education sciences

MDPI

Article

Hybrid Approach to Predicting Learning Success Based on Digital Educational History for Timely Identification of At-Risk Students

Tatiana A. Kustitskaya, Roman V. Esin, Yuliya V. Vainshtein and Mikhail V. Noskov *

School of Space and Information Technology, Siberian Federal University, 660041 Krasnoyarsk, Russia; tkustitskaya@sfu-kras.ru (T.A.K.); resin@sfu-kras.ru (R.V.E.); yweinstein@sfu-kras.ru (Y.V.V.)
* Correspondence: mvnoskov@yandex.ru

Abstract: Student retention is a significant challenge for higher education institutions (HEIs). The fact that a considerable number of dropouts from universities are primarily due to academic underperformance motivates universities to develop learning analytics tools based on models for predicting learning success. However, the scalability of such models is limited since students' academic performance and engagement, as well as the factors influencing them, are largely determined by the educational environment. The article proposes a hybrid approach to forecasting success in completing an academic semester, which involves creating a set of predictive models. Some of the models use historical student data, while others are intended to refine the forecast using current data on student performance and engagement, which are regularly extracted from available sources. Based on this approach, we developed an ensemble of machine learning models and the Markov-process model that simultaneously address the tasks of forecasting success in mastering a course and success in completing a semester. The models utilize digital footprint data, digital educational history, and digital personality portraits of students extracted from the databases of Siberian Federal University, and the resulting ensemble demonstrates a high quality of the forecast. The proposed approach can be utilized by other HEIs as a framework for creating mutually complementary forecasting models based on different types of accessible educational data.

Keywords: learning analytics; student retention; learning success prediction; at-risk students; digital footprint; digital educational history; digital personality portrait; LMS Moodle

check for updates

Citation: Kustitskaya, T.A.; Esin, R.V.; Vainshtein, Y.V.; Noskov, M.V. Hybrid Approach to Predicting Learning Success Based on Digital Educational History for Timely Identification of At-Risk Students. *Educ. Sci.* **2024**, *14*, 657. https://doi.org/10.3390/educsci14060657

Academic Editor: Menucha Birenbaum

Received: 24 April 2024
Revised: 10 June 2024
Accepted: 12 June 2024
Published: 17 June 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

1.1. Student Retention

Student retention is a significant challenge for universities worldwide [1]. The expulsion of a student represents not only the loss of a potential specialist for society and an economic loss for the university, but also, possibly, a personal tragedy for the student [2–4]. Unfortunately, the rate of student expulsion is high in many universities. For instance, the OECD reports that the average bachelor's graduation rate for students under thirty in OECD countries is 34%. In particular, only 23% of such students complete a bachelor's degree program in Austria, 30% in Turkey, 33% in Germany, 34% in Canada, 46% in the UK, and 51% in Australia [5].

According to the statistical data of the Ministry of Science and Higher Education of Russia, in 2019, the number of first-year students in full-time bachelor's programs at state universities in Russia was 412,142. However, in 2023, the number of graduates from these programs was 287,654. The resulting graduation rate based on the theoretical duration of bachelor's programs is 69.79%, which is an optimistic estimate, as it includes students who enrolled before 2019 but completed their studies later than the planned 4-year period due to academic leave or expulsions followed by reinstatements [6].

1.2. Formulations of the Task of Forecasting Learning Success

Early forecasting of a student's learning success is a key tool in addressing the issue of student attrition, as such forecasts enable timely interventions to help students overcome academic challenges and achieve learning success [7,8]. The clear value of solving this task is the reason for numerous research studies and scientific articles dedicated to this issue [9–11].

In these studies, the task of forecasting learning success is primarily viewed as predicting student attrition in the next semester or predicting the success of mastering a specific course. However, we believe that a broader range of problem formulations should be considered, determined by the peculiarities of national legislation or the interests of various education stakeholders, as stated in reference [12]. For university administration, a generalized forecast is important, such as predicting whether a student will be expelled in the next semester [13]. For a course instructor, the forecast for the success of mastering the course is of interest [14]. For a student, perhaps the most crucial aspect is the forecast for successfully completing the entire academic session, partly because in some countries, including Russia, a student may receive a scholarship if they pass all the exams in the semester [15].

1.3. Literature Review of the Data and Machine Learning Algorithms for Predicting Learning Success

Machine learning models are increasingly being used to forecast academic success, as they often demonstrate higher forecasting accuracy than traditional statistical methods. Many studies utilize methods based on decision trees and their ensembles, while models based on neural networks have also shown good results [16].

For example, in reference [17], the task of forecasting overall academic performance is addressed using data on students' socio-economic status and entrance exam results at a Chinese university, with a four-layer Artificial Neural Network (ANN) being used as the forecasting model. In the study [18], three-layer ANNs are also employed solely for predicting completion of an educational program (forecasting the risk of long-term dropout). The Graduate Grade Point Average (GGPA), the average performance score for the first semester of study, and the year of enrollment at the university were identified as the most important predictors of academic performance. In some cases, researchers forecast the passing score of GGPA based on enrollment data (age, gender, entrance exam data, region/ethnicity), and grades for each of the four years of study [19].

In reference [20], a two-stage hybrid system for forecasting learning success is described, where in the first stage, the probability of passing or failing a course is predicted, and in the second stage, a multi-class classification task is solved to refine the potential grade obtained by the student. In reference [21], the task is to predict applicants' early academic performance using the admission data (school grade average, Scholastic Assessment Test score (SAT), and General Aptitude Test score). The author utilizes four machine learning methods (ANN, Decision Tree (DT), Support Vector Machine (SVM), and Naive Bayes (NB)), and neural networks show the best result. A similar solution, but for a medical university in Iran, is presented in [22].

In the study [23], authors identify the key machine learning algorithms for predicting learning success in different variations of this task—DT, ANN, SVM, and NB. Additionally, the authors determine that academic, demographic, internal assessment, and family/personal attributes are the key predictive features. These feature sets are also used in reference [12].

Research studies often focus on early forecasting within one or several courses; for example, in reference [24], authors work on the comparative study aiming to find the best combinations of models and datasets. They extracted educational data from an introductory programming course in the learning management system (LMS) Moodle. The authors used thirteen combinations of data sets (based on three types of student interactions: cognitive presence, social presence, and teaching presence) along with five classification algorithms

(k-Nearest Neighbor, Multilayer Perceptron, NB, AdaBoost, and Random Forest (RF)). In [25], the study is devoted to forecasting academic performance in courses of physics, calculus, and programming. The work uses a Bayesian network to predict the students' grades in three major courses, based on existing feature descriptions of students, including demographic and academic variables. The task of predicting learning success in specific medical courses, taught using blended learning, is addressed in reference [26]. The authors note that the obtained predictive models can be considered portable only for courses that are homogeneous in terms of institutional settings, discipline, nominal learning design, and course size. However, differences in the implementation of the pedagogical model negatively affect the models' predictive power.

Data obtained from LMSs are often used for academic success prediction. For example, in the study [27], an early forecasting system is based on student interaction data from LMS Moodle. In reference [27], predictors such as activity in the online environment, attendance, task completion productivity, and age are used. The work [28] is devoted to reviewing current research studies that analyze data of online learners to predict their outcomes in various prediction task scenarios (prediction of diploma attainment, prediction of grades for disciplines, identification of at-risk students, student attrition/retention forecasts).

Summarizing the results of the researchers, we can conclude that the following features have the strongest predictive power for learning success prediction in various studies: average grade, earned credits, and gender [29]; emotional, demographic, academic, students' motivation [30], and general data [31]; learning data in the online environment and records from the LMS [32]. However, when using personal characteristics data and data from the LMS, it is important to consider that self-reported questionnaire data, intended for predicting academic success, may not be as objective compared to LMS data [33], and at the same time, demographic data and entrance exam data are more reliable for predicting learning success than LMS data [34].

