Table 1. Attributes used grouped by class of attribute.

Class of Attribute	Attribute	Type
Demographic data	Marital status	Numeric/discrete
	Nationality	Numeric/discrete
	Displaced	Numeric/binary
	Gender	Numeric/binary
	Age at enrollment	Numeric/discrete
	International	Numeric/binary
Socioeconomic data	Mother's qualification	Numeric/discrete
	Father's qualification	Numeric/discrete
	Mother's occupation	Numeric/discrete
	Father's occupation	Numeric/discrete
	Educational special needs	Numeric/binary
	Debtor	Numeric/binary
	Tuition fees up to date	Numeric/binary
	Scholarship holder	Numeric/binary
Macroeconomic data	Unemployment rate	Numeric/continuous
	Inflation rate	Numeric/continuous
	GDP	Numeric/continuous
Academic data at enrollment	Application mode	Numeric/discrete
	Application order	Numeric/ordinal
	Course	Numeric/discrete
	Daytime/evening attendance	Numeric/binary
	Previous qualification	Numeric/discrete
Academic data at the end of 1st semester	Curricular units 1st sem (credited)	Numeric/discrete
	Curricular units 1st sem (enrolled)	Numeric/discrete
	Curricular units 1st sem (evaluations)	Numeric/discrete
	Curricular units 1st sem (approved)	Numeric/discrete
	Curricular units 1st sem (grade)	Numeric/continuous
	Curricular units 1st sem (without evaluations)	Numeric/discrete
Academic data at the end of 2nd semester	Curricular units 2nd sem (credited)	Numeric/discrete
	Curricular units 2nd sem (enrolled)	Numeric/discrete
	Curricular units 2nd sem (evaluations)	Numeric/discrete
	Curricular units 2nd sem (approved)	Numeric/discrete
	Curricular units 2nd sem (grade)	Numeric/continuous
	Curricular units 2nd sem (without evaluations)	Numeric/discrete
Target	Target	Categorical

3. Materials and Methods

This section describes the process that was followed for building the dataset and also presents a brief exploratory data analysis highlighting some relevant issues that may help other researchers quickly get their hands on the dataset and work with it, such as the imbalanced nature of data, the multicollinearity found in the features, and the results of permutation feature importance using the most used algorithms in similar problems shown in the literature.

3.1. Data Preprocessing

The data are collected in three different formats: (i) as Microsoft Access databases from CNAES; (ii) as comma-separated values (CSV) files from the AMS; and (iii) as manual data collected from the site of PORDATA concerning macroeconomics data.

Apart from the data received from CNAES, which are processed through a Visual Basic for Applications (VBA) program in a Microsoft Windows system, all the other code (in Python) runs on the Ubuntu operating system on an NVIDIA DGX Station computer with 2 CPU Intel Xeon E5-2698V4 with 20 core 2.2 GHz, 256 GB of memory, and 4 NVIDIA Tesla V100 GPU. This same computer was also used for training the machine learning

Acknowledgments: The authors would like to thank the Polytechnic Institute of Portalegre for providing support for this project, particularly to the Academic Services Department for providing the data and explaining the attributes used.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AMS	Academic Management System
CATBOOST	CatBoost
CSV	Comma-separated values
DGES	Direção Geral do Ensino Superior
DPO	Data Protection Officer
GDPR	General Data Protection Regulation
LIGHTGBM	Light Gradient Boosting Machine
PAE	Enterprise Application Platform
RF	Random Forest
XGBOOST	Extreme Gradient Boost

Appendix A

Table A1. Marital status values.

Attribute	Values
Marital status	1—Single 2—Married 3—Widower 4—Divorced 5—Facto union 6—Legally separated

Table A2. Nationality values.

Attribute	Values
Nationality	1—Portuguese 2—German 3—Spanish 4—Italian 5—Dutch 6—English 7—Lithuanian 8—Angolan 9—Cape Verdean 10—Guinean 11—Mozambican 12—Santomean 13—Turkish 14—Brazilian 15—Romanian 16—Moldova (Republic of) 17—Mexican 18—Ukrainian 19—Russian 20—Cuban 21—Colombian

Table A3. Application mode values.

Attribute	Values
Application mode	1—1st phase—general contingent 2—Ordinance No. 612/93 3—1st phase—special contingent (Azores Island) 4—Holders of other higher courses 5—Ordinance No. 854-B/99 6—International student (bachelor) 7—1st phase—special contingent (Madeira Island) 8—2nd phase—general contingent 9—3rd phase—general contingent 10—Ordinance No. 533-A/99, item b2) (Different Plan) 11—Ordinance No. 533-A/99, item b3 (Other Institution) 12—Over 23 years old 13—Transfer 14—Change in course 15—Technological specialization diploma holders 16—Change in institution/course 17—Short cycle diploma holders 18—Change in institution/course (International)

Table A4. Course values.

