





Review

# Predicting Student Performance in Introductory Programming Courses

João P. J. Pires <sup>1</sup>, Fernanda Brito Correia <sup>1</sup>, Anabela Gomes <sup>1,2</sup>, Ana Rosa Borges <sup>1</sup> and Jorge Bernardino <sup>1,2,\*</sup>

<sup>1</sup> Coimbra Institute of Engineering—ISEC, Polytechnic University of Coimbra, Rua da Misericórdia, Lagar dos Cortiços, S. Martinho do Bispo, 3045-093 Coimbra, Portugal; a21280231@isec.pt (J.P.J.P.); fernanda@isec.pt (F.B.C.); anabela@isec.pt (A.G.); arborges@isec.pt (A.R.B.)

<sup>2</sup> Centre for Informatics and Systems of the University of Coimbra, Polo II, Pinhal de Marrocos, 3030-290 Coimbra, Portugal

\* Correspondence: jorge@isec.pt

**Abstract:** The importance of accurately predicting student performance in education, especially in the challenging curricular unit of Introductory Programming, cannot be overstated. As institutions struggle with high failure rates and look for solutions to improve the learning experience, the need for effective prediction methods becomes critical. This study aims to conduct a systematic review of the literature on methods for predicting student performance in higher education, specifically in Introductory Programming, focusing on machine learning algorithms. Through this study, we not only present different applicable algorithms but also evaluate their performance, using identified metrics and considering the applicability in the educational context, specifically in higher education and in Introductory Programming. The results obtained through this study allowed us to identify trends in the literature, such as which machine learning algorithms were most applied in the context of predicting students' performance in Introductory Programming in higher education, as well as which evaluation metrics and datasets are usually used.

**Keywords:** higher education; introductory programming; machine learning; student's performance prediction



**Citation:** Pires, J.P.J.; Brito Correia, F.; Gomes, A.; Borges, A.R.; Bernardino, J. Predicting Student Performance in Introductory Programming Courses. *Computers* **2024**, *13*, 219. <https://doi.org/10.3390/computers13090219>

Academic Editors: Stamatios Papadakis and Stelios Xinogalos

Received: 4 July 2024

Revised: 13 August 2024

Accepted: 2 September 2024

Published: 5 September 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

Introductory Programming (IP) continues to be a challenge for higher education (HE) students, with high failure rates [1,2]. Different studies have attempted to analyze and identify possible difficulties associated with learning to program, including aspects such as the complex syntax of programming languages [3] and conceptual knowledge related to how the language itself works [4]. The difficulties reported by students and teachers in these studies included understanding programming structures; dividing functionality into procedures; finding bugs in one's own program; and understanding variables, functions, loop structures, recursion, arrays, pointers and references, structured data types, abstract data types, and input/output handling. Several reasons have been identified to explain these difficulties [5], namely (i) teaching methods; (ii) study methods; (iii) students' abilities and attitudes; (iv) the nature of programming; and (v) psychological effects. This has motivated research into the factors that influence the learning process and how to overcome them [6]. It is possible to find references in the literature to several other factors that influence the learning experience of programming, such as the ability to think abstractly [7], previous experience of programming [8,9], the effect of teaching approaches that take into account students' motivational strategies and learning preferences and styles [10,11], mathematics skills [12], or problem-solving skills [13].

The significant problem of student failure underlines the need for effective interventions, as outlined in several studies that have been carried out with the aim of analyzing and providing possible solutions to this problem [14]. One of the solutions often analyzed

by researchers focuses on analyzing and improving the usual methods of teaching programming in HE [15,16]. Numerous tools and strategies were also developed to contribute to the solution of this problem, and although each one has its merits, none of them has yet completely solved the problem [16]. Among the various strategies studied, the use of methods to predict student performance stands out as a possible solution [17,18]. While several metrics have been used to evaluate student performance, Quille and Bergin [19] introduced the PreSS metric, which focuses specifically on predicting student success in IP courses. The authors propose a relevant approach by developing a prediction model called PreSS, which has been validated and refined over 13 years. The paper discusses the challenges of predicting student success in CS1, such as high attrition rates and the difficulty of identifying students at risk. The authors found that PreSS can predict student success with 71% accuracy and that it can be used to develop interventions to improve student performance. Although these results are positive, 71% is still not ideal and highlights the difficulties that remain in finding solutions.

In this study, student performance is defined as their proficiency in IP, measured through a combination of grades in formal assessments (such as exams and assignments), participation in practical activities, and performance on coding tasks. This definition is consistent with the literature, which emphasizes the importance of assessing not only theoretical knowledge, but also students' practical programming skills [20].

This study delves into this area with the aim of providing a comprehensive and systematic review of the existing literature on the methodologies used to predict student performance in HE, particularly in programming, focusing on machine learning (ML) as a prediction method. It emphasizes the application of these ML algorithms, identifying which algorithms have been used and their performance, and analyzing which evaluation metrics and datasets have been used. ML algorithms are extremely useful techniques that are commonly used to develop models for predicting student performance [21].

Data mining is the extraction of important and potentially useful information or trends from data. Educational data mining (EDM) is the application of data mining techniques in the educational context with educational data. ML is one of the data mining techniques that is widely used in the field of EDM [22,23]. The emphasis on ML in this study is due to its ability to handle large datasets, discover complex patterns, and make predictions with high accuracy. There are other techniques, such as traditional statistical methods, that can also be used in the process of studying student performance. However, although statistical methods are well established and easy to interpret, they are not able to capture complex patterns in the data as effectively as ML techniques. On the other hand, ML can offer greater accuracy and adaptability, with the slight disadvantage of requiring more computational resources and specialized knowledge to implement.

The area of IP has a profound significance for aspiring computer engineers, serving as the fundamental basis on which their entire educational and professional path is built. Aptitude in programming is not just a prerequisite but an essential skill that forms the core of a future engineer's capabilities. This skill lays the foundations for understanding complex algorithms, problem-solving methodologies, and the intrinsic aspects of software development [24]. This systematic literature review aims to address the gap in understanding which ML algorithms should be applied and which are most effective in predicting student performance in IP, by analyzing their performance and examining the evaluation metrics and datasets used. In addition to other similar studies [25], this paper contributes to the field by providing a clear overview of the current state of research and identifying the most promising approaches and areas for future research, which will help in decision-making processes for the further development of the field and the search for solutions.

The remainder of this paper is structured as follows. Section 2 describes the methodology used to conduct this systematic review, highlights the methods identified, and presents the algorithms specified in related works. Section 3 identifies the studies found that address the problem under study. Section 4 summarizes the results and Section 5 presents the answers to the research questions. Section 6 presents the main threats to the validity of

this study and Section 7 identifies future research directions. Finally, Section 8 presents the conclusions.

## 2. Methodology

This study consists of a review of the current literature to contribute to identifying current trends and major difficulties and to present a comparative analysis of some of the existing studies in the context of predicting student performance in IP courses in HE, using the Kitchenham methodology [26]. This methodology describes some guidelines for developing systematic reviews. The following steps were applied: (a) definition of research questions; (b) selection of data sources; (c) definition of the search string; (d) definition of preconditions; (e) definition of inclusion and exclusion criteria; and (f) data extraction and analysis.

**Research Questions**—For this study, three research questions (RQ) were defined as follows:

RQ1: What were the most-used ML algorithms proposed by the researchers for predicting students' performance in HE in IP?

RQ2: Which datasets were used, and which evaluation metrics were considered most appropriate for measuring the performance of predictive models in HE in IP?

RQ3: Which ML algorithms seem to better predict students' performance in HE in IP?