1.4. Research Aims of the Study

The literature review shows that despite the presence of numerous examples in modern studies of solving the task of predicting learning success, direct transfer of these models to a specific context may not be sufficiently effective. This is due to the variety of factors influencing learning success, such as the characteristics of educational institutions, the specifics of curricula, the structure and content of the existing databases, the volume of information stored in them, and others. Each HEI has its unique features, including those related to the national education system and the level of informatization, which may not always be accounted for by existing models. Therefore, to achieve accurate and reliable results, it is necessary to create custom models that address the task of predicting learning success in the formulations that are of interest to specific stakeholders, taking into account the specifics of the particular educational environment and the characteristics of the analyzed data.

We believe that many universities can address the task of predicting learning success simultaneously in several formulations, and it is possible to use different types of educational data to obtain forecasts that complement each other.

In the study, we consider two types of learning success and failure:

- *learning success of mastering the course* means passing the course, while failure means failing the course;
- *learning success of completing the semester* is passing all the courses of the curriculum in the semester, while failure is failing at least one of the courses.

In accordance with these formulations of learning success, we set the following **research aims**:

RA 1: Develop a hybrid approach to forecasting the success of learning in a university, making it possible to solve this task in two aspects—predicting the success of mastering a course and the success of completing a semester.

RA 2: Develop an ensemble of forecasting models using educational data to implement this approach and assess its accuracy and applicability.

These aims are considered separately as the results of achieving these goals will be scalable to varying degrees. The hybrid approach to forecasting (RA 1) can be implemented in any HEI with a sufficient level of digitalization of the educational process. At the same time, the specific forecasting model based on this approach (RA 2) is aimed at implementation specifically at SibFU and may only be partially applicable to other universities.

2. Materials and Methods

2.1. Special Aspects of the Educational Process in Russia and Local Regulatory Acts of Siberian Federal University

The choice of forecasting task formulations in our study is explained by the peculiarities of Russian legislation. A student having an unsatisfactory mark has two attempts to rectify the situation in the course of up to one year. Every educational institution can set down the dates of academic debt elimination independently. For instance, in Siberian Federal University (SibFU), a student can make attempts to pass the failed course during the next two semesters after the examination period. Thus, Russian higher education sees cases of students passing all exams in the current examination period, but being expelled in the view of the results of the previous examination period.

The organization of studies is regulated by federal enactments and regulatory documents of educational institutions.

Federal Law of the Russian Federation of December No. 273 "On Education in the Russian Federation" (supplemental and amended, coming into effect on 1 January 2024) [35] establishes legal, organizational, and economic bases for education in the Russian Federation and regulates social relations appearing in education in the context of realizing the right to education, providing state assurance of human rights and freedoms in the field of education, and creating conditions for the realization of the right to education. According to Part 2 of Article 13 of the Law on Education, various educational technologies, including distance learning technologies and electronic learning, are used to implement educational programs.

The procedure of organization and implementation of educational process for Bachelor's, Specialist, and Master's degree programs is established by Decree of the Ministry of Higher Education and Science of the Russian Federation No. 245 [36].

The development and approval of educational programs and the requirements to their content are under the jurisdiction of the HEIs and are regulated by local legislation. In SibFU, there are a number of local normative acts that regulate the organization of educational process and the use of electronic learning and distance learning technologies.

The Regulation on Student Current Assessment and Interim Attestation [37] establishes the procedure of control of student current academic performance and interim attestation (taking tests, exams, term projects etc.).

The Regulation on Electronic Information Educational Environment of SibFU [38] defines the aim, objectives, structure, and operating procedures of the electronic information educational environment of the university, listing the tools for running classes eligible for distance learning—video conference services for webinars for synchronous online learning and the platform of electronic learning "e-Courses" powered by Moodle for asynchronous online learning [39]. This regulation determines dates and locations of students' digital footprints obtained from the capture of educational process development.

The Regulation on Electronic Learning and Distance Learning Technology Implementation determines the conditions and requirements of the educational process with the use of electronic learning and distance learning technologies, and sets requirements to distance learning technologies as well [40].

SibFU has various proportions of the use of electronic education in various courses of study—they can be taught offline, in blended format, online, or in hybrid format. In some cases, academic courses may not be accompanied by electronic courses and that is why the

current educational process leaves no digital footprint in the LMS. Currently, electronic courses accompany around 50% of all academic courses taught in SibFU.

2.2. Educational Data Governance Policies and Data Utilized in the Study

It is necessary to rigidly follow ethical and safety principles when researching education data processing and use. The following legal acts in the realm of information security provide the basis for the use of the obtained information: General Data Protection Regulation (GDPR) [41]; Federal Law of the Russian Federation No. 149. "On Information, Information Technologies and Protection of Information" [42]; and Federal Law of the Russian Federation No. 152. "On Personal Data" [43]. These statutory documents regulate personal data processing, specify requirements to security, and confer rights to data subjects.

Furthermore, additional regulations of the university, such as the Personal Data Regulation [44] and the "e-Courses" Electronic Learning Environment Acceptable Use Policy accompanied by informed consent from students [45], play an important role in providing security and protection of student data and observing ethical principles.

The Personal Data Regulation establishes the procedure of processing and protecting information of all personal data subject categories, the scope of personal data, the rules of accounting, managing, and transferring data, and data access, as well as the measures aimed at leakage protection and unauthorized data access.

The Acceptable Use Policy determines the rules and responsibility of data use in the realms of educational process as well as the "inappropriate use" category, which is prohibited without any exceptions.

In our study to train predictive models, we downloaded the data spreadsheets containing information about students of 2018–2023 academic years from the university online environment database. These data included four types of data:

- general educational data about students including personal information, training program, academic group code, year of training and so on;
- student grade book data that contains student academic performance information including results of exams, tests, term projects, and other academic activities;
- data on changes in student status containing information on transferring students between groups, educational program change, taking academic leave, and other changes;
- LMS Moodle activity data containing information on student activity in the university electronic environment including task and test completion, studying theory, forum communication, electronic learning course element reference, and other interactive operations.

The whole volume of the information provided was split into five academic years: 2018/2019, 2019/2020, 2020/2021, 2021/2022, and 2022/2023, each one of them storing the data on the students for the corresponding period.

2.3. Digital Profile of a Student

The digital profile components represented in Figure 1 were created to analyze the learning outcomes and academic success of students, their personal traits, electronic environment activity and its dynamic pattern, student grade book data, and student data on changes in student status. Firstly, we consider the characteristics of a student, which tend to change dynamically in the course of the learning process with their significance changing as long as learning experience is gained. The current state of these characteristics can be described by the *digital footprint*, while the previous states can be described by the *digital educational history*. Secondly, we examine more static personal characteristics and refer to them as the *digital personality portrait*.

Based on general educational data about students, the digital personality portrait of a student was created, which represents an assembly of digital representations of personal data, sociodemographic, and other characteristics of a student [46].

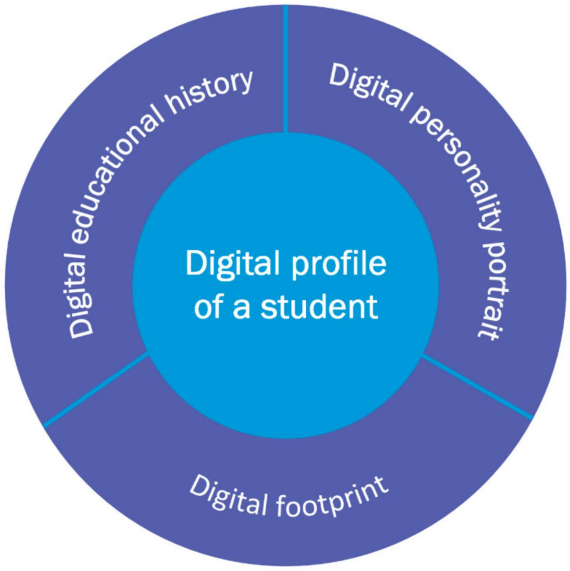


Figure 1. Components of the digital education profile of a student.