Attribute	Values
Course	1—Biofuel Production Technologies 2—Animation and Multimedia Design 3—Social Service (evening attendance) 4—Agronomy 5—Communication Design 6—Veterinary Nursing 7—Informatics Engineering 8—Equiniculture 9—Management 10—Social Service 11—Tourism 12—Nursing 13—Oral Hygiene 14—Advertising and Marketing Management 15—Journalism and Communication 16—Basic Education 17—Management (evening attendance)

Table A5. Previous qualification values.

Attribute	Values
Previous qualification	1—Secondary education 2—Higher education—bachelor's degree 3—Higher education—degree 4—Higher education—master's degree 5—Higher education—doctorate 6—Frequency of higher education 7—12th year of schooling—not completed 8—11th year of schooling—not completed

Table A5. *Cont.*

Attribute	Values
	9—Other—11th year of schooling 10—10th year of schooling 11—10th year of schooling—not completed 12—Basic education 3rd cycle (9th/10th/11th year) or equivalent 13—Basic education 2nd cycle (6th/7th/8th year) or equivalent 14—Technological specialization course 15—Higher education—degree (1st cycle) 16—Professional higher technical course 17—Higher education—master's degree (2nd cycle)

Table A6. Mother's and Father's values.

Attribute	Values
Mother's qualification Father's qualification	1—Secondary Education—12th Year of Schooling or Equivalent 2—Higher Education—bachelor's degree 3—Higher Education—degree 4—Higher Education—master's degree 5—Higher Education—doctorate 6—Frequency of Higher Education 7—12th Year of Schooling—not completed 8—11th Year of Schooling—not completed 9—7th Year (Old) 10—Other—11th Year of Schooling 11—2nd year complementary high school course 12—10th Year of Schooling 13—General commerce course 14—Basic Education 3rd Cycle (9th/10th/11th Year) or Equivalent 15—Complementary High School Course 16—Technical-professional course 17—Complementary High School Course—not concluded 18—7th year of schooling 19—2nd cycle of the general high school course 20—9th Year of Schooling—not completed 21—8th year of schooling 22—General Course of Administration and Commerce 23—Supplementary Accounting and Administration 24—Unknown 25—Cannot read or write 26—Can read without having a 4th year of schooling 27—Basic education 1st cycle (4th/5th year) or equivalent 28—Basic Education 2nd Cycle (6th/7th/8th Year) or equivalent 29—Technological specialization course 30—Higher education—degree (1st cycle) 31—Specialized higher studies course 32—Professional higher technical course 33—Higher Education—master's degree (2nd cycle) 34—Higher Education—doctorate (3rd cycle)

Table A7. Mother's and Father's occupation.

Attribute	Values
Mother's occupation Father's occupation	1—Student 2—Representatives of the Legislative Power and Executive Bodies, Directors, Directors and Executive Managers 3—Specialists in Intellectual and Scientific Activities 4—Intermediate Level Technicians and Professions 5—Administrative staff 6—Personal Services, Security and Safety Workers, and Sellers 7—Farmers and Skilled Workers in Agriculture, Fisheries, and Forestry 8—Skilled Workers in Industry, Construction, and Craftsmen 9—Installation and Machine Operators and Assembly Workers 10—Unskilled Workers 11—Armed Forces Professions 12—Other Situation; 13—(blank) 14—Armed Forces Officers 15—Armed Forces Sergeants 16—Other Armed Forces personnel 17—Directors of administrative and commercial services 18—Hotel, catering, trade, and other services directors 19—Specialists in the physical sciences, mathematics, engineering, and related techniques 20—Health professionals 21—Teachers 22—Specialists in finance, accounting, administrative organization, and public and commercial relations 23—Intermediate level science and engineering technicians and professions 24—Technicians and professionals of intermediate level of health 25—Intermediate level technicians from legal, social, sports, cultural, and similar services 26—Information and communication technology technicians 27—Office workers, secretaries in general, and data processing operators 28—Data, accounting, statistical, financial services, and registry-related operators 29—Other administrative support staff 30—Personal service workers 31—Sellers 32—Personal care workers and the like 33—Protection and security services personnel 34—Market-oriented farmers and skilled agricultural and animal production workers 35—Farmers, livestock keepers, fishermen, hunters and gatherers, and subsistence 36—Skilled construction workers and the like, except electricians 37—Skilled workers in metallurgy, metalworking, and similar 38—Skilled workers in electricity and electronics 39—Workers in food processing, woodworking, and clothing and other industries and crafts 40—Fixed plant and machine operators 41—Assembly workers 42—Vehicle drivers and mobile equipment operators 43—Unskilled workers in agriculture, animal production, and fisheries and forestry 44—Unskilled workers in extractive industry, construction, manufacturing, and transport 45—Meal preparation assistants 46—Street vendors (except food) and street service providers

Table A8. Gender values.