**Data Sources**—The papers analyzed in this study were the result of searches in the search engine Google Scholar.

**Search String**—The search string used was as follows: ("introduction to programming" OR "programming learning" OR "learning programming" OR "learn to program" OR CS1) AND (novice OR freshman OR beginner) AND ("higher education") AND (aptitude OR performance OR skill OR ability OR proficiency OR facility OR talent) AND ("predict") AND ("machine learning") AND—"video" AND—"ChatGPT".

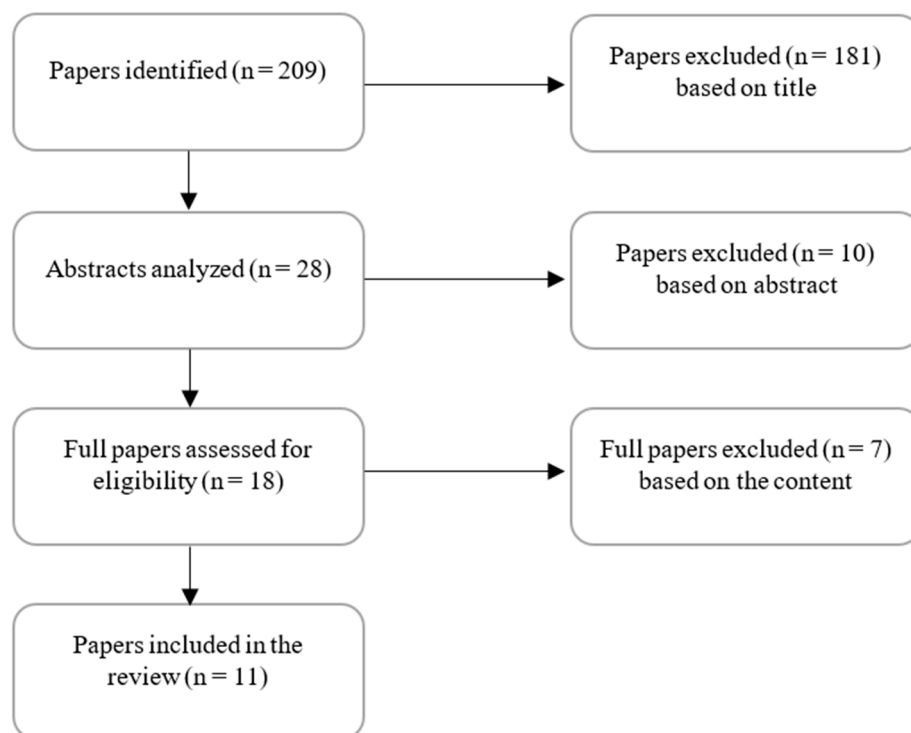
**Preconditions**—Two conditions were met as follows: publications from 2015 to 2024 and written in English.

**Inclusion and Exclusion Criteria**—To select interesting studies, the following inclusion criteria were defined: (i) responses to the research questions established. The exclusion criteria were as follows: (e1) the paper does not refer to IP; (e2) the paper does not focus on HE; (e3) the paper does not refer to the use of ML; (e4) the paper is not available; and (e5) studies others than papers were not considered (like theses and books).

**Data Extraction**—Initially, the search string was built to find papers related to this problem. Next, some preconditions were appointed for selecting the papers, and then the inclusion and exclusion criteria were defined. Once this was finished, the phase of collecting papers began. Based on the search string and the preconditions, 209 records were identified, and by checking the title, 28 papers ended up remaining. Next, the abstracts of these selected papers were analyzed and checked for their compliance with the proposed research questions and defined criteria for the remaining 18 papers. These 18 papers were read in full, taking into account the research questions and the defined criteria. The search ended with 11 papers to be analyzed. Figure 1 shows the full literature review process steps.

The aim of this study is to explore the different methods that shape the scenario of predicting students' performance in HE institutions, with particular emphasis on the field of IP and focusing on ML algorithms.

The following section gives an overview of the types of datasets used in the state of the art in this specific field and the evaluation metrics. The aim of this analysis is to understand existing strategies, inspire future innovative approaches, and improve forecast methods' accuracy and usefulness.



**Figure 1.** Literature review process steps.

### 2.1. Datasets

Proper data collection plays a critical role in building a solid foundation for research. This section examines the different datasets used in the papers presented above. Table 1 presents the datasets used in the analyzed papers.

**Table 1.** Datasets used in the analyzed papers.

Datasets	Papers
Questionnaires	[27–29]
Academic records	[27,28,30–36]
Personal data	[30,32,34,35,37]

In the papers, data were collected from different datasets:

- Questionnaires

Questionnaires consist of a set of questions that, in the context of the subject under study and, as analyzed in the related works, may or may not include questions directly related to programming. The answers to these questionnaires make it possible to extract important information about the students and their theoretical and practical skills.

- Academic Records

Class attendance, course averages, test scores, practical work, and other assessment components are perhaps the most important data for correctly predicting students' performance. In the analyzed papers, using academic records from previous years to train the model and then using academic records from the current year to predict performance was widespread.

- Personal Data

Another type of data commonly used were personal data. These data include information such as age, gender, nationality, and programming experience. These data can identify some patterns that may be associated with academic performance in IP.

## 2.2. Evaluation Metrics

The correct choice and interpretation of evaluation metrics is essential when analyzing the performance of a model. This section consists of a description of the metrics identified in the related works, which provide information about a model's predictive ability, and are used here to predict students' academic performance in IP in HE. The most important metrics that have been identified are the following:

- Accuracy-is the most common evaluation metric. It works as an indicator of the model's overall performance by determining the percentage of correctly classified instances [38].
- Precision-is an extremely useful metric, in situations where false positives are more important than false negatives because it allows us to evaluate the model and determine what percentage of instances classified as positive are positive [38].
- Recall-is a metric that can be very useful when false negatives are more important than false positives, as it allows us to calculate the percentage of true positive instances that have been correctly classified [38].
- F1-Score-is created when the precision and recall metrics are both critical. This metric combines precision and recall through a harmonic mean, so by maximizing the value of the F1-Score, we simultaneously maximize the precision and recall [38].
- Specificity-is the metric that allows calculating the percentage of truly negative instances that have been classified as negative; in other words, it assesses the ability of the model to correctly identify negative cases [38].
- Sensitivity-is the metric that allows calculating the percentage of truly positive instances that have been classified as positive [38].
- Root Mean Square Error (RMSE)-is commonly used to evaluate model performance. This metric measures the average squared difference between the values predicted by the model and the actual values. The lower the RMSE value, the better the model's performance [38].

## 3. Related Work

The following provides a detailed overview of the different methodologies identified in this systematic review, highlighting the variety of datasets, algorithms, and evaluation metrics that represent the state of the art in this specific field.

Over the last few years, many studies have been conducted on predicting students' performance in HE, with a particular focus on IP. This section presents the analysis of the papers obtained through the methodology described.

The study carried out in [27] presents a ML model that attempts to predict students at risk of poor performance. The algorithm used in this research was Gradient Boosting, which was trained on students' data obtained from questionnaires and class works conducted during the first two weeks of the semester, with a total of 471 students. The model's expected output is a number between 0 and 1, representing the probability that a given student is at risk of performing poorly.

The study presents a promising approach to predicting student performance with a relatively large database, using a good variation of assessment metrics, and achieving significantly positive accuracy and specificity results. However, there are some significant limitations, such as low recall and precision, which indicate that the model may not adequately identify all students at risk. Improvements can be made by extending the data collection period, increasing the diversity of the sample, and integrating more temporal and contextual data to increase the robustness and generalizability of the model. Table 2 is a summary of the datasets and the algorithm used, along with the metric values obtained.