Based on the student grade book data and the data on changes in student status, a digital educational history for every individual was arranged. We imply the digital educational history to be multidimensional structured dynamic data on student academic activity and their learning outcome [46,47].

Based on the LMS Moodle activity data, a student digital footprint was created, which contains current learning characteristics of a student within the electronic course: the grade points, clicking internal links, time spent on course pages, etc. [48,49].

The application of the complex approach encapsulating the use of the digital personality portrait, the digital educational history, and the digital footprint of the student enables a comprehensive representation of every student. It allows the consideration of multiple characteristics of personality and learning activity, which makes it possible to create accurate models of learning success prediction and to enhance the efficiency of the learning process.

2.4. Methods

This study is based on the system approach, which contributed to considering academic success as an integral system, making it possible to take into consideration a wide range of interrelating factors that have influence on learning and to elaborate a hybrid approach for forecasting.

In this study, we extensively use the principles of learning analytics, as well as the principles of learning individualization and education digitalization. Focusing on learning analytics makes it possible to conduct a thorough analysis of educational process data, to reveal implicit patterns, and to make interpreted decisions based on the obtained results. The principles of learning individualization and education digitalization are the key factors in developing the model of learning success prediction. They regard each student as the subject of educational activity characterized by a set of data related to the individual learning process. These personalized data are used to create an effective forecasting model.

Various methods were used to develop the predictive models based on the digital history. The models of learning success prediction relying on the digital educational history and the current characteristics of the digital footprint included bagging and boosting algorithms such as RF (scikit-learn 1.2.2), XGBoost (xgboost 1.7.3), CatBoost (catboost 1.0.6), and LightGBM (lightgbm 2.2.3). The choice of these algorithms is motivated by several reasons:

- these algorithms have proven to be effective for supervised learning on tabular data;
- logical algorithms usually effectively address multicollinearity issues, which arise in our case as digital educational history data are interdependent;
- they allow for the analysis of the importance of the features, enhancing the interpretability of the prediction results.

To leverage the strengths of each model and obtain more stable forecasts, we also develop the weighted average ensemble of the mentioned models.

To simulate academic performance in various courses we use the special case of the continuous-time Markov processes—the birth-and-death process. To model the Markov process, we assume that at each moment in time, the student is determined by the level of proficiency in the course material, which can be assessed on a four-point scale. The transition from one state to another occurs under the influence of the intensity of the process of obtaining information and assimilating information. These intensities represent the parameters of the birth-and-death process and are estimated on historical data. This approach is described in detail in the research papers [50–53].

The probabilities of the states are calculated by solving the Kolmogorov equations. After calculating the probabilities p(grade = k) for k = 2, 3, 4, 5, the expected value of a student's grade for a course on the four-point scale is computed using the formula

$$E(\text{grade}) = 2 \cdot p(\text{grade} = 2) + 3 \cdot p(\text{grade} = 3) + 4 \cdot p(\text{grade} = 4) + 5 \cdot p(\text{grade} = 5). \quad (1)$$

The final model, which implements the hybrid approach, is created by combining several models that operate on different datasets. To implement this combination, we build a meta-model using stacking.

The development and training of the models based on the data analysis algorithms included exploratory data analysis; imputation of missing values; creation, extraction, and selection of features; assessment of training quality; and assessment of feature importance.

3. Results

3.1. Hybrid Approach to Forecasting the Success of Learning

Since the digitalization of the educational process in SibFU is being carried out gradually, the university began to record different components of the digital profile at different times. Specifically, data necessary for creating indicators of the digital personal portrait and educational history have been stored in databases since 2016. At the same time, data of the digital footprint in a form that allows assessment of academic performance dynamics are only available from 2022.

It would be possible to train the prediction model using data collected only from 2022 onwards, but this may lead to a problem of model underfitting, caused by the following factors:

- insufficient sample size compared to the significant number of characteristics of the digital profile;
- underrepresentation (limited variability) of data in terms of observing the impact of the time factor. If the model is trained only on data from the 2022–2023 academic year, only one educational cycle will be included in the training.

Addressing the task of predicting student attrition, we need to create a model for forecasting Learning Success in completing the Semester (model LSSp). It is quite logical to build this model based on data on the previous academic performance of students, as well as on their digital personality portrait. However, identifying at-risk students is not enough to solve the issue of student retention.

In order to provide maximum support to these students, it is necessary to thoroughly understand the nature of the difficulties they are facing. In particular, it is highly desirable to identify in which specific courses the difficulties have arisen, i.e., to solve the task of predicting academic success for each course. Such models for predicting success based

on digital footprint data in the electronic educational environment has been previously developed and tested in the educational process at SibFU [47,49].

The model for forecasting Learning Success of mastering a Course (model LSC) is the model of the birth-and-death process [50]. It demonstrated high accuracy in predicting the success of learning in subjects accompanied by e-courses. At the same time, it cannot be used for subjects not accompanied by e-courses which are quite common (we discussed heterogeneity in the provision of the educational process with electronic courses in Section 2.1).

As a result of analyzing the above-mentioned features of the institutional environment, we have decided to create an ensemble of models. One of the models will be trained on a large volume of data from the digital educational history and the digital personality portrait, while the other will be trained on a smaller volume of data from the LMS digital footprint. Such a two-tiered forecasting organization will allow us to provide a forecast of academic success even in the case of a complete absence of digital footprint in the LMS.

It is obvious that the number of academic debts is mainly determined by the success of learning in each of the studied courses in the semester. However, due to the applicability issues of model LSC described above we do not receive a forecast of academic success for every course of the considered semester for the majority of students. This means that we require additional characteristics of students' current learning behavior to predict academic success for the semester.

Many of the electronic courses on the "e-Courses" platform are not officially assigned to the courses of curricula, but they are to varying degrees related to the study process in those courses. Some academics use the electronic environment only to provide students with learning resources (lectures, lecture presentations, links to books or videos); they do not include any assessment tasks in their e-courses. Additionally, students may subscribe to several e-courses related to one course of the curriculum, and the names of e-courses may differ from the names of courses in the curriculum. Overall, it is difficult to correlate the set of electronic courses to which a student is subscribed with the set of courses they are studying in the current semester.

However, students' activity in the online learning environment is an important characteristic of their learning behavior, even if this activity cannot be associated with courses of the curriculum. If a student regularly accesses learning resources, this means at least that they have not abandoned their studies.

All the points discussed above lead us to create a hybrid approach, which is based on available educational data from various sources and solves the task of predicting learning success in two formulations. The model based on such an approach should utilize both previous and current digital education history, as well as digital personality portraits. We plan to use as predictors the following groups of variables:

- the forecast made by model LSSp based of previous educational history;
- the forecasts (made by model LSC) of success in mastering all courses accompanied by the e-courses a student is subscribed to in this semester (instead of forecasts for success for all courses of the curriculum, which we are often not able to receive);
- characteristics of activity and performance of students on the "e-Courses" platform.

The scheme of the ensemble model based on the introduced approach is presented in Figure 2.

