank in the class (percentile)
Gender ratio (Num)	% of women in the class
Authorized days of absences (Num)	Number of days missed (authorized)
Authorized units of absences (Num)	Number of classes missed. 1 class = 1 unit. (authorized)

Unauthorized days of absences (Num)	Number of days missed (unauthorized)
Unauthorized units of absences (Num)	Number of classes missed. 1 class = 1 unit. (unthorized)
MCaRtable (Bool)	Whether the student benefits from the Cartable program, which distributed 1 million backpacks throughout the country.
Tayssir (Bool)	Whether the student benefits from Tayssir, which is a monetary compensation for students to stay at school.
Scholarship ID (Cat)	Scholarship situation
Gender ID (Cat)	1 for male, 2 for female
Citizenship (Cat)	Citizenship
Place of birth (Cat)	Place of birth
Handicap ID (Cat)	Handicap ID
Type of preschool (Cat)	Type of preschool (religious, modern, none)
Profession of the father (Cat)	Profession of the father
Profession of the mother (Cat)	Profession of the mother

Number of students in the class (Num)	Number of students in the class
Number of female students (Num)	Number of female students in the class
Class GPA	Mean grade of the class
Region (Cat)	Region name
Province name (C at)	Name of the province
School nature (Cat)	Nature of the school
School type (Cat)	School type
INDHcom (Bool)	Town (National Human Development Initiative)
INDHquart (Bool)	Neighborhood (National Human Development Initiative (INDH)
Number of affiliated schools (Num)	Number of affiliated schools
Number of supervisors (Num)	Number of supervisors in the administration of the school
Modern Preschool (Bool)	Modern preschool
Regular Primary School (Bool)	Regular primary school

Regular Primary School with religious education(Bool)	Regular primary school with religious education
Regular middle school (Bool)	Regular middle school
Regular middle school with religious education (Bool)	Regular middle school + religious education
Regular High School (Bool)	Regular high school
Regular high school + religious education (Bool)	Regular high school + religious education
Technical School (Bool)	Technical school
Exists Higher School prepratory classes (Bool)	Higher School Preparatory Classes

Exists Senior techologist's certificate (Bool)	Senior technologist's certificate
Institutional Management Board (Bool	Institutional Management Board
Pedagogical advice (Bool)	Pedagogical advice
Teaching Council (Bool)	Teaching Council
Class Council (Bool)	Class Council
Parent Association (Bool)	Parent Association
Sports Association (Bool)	Sports Association
School Cooperation Association (Bool)	School Cooperation Association
Partnership (Bool)	Establishment has a partnership with external organizations
Health Club (Bool)	Health club
Listening center (Bool)	Listening center (educational and psychological support)
AssEcolReussite (Bool	Modified program

Tayssir (Bool)	Cash transfer program
Environment club (Bool)	Environment club
Hard construction (concrete) (Bool)	Hard construction (concrete)
Prefab construction (Bool)	Prefab construction
Const_Autre	Other construction
Hard closure (Bool)	Hard closure
Partial closure (Bool)	Partial closure
Fence closure (Bool)	Fence closure
Wooden closure (Bool)	Wooden closure
Other closure (Bool)	Other closure
No closure (Bool)	No closure
School surface area (Num)	School surface area
Courtyard surface area (Num)	Courtyard surface area
Green area (Bool)	Green area

Extension surface area (Num)	Extension surface area
Canteen (Bool)	Canteen
Canteen capacity (Num)	Canteen capacity (Num)
Cafeteria (Bool)	Cafeteria
Number of seats in canteen (Num)	Number of seats in canteen
School residence (Bool)	School residence
Residence capacity (Num)	Residence capacity
Exists internet (Bool)	Internet available
Exists professional training (Bool)	Professional training
Informal education available (Bool)	Informal education available