The work proposed in [28] applies four ML algorithms to create predictive models capable of predicting possible students at risk of failing an IP course. The algorithms used were Decision Tree, K-Nearest Neighbour (KNN), Naïve Bayes, and Support Vector Machine (SVM). The data used to train the models were obtained from an IP course where features were the average entry grade in the course, the result of a questionnaire, practical

assessments, and other class assessments of the full semester. All this information together relates to the four semesters of the academic years 2018/2019 and 2019/2020, with a total of 244 records. The accuracy, precision, and recall metrics were used to evaluate the results of the models. Both Decision Tree and SVM achieved the highest performance, but Decision Tree was twice as fast as SVM at running the model.

**Table 2.** Datasets, algorithms, and metric values used in [27].

Dataset	Algorithm	Metric	Value
Academic records Questionnaires	Gradient Boosting	Accuracy	82%
		Recall	57%
		Precision	56%
		Specificity	88%

The study shows impressive results with high levels of precision, recall, and accuracy in various algorithms. The sample size seems to be average and realistic. The reason for such positive results may be associated with the fact that the dataset uses information from the entire semester to predict the final grade, thus having a greater number of assessment components already performed, making it easier to predict the final grade. This study does not seem to be recommended if an early prediction is needed to help at-risk students, as it requires the full semester's information. Table 3 is a summary of the datasets and algorithms used, along with the metric values obtained.

**Table 3.** Datasets, algorithms, and metric values used in [28].

Dataset	Algorithm	Metric	Value
Academic records Questionnaires	Decision Tree	Accuracy	99.18%
		Precision	100%
		Recall	98.7%
	KNN	Accuracy	98.36%
		Precision	99.3%
		Recall	100%
	Naïve Bayes	Accuracy	96.72%
		Precision	97.4%
		Recall	97.4%
	SVM	Accuracy	99.18%
		Precision	100%
		Recall	98.7%

ML techniques are also used to identify high and low performers in the study carried out in [30]. The data considered for the models are from one semester of an IP course at the University of Helsinki with a total of 86 students. Age, gender, grade point average, and programming experience are some information available in the data used. The intended results were the prediction of the final grade of the course. The algorithms used were ADTree, Bayesian Network, Conjunctive Rule, Decision Stump, Decision Table, J48, Naïve Bayes, and PART. The authors tried to predict the students' exam results to understand which algorithm was the best between the models with the best accuracy, Random Forest, and Decision Stump. Random Forest obtained a 90% accuracy and Decision Stump obtained 83%.

Overall, the study showed positive results for practically all the algorithms used. However, the dataset used is very limited, which introduces potential bias due to its reduced diversity. In addition, the fact that the authors relied solely on accuracy as the only evaluation metric may have limited their understanding of the model's performance. In the context of the problem, the use of other metrics, such as specificity and sensitivity, can be very useful in understanding whether or not you are dealing with a case where the

models are unable to generalize. Table 4 gives a summary of the datasets and algorithms used, along with the metric values obtained.

**Table 4.** Datasets, algorithms, and metric values used in [30].

Dataset	Algorithm	Metric	Value
Academic records Personal data	ADTree	Accuracy	83%
	Bayesian Network	Accuracy	76%
	Conjunctive Rule	Accuracy	86%
	Decision Stump	Accuracy	90%
	Decision Table	Accuracy	76%
	J48	Accuracy	89%
	Naïve Bayes	Accuracy	86%
	PART	Accuracy	76%
	Random Forest	Accuracy	90%

The work carried out in [31] aimed to identify students with learning difficulties at an earlier stage to give them more attention in the form of extra class hours. Identifying students with difficulties was performed by predicting the probability of completing the course using an ML model. Data from 181 students, from the previous year, were used as training data for the model. These data contain the grades obtained in three teaching activities that were carried out every year during the first five weeks of the course: seminar, quiz, and practical work. The problem presented was a binary classification, for which the authors decided to use the Logistic Regression algorithm, where the result was “yes” (1) or “no” (0) to know if the student would pass or fail. The precision, recall, and F1-Score metrics were used to check the quality of the model.

The study presents a relatively different approach to identifying students with learning difficulties early. The high recall rate is a major strength, ensuring that most students who pass are identified. However, the moderate precision may indicate difficulties in predicting students who are unlikely to pass. The sample size is relatively small and may therefore represent a conditioned sample. Possible improvements would be to expand the sample and explore different algorithms or combinations of algorithms to improve model precision. A summary of the datasets and the algorithm used, together with the metric values obtained, is presented in Table 5.

**Table 5.** Datasets, algorithms, and metric values used in [31].

Dataset	Algorithm	Metric	Value
Academic records	Logistic Regression	Precision	67%
		Recall	92%
		F1-Score	77%

Three ML algorithms to predict potential student grades at an early stage in the semester were used in the study of [32]. The outcome to be expected was a value from 0 to 100 that indicates the student’s final grade. With this study, the authors aim to help students identify their likely grades and, if necessary, modify their academic behavior. The data used in the study were collected at the University of Buraimi. The dataset consists of information on 50 students and includes attributes such as gender, first-test grade, and attendance, among others. The performance of the algorithms used was evaluated, based on the accuracy metric and F1-Score. The three algorithms were Decision Table, J48, and Naïve Bayes.

The study presents good results for the three algorithms used. However, the extremely small sample size may result in the models not being able to generalize, which is the main limitation of this study as ML algorithms trained on small datasets may not achieve satisfactory results. It was recommended to use metrics such as specificity or sensitivity to understand whether this generalization capacity exists. The exploration and addition of new data would add further relevance to the study results. Table 6 is a summary of the datasets and algorithms used, along with the metric values obtained.

**Table 6.** Datasets, algorithms, and metric values used in [32].

Dataset	Algorithm	Metric	Value
Academic records Personal data	Decision Table	Accuracy	83%
		F1-Score	79%
	J48	Accuracy	88%
		F1-Score	88%
Naïve Bayes	Accuracy	84%	
	F1-Score	83%	

The study of [33] presents a model developed to try to identify in advance students at risk of failing an IP course by predicting the final exam grade. It used data from assignments and class exercises carried out in the first two weeks of the semester, and these data corresponded to the years 2016, 2017, and 2018. The developed model used the Random Forest algorithm, where the 2016 data with a total of 93 records were used for training, the 2017 data with 94 records were used for validation, and the 2018 data with 102 records were used for testing. The results obtained for Random Forest show a low overall accuracy, but a good sensitivity in predicting students that are at risk. According to the authors, the results obtained were not the best due to the lack of balance in the dataset used.

The authors acknowledge the limitations of the study itself, pointing to the inability of the model to learn on an unbalanced dataset. However, it should be noted that the model appears to have a good ability to identify students who are at risk. Nevertheless, it later fails to predict students who are not at risk. A limitation of the study is the small size of the sample used to train the model. Table 7 is a summary of the datasets and the algorithm used, together with the metric values obtained.

**Table 7.** Datasets, algorithms, and metric values used by [33].

Dataset	Algorithm	Metric	Value
Academic records	Random Forest	Accuracy	60%
		Sensitivity	77%

Different neural networks to identify at-risk students were tested in the work presented in [34]. The study used historical students' data over the years in IP courses in HE. These data contained personal information and records of different assessment grades given during the semester over the last 7 years. In total, the dataset contained information of 592 students. The authors tested 25 different network topologies and concluded that the Probabilistic Neural Network (PNN) obtained the best results.