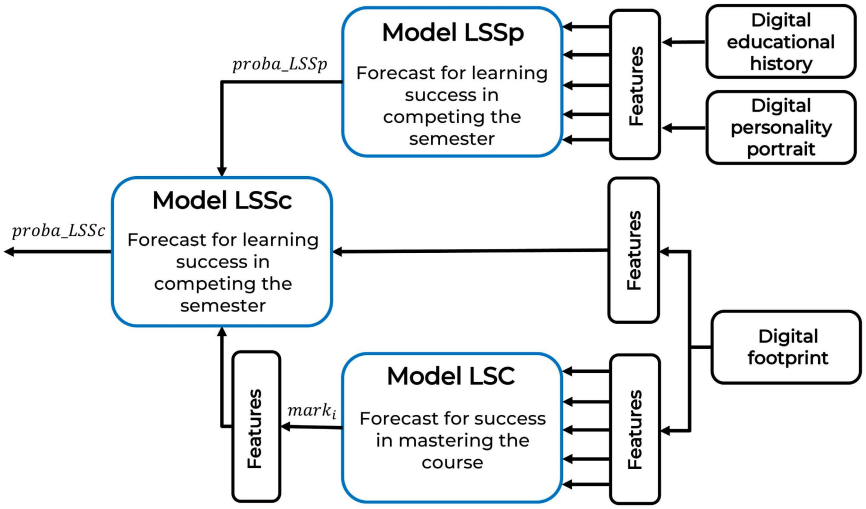


Figure 2. The scheme of the hybrid forecast.

Next, we describe each of the components of the hybrid forecast in more detail.

3.2. Model LSSp—Model for Forecasting Learning Success in Completing the Semester Based on Previous Educational History

Model LSSp is designed to predict learning success for completing the current semester using data from the digital educational history of previous semesters and the digital personal portrait of a student. We consider this problem as a binary classification task where the target variable equals 1 if the student fails at least one of the courses of the current semester, and equals 0 if the student successfully passes all the courses. Class 1 is the priority class in our case, as students with potential academic debts are those who need timely assistance and support in the educational process. To solve this task correctly, it is necessary not only to use relevant data about students but also to take into account the institutional conditions of the educational process, as well as the regulatory local acts of the university.

The model is based on the following data from the electronic environment of SibFU:

- General educational data about students, including a unique Student Code, general information about students (gender, citizenship, current student status), information about study plans and educational programs, individual study plans, important dates (date of birth, year of enrollment), and other educational data (institute of study, study group, form of education, benefits).
- Student grade book data, including Student Code, information about learning courses, the corresponding years and semesters of study, type of intermediate certification for each course in the study plan, dates of intermediate certification and grades, data on retakes, information about teachers.
- Data on changes in student status. By a change in student status, we mean transfers (to another university or to another specialty), taking an academic leave and returning from it, expulsion, and reinstatement to studies. The dataset provides data on orders of changes in student status. It includes Student Code, previous and current student status, number and date of the corresponding order, previous and current study plans, study groups, education levels, forms of education, years of study, fields of study, reasons and dates of academic leaves, expulsions, transfers, and reinstatements.

Based on the provided data, we prepared a dataset for developing a machine learning predictive model, which includes the target variable and the feature description of students. The features were extracted from the aforementioned general educational data, student grade book data, and data on changes in student status merged using the unique student identifier (Student Code). The feature extraction included the data preprocessing stage and the stage of constructing the students' digital educational history.

3.2.1. Data Preprocessing

Preprocessing of general educational data. We merged data from different academic years into one dataframe and conducted the following feature preprocessing: nominal features remained unchanged, binary features were converted into numerical format, categorical variables were encoded using One Hot Encoding. We also added two new features—Age (calculated at the end of the current semester using Date of Birth) and Group of Specialties (derived from student specialty data). In total, we obtained 73,394 student records.

Preprocessing of student grade book data. The grade book data provides comprehensive information on the results of intermediate assessment of students in each semester of study up to the current moment and constitutes the "learning outcome" part of the educational history. Based on the available grade book entries for each student in each semester of past study, the following indicators were grouped and aggregated: the target variable (presence of at least one academic debt), data on the workload per session (number of disciplines, distribution of disciplines by different types of intermediate assessment), data on debts per session, average grade per session, number of retakes, distribution of debts by different types of intermediate assessment (exams or credits), and average grade after retakes. Additionally, we calculated relative indicators of the number of failed disciplines per session, for retakes. In total, after aggregation, 314,237 records were obtained.

Preprocessing of data on changes in student status. We extracted information about types of changes in student status from the variable Title of Order. Using this information, all orders were divided into four categories: transfer, granting of academic leave, reinstatement, expulsion. Based on the obtained categories, the following features were created: Total Number of Transfers, Total Number of Academic Leaves, Total Number of Reinstatements, and Total Number of Expulsions, as well as Reason and Date of the four most recent records in each category.

Formation of a digital educational history. After merging the prepared features from general educational data, student grade book data, and data on changes in student status for every academic year, we conducted a retrospective merge to compile the final dataset. The main idea of this merging was to create a generalized educational history for each student over the entire period of study and to segment this educational history into three-semester intervals. This approach allows for the creation of the generalized educational history over the entire period of study, providing a more comprehensive understanding of the student's academic path. Additionally, dividing the educational history into three-semester intervals facilitates the analysis and understanding of students' academic performance and learning dynamics, and enables the identification of trends and peculiarities in their learning.

The reason for choosing three-semester intervals is related to the procedure for expelling students due to academic underperformance at SibFU. A student is actually expelled in the third semester after receiving failing grades for the semester if they also fail retakes in the following two semesters. Therefore, utilizing data on academic performance from the three previous semesters enables us to consider information about all the academic debts that the student may have accumulated up to the current moment, without having been expelled yet. This retrospective segmentation of the data also increases the number of records for more effective training of the forecasting models.

For first-year students in this three-semester academic history, there was no information on academic performance in all or some of the preceding semesters because these students had not yet enrolled in the university at that time. All missing values of this type were filled with zeros.

The resulting dataset for training model LSSp contains 162,567 records. The target variable is set to 1 for students with one or more academic debts in the current semester, and 0 for students who received passing grades for the semester. The dataset can be considered balanced as 40.6% of the records belong to class 1. The total number of predictors for the model is 74.

3.2.2. Model Training and Validation

We trained and validated four algorithms to predict learning success in completing the semester: RF, XGBoost, CatBoost, LightGBM. They all demonstrate comparable predictive quality but operate differently, thereby capturing different non-linear dependencies in the data. Therefore, building a meta-model based on these algorithms allows for further enhancement of the prediction quality.

For each of the four models, we performed the hyperparameters' selection. The resulting weighted prediction of all four models was used as the meta-model. This involved optimizing the coefficients through maximizing the weighted F-score metric. The final prediction of the LSSp model is calculated as follows:

$$\text{proba_LSSp} = 0.35 \cdot \text{proba_RF} + 0.32 \cdot \text{proba_XGBoost} + 0.13 \cdot \text{proba_Catboost} + \\ + 0.2 \cdot \text{proba_LightGBM}$$

The prediction quality was evaluated on a validation dataset, which consisted of 30% of the original dataset and was not used during model training. The quality metrics for the trained models are presented in Table 1.

Table 1. Quality Metrics for the model LSSp.

Model	Quality Metrics			
	Accuracy	**Precision**	**Recall**	**Weighted_F**
RF	0.8051	0.7741	0.7091	0.7602
XGBoost	0.8135	0.7518	0.7374	0.7673
CatBoost	0.8143	0.7783	0.7349	0.7692
LightGBM	0.8148	0.7764	0.7401	0.7689
model LSSp				0.7793

3.3. Model LSC—Model for Forecasting Learning Success in Mastering a Course

In our study, the only source of current educational history data is the digital footprint from the "e-Courses" platform. To characterize a student's learning behavior and academic performance in a learning course we extract his or her digital footprint data from the corresponding e-course on a weekly basis.