Attribute	Values
Gender	1—male 0—female

Table A9. Attendance regime values.

Attribute	Values
Daytime/evening attendance	1—daytime 0—evening

Table A10. Yes/No attributes.

Attribute	Values
Displaced Educational special needs Debtor Tuition fees up to date Scholarship holder International	1—yes 0—no

References

1. Behr, A.; Giese, M.; Teguim Kamdjou, H.D.; Theune, K. Motives for Dropping out from Higher Education—An Analysis of Bachelor's Degree Students in Germany. *Eur. J. Educ.* **2021**, *56*, 325–343. [CrossRef]
2. Kehm, B.M.; Larsen, M.R.; Sommersel, H.B. Student Dropout from Universities in Europe: A Review of Empirical Literature. *Hungarian Educ. Res. J.* **2020**, *9*, 147–164. [CrossRef]
3. Atchley, W.; Wingenbach, G.; Akers, C. Comparison of Course Completion and Student Performance through Online and Traditional Courses. *Int. Rev. Res. Open Distance Learn.* **2013**, *14*, 104–116. [CrossRef]
4. Quinn, J. *Dropout and Completion in Higher Education in Europe among Students from Under-Represented Groups*; An Independent report authored for the NESET network of experts; European Commission: Brussels, Belgium, 2013.
5. Namoun, A.; Alshanqiti, A. Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review. *Appl. Sci.* **2020**, *11*, 237. [CrossRef]
6. Saa, A.A.; Al-Emran, M.; Shaalan, K. Mining Student Information System Records to Predict Students' Academic Performance. *Adv. Intell. Syst. Comput.* **2020**, *921*, 229–239. [CrossRef]
7. Akçapınar, G.; Altun, A.; Aşkar, P. Using Learning Analytics to Develop Early-Warning System for at-Risk Students. *Int. J. Educ. Technol. High. Educ.* **2019**, *16*, 40. [CrossRef]
8. Daud, A.; Lytras, M.D.; Aljohani, N.R.; Abbas, F.; Abbasi, R.A.; Alowibdi, J.S. Predicting Student Performance Using Advanced Learning Analytics. In Proceedings of the 26th International World Wide Web Conference 2017, WWW 2017 Companion, Perth, Australia, 3–7 April 2017; pp. 415–421. [CrossRef]
9. Martins, M.V.; Tolledo, D.; Machado, J.; Baptista, L.M.T.; Realinho, V. Early Prediction of Student's Performance in Higher Education: A Case Study. *Adv. Intell. Syst. Comput.* **2021**, *1365*, 166–175. [CrossRef]
10. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-Sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
11. He, H.; Bai, Y.; Garcia, E.A.; Li, S. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. In Proceedings of the International Joint Conference on Neural Networks, Hong Kong, China, 1–8 June 2008; pp. 1322–1328. [CrossRef]
12. Chen, C.; Liaw, A.; Breiman, L. Using Random Forest to Learn Imbalanced Data. *Univ. Calif. Berkeley* **2004**, *110*, 1–12.
13. Liu, X.Y.; Wu, J.; Zhou, Z.H. Exploratory Undersampling for Class-Imbalance Learning. *IEEE Trans. Syst. Man Cybern. Part B Cybern.* **2009**, *39*, 539–550. [CrossRef]
14. Maclin, R.; Opitz, D. An Empirical Evaluation of Bagging and Boosting. In Proceedings of the National Conference on Artificial Intelligence, Providence, RI, USA; 1997; pp. 546–551.
15. Hido, S.; Kashima, H.; Takahashi, Y. Roughly Balanced Bagging for Imbalanced Data. *Stat. Anal. Data Min.* **2009**, *2*, 412–426. [CrossRef]
16. Wang, S.; Yao, X. Diversity Analysis on Imbalanced Data Sets by Using Ensemble Models. In Proceedings of the 2009 IEEE Symposium on Computational Intelligence and Data Mining, Nashville, TN, USA, 30 March–2 April 2009; pp. 324–331. [CrossRef]
17. Saarela, M.; Jauhiainen, S. Comparison of Feature Importance Measures as Explanations for Classification Models. *SN Appl. Sci.* **2021**, *3*, 272. [CrossRef]