The results of this study are very positive. The sample size is one of the largest compared to all the papers analyzed so far. Although in the end the authors only use the accuracy metric, the paper presents the corresponding confusion matrix of the neural network, which allows us to verify the good ability of the model to classify both students at risk and those who are not at risk. The dataset also presents a good balance of classes. This study shows the potential of neural networks. A summary of the datasets and the algorithms used, together with the metric value obtained, is given in Table 8.



**Table 8.** Datasets, algorithms, and metric values used by [34].

Dataset	Algorithm	Metric	Value
Academic records Personal data	PNN	Accuracy	91%

In [29], the study applied different algorithms for predicting students' performance in programming in HE. The data were obtained from the University of Madras, India, and include information about the students collected through a questionnaire. These data were collected as part of an introductory Python programming course in the first semester of the first year of the course. In total, the data contained records for 490 students. In this study, five different algorithms were tested and evaluated using the recall, precision, F1-Score, and accuracy metrics. The algorithms were the following: Decision Tree, KNN, Naïve Bayes, Random Forest, and SVM. Overall, Naïve Bayes obtained the best results.

The study presents a comprehensive and robust analysis using a variety of machine learning algorithms to predict student performance. The evaluation metrics used are relevant. The sample size also appears to be representative and complete. Considering that the study only uses data from a questionnaire, the results are very good. Table 9 summarizes the datasets and algorithms used, and the metric values obtained.

**Table 9.** Datasets, algorithms, and metric values used in [29].

Dataset	Algorithm	Metric	Value
Questionnaires	Decision Tree	Accuracy	85.10%
		Precision	86.57%
		Recall	84.89%
		F1-Score	85.72%
	KNN	Accuracy	84.28%
		Precision	85.82%
		Recall	83.46%
		F1-Score	84.62%
	Naïve Bayes	Accuracy	91.02%
		Precision	92.21%
		Recall	90%
		F1-Score	91.09%
	Random Forest	Accuracy	87.55%
		Precision	86.06%
		Recall	88.60%
		F1-Score	87.31%
SVM	Accuracy	88.77%	
	Precision	89.79%	
	Recall	88%	
	F1-Score	88.88%	

The work carried out in [37] applies the J48, Multilayer Perceptron, Naïve Bayes, REPTree, and Sequential Minimal Optimization algorithms to predict students' academic performance in an IP course. The data for the models were collected from a C programming course at the University of Madras, India. The dataset contains 300 records with demographic information about the students. The results obtained were based on accuracy. Out of these algorithms, the Multilayer Perceptron showed the best performance.

Although this study concluded that the five models achieved high accuracy in predicting student performance, relying on accuracy as the sole evaluation metric may have a limited understanding of model performance. More metrics are needed to better under-

stand whether the results are good. Good accuracy does not always mean good prediction. Table 10 provides an overview of the datasets, algorithms, and metrics that were collected.

**Table 10.** Datasets, algorithms, and metric values used in [37].

Dataset	Algorithm	Metric	Value
Personal data	J48	Accuracy	92.03%
	Multilayer Perceptron	Accuracy	93.23%
	Naïve Bayes	Accuracy	84.46%
	REPTree	Accuracy	91.03%
	Sequential Minimal Optimization	Accuracy	90.03%

The study in [35] takes a different approach from the previous work. Here, a deep learning model was developed and used data, such as the student's identity, professional skills, and academic records, including the code developed by the IP courses in HE students. The main goal is to predict the student's performance in the final exam; it is a regression problem. The authors ran several tests to decide which architecture to use and concluded that a Deep Neural Network (DNN) with four fully connected layers was a good architecture. In this DNN, the input layer had 128 neurons, the second layer had 64, the third layer had 8, and the output layer had only 1. The activation function used was the Rectified Linear Unit (ReLU), and the evaluation metric selected was the RMSE. The results were 12.68 RMSE for the DNN. Finally, a comparison was made with six other algorithms (SVM, Bayesian Ridge, Random Forest, Extra Trees, Gradient Boosting, and Decision Tree), where the SVM stood out as having the best performance, although still below the DNN, with an RMSE of 14.07.

This study explores a new aspect of predicting student performance and shows the great potential of deep learning for the context. A negative point identified was the failure to clarify the size of the dataset used to train the models. The results are very promising and consistent with the literature, suggesting that the way forward could be through deep learning, as these architectures show constant improvements in performance. Table 11 provides a summary of the datasets and the algorithm used, together with the metric values obtained.

**Table 11.** Datasets, algorithms, and metric values used in [35].

Dataset	Algorithm	Metric	Value
Academic records	DNN	RMSE	12.68
Personal data	SVM	RMSE	14.07

Finally, in [36], the study applied the ID3 and J48 classification algorithms to analyze students' performance in an IP course. The data used to develop the models came from the first semester of an IP course that ran for three months, from September to November 2017, at Usmanu Danfodiyo University in Nigeria. The data consisted of 239 instances with information on class attendance, test grades, and completed assignments. The results of the models were evaluated based on four metrics: accuracy, precision, recall, and F1-Score. The results of both algorithms were similar, so to clarify what was the best one, the authors analyzed the number of instances that were correctly classified; J48 obtained 208 and ID3 obtained 204 out of 239 total instances, so the authors concluded that J48 slightly outperformed ID3.

This study achieved very positive results for both models. The metrics used are relevant and sufficient to understand the quality of the models' predictions, and the sample used is of an acceptable size. Improvements can be made by extending the dataset. A

summary of the datasets and algorithms used, together with the metric values obtained, is shown in Table 12.

**Table 12.** Datasets, algorithms, and metric values used in [36].

Dataset	Algorithm	Metric	Value
Academic records	ID3	Accuracy	85.4%
		Precision	84.8%
		Recall	86%
		F1-Score	85%
	J48	Accuracy	87%
		Precision	80%
		Recall	87%
		F1-Score	82.8%

#### 4. Results

This section summarizes the results, identifying the main findings and trends found from the literature reviewed, the models and algorithms used, the most-used evaluation metrics and datasets, and the algorithms that gave the best results including the deep learning approach. The identification of patterns provides a more comprehensive overview of the characteristics studied that are most relevant in predicting students' academic success in IP in HE.

In addition to summarizing the key findings, this section highlights the contributions that promise to catalyze future progress in this important area.

##### 4.1. Models

The application of ML models has emerged as a fundamental approach to obtain valuable details about students' performance in IP courses in HE. Table 13 lists the ML models used in the analyzed papers. Each model is presented along with the references for the specific papers in which they were implemented. In total, twenty different algorithms were used in all the analyzed papers. Some of these algorithms were used in more than one paper, and there were studies that applied and compared multiple models.

**Table 13.** Models used in analyzed papers.

Algorithms	Papers
ADTree	[30]
Bayesian Network	[30]
Conjunctive Rule	[30]
Decision Stump	[30]
Decision Table	[30,32]
Decision Tree	[28,29]
DNN	[35]
Gradient Boosting	[27]
ID3	[36]
J48	[30,32,36,37]
KNN	[28,29]
Logistic Regression	[31]
Multilayer Perceptron	[37]
Naïve Bayes	[28–30,32,37]
PART	[30]
PNN	[34]
Random Forest	[29,30,33]
REPTree	[37]
Sequential Minimal Optimization	[37]
SVM	[28,29,35]

#### 4.2. Most-Used Algorithms

The presentation of the results made it possible to see some patterns and extract some important knowledge. It was possible to identify that the algorithms J48, Naïve Bayes, Random Forest, and SVM tend to be the most used in three or more papers. Other algorithms such as Decision Tree, Decision Table, and KNN are also used, although less frequently, in only two papers. However, although simpler algorithms such as Random Forest and Naïve Bayes are often used, it is important to be aware that more complex algorithms such as DNNs, even if they are used less frequently, can provide superior prediction performance. The paper of [35] that uses a DNN demonstrate this, as does the paper of [34] that uses a PNN on a big dataset and obtains some of the best results identified.