We define the current academic performance of a student in an e-course as the overall score in the gradebook of the e-course at a given moment in time. We suggest defining activity and effectiveness in terms of clicks of the student in the e-course. For this purpose, all user actions in the electronic course are classified into one of two categories:

- active clicks refer to students' interactions with reading materials, viewing the main course page, checking their grades, reviewing previous test attempts, studying the course glossary, and clicking on hyperlinks;
- effective clicks are any students' interactions with the e-course that change its content—writing on forums, submitting assignments, adding entries to the glossary, participating in polls, and submitting test attempts.

The above-mentioned components of the current educational history can be considered universal predictors of learning success in mastering the course, as they do not depend on the structure of the e-course. The standardized number of active clicks, number of effective clicks, and overall grade are utilized by model LSC to predict students' learning success in each of the courses offered in the semester. Normalization plays a crucial role in data preprocessing in this scenario, given the diverse nature of data collected during the learning process. Additionally, normalization is important due to the significant variability in academic performance, activity, and effectiveness metrics across different e-courses.

Model LSC is designed to predict the probability of obtaining a particular grade in a course based on these data. The calculation of probabilities in the model is carried out taking into account the personalized function introduced into the model. The function is calculated based on current academic performance, activity, and effectiveness of students

in the e-course. These indicators were selected based on the analysis of various approaches to defining the essence and outcomes of learning [54–56].

We calculate the expected grade $E(grade)$ using (1), and classify the student's performance in the course as success or failure using the rule

$$y = \begin{cases} failure, & E(grade) < \alpha \\ success, & E(grade) \geq \alpha, \end{cases}$$

where α is a threshold.

The parameter α selection and the forecasting quality assessment were conducted using the digital educational history data of 1788 students from the School of Space and Information Technology at SibFU. For each student during the academic year 2022–2023, a digital footprint from the "e-Courses" platform was collected, as well as data on the students' semester grades in courses taught using "e-Courses". The number of such courses taught in a blended learning format ranged from 3 to 10 for each student. Thus, the obtained dataset contained 13,312 records. Each record presented information about the learning behavior and performance of a student in a course and the outcome—success in the course (denoted as class 0) or failure (denoted as class 1). Approximately 30.2% of the records in the dataset belonged to class 1.

The dataset was split into two subsets. We selected α on the first subset, and the obtained value was $\alpha = 2.6$. The second subset was used for the model quality assessment. As shown in Table 2, the quality on the test set was quite high.

Table 2. Quality Metrics for model LSC.

Week	Quality Metrics			
	Accuracy	**Precision**	**Recall**	**Weighted_F**
7	0.7485	0.9248	0.7660	0.7932
11	0.8086	0.9341	0.8333	0.8517
14	0.8615	0.9402	0.8936	0.9026
18	0.8857	0.9363	0.9285	0.9300

3.4. Model LSSc—Model for Forecasting Learning Success for Completing the Semester Based on Current Educational History

From Table 1, it is evident that the forecast quality of model LSSp is quite low, with a weighted F-score of 0.78. Since we assume that the educational situation and behavior of students in the current semester significantly contribute to their successful completion of the semester, we need to add predictors to the model based on the students' current educational history. If this hypothesis is true, the forecast quality of the resulting model should increase.

In order to select features that fully describe students' learning behavior on "e-Courses", we single out the following categories among all e-courses a student is subscribed to:

- assessing e-courses (in which course grades changed at least once for one of the subscribed users last week).
- frequently visited e-courses (which were accessed by their subscribers at least 50 times last week).

This categorization allows us to assess roles of particular e-courses in a student's learning. If grades are regularly assigned to at least some students in a specific e-course, it is likely that the corresponding subject is being taught using blended learning technology for all students subscribed to the e-course. Therefore, the assigned grades themselves, the regularity of their updates, and the student's regularity of accessing the course are important characteristics of their learning success in that subject.

If there is no regular assessment of students in the particular e-course, it is likely that the course is only used to provide students with study materials. The student's grades in

such e-courses are unlikely to be relevant indicators of their learning success in the course. However, if other users regularly access this e-course, it is likely that the materials from it are actively used in teaching. Therefore, the number of times a student accesses this e-course can provide important information about their learning in the corresponding subject.

However, initially, we did not exclude courses with low activity from consideration since the statements above were merely hypotheses. We further apply a data-driven approach to select significant predictors—if a predictor proves to be insignificant for a model trained on real data, then we exclude it. Initially, we considered the following list of current educational history characteristics as predictors for model LSSc:

- week of semester;
- semester (spring or fall);
- year of study;
- number of e-courses a student is subscribed to (total number of e-courses, number of assessing e-courses, and number of frequently visited e-courses);
- average grades (for all e-courses, for assessing e-courses, for frequently visited e-courses);
- number of active clicks (for all e-courses, for assessing e-courses, for frequently visited e-courses);
- number of effective clicks (for all e-courses, for assessing e-courses, for frequently visited e-courses);
- averages of the forecasts made by model LSC (for all e-courses, for assessing e-courses, for frequently visited e-courses).

For training model LSSc, we used educational data obtained over two semesters of the 2022–2023 academic year. These data included the following:

- model LSSp forecasts for both academic semesters for all students studying at the university at that time (for 26,732 students);
- final grades of students in subjects studied in the respective semesters (for 26,732 students);
- digital footprint data from "e-Courses" for both semesters (for 16,476 learners in the autumn semester and 15,061 learners in the spring semester);
- weekly performance forecasts obtained using model LSC (for 16,476 students in the autumn semester and 15,061 students in the spring semester).

After removing data on students not covered by online learning, we received a dataset with 31,537 rows of educational data (if a person studied in both the spring and fall semesters, then the corresponding data form two rows in the dataset). The resulting dataset was nearly balanced, with 40.44% of students belonging to class 1—each of them having at least one academic debt in the corresponding semester.

We trained the XGBoost classifier with default hyperparameters while simultaneously conducting feature selection using the feature importance metric. The resulting list of the most significant predictors includes the following:

- the forecast of academic success for the semester made by model LSSp;
- week of semester;
- semester (spring or fall);
- year of study;
- total number of e-courses;
- number of assessing e-courses;
- number of frequently visited e-courses;
- average grades for assessing e-courses;
- average grades for frequently visited e-courses;
- number of active clicks for all e-courses;
- number of effective clicks for assessing e-courses;
- number of effective clicks for frequently visited e-courses;
- averages of the forecasts made by model LSC for assessing e-courses;
- averages of the forecasts made by model LSC for frequently visited e-courses.

Next, we divided the original dataset into training and testing sets 15 times. We trained the model on the training sets and calculated quality metrics on the testing sets. Figure 3 shows the average values of the classifier's quality metrics on the testing sets.

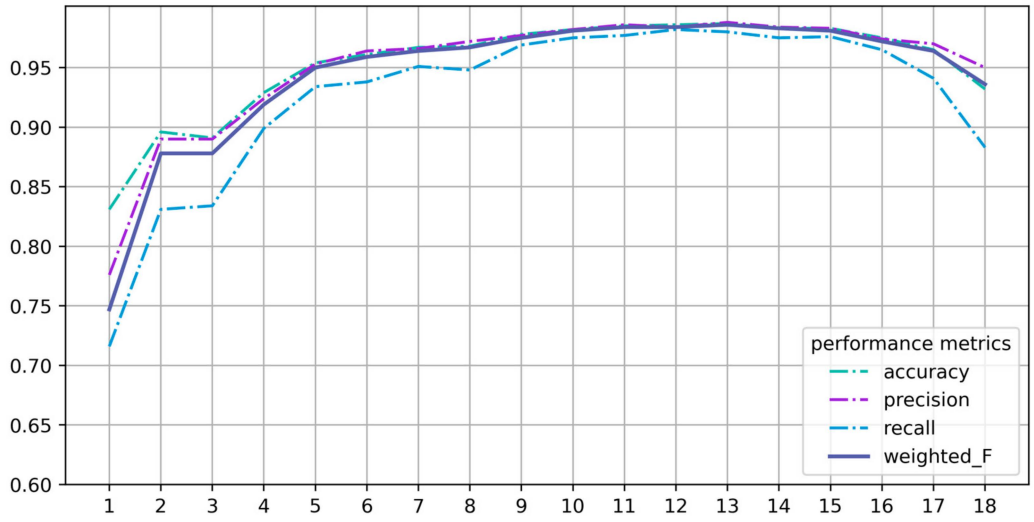


Figure 3. The dynamics of the forecasting quality of model LSSc throughout the semester.

Starting from the second week, all quality metrics of model LSSc exceed the quality metrics of model LSC. By the middle of the semester, the metrics increase and between weeks 7 and 16, they exceed 0.95. From the second to the last week of the semester, the value of the key metric for us, the weighted F-score, ranges from 0.875 to 0.98.

4. Discussion

4.1. Access to Forecasts

At the current stage of its development, the forecasting service Pythia [57,58], which incorporates the described predictive models, is currently in a pilot mode and provides forecasts exclusively to university management, heads of educational programs, heads of departments, and student office employees.

Students do not have access to forecast for several reasons. Firstly, we believe that simply informing students about the high risk of failure in their studies can be discouraging for learners—it is essential for students to simultaneously receive information on how they can improve the situation. Such guidance can be offered through learning analytics dashboards or recommender systems and should be transparent to ensure trustworthiness [59]. Secondly, as learners can base decisions about their future studies on such forecasts, all ethical considerations regarding the provision of these forecasts should be thoroughly examined. The requirements of security, accountability, and robustness to data shift should be met before offering forecasts and corresponding recommendations to students.

Furthermore, it is important to study the impact of forecasts and recommendations on student's self-regulation and motivation. These are objectives for future research.

4.2. Impact of the Models on Educational Quality at SibFU

The application of the forecasting models in the educational process allowed for real-time dynamic monitoring of students' academic performance to provide them with timely pedagogical assistance. The forecasts from the models formed the basis for classifying students into groups of low, medium, and high risk of academic underachievement.

At the School of Space and Information Technology of SibFU, a pilot study was conducted on the development and application of support measures for students from medium and high-risk groups. Several pedagogical scenarios were designed to support students from the at-risk groups. The scenarios involve:

- identifying problems faced by students through surveys and questionnaires;
- developing individual recommendations;
- providing individual mentorship and counseling support;
- incorporating additional educational resources and materials into learning;
- providing study consultations and extra classes, etc.

Pedagogical support for each student in the at-risk groups was accompanied by regular monitoring of the student's progress, changes in the forecast of their learning success, and an assessment of the adequacy of the pedagogical measures implemented.

The number of students involved in the pilot study in the 2022–2023 academic year was 2355 individuals. For these students, the developed models were used to identify at-risk students. The described pedagogical support measures were applied to work with students from the at-risk groups. By the end of the academic year, the percentage of students expelled from the institute due to academic underachievement was 8.49%. In comparison in the previous academic year, the number of students was 2261 and 10.61% were expelled due to learning failure.

However, it is unreasonable to claim that the increase in student retention by 2.12% occurred solely due to the use of predictive models and pedagogical support measures, since many other factors can influence academic success, and these factors were not considered in the study. Additional research is necessary to determine and analyze the causal relationships between the increase in student retention rates and the implementation of the developed models and support measures.

4.3. Scalability of Research Results

The study was conducted on a large sample, including students from all courses and programs of a major federal university. However, the applicability of the obtained results to predicting learning success in other HEIs should be investigated separately.

Due to the fact that the proposed ensemble of forecasting models is built taking into account both the peculiarities of legislation in the field of education and local regulatory acts of the university, as well as using educational data sources specific to a particular university, direct transfer of the model for use in another educational environment is unlikely to be feasible.

First of all, the composition of predictors in the models may change due to the following reasons:

- the preceding digital educational history may include academic performance for a different number of semesters, and data on learning outcomes at the previous educational level, such as USE in Russia, SAT or ACT in the USA, etc.;
- there may be a different set of characteristics in the digital portrait, for example, psychological characteristics of students may be added;
- the composition of predictors from the current educational history is largely determined by the form of education and the LMS used by the university.

It is a debatable question whether forecasting based on current educational history can be effective if conducted not in a blended learning environment, but in a traditional face-to-face learning environment. A crucial factor for the effectiveness of such a forecast will be the level of digitalization of the educational process—whether there is an electronic journal, and whether student academic activity is tracked in any way.

However, we believe that the proposed approach of incorporating previous education history of a student to forecast their learning success in the semester, supplemented by regular in-semester updates of the forecast using current educational history, can be implemented in most HEIs with well-developed information infrastructure.

4.4. Stability of Models' Performance

Currently, the quality of the ensemble model has been analyzed on the test dataset formed from educational data of students from the same semesters of study as in the

training dataset. However, it can be problematic to obtain the same good quality of the forecast on new data.

Although the test dataset can be considered to be IID (Independently and Identically Distributed) due to the randomness of the training-test splitting, good performance on the test dataset does not guarantee reliability of the models in reference [60]. Educational data, like other personal data, are subject to the influence of time and the data shift effect. Among the factors influencing the data under study, we can include the change in the level of provision of academic disciplines with electronic courses, which will inevitably occur, as digitization of the educational process is one of the priority tasks at the university. The analysis of the models' stability is one of the objectives of future research.

In addition, we assume that applying supportive measures to students can lead to changes in the conditional distributions of model variables. Therefore, it will be necessary to add new predictors to the model that provide information about the support measures. Currently, student offices record the dates and methods of contacting at-risk students. This information will be used in the formation of training datasets during the retraining of the model.

One of the difficulties we will encounter during the retraining of models is that true class labels are updated infrequently—about once every six months. In order to promptly identify problems with the quality of forecasting during the monitoring of model performance, it is necessary to identify one or more variables strongly associated with the response (for example, having a high correlation with it), the values of which can be obtained at any time during the academic semester.

Considering all of the above, we intend to incorporate the following aspects in future work on predictive models in production:

1. Assessment of forecasting quality on new data.
2. Analysis of the models' reliability by assessing their robustness and the speed of obsolescence.
3. Study of the data shift, addressing this issue.
4. Weekly monitoring of the models' performance.
5. Retraining of the model on new data, possibly including new predictors.

5. Conclusions

This paper presents the hybrid approach for predicting learning success in completing the current semester based on educational data gathered from various university sources. The comprehensive approach involves the use of digital personality portrait, digital educational history, and digital footprint of students, providing a holistic representation of each student, while considering their personal traits, academic performance, and learning behavior.

According to the presented approach, the system of forecasting academic success can be envisioned as a construction set consisting of several components. The first block represents a model for predicting learning success in completing the current semester based on the student's previous digital educational history and the digital personality portrait. The second block utilizes the model for predicting success in mastering the current courses of the curriculum, based on the student's digital footprint in the LMS. The third block is built on the basis of the student's current digital educational history to predict their success in completing the current semester. By combining these three blocks, the system provides accurate and comprehensive predictions of students' learning success in the current academic period.

One of the advantages of the hybrid approach is the ability to obtain predictions for learning success even in the complete absence of digital footprint data in the LMS. In such cases, the forecast will be based on the student's previous digital educational history and digital personality portrait. In future research, we plan to expand the scope of data included in the personality portrait, incorporating psychological characteristics, which will enable users to predict academic success more accurately.