#### 4.3. Most-Used Evaluation Metrics

It was also found that a wide range of evaluation metrics, including precision, recall, and F1-score, are extremely useful and used in this context. However, none of these metrics come close to the levels of use of the accuracy metric, which was used in nine of the eleven papers in the literature review. The predominant use of the accuracy metric can have a significant impact on the way predictive models are evaluated. This suggests that many researchers prefer a simple and straightforward metric, possibly due to its ease of interpretation. However, this may also lead to an underestimation of the limitations of accuracy, particularly in unbalanced datasets where the metric can be misleading.

#### 4.4. Most-Used Dataset

Among the datasets identified, academic records proved to be the most frequently used data for predicting performance in IP in HE, being reported in nine of the eleven papers analyzed. This makes sense since past records of a student's grades allow many conclusions to be drawn about the level of a student's skills and competences.

#### 4.5. Datasets Analysis

When analyzing the different papers selected, constant variations were identified in the types of datasets used (academic records, personal data, and questionnaires) as well as in the size of the dataset itself. While academic records are the most common data source in the reviewed studies, it is important to analyze whether other types of data, such as questionnaires and personal data, can contribute to the accuracy of predictions. More detailed analysis of the datasets is therefore needed to understand the extent to which these variations have an influence on the results.

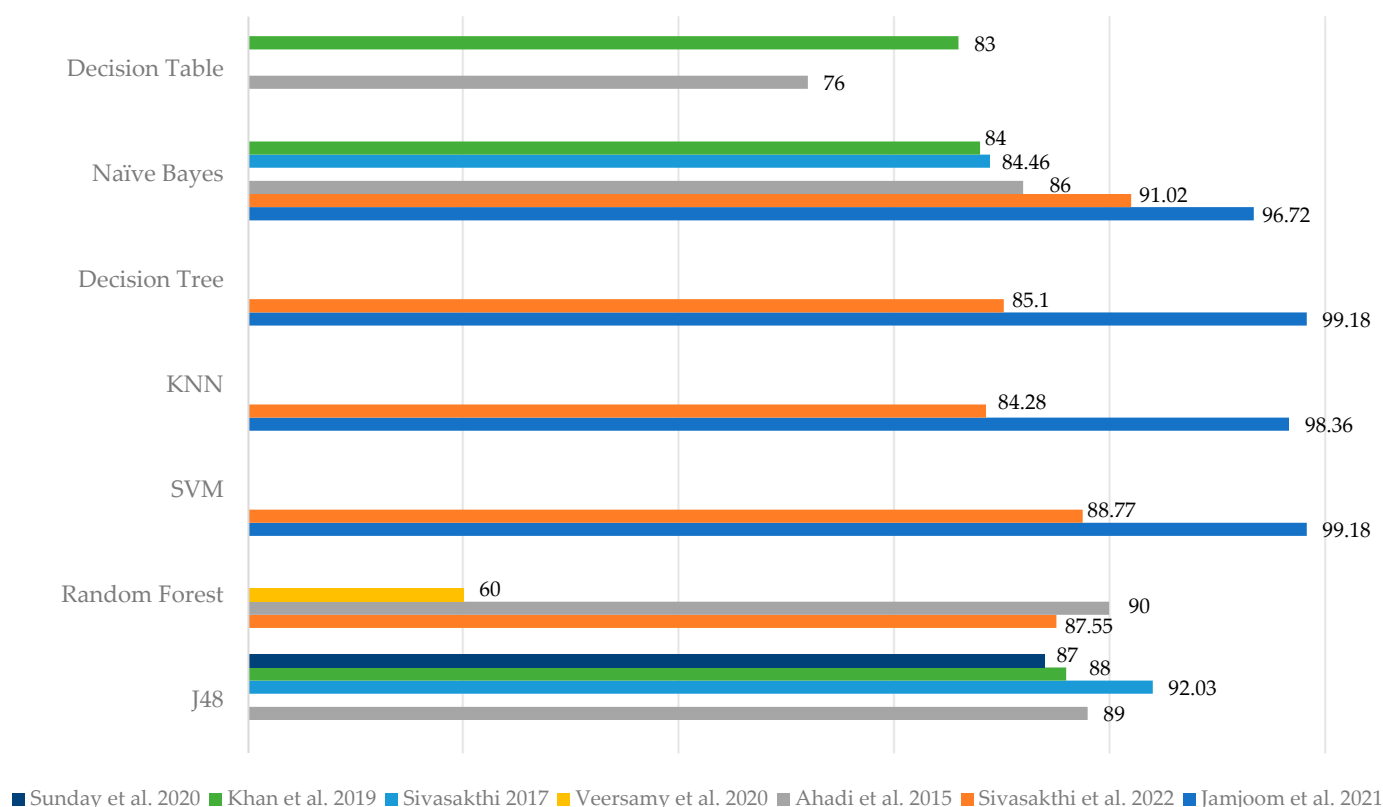
Comparing the results of predictions based on different data sources, there is no clear evidence that one particular data source leads to better performance than the others. This can be analyzed in the studies carried out in [29,36,37], where each study uses a different data source: questionnaires in [29], academic records in [36], and personal data in [37]. In these three studies, the models were able to achieve good results, regardless of the variation in the type of data. However, it was found that these results can be just as good or even better when several data sources are used together, as can be seen in the studies of [28,34]. In these studies, the use of two types of data resulted in some of the best performances identified in the studies analyzed. This suggests that integrating different types of data can improve the ability to predict student performance. Further research should explore the potential benefits of integrating multiple data sources to improve predictive accuracy.

An analysis of the size of the datasets used to train the different models and the respective quality of the predictions obtained shows that the algorithms with the best performance were in most cases used with the largest datasets, as observed in the studies of [28,29,34,37], where the datasets always have more than 200 instances. This suggests that factors such as data quality may also play an important role in influencing predictive accuracy. In the studies analyzed, there were cases where small datasets achieved good performance [32], but with models trained on so little data, there may be an inability of the

model to generalize to new instances, related to the inability of the model to represent all the variability and details of the population.

#### 4.6. Algorithms with Best Results

Among all the models observed, some of the ML algorithms that showed the best results were the SVM and the Decision Tree, which, in the paper of [28], using questionnaires and academic records datasets, achieved an accuracy of 99.18%. Also noteworthy were Naïve Bayes and J48, which always gave very positive results. Figure 2 shows the accuracy of the models used in the analyzed papers. It is important to note that Figure 2 only shows the cases of algorithms that have been used more than once so that there is a point of comparison and that were evaluated using the accuracy metric.



**Figure 2.** Accuracy of the most-used models. The references used are: Jamjoom et al. 2021 [28], Sivasakthi et al. 2022 [29], Ahadi et al. 2015 [30], Khan et al. 2019 [32], Veersamy et al. 2020 [33], Sunday et al. 2020 [36] and Sivasakthi 2017 [37].

Based on the models in Figure 2 and the corresponding accuracy results, an ANOVA test was performed to determine any significant statistical differences. This test was performed using the IBM SPSS Statistics tool. The ANOVA results show that the  $p$ -value between the groups was 0.395, indicating that there is no evidence of statistically significant differences between the group means. The F-statistic value confirms this fact, as it reaches a value of 1.137, indicating that the differences between the group means are not large enough to be considered statistically significant.