Based on the hybrid approach, we developed base-models LSSp, LSC, and meta-model LSSc for forecasting learning success. They were trained and validated on educational data from SibFU databases. Model LSSc shows good quality of prediction on test sets—the average weighted F-score ranges from 0.875 to 0.98 starting from the second week of the semester.

Models LSSp, LSC, and LSSc are implemented in the academic performance forecasting service Pythia at SibFU and currently are undergoing validation on new data. For further successful integration of the model into the educational process, a well-developed information infrastructure and continuous monitoring of the model's operation are essential. This will enable the timely identification of any potential issues or discrepancies between the model's results and real-world practice, facilitating prompt adjustments to the model.

Based on the hybrid approach, we developed the forecasting system, which is effective in predicting students' academic performance and timely identifying at-risk students. Thus, we believe that other HEIs with well-developed information infrastructure will benefit from utilizing this approach for creating their systems of forecasting learning success and improving student retention.

Author Contributions: Conceptualization, M.V.N., Y.V.V., T.A.K. and R.V.E.; software, T.A.K. and R.V.E.; validation, R.V.E. and T.A.K.; formal analysis, R.V.E., T.A.K. and Y.V.V.; data curation, T.A.K. and R.V.E.; writing—original draft preparation, T.A.K., R.V.E., Y.V.V. and M.V.N.; writing—review and editing, R.V.E. and T.A.K.; supervision, M.V.N.; project administration, Y.V.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of the School of Space and Information Technology of Siberian Federal University (10/2, 28 May 2021).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The samples of depersonalized data used in this study can be provided upon appropriate request.

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

References

1. Tight, M. Student retention and engagement in higher education. *J. Furth. High. Educ.* **2020**, *44*, 689–704. [CrossRef]
2. Tinto, V. Dropout from higher education: A theoretical synthesis of recent research. *Rev. Educ. Res.* **1975**, *45*, 89–125. [CrossRef]
3. Lau, L.K. Institutional factors affecting student retention. *Education* **2003**, *124*, 126.
4. Varol, S.; Catma, S. Assessing the impact of a distance-based spatial factor on retention in the US colleges. *Educ. Sci.* **2021**, *11*, 508. [CrossRef]
5. Birliği, E.İ.; Teşkilatı, K. *Education at a Glance 2021: OECD Indicators*; OECD Publishing: Paris, France, 2021.
6. The Ministry of Science and Higher Education of Russia. Available online: https://minobrnauki.gov.ru/action/stat/highed/ (accessed on 14 April 2024).
7. Delen, D. A comparative analysis of machine learning techniques for student retention management. *Decis. Support Syst.* **2010**, *49*, 498–506. [CrossRef]
8. Addison, L.; Williams, D. Predicting student retention in higher education institutions (HEIs). *High. Educ. Ski. Work.-Based Learn.* **2023**, *13*, 865–885. [CrossRef]
9. Barbera, S.A.; Berkshire, S.D.; Boronat, C.B.; Kennedy, M.H. Review of undergraduate student retention and graduation since 2010: Patterns, predictions, and recommendations for 2020. *J. Coll. Stud. Retent. Res. Theory Pract.* **2020**, *22*, 227–250. [CrossRef]
10. Behr, A.; Giese, M.; Teguim Kamdjou, H.D.; Theune, K. Dropping out of university: A literature review. *Rev. Educ.* **2020**, *8*, 614–652. [CrossRef]
11. Arizmendi, C.J.; Bernacki, M.L.; Raković, M.; Plumley, R.D.; Urban, C.J.; Panter, A.; Greene, J.A.; Gates, K.M. Predicting student outcomes using digital logs of learning behaviors: Review, current standards, and suggestions for future work. *Behav. Res.* **2023**, *55*, 3026–3054. [CrossRef]
12. Alam, A.; Mohanty, A. Predicting Students' Performance Employing Educational Data Mining Techniques, Machine Learning, and Learning Analytics. In Proceedings of the International Conference on Communication, Networks and Computing, Gwalior, India, 8–10 December 2022; Springer: Cham, Switzerland, 2022; pp. 166–177.

13. Romine, K.D.; Baker, R.M.; Romine, K.A. The relationship between administrative intensity and student retention and success: A three-year study. *Educ. Sci.* **2018**, *8*, 159. [CrossRef]
14. Tait, J. The tutor/facilitator role in student retention. *Open Learn. J. Open Distance e-Learn.* **2004**, *19*, 97–109. [CrossRef]
15. Dixon, W.J. *Predicting Student Retention Using Scholarship and Grant Aid*; Liberty University: Lynchburg, VA, USA, 2018.
16. Sghir, N.; Adadi, A.; Lahmer, M. Recent advances in Predictive Learning Analytics: A decade systematic review (2012–2022). *Educ. Inf. Technol.* **2023**, *28*, 8299–8333. [CrossRef] [PubMed]
17. Lau, E.; Sun, L.; Yang, Q. Modelling, prediction and classification of student academic performance using artificial neural networks. *SN Appl. Sci.* **2019**, *1*, 982. [CrossRef]
18. Gafarov, F.M.; Rudneva, Y.B.; Sharifov, U.Y. Predictive Modeling in Higher Education: Determining Factors of Academic Performance. *High. Educ. Russ.* **2023**, *32*, 51–70. (In Russian) [CrossRef]
19. Yakubu, M.N.; Abubakar, A.M. Applying machine learning approach to predict students' performance in higher educational institutions. *Kybernetes* **2022**, *51*, 916–934. [CrossRef]
20. Huang, C.; Zhou, J.; Chen, J.; Yang, J.; Clawson, K.; Peng, Y. A feature weighted support vector machine and artificial neural network algorithm for academic course performance prediction. *Neural Comput. Appl.* **2023**, *35*, 11517–11529. [CrossRef]
21. Mengash, H.A. Using data mining techniques to predict student performance to support decision making in university admission systems. *IEEE Access* **2020**, *8*, 55462–55470. [CrossRef]
22. Amirhajlou, L.; Sohrabi, Z.; Alebouyeh, M.R.; Tavakoli, N.; Haghighi, R.Z.; Hashemi, A.; Asoodeh, A. Application of data mining techniques for predicting residents' performance on pre-board examinations: A case study. *J. Educ. Health Promot.* **2019**, *8*, 108. [CrossRef]
23. Alsariera, Y.A.; Baashar, Y.; Alkawsi, G.; Mustafa, A.; Alkahtani, A.A.; Ali, N. Assessment and evaluation of different machine learning algorithms for predicting student performance. *Comput. Intell. Neurosci.* **2022**, *2022*, 4151487. [CrossRef]
24. Buschetto Macarini, L.A.; Cechinel, C.; Batista Machado, M.F.; Faria Culmant Ramos, V.; Munoz, R. Predicting students success in blended learning—Evaluating different interactions inside learning management systems. *Appl. Sci.* **2019**, *9*, 5523. [CrossRef]
25. Sharabiani, A.; Karim, F.; Sharabiani, A.; Atanasov, M.; Darabi, H. An enhanced bayesian network model for prediction of students' academic performance in engineering programs. In Proceedings of the 2014 IEEE global engineering education conference (EDUCON), Istanbul, Turkey, 3–5 April 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 832–837.
26. Jovanović, J.; Saqr, M.; Joksimović, S.; Gašević, D. Students matter the most in learning analytics: The effects of internal and instructional conditions in predicting academic success. *Comput. Educ.* **2021**, *172*, 104251. [CrossRef]
27. Bravo-Agapito, J.; Romero, S.J.; Pamplona, S. Early prediction of undergraduate Student's academic performance in completely online learning: A five-year study. *Comput. Hum. Behav.* **2021**, *115*, 106595. [CrossRef]
28. Alhothali, A.; Albsisi, M.; Assalahi, H.; Aldosemani, T. Predicting student outcomes in online courses using machine learning techniques: A review. *Sustainability* **2022**, *14*, 6199. [CrossRef]
29. Kocsis, Á.; Molnár, G. Factors influencing academic performance and dropout rates in higher education. *Oxf. Rev. Educ.* **2024**, 1–19. [CrossRef]
30. Llauró, A.; Fonseca, D.; Villegas, E.; Aláez, M.; Romero, S. Improvement of Academic Analytics Processes Through the Identification of the Main Variables Affecting Early Dropout of First-Year Students in Technical Degrees. A Case Study. *Int. J. Interact. Multimed. Artif. Intell.* **2023**, 1–12. [CrossRef]
31. Kukkar, A.; Mohana, R.; Sharma, A.; Nayyar, A. Prediction of student academic performance based on their emotional wellbeing and interaction on various e-learning platforms. *Educ. Inf. Technol.* **2023**, *28*, 9655–9684. [CrossRef]
32. Alyahyan, E.; Düştegör, D. Predicting academic success in higher education: Literature review and best practices. *Int. J. Educ. Technol. High. Educ.* **2020**, *17*, 3. [CrossRef]
33. Tempelaar, D.; Rienties, B.; Nguyen, Q. Subjective data, objective data and the role of bias in predictive modelling: Lessons from a dispositional learning analytics application. *PLoS ONE* **2020**, *15*, e0233977. [CrossRef] [PubMed]
34. Tempelaar, D.T.; Rienties, B.; Giesbers, B. In search for the most informative data for feedback generation: Learning analytics in a data-rich context. *Comput. Hum. Behav.* **2015**, *47*, 157–167. [CrossRef]
35. Federal Law of the Russian Federation No. 273 "On Education in the Russian Federation". Available online: https://www.consultant.ru/document/cons_doc_LAW_140174/ (accessed on 14 April 2024).
36. Decree of the Ministry of Higher Education and Science of the Russian Federation No. 245 "On Approval of the Procedure for Organizing and Implementing Educational Activities According to Higher Education Program—Bachelor's Programs, Specialist Programs, Master's Programs". Available online: https://www.consultant.ru/document/cons_doc_LAW_393023/ (accessed on 14 April 2024).
37. Siberian Federal University Regulation on Student Current Assessment and Interim Attestation. Available online: https://about.sfu-kras.ru/node/8809 (accessed on 14 April 2024).
38. Siberian Federal University Regulation on Electronic Information Educational Environment of SibFU. Available online: https://about.sfu-kras.ru/node/10065 (accessed on 14 April 2024).
39. Moodle. Available online: https://moodle.org/ (accessed on 14 April 2024).
40. Siberian Federal University Regulation on Electronic Learning and Distance Learning Technology Implementation. Available online: https://about.sfu-kras.ru/node/9739 (accessed on 14 April 2024).