#### 4.7. Deep Learning Approach

While traditional ML techniques are most commonly used to predict student performance, the application of deep learning methods is still in its early stages. The relative newness and complexity of deep learning may be factors contributing to its limited adoption. In the paper of [35], a DNN was used to predict student performance using personal data and academic records. The DNN performed better than six other algorithms, includ-

ing the widely used SVM. This suggests that DNNs, with their ability to model complex, non-linear relationships in data, have significant potential in this area. The conclusions of the paper of [35] provide a valuable direction for future research in this area.

#### 4.8. Challenges and Gaps

This section highlights the major challenges and limitations identified in the literature review. The following difficulties have a direct impact on the quality of the study of predicting student performance in IP in HE:

- **Data Quality Assurance**

The foundation of any predictive model is the quality of the underlying data. Ensuring the relevance and accuracy of the data is critical to the success of a study [39]. Data quality assurance involves a careful process of data collection and pre-processing [40]. Given the complexity of educational data, which often include multiple sources and formats, a robust data assurance strategy is essential. Achieving balance in datasets is also a significant challenge, as imbalances can affect results and compromise the predictive ability of the model.

To address data quality concerns, we propose the implementation of two possible solutions: (i) Advanced data cleaning techniques: the implementation of sophisticated data cleaning methods, such as the detection of outliers (an observation that is very different from the others in the series or is inconsistent) and the handling of missing values (one or more missing instances in the data) through imputation techniques (to handle missing data by filling in or estimating the missing values instead of discarding the incomplete records), can improve data quality. (ii) Balancing techniques: applying techniques such as SMOTE (Synthetic Minority Over-sampling Technique) to resolve class imbalances and increase the predictive power and generalizability of the model.

- **Selection of the Algorithms**

The abundance of available ML algorithms presents a significant challenge in the selection process. While the diversity of options is beneficial, it is difficult to identify the most appropriate algorithm for the given context. Ideally, a comprehensive approach would involve testing multiple algorithms to identify the most effective predictive model in IP in HE. However, time constraints often limit this exhaustive exploration. Finding a balance between exhaustive exploration and time efficiency becomes a critical aspect of the research process. Researchers must carefully weigh the necessary constraints and adopt strategies that allow for a careful selection of algorithms without compromising the overall study schedule. To help with this, research has been undertaken to develop strategies for selecting the most appropriate algorithm for specific tasks [41].

In order to mitigate the challenges of algorithm selection, this study also acts as an attempt to find a solution by carrying out a comparative analysis of different algorithms used in the field of EDM. This approach will allow for a more informed decision to be made about the most appropriate algorithm for predicting student performance.

Another viable solution to this problem, if there are no time constraints, is to apply cross-validation techniques to a group of algorithms. The use of cross-validation methods ensures a robust evaluation and helps to select the model with the best-performing model from the group.

- **Evaluation Metrics Application**

The selection of appropriate evaluation metrics is a decision that depends on the specific problem at hand [42]. The literature review revealed cases where studies were based on a single metric, potentially limiting the understanding of a model's performance. Multiple metrics are recommended to provide a comprehensive view of the predictive model's strengths and weaknesses. Researchers need to match the metrics chosen to the complexity of the educational context and the nature of the prediction task. This will ensure an evaluation that goes beyond a superficial assessment and provides a deeper

understanding of the model's effectiveness. The simplest solution to this problem is to apply multi-metric evaluation, which uses a range of metrics, such as accuracy, precision, recall, and F1-score, to provide a more comprehensive assessment of model performance.

By addressing these challenges, researchers can strengthen the foundations of their studies and improve the reliability and applicability of ML techniques in predicting student performance in the dynamic landscape of HE, particularly in IP.

## 5. Discussion

At the beginning of the research, three research questions were identified, which will be answered below.

RQ1: What were the most-used ML algorithms proposed by the researchers for predicting students' performance in HE in IP? The literature review identified several algorithms applicable to the context of this study. However, four algorithms stood out and were used in multiple papers. These algorithms were Naïve Bayes, J48, Random Forest, and SVM, used in five, four, three, and three different papers, respectively. As mentioned above, when presenting the most commonly used algorithms, the fact that these algorithms are widely used does not always mean that they are the best algorithms for this type of problem, as has been shown with approaches using neural networks such as DNN and PNN, which have achieved extremely positive results.

RQ2: Which datasets were used, and which evaluation metrics were considered most appropriate for measuring the performance of predictive models in HE in IP? Regarding the datasets, academic records were the most used; with those, it is possible to observe the student's past performance and understand their ability and academic commitment over time. The fact that it is used most often does not mean that it produces better results, since the analysis carried out showed that the best results were obtained by using several data sources. It is worth noting, however, that in these cases of multiple use, the type of data used were always academic records, which highlights their relevance and impact in the context of prediction. As for the metrics, after the analysis of the papers, it became clear that there was no perfect answer to this question. There is no ideal metric. The choice of metrics always depends on the context in which people work. However, it was possible to conclude that more than one metric should be chosen, usually accuracy and F1-Score. Accuracy is preferred because of its simplicity and ease of interpretation. However, its limitation in unbalanced datasets, where it can be misleadingly high, necessitates the use of additional metrics. The F1-Score, which combines precision and recall into a single metric, is particularly valuable in scenarios where the balance between false positives and false negatives is critical. By using a combination of metrics, it is possible to gain a more complete understanding of model performance. The type of problem (classification or regression) to be applied also influences the metric to be chosen. For example, for regression and deep learning problems, the use of the RMSE metric is recommended.

RQ3: Which ML algorithms seem to better predict students' performance in HE in IP? Twenty algorithms were identified in the papers analyzed. To understand which could be the best, only those algorithms that were applied more than once were considered to compare them, by calculating the average accuracy of the algorithm. This way, it can be concluded that the algorithm with the best average accuracy was the SVM, which was applied in three studies (only in two of those was the accuracy metric used) and obtained an average accuracy of 93.97% using both academic records and questionnaire datasets. The fact that there are not yet many well-established studies on the application of deep learning means that the results found could not be used in this comparison. However, the potential demonstrated in papers such as [35] is great and suggests that this could be an opportunity for future research.

In the literature reviewed, there were situations where the amount of data used were very small, as well as cases of unbalanced datasets, directly affecting the quality of the model, as explicitly stated by the respective authors. In some of the papers analyzed, there

were cases of the use of little-known algorithms that do not appear in any other study, demonstrating that they may not be the most appropriate.

## 6. Threats to Validity

In this section, we present the main threats to the validity of this work and discuss mitigation strategies. This study, although comprehensive in its review of methodologies for predicting student performance in IP courses, has some limitations that need to be acknowledged.

Firstly, the scope of our literature review is limited by the selection criteria and databases used to identify relevant studies. Although we have attempted to include relevant studies, there is always the possibility that some relevant research may have been missed due to the limitations of our search parameters, such as the search string, the date of publication, or the availability of publications in certain databases.

Secondly, the review focuses specifically on the use of ML algorithms to predict student performance. Although ML techniques are a powerful tool, this focus inherently excludes other potentially valuable methodologies that could also contribute to understanding and improving student outcomes.

Finally, the conclusions reached in this study depend on the performance of the ML algorithms in the papers analyzed. The performance of these algorithms depends on the quality and characteristics of the datasets used, as well as on the evaluation steps applied. The variability in the granularity of the data and the attributes included in the datasets can significantly influence the results and the generalizability of the conclusions. The evaluation methods used and reported are the only way to analyze the results of the papers. Cases where only metrics such as accuracy are used, without further information presented, can sometimes be misleading and directly affect the conclusions reached in this study. Recognizing this type of limitation is crucial to interpreting the conclusions and results of this study.