41. General Data Protection Regulation (EU GDPR). Available online: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679 (accessed on 14 April 2024).
42. Federal Law of the Russian Federation No. 149 "On Information, Information Technologies and Protection of Information". Available online: https://www.consultant.ru/document/cons_doc_LAW_61798/ (accessed on 14 April 2024).
43. Federal Law of the Russian Federation No. 152 "On Personal Data". Available online: https://pd.rkn.gov.ru/docs/Federal_Law_On_personal_data.doc (accessed on 14 April 2024).
44. Siberian Federal University Personal Data Regulation. Available online: https://about.sfu-kras.ru/docs/9214/pdf/928460 (accessed on 14 April 2024).
45. Siberian Federal University Informed Consent from Students About Using the Electronic Learning Environment. Available online: https://edu.sfu-kras.ru/node/5032 (accessed on 14 April 2024).
46. Esin, R.V.; Zykova, T.V.; Kustitskaya, T.A.; Kytmanov, A.A. Digital educational history as a component of the digital student's profile in the context of education transformation. *Perspect. Sci. Educ.* **2022**, *59*, 566–584. (In Russian) [CrossRef]
47. Kustitskaya, T.A.; Esin, R.V.; Kytmanov, A.A.; Zykova, T.V. Designing an Education Database in a Higher Education Institution for the Data-Driven Management of the Educational Process. *Educ. Sci.* **2023**, *13*, 947. [CrossRef]
48. Balyakin, A.; Mamonov, M.; Nurbina, M.; Taranenko, S. Digital Footprint and Education: Some Remarks. In *Perspectives and Trends in Education and Technology: Selected Papers from ICITED 2021*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 485–493.
49. Esin, R.V.; Kustitskaya, T.A.; Noskov, M.V. Predicting academic performance in a course by universal features of LMS Moodle digital footprint. *Inform. Educ.* **2023**, *38*, 31–41. (In Russian) [CrossRef]
50. Noskov, M.; Somova, M.; Fedotova, I. Management of the success of student's learning based on the markov model. *Inform. Educ.* **2018**, *10*, 4–11. (In Russian) [CrossRef]
51. Tolmachev, A.; Sinitsyn, E.; Astratova, G. Mathematical Modeling and Forecasting of Student's Academic Performance on Massive Online Courses. In Proceedings of the 2nd International Scientific and Practical Conference on Digital Economy (ISCDE 2020), Yekaterinburg, Russia, 5–6 November 2020; Atlantis Press: Amsterdam, The Netherlands, 2020; pp. 121–127.
52. Noskov, M.; Vainshtein, Y.; Somova, M.; Fedotova, I. Prognostic model for assessing the success of subject learning in conditions of digitalization of education. *RUDN J. Informatiz. Educ.* **2023**, *20*, 7–19. (In Russian) [CrossRef]
53. Noskov, M.; Fedotova, I.; Somova, M.; Vainshtein, Y. Mathematical model for early prediction of learning success. In *AIP Conference Proceedings*; AIP Publishing: College Park, MD, USA, 2024.
54. Hussain, M.; Zhu, W.; Zhang, W.; Abidi, S.M.R. Student engagement predictions in an e-learning system and their impact on student course assessment scores. *Comput. Intell. Neurosci.* **2018**. [CrossRef] [PubMed]
55. Galimova, E.G.; Konysheva, A.V.; Kalugina, O.A.; Sizova, Z.M. Digital Educational Footprint as a Way to Evaluate the Results of Students' Learning and Cognitive Activity in the Process of Teaching Mathematics. *EURASIA J. Math. Sci. Technol. Educ.* **2019**, *15*, em1732.
56. Tamada, M.M.; Giusti, R.; de Magalhães Netto, J.F. Predicting student performance based on logs in Moodle LMS. In Proceedings of the 2021 IEEE Frontiers in Education Conference (FIE), Lincoln, NE, USA, 13–16 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–8.
57. Siberian Federal University Forecasting Service Pythia. Available online: https://p.sfu-kras.ru (accessed on 14 April 2024).
58. Santyev, E.A.; Zakharyin, K.N.; Shniperov, A.N.; Kurchenko, R.A.; Shefer, I.A.; Vainshtein, Y.V.; Somova, M.V.; Fedotova, I.M.; Noskov, M.V.; Kustitskaya, T.A.; et al. "Pythia" Academic Performance Prediction System. 2023. Available online: https://www.elibrary.ru/download/elibrary_56002717_37844521.PDF (accessed on 14 April 2024).
59. Baneres, D.; Guerrero-Roldán, A.E.; Rodríguez-González, M.E.; Karadeniz, A. A predictive analytics infrastructure to support a trustworthy early warning system. *Appl. Sci.* **2021**, *11*, 5781. [CrossRef]
60. Araujo Santos, V.C.; Cardoso, L.; Alves, R. The quest for the reliability of machine learning models in binary classification on tabular data. *Sci. Rep.* **2023**, *13*, 18464. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.