## 7. Future Research Directions

In this section, we suggest future research directions for predicting student performance in IP courses in HE.

- Use of deep learning techniques

The literature review revealed a significant gap in the use of deep learning techniques for predicting performance in IP in HE. Although several ML algorithms have been identified, the specific application of deep learning in this educational context remains vastly underutilized. Thus, identifying and exploiting strategies that fully utilize the capabilities of DNNs may represent the next significant step in the evolution of students' performance prediction.

The emerging field of deep learning has demonstrated success in applications ranging from computer vision to natural language processing. The application of these techniques in the education domain remains largely unexplored. By employing more complex models and deeper layers, we can capture richer representations of the patterns present in students' data.

To advance in this area, it is essential to develop a deeper understanding of the specific challenges faced when applying deep learning techniques in an educational context. One obstacle is the need to deal with often complex and diverse datasets, which requires effective pre-processing strategies. In addition, the interpretability of deep learning models can be challenging, and it is crucial to develop approaches that make model decisions more understandable and explainable.

- Integration of various data sources

Another promising direction for future research is to explore using different types of data to predict performance in IP courses in HE. In the analyzed papers, the datasets mainly consist of assessment grades, personal data, or quiz answers. There are also a few studies



that work with the code developed by the students. The development of models that take into consideration the code developed could provide a more comprehensive understanding of student engagement and learning patterns.

The use of aptitude tests in IP could also be a promising strategy for using the obtained results to predict students' success when they start an IP course.

Analyzing different types of information can offer a different view of students' mental and behavioral aspects, allowing for the development of predictive models that do not rely only on traditional metrics.

These approaches have their own challenges, including the complexities of data integration, the development of sophisticated algorithms capable of processing multiple types of data, and the resolution of privacy issues related to the collection and analysis of different kinds of data.

## 8. Conclusions

Through this review of the available literature, this study examined the methodologies for predicting student performance in IP in HE, with a particular focus on analyzing the ML algorithms applied and the evaluation metrics used. This study describes a systematic literature review in which several papers were selected and analyzed. The results obtained provided valuable insights into the effectiveness of these approaches in an educational context.

The results of this study suggest that the use of ML algorithms, such as SVM, can be a valuable tool for predicting student performance in IP in HE. These algorithms can help identify students at risk of failing, allowing for targeted interventions and improved educational support. However, it is important to note that the performance of the models is directly dependent on the quality of the data used to train and test the model, suggesting that a process of data collection and analysis is essential for successful implementation.

We believe that this study makes a significant contribution to understanding which ML algorithms are used for predicting student performance in IP in HE and which evaluation metrics and data sources are most commonly used. The search for effective solutions to the high failure and dropout rates in IP continues and with the results obtained in this study we hope to assist future research in the process of choosing which algorithms to use in developing a possible solution, as well as the evaluation metrics to use and the types of data to explore that are best suited to this type of problem. By systematically comparing algorithms, metrics, and datasets, we have highlighted their relative strengths and limitations, and thereby clarified their practical utility and potential effectiveness in predicting student performance.

Naïve Bayes and J48 were found to be the most commonly used ML algorithms, with SVM achieving the highest average accuracy. However, deep learning techniques such as DNN and PNN showed significant potential, suggesting a promising direction for future research. In terms of datasets and evaluation metrics, academic records were the most commonly used, often in combination with other data sources to improve results. While accuracy was the most commonly used metric, this study highlighted the need for a combination of metrics to comprehensively assess model performance.

In the quest to improve the prediction of student performance in IP in higher education, some promising directions for future research have been identified. The application of deep learning techniques, which are still relatively unexplored in educational contexts, could be the next step in improving the accuracy of predictions, as well as exploring the use of different types of data, such as student-developed code and the use of aptitude tests, which could provide richer datasets and more nuanced insights into student performance.

**Author Contributions:** Conceptualization, F.B.C., A.G. and A.R.B.; methodology, F.B.C., A.G., A.R.B. and J.B.; validation, F.B.C., A.G., A.R.B. and J.B.; formal analysis, F.B.C., A.G. and A.R.B.; investigation, J.P.J.P.; resources, F.B.C., A.G. and A.R.B.; data curation, F.B.C., A.G. and A.R.B.; writing—original draft preparation, J.P.J.P.; writing—review and editing, F.B.C., A.G., A.R.B. and J.B.; supervision,

F.B.C., A.G. and A.R.B.; project administration, F.B.C., A.G., A.R.B. and J.B.; funding acquisition, J.B. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

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

## References

1. Watson, C.; Li, F.W. Failure rates in introductory programming revisited. In Proceedings of the 2014 Innovation and Technology in Computer Science Education Conference, Uppsala, Sweden, 21–25 June 2014. [\[CrossRef\]](#)
2. Bennedsen, J.; Caspersen, M.E. Failure rates in introductory programming. *ACM Inroads* **2019**, *10*, 30–36. [\[CrossRef\]](#)
3. Lahtinen, E.; Ala-Mutka, K.; Järvinen, H.-M. A study of the difficulties of novice programmers—12 years later. *ACM SIGCSE Bull.* **2005**, *37*, 14–18. [\[CrossRef\]](#)
4. Qian, Y.; Lehman, J. Students' misconceptions and other difficulties in introductory programming. *ACM Trans. Comput. Educ.* **2017**, *18*, 1–24. [\[CrossRef\]](#)
5. Gomes, A.; Mendes, A.J.N. Learning to program—difficulties and solutions. In Proceedings of the International Conference on Engineering Education—ICEE 2007, Coimbra, Portugal, 3–7 September 2007.
6. Jenkins, T. On The Difficulty of Learning To Program. In Proceedings of the 3rd Annual LTSN-ICS Conference, Loughborough University, Loughborough, UK, 27–29 August 2002.
7. Bennedsen, J.; Caspersen, M.E. Abstraction ability as an indicator of success for learning object-oriented programming? *ACM SIGCSE Bull.* **2006**, *38*, 39–43. [\[CrossRef\]](#)
8. Byrne, P.; Lyons, G. The effect of student attributes on success in programming. *ACM SIGCSE Bull.* **2001**, *33*, 49–52. [\[CrossRef\]](#)
9. Luxton-Reilly, A.; Simon, Albluwi, I.; Becker, B.A.; Giannakos, M.; Kumar, A.N.; Ott, L.; Paterson, J.; Scott, M.J.; Sheard, J.; et al. Introductory programming: A systematic literature review. In Proceedings of the ITiCSE '18: 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education, Larnaca, Cyprus, 2–4 July 2018.
10. Gomes, A.; Mendes, A. A study on student's characteristics and programming learning. In Proceedings of the ED-MEDIA 2008—World Conference on Educational Multimedia, Hypermedia & Telecommunications, Vienna, Austria, 30 June 2008.
11. Gomes, A.; Mendes, A. A teacher's view about introductory programming teaching and learning: Difficulties, strategies and motivations. In Proceedings of the 2014 IEEE Frontiers in Education Conference (FIE), Madrid, Spain, 22–25 October 2014. [\[CrossRef\]](#)
12. Tomai, E.; Reilly, C.F. The impact of math preparedness on introductory programming (CS1) success (abstract only). In Proceedings of the SIGCSE '14: The 45th ACM Technical Symposium on Computer Science Education, Atlanta, GA, USA, 5–8 March 2014. [\[CrossRef\]](#)
13. Lishinski, A.; Yadav, A.; Enbody, R.; Good, J. The influence of problem solving abilities on students' performance on different assessment tasks in CS1. In Proceedings of the SIGCSE '16: The 47th ACM Technical Symposium on Computing Science Education, Memphis, TN, USA, 2–5 March 2016; pp. 329–334. [\[CrossRef\]](#)
14. Jokhan, A.; Chand, A.A.; Singh, V.; Mamun, K.A. Increased Digital Resource Consumption in Higher Educational Institutions and the Artificial Intelligence Role in Informing Decisions Related to Student Performance. *Sustainability* **2022**, *14*, 2377. [\[CrossRef\]](#)
15. Sobral, S.R. Strategies on Teaching Introducing to Programming in Higher Education. In *Advances in Intelligent Systems and Computing*; Springer: Cham, Switzerland, 2021. [\[CrossRef\]](#)
16. Gomes, A.; Mendes, A.J.; Marcelino, M.J. Computer Science Education Research—An overview and some proposals. In *Innovative Teaching Strategies and New Learning Paradigms in Computer Programming*; Queiroz, R., Ed.; IGI-Global: Hershey, PA, USA, 2015; pp. 1–29.
17. Köhler, J.; Hidalgo, L.; Jara, J.L. Predicting Students' Outcome in an Introductory Programming Course: Leveraging the Student Background. *Appl. Sci.* **2023**, *13*, 11994. [\[CrossRef\]](#)
18. Liu, Y.; Fan, S.; Xu, S.; Sajjanhar, A.; Yeom, S.; Wei, Y. Predicting Student Performance Using Clickstream Data and Machine Learning. *Educ. Sci.* **2023**, *13*, 17. [\[CrossRef\]](#)
19. Quille, K.; Bergin, S. CS1: How will they do? How can we help? A decade of research and practice. *Comput. Sci. Educ.* **2019**, *29*, 254–282. [\[CrossRef\]](#)
20. Shahiri, A.M.; Husain, W.; Rashid, N.A. A Review on Predicting Student's Performance Using Data Mining Techniques. *Procedia Comput. Sci.* **2015**, *72*, 414–422. [\[CrossRef\]](#)
21. Alhothali, A.; Albsisi, M.; Assalahi, H.; Aldosemani, T. Predicting Student Outcomes in Online Courses Using Machine Learning Techniques: A Review. *Sustainability* **2022**, *14*, 6199. [\[CrossRef\]](#)
22. Silva, C.; Fonseca, J. Educational data mining: A literature review. In *Advances in Intelligent Systems and Computing*; Springer: Cham, Switzerland, 2017. [\[CrossRef\]](#)

23. Bachhal, P.; Ahuja, S.; Gargish, S. Educational Data Mining: A Review. *J. Physics Conf. Ser.* **2021**, *1950*, 012022. [[CrossRef](#)]
24. Scherer, R.; Siddiq, F.; Viveros, B.S. The cognitive benefits of learning computer programming: A meta-analysis of transfer effects. *J. Educ. Psychol.* **2019**, *111*, 764–792. [[CrossRef](#)]
25. Sobral, S.; Oliveira, C. Predicting Students' Performance in Introductory Programming Courses: A Literature Review. In Proceedings of the 15th International Technology, Education and Development Conference, Online, 8–9 March 2021; pp. 7402–7412.
26. Kitchenham, B. *Procedures for Performing Systematic Reviews*; Keele University: Keele, UK; National ICT Australia: Sydney, Australia, 2004; Volume 33.
27. Silva, M.; Shaffer, E.G.; Nytko, N.; Amos, J.R. A case study of early performance prediction and intervention in a computer science course. In Proceedings of the ASEE Annual Conference and Exposition, Conference Proceedings, Virtual Online, June 2020. [[CrossRef](#)]
28. Jamjoom, M.; Alabdulkreem, E.; Hadjouni, M.; Karim, F.; Qarh, M. Early prediction for at-risk students in an introductory programming course based on student self-efficacy. *Informatica* **2021**, *45*, 1–9. [[CrossRef](#)]
29. Sivasakthi, M.; Pandiyan, M. Machine Learning Algorithms to Predict Students' Programming Performance: A comparative Study. *J. Univ. Shanghai Sci. Technol.* **2022**, *24*, 1–8.
30. Ahadi, A.; Lister, R.; Haapala, H.; Vihavainen, A. Exploring machine learning methods to automatically identify students in need of assistance. In Proceedings of the ICER '15: International Computing Education Research Conference, Omaha, NE, USA, 9–13 July 2015.
31. Đambić, G.; Krajcar, M.; Bele, D. Machine learning model for early detection of higher education students that need additional attention in introductory programming courses. *Int. J. Digit. Technol. Econ.* **2016**, *1*, 1–11.
32. Khan, I.; Al Sadiri, A.; Ahmad, A.R.; Jabeur, N. Tracking student performance in introductory programming by means of machine learning. In Proceedings of the 2019 4th MEC International Conference on Big Data and Smart City (ICBDSC), Muscat, Oman, 15–16 January 2019. [[CrossRef](#)]
33. Veerasamy, A.K.; D'Souza, D.; Apiola, M.-V.; Laakso, M.-J.; Salakoski, T. Using early assessment performance as early warning signs to identify at-risk students in programming courses. In Proceedings of the 2020 IEEE Frontiers in Education Conference (FIE), Uppsala, Sweden, 21–24 October 2020. [[CrossRef](#)]
34. Cooper, C. Using Machine Learning to Identify At-risk Students in an Introductory Programming Course at a Two-year Public College. *Adv. Artif. Intell. Mach. Learn.* **2022**, *2*, 407–421. [[CrossRef](#)]
35. Shen, G.; Yang, S.; Huang, Z.; Yu, Y.; Li, X. The prediction of programming performance using student profiles. *Educ. Inf. Technol.* **2023**, *28*, 725–740. [[CrossRef](#)]
36. Sunday, K.; Ocheja, P.; Hussain, S.; Oyelere, S.S.; Samson, B.O.; Agbo, F.J. Analyzing student performance in programming education using classification techniques. *Int. J. Emerg. Technol. Learn. (ijET)* **2020**, *15*, 127–144. [[CrossRef](#)]
37. Sivasakthi, M. Classification and prediction based data mining algorithms to predict students' introductory programming performance. In Proceedings of the 2017 International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, India, 23–24 November 2017; pp. 346–350.
38. Naser, M.Z.; Alavi, A.H. Error Metrics and Performance Fitness Indicators for Artificial Intelligence and Machine Learning in Engineering and Sciences. *Arch. Struct. Constr.* **2021**, *3*, 499–517. [[CrossRef](#)]
39. Sessions, V.; Valtorta, M. The effects of data quality on machine learning algorithms. In Proceedings of the 2006 International Conference on Information Quality, ICIQ 2006, Cambridge, MA, USA, 10–12 November 2006.
40. Realinho, V.; Machado, J.; Baptista, L.; Martins, M.V. Predicting Student Dropout and Academic Success. *Data* **2022**, *7*, 146. [[CrossRef](#)]
41. Sala, R.; Zambetti, M.; Pirola, F.; Pinto, R. How to select a suitable machine learning algorithm: A feature-based, scope-oriented selection framework. In Proceedings of the Summer School "Francesco Turco", Palermo, Italy, 12–14 September 2018.
42. Liu, Y.; Zhou, Y.; Wen, S.; Tang, C. A Strategy on Selecting Performance Metrics for Classifier Evaluation. *Int. J. Mob. Comput. Multimedia Commun.* **2014**, *6*, 20–35. [